ABSTRACT

WANG, XIN. Statistical Methods for Gene-Gene Interaction: Detections and Classifications. (Under the direction of Dr, Jung-Ying Tzeng and Dr. Daowen Zhang).

Current focus of genetic association studies for complex disease has been shifted from assessing the genetic main effect to interaction effect among genes. Gene-Gene interactions (GxG) are believed to play an important role in complex diseases. Detecting GxG would help us to reveal the underlying mechanisms of complex disease, explain the missing heritability and understand the inconsistency among different studies.

Many proposed GxG methods considered interactions among SNPs instead of interactions among genes. We believe that there are several advantages to assess GxG at gene level instead of SNP level. Also, when the number of candidate genes increases, the corresponding number of GxG increases dramatically. Thus, a follow-up question is how to efficiently map GxG among large gene set. Using exhausting search would be time consuming and loss power. Reducing the searching space is a promising solution to find the casual GxG. Once we find the important disease genes, how to use them to predict patients' status with probability estimates is another interesting topic in both statistics and bioinformatics research.

In chapter 2, I describe the method that using similarity regression to map GxG between 2 candidate genes at gene level. The simulation study and real data analysis show that the proposed method has a strong and robust power in detecting GxG compared to several other methods. In chapter 3, I give a detailed description for the L1 penalty regression model used to find the causal GxG among large gene set. This model successfully incorporates biological information into the statistical model so that the results have both biological and statistical support. In chapter 4, I show the method that solves the multi-class soft classification problem

with support vector machine (SVM). This method is developed under a fast yet efficient framework. The simulation study shows that it has a good classification accuracy when the underlying probability functions are complicated.

Statistical Methods for Gene-Gene Interaction: Detections and Classifications

by
Xin Wang

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

Bioinformatics and Statistics

Raleigh, North Carolina

2013

APPROVED BY:

Dr. Jung-Ying Tzeng Co-Chair of Advisory Committee	Dr. Daowen Zhang Co-Chair of Advisory Committee
Dr. Alison Motsinger-Reif	Dr. Nadia Singh
Dr. Yichao Wu	

DEDICATION

To my parents and my wife!

BIOGRAPHY

Xin Wang was born in Jiangsu, the People's Republic of China. He received his Bachelor of Science degree in Plant Technology from Shanghai Jiao Tong University in 2008. During his undergraduate study, he found his interest and enthusiasm in building statistical model to analysis genetic data. In 2008, he entered the bioinformatics Ph.D. program and pursued a Ph.D. co-major in bioinformatics and statistics at the North Carolina State University (NCSU). During his study at NCSU, he worked under the direction of Dr. Jung-Ying Tzeng and Dr. Daowen Zhang. He focused on his research on method development for gene-gene interaction study and classification problem.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor Dr. Jung-Ying Tzeng, for her influence both academically and personally. She is always patient and provides very helpful suggestions for the problems I encountered during my research. I also thank Dr. Daowen Zhang for his advice on my research. I also would like to thank Dr. Zhao-Bang Zeng for his extremely generous support in the last several years. I would like to thank my other committee members. Dr. Alison Motsinger-Reif is always happy to explain the questions I have in my research, Dr. Yichao Wu helps me a lot when I started my research and Dr. Nadia Singh gives me lots of help in understanding the biological knowledge.

I also appreciate the support from the staffs at the Bioinformatics Research Center. Siarra Dickey helped address many issues I had about the graduate school regulations. Chris Smith helped solve many HPC problems I encountered during my research. Kevin Dudley gave me many support for IT issues. Many thanks to all the nice people I met here! Without their kind help, I would not be able to complete my doctoral study so smoothly.

A special thanks to Zhi Wang, Jing Zhao, Kuangyu Wang, Yuelong Guo, Ronglin Che, Wenjing Lu, ‘Ginger’ Monnat Pongpanich, Oyindamola Oki, Gunjan Hariani, and Alexander Griffing for all their help and wonderful friendship.

Finally, I would like to express my special thanks to my wife, Qiaoqiao Liu. Without her sincerely love, I would never be who I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Definition of Gene-Gene interaction (GxG) and its importance

There has been a long-standing interest in the investigation of interactions in genetics, including gene-environment and gene-gene interactions, based on the assumption that they play an important role in the etiology of complex diseases or traits. Biological interaction means physical interaction among biomolecules in gene regulatory networks and biochemical pathways. GxG makes the effect of a gene on a phenotype be dependent on one or more other genes (Bateson, 1909; Moore and Williams, 2005). From the statistical view, interaction means deviation from additivity in a linear model that describes the relationship between multilocus genotype and phenotype variation at the population level (Fisher, 1918). Although there were debates about the relationship between biological interaction and statistical interactions, evidences showed that statistical interactions and biological interactions can converge to the same scientific process (Bush et al., 2009). For example, Bridge used statistical model to identify genes with interaction effects on *Drosophila* eye color (Bridge, 1919), and the corresponding biological mechanism that depicts how these genes influence biological pathways was understood many years later (Lloyd et al., 1998). Thus, the statistical interaction evidences can be used to infer the biological interactions.

Investigating GxG would benefit us in both of biological and statistical view. From biological views, detecting GxG would help us reveal the underlying process of complex disease, such as hypertension, cancer, diabetes, and psychiatric disorders (Lin et al., 2013; Pillai et al., 2013; Koh-Tan et al., 2013; Howson et al., 2012; Ziyab et al., 2013). The developments of complex disease are believed to involve complicated biological mechanism at different levels, such as gene regulatory networks and protein-protein interaction. Among these mechanisms, a range of genes and/or their products are involved such as coding and non-coding RNA, protein and metabolites. They are working together instead of working independently, e.g., in gene regulation model, a protein product from one gene can control the expressions of other genes (Zuk et al., 2012). These genes may have no main effects but strong interaction effect. Thus, ignoring GxG effect would miss these important genes and cause inconsistent findings among different studies. For example, only gene apolipoprotein E4 was found to be important to the Alzheimer's Disease (AD) using single locus method (Corder et al., 1993; Saunders et al., 1993; Strittmatter et al., 1993). Years later, 3 more genes were found to be important to AD using GxG detection method and the corresponding mechanisum describing how these genes interact with each other was found accordingly (Combarros et al., 2008). Diabetes is another complex disease that is believed involving multiple interacting genes (Florez et al., 2003). For instance, GxG have been found between loci on chromosomes 2 and 15, as well as between loci on chromosomes 1 and 10, in patients with type II diabetes (Cox et al., 1999; Wiltshire et al., 2006).

Understanding GxG may also help to uncover the missing heritability (Marchini et al., 2005; Evans et al., 2006) and explain the inconsistent findings from main-effect analyses (Hirschohorn et al., 2002). Even if GxG explains only a tiny fraction of "missing heritability", the importance of revealing the specific interactions that underlie that fraction would also give us the unique type of biological insight. From such insight, we can having a better understanding of biological mechanisms on both gene level and pathway level.

Current methods for GxG detections

Many methods have been proposed for GxG detection. They can be divided into 2 groups according to whether they consider the interaction at the SNP level or gene level.

SNP-based GxG method.

The most straight forward way of detecting SNP-SNP interaction is to build a regression-based model with phenotype and genotype to test the interaction effect. A similar approach can be used for case-control study, in which case logistic rather than linear regression is used. Although regression based tests of interaction seems the most natural way, it can only capture of the interaction effect between 2 SNPs. Kooperberg et al. (2001) proposed the logic regression which uses the indicator of SNP sets rather than the SNPs' genotypes as predictors. Thus the interaction effect among many SNPs can be included into the model by the product of 2 predictors. This method has shown higher power in detecting interaction effect when higher-order SNP-SNP interactions exist. However, the construction of SNP set is arbitrary.

An SNP set may contain just 2 SNPs or all available SNPs. One potential disadvantage of logic regression is that it needs to try all possible SNP combinations and this can be time consuming. Similarly, multivariate adaptive regression splines (MARS) (Friedman 1991; Cook et al., 2004) uses SNP set indicator to represent SNP information rather than using the genotype values of 0, 1, and 2. MARS can also select variables so it is more efficient to find the important GxG. Lin et al. (2008) showed that MARS has a better performance in finding GxG than traditional logistic regression.

Bayesian model can also be applied to detect the GxG. Zhang and Liu (2007) proposed Bayesian Epistasis Association Mapping (BEAM) method to find causal genes for age-related macular degeneration. In BEAM, SNPs are divided into 3 groups: 1) a group contains the SNPs with no effect to trait; 2) a group contains SNPs with main effect; and 3) a group contains SNPs with interaction effects. Prior distributions are specified for the group membership of each SNPs and the relevant parameters. Combining prior distributions with the likelihood, a posterior distribution can be generated for statistical inference using MCMC simulation. Zhang and Liu (2007) have shown that BEAM is more powerful than stepwise logic regression in GxG mapping. One possible limitation of BEAM it that it assumes that the SNP correlation can be directly inferred from the MCMC. The SNPs are usually highly correlated in the same gene but the SNPs between different genes may be independent. Thus, without considering the gene information, such assumption may cause power loss when analyzing dense SNP data.

Besides the regression models, recursive partitioning approaches are also used to detect GxG interaction by building classification trees. For example, Breiman et al. (1984) proposed classification and regression tree (CART) to find the important genes for the trait. Each SNP can be used as a node in the classification tree to split the population. Important SNPs are first selected and set (?) as parental nodes since they have better abilities to differentiate the observations. Less important SNPs are selected later and set (?) as child nodes to further grow the tree. The tree stops growing until all SNPs are used. The importance of each SNP is measured through the process of tree building.

Instead of a single tree, using an ensemble of trees may significantly improve the classification accuracy. A popular ensemble tree approach is Random Forest (RF). Unlike CART that uses all available SNPs to build one tree, RF first forms many small SNP subsets by selecting different SNP combinations. Each subset is then used to build a classification tree and measure the importance of the corresponding SNPs. Different SNP subsets may have different classification accuracy, depending on whether it includes the causal SNPs. RF summarizes the importance scores weighted by the accuracy and form an overall importance factor for each SNP. RF can measure SNPs importance fast because the algorithm can run in parallel and each small SNP subset can be done quickly. However, just like the CART, RF can only measure the importance of the SNPs but it is difficult to directly reveal the interaction pattern between SNPs.

Many other data-mining approaches have been proposed for mapping GxG interactions, such as genetic programming (Nunkesser et al., 2007), neural networks (Motsinger et al., 2006; Motsinger et al., 2008), pattern-mining (Li et al., 2007; Long et al., 2009), and multifactor dimensionality reduction (MDR) (Ritchie et al., 2001; Hahn et al., 2003; Moore, 2004). MDR was first proposed for case-control data, but later it was extend to quantitative data and adjustment for covariates (Lee et al., 2007; Lou et al., 2007). Similar to random forest, MDR also consider small SNP subsets to find the GxG pattern. Each SNP subset has a pre-specified set size, e.g., each subset contain *n* SNPs. MDR classifies the genotype combinations of each *n*-SNP subset into "high risk" and "low risk" by the ratio of case number and control number under the corresponding genotype combination. An *n*-SNP set may have many possible genotype combinations but MDR reduce them into 2 possible value. Cross validation is then used to find the best SNP subset with the minimum classification error. MDR has been used to identify GxG in many complex disease, such as breast cancer (Ritchie et al., 2001), type 2 diabetes (Cho et al., 2004), rheumatoid arthritis (Julia et al., 2007) and coronary artery disease (Tsai et al., 2007).

Gene-based GxG method

Instead of studying SNP-SNP interactions, many approaches have been proposed to detect GxG at gene level. Compared to the SNP based methods, the gene based methods may have the following advantages. First， genes are the basic units in the biological mechanism.

Hence the gene-level results can be more biologically insightful, easier to interpret, more informative in revealing underlying mechanisms. Second, most genetic variants have different allele frequencies, LD structure, and heterogeneity across diverse human populations while the gene itself is highly consistent across populations. Thus, gene-based method may lead to more consistent results across different studies. Third, modeling the multi-SNPs information within a gene also incorporates the LD among SNPs in the downstreaming analysis. Forth, the polygenic nature of the complex diseases suggests moderate effect size for individual variants. SNPs in aggregate tend to result in more detectable main effects due to the amplification of individual moderate effects Finally, via appropriate dimension reduction to summarize the multi-SNP information, gene-level GxG methods are able to use less degrees of freedom, which further help to gain power improvement over the SNP-level analyses.

The key step in gene based method is to summarize the genetic information from SNP level to gene level. Chatterjee et al. (2006) proposed Tukey's 1-df method to investigate the GxG between two candidate genes. It summarizes each gene's information by the sum of the main effects from the SNPs within the corresponding gene and then uses the product of the two sums as the GxG term. Thus, the model only involves one interaction at the gene level instead of many SNP-level interactions. This method has shown great power improvement in detecting GxG. Motivated by this idea, Wang et al. (2009) improved Tukey's 1-df method with two other ways of summarizing gene information. One way is to use the 1st principle component from Principle Component analysis (PCA) to summarize genetic information so that the LDof SNPs within a gene is taken into account. The other way is to use the 1st lead component from Partial Least Square (PLS) analysis to summarize gene information so that not only the LD information but also the correlations between gene and trait are considered. From the simulation study, the PCA or PLS based method have a better performance than the Tukey's 1-df method, especially when the causal SNP has no or little marginal effect.

GxG detection for large gene set

Most of the methods are available for studying interactions among two or a few genes. However, for complex traits, it is often to have a list of many candidate genes to explore GxG. Even with a moderate size gene set, there can be a huge number of GxG terms even at gene level, e.g., a set of 10 genes would lead to 45 pairwise GxG terms. If the method was proposed for detecting SNP-SNP interaction, the number of corresponding tests would increase exponentially. Several methods have been developed to speed up the exhausting searching using different algorithms (Hemani et al., 2011; Kam-Thong et al., 2011; Schupbach et al., 2010). These methods are confined to either binary disease or quantitative traits and several are designed specifically for computers equipped with particular graphical processing units. However, another problem for all pair-wised interaction searching is power loss due to multiple testing. An alternative solution is to reduce search space of GxG by filtering out potentially unimportant genes (Richie, 2011). In current practice, the GxG search space is reduced either in a trait-supervised fashion or using prior biological information.

A widely used trait-supervised method for reducing the search space is the 2-stage method (screen and clean). For the screen step, it would first apply main-effect association tests on each gene|SNP to remove unimportant ones and then model interactions among the remaining ones (Wu et al., 2010). Two interaction mechanisms for Ayotrophic Lateral Sclerosis (ALS) have been identified by such method (Sha et al., 2009). However, filtering out genes/SNPs through main-effect screening would have low power if the casual genes only have strong interaction effect but no main effect.

Recently, some improvements have been made for the 2-stage approach. Boolean Operation-based Screening and Testing (BOOST) and its modified version GBOOST (Wan et al., 2010; Yung et al., 2011) first examines all two-locus interactions in the screening step where promising SNP pairs are determined through a Kulback–Leibler divergence screen. In the following testing stage, likelihood ratio and χ^2 tests are performed to check if an interactive effect is significant.

Some other 2-stage based methods utilize non-parametric approaches to screen and test GxG. Relief (Robnik-Sikonja and Kononenko, 2003) and Tuned Relief (TuRF) (Moore and White, 2007) use the nearest neighbor method to screen the important genes. For each individual, his/her nearest neighbor is defined as the one has the highest genetic similarity with him/her. Note that the two individuals in one pair may have totally different outcomes, e.g., control vs. case. If a gene is important to the trait, sharing the same genotype would make the paired individuals have no or little trait difference and vice versa. Reliefs sums up all the weighted trait differences to test whether one gene is important to the trait.

Two-stage RF-MARS (TRM) uses RF to screen for important genes and MARS to test the GxG significance. An advantage of TRM is that both RF and MARS automatically select the suitable model based on the data, this feature makes the interaction search more effectively and efficiently.

Incorporating the Biological information into the statistical model is another way to reduce the searching space. Many Bayesian frameworks use biological knowledge as prior information to facilitate finding GxG. Frank et al. (2006) proposed Prioritizer which summarizes the information from multiple sources to generate a “global gene network”. According to this network, each gene is prioritized differently according to its gene function. The assumption behind the approach is disease genes are usually functionally related. Emphasizing the connections between genes may enable us to find more true disease genes when analyzing susceptibility ones. Province Borecki (2008) also proposed a Bayesian resampling method which try to capture the genes with little marginal independent effect but strong interaction effect by using the biological information as a guide in variable selection.

Instead of directly detecting the significant important GxG, Biofilter (Bush et al., 2009) builds the list of potential important genes based on database such as KEGG, Protein interaction database (PID), Biocarter etc. Its underlying rationale is that if more biological evidences exist to support the interactions among a group of genes, the corresponding statistical evidence for GxG is more credible. Biofilter uses an implication index, which is the number of databases supporting certain GxG, to quantify the strength of biological support. If no databases provide support to certain GxG, it would be removed from the search space. Recent studies (Pendergrass et al., 2013; Turner et al., 2011; Bush et al., 2011) showed that Biofilter can effectively reduce the GxG search space and result in biologically meaningful GxG findings. However, directly filtering out genes without incorporating trait information can be too arbitrary. It is not trait-specific and may limit the chance of finding novel GxG.

There are several advantages to perform statistical analyses coupled with biological guidance. It leads to credible findings with both biological and statistical supports. The results may have higher chances to shed insights in forming follow-up biological hypotheses for further cellular and molecular studies.

Multiclass soft classification using Support vector machine

The genes found by GxG mapping methods can be used to understand the underlie mechanism of complex diseases. Sometimes people are also interested in using these important genes to predict new patient’s disease status, e.g., in cancer diagnosis, even for the same type of tumors, it is usually critical to divide them into several subgroups based on their histopathological type, grade, stage, and genetic information. The knowledge of a specific subtype helps to tailor the treatment approaches and dose levels for increased efficacy and drug sensitivity, low toxicity, and the best outcome. Generally speaking, it's a multiclass classification problem.

Depending on what the ultimate goal is, classification can be generally divided into hard classification and soft classification. In hard classification, one is only interested in estimating a classification rule (or classifier) which shall be used to assign a label to a new input vector. Popular examples of hard classifiers include support vector machines, nearest neighbor classifiers, and classification trees. On the other hand, the goal of soft classification is to estimate the conditional probabilities of the response belonging to different subclasses. The probability functions are usually more complex than the classification boundary, so in some sense soft classification aims to solve a more difficult problem than hard classification. However, the probability estimates from soft classification provide valuable measure of uncertainty in classification and hence more informative for decision makings.

Traditional probability estimation methods are based on either regression techniques such as multiple logistic regression, or the density estimation approach such as linear (or quadratic) discriminant analysis (LDA or QDA).

The traditional SVMs have shown high classification accuracy for many applications in assorted scientific areas such as cancer diagnosis, handwritten digits recognition, junk email detection. However, it does not directly featured with soft classification.

Recently, some methods have been proposed to use SVM to estimate the probabilities for multi-class problem. Wang et al. (2008) demonstrated that soft classification can be achieved by training a series of weighted SVM and then aggregating decision rules to form conditional class probabilities. Wu et al. (2010) generalized this method from the binary case to the multiclass case by training weighted multiclass classifiers. However, the number of weighted multi-category classifiers to be trained increases exponentially fast when the number of classes gets larger or the weight grid becomes finer. The computational cost increases dramatically with the number of classes. In addition, when the overall classification problem changes by adding another class, the results for the original problem cannot be used any more and one needs to start over by training all weighted hard classifiers.

Topic Addressed in this dissertation.

In Chapter 2, I give a detailed description of the method using similarity regression to detect GxG for two genes at gene level. 3 statistics are constructed for the corresponding tests: interaction test, joint test and conditional main effect test. Simulation and real data analysis are conducted to compare the proposed method with several other methods. Their distributions are then derived using variance component models. The results from simulation study shows that in most cases, the proposed method has a better performance than the other methods. Although in the real data analysis the GxG between the two candidate genes is not significant, one candidate gene is found to be significant important using the conditional main effect test.

In Chapter 3, I give a comprehensive description of how to find GxG among a large number of candidate genes. I use a new penalized L1 regression model that incorporates both biological information and supervision from traits. Specifically, I first apply the principal component (PC) analysis to summarize the multi-SNP genotypes and SNPxSNP design matrix at gene level, and perform gene selections for important main and interaction effects using L1 penalty regression model. The penalty incorporates supports from known pathways related to the trait and trait-supervised adaptive weights. Simulations and real data analysis are used to demonstrate the utility of the pathway-guided penalized regression for GxG identification.

In Chapter 4, I describe the method of using SVM to solve the multi-class classification problem with probability estimate. I extend the method from binary case to multi-class case in an simple yet efficient way. Simulation and real data are conducted to compare the proposed method and other commonly used methods. The result from the simulation shows that the proposed method has a good and robust performance in many scenarios, especially when the underlying probability function has a complicated form, the proposed method has the best performance.

References

Agresti, A. and Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. The American Statistician, 52 119-126.

Bateson W. Mendel's Principles of Heredity. Cambridge University Press, Cambridge, UK (1909).

Breiman L, Freidman JH, Olshen RA, Stone CJ. Classification and regression trees. Chapman and Hall/CRC; New York: 1984

Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia. 2004;47:549–554.

Combarros, O., et al. Epistasis in sporadic Alzheimer's disease. Neurobiology of Aging, e-publication ahead of print (2008)

Corder, E. H., et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261, 921–923 (1993)

Cook NR, Zee RY, Ridker PM (2004) Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. Stat Med 23:1439–1453

Cox, N. J., et al. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nature Genetics 21, 213–215 (1999)

Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. 52, 399–433 (1918)

Florez, J. C., et al. The inherited basis of diabetes mellitus: Implications for the genetic analysis of complex traits. Annual Review of Genomics and Human Genetics 4, 257–291 (2003).

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–66

Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics. 2003;19:376–382.

Hemani G, Theocharidis A, Wei W, Haley C. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics. 2011;27:1462–1465.

Julia A, Moore J, Miquel L, Alegre C, Barcelo P, Ritchie M, Marsal S. Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. Genomics. 2007;90:6–13.

Kam-Thong T, Czamara D, Tsuda K, Borgwardt K, Lewis CM, Erhardt-Lehmann A, Hemmer B, Rieckmann P, Daake M, Weber F, et al. EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. Eur. J. Hum. Genet. 2011;19:465–471.

Lee SY, Chung Y, Elston RC, Kim Y, Park T. Log-linear model based multifactor-dimensionality reduction method to detect gene-gene interactions. Bioinformatics. 2007;23:2589–2595. [PubMed]

Li Z, Zheng T, Califano A, Floratos A.. Pattern-based mining strategy to detect multi-locus association and gene environment interaction. BMC Proceedings. 2007

Lin HY, Wang W, Liu YH, Soong SJ, York TP, Myers L, Hu JJ. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer. J Hum Genet. 2008;53(9):802-11.

Long Q, Zhang Q, Ott J. Detecting disease-associated genotype patterns. BMC Bioinformatics. 2009;10

Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, D LM. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet. 2007;80:1125–1137.

Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays. 2005 Jun;27(6):637-46.

Moore JH. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn. 2004;4:795–803.

Motsinger A, Lee S, Mellick G, Ritchie M. GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics. 2006;7:39.

Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol. 2008;32:325–340.

Nunkesser R, Bernholt T, Schwender H, Ickstadt K, Wegener I. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. Bioinformatics. 2007;23:3280–3288.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001;69:138–147.

Saunders, A. M., et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology 43, 1467–1472 (1993)

Schupbach T, Xenarios I, Bergmann S, Kapur K. FastEpistasis: a high performance computing solution for quantitative trait epistasis. Bioinformatics. 2010;26:1468–1469.

Strittmatter, W. J., et al. Apolipoprotein E: High-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proceedings of the National Academy of Sciences 90, 1977–1981 (1993).

Tsai CT, Hwang JJ, Ritchie MD, Moore JH, Chiang FT, Lai LP, Hsu KL, Tseng CD, Lin JL, Tseng YZ. Renin-angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene-gene interaction. Atherosclerosis. 2007;195:172–180.

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, Yu W. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am.J. Hum. Genet. 2010;87:325–340.

Wang, J., Shen, X. and Liu, Y. (2008). Probability estimation for large margin classifiers. Biometrika, 95 149-167.

Wiltshire, S., et al. Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21–25 and 10q23–26 in northern Europeans. Annals of Human Genetics 70, 726–737 (2006)

Wu, Y., Zhang, H. H. and Liu, Y. (2010). Robust model-free multiclass probability estimation. Journal of the American Statistical Association., 105 424-436.

Yung LS, Yang C, Wan X, Yu W. GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. Bioinformatics. 2011;27:1309–1310.

Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nat Genet. 2007;39:1167–1173.

Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012 Jan 24; 109(4):1193-8.

Chapter 2

Gene-Gene Interactions Using Gene-Trait Similarity Regression

Introduction

Gene-Gene Interactions (GxG) are believed to be ubiquitous in biology system (Moore, 2003; Carlborg and Haley, 2004) and to play an important role in gene regulation, signal transduction, biochemical networks, and many other physiological and developmental pathways (Moore, 2003; Greenspan, 2001; Phenix et al., 2013; Barkoulas et al., 2013). Identifying GxG may improve the ability to find the susceptible genes and provide insights into biological mechanisms for complex disease, such as Alzheimer’s disease, diabetes, cardiovascular and cancer (Lin et al., 2013; Pillai et al., 2013; Koh-Tan et al., 2013; Howson et al., 2012; Ziyab et al., 2013). It may also help to explain the missing heritability and the inconsistent findings in many genetic association studies (Marchini et al., 2005; Evans et al., 2006).

Many statistical methods have been proposed for studying GxG and have successfully detected some important GxG for complex disease (Zhang and Bonney, 2000; Sheriff and Ott, 2001; Kooperberg and Ruczinski, 2005; Cordell, 2009; Yee et al., 2013). However, most methods consider pair-wised SNP-SNP interactions. There has been a growth in popularity to consider GxG at gene level due to several reasons (Jorgenson and Witte, 2006; Neale and

Sham, 2004; Wang et al., 2009). First，genes are the basic units in the biological mechanism and SNPs within a gene tend to work concordantly. Thus gene-level results may be more biologically insightful, easier to interpret or more informative in revealing underlying mechanisms. Second, the correlations between SNPs are considered when modeling multi-SNP information. Third, SNPs in aggregate tend to result in more detectable main effects due to the amplification of individual moderate effects. We expect a similar amplification effect to exist for GxG effects. Finally, because gene-level GxG methods tend to use less degree of freedom, it is expected to gain potential more power than SNP level analysis.

One key task in model GxG effects at gene level is to aggregate the multi-SNP information at gene level. Chatterjee et al. (2006) proposed Tukey’s 1-df method to investigate the GxG between two candidate genes. The method first obtain the gene-level information by taking the genotype sums from each SNP of a gene, and then uses the product of the two genotype sums as the GxG interaction variable. Consequently, the model only uses one parameter for assessing the gene-level GxG effects. Wang et al. (2009) improved this method by obtaining gene-level information using the principal component analysis (PCA) and partial least square (PLS). In PCA-based methods, the first component from PCA is used to summarize the multi-SNP information and accounts for the linkage disequilibrium (LD) among SNPs. In PLS, the first component from PLS is used to to summarize the multi-SNP information so that not only the LD information but also the correlation between genes and traits are considered. From the simulation study, the PCA or PLS based method have a better performance than the Tukey’s 1-df method, especially when the causal SNP has no or little marginal effect.

Here we propose a new method using the gene-trait similarity regression (SimReg) to detect GxG at gene level. This regression model correlates trait similarity with gene level genetic similarity. Compare to the PCA/PLS based methods, the proposed method provide the following improvement. First, PCA/PLS tend to suffer from information lose in the dimension reduction process by using only the first component, or encounter curse of dimensionality if multiple components are used. In contrast, SimReg incorporates all SNP information within a gene while performing GxG test with a small degrees of freedom. Second, first component from PCA or PLS is actually a sum of weighted SNP genotypes. The interaction term formed by multiplying the two linear combinations would only capture limited forms of non-additive effect. In contrast, as we describe in the method section, SimReg is capable to model a variety of effects by selecting different metrics to quantify gene similarity. Finally, the GxG test under the SimReg framework is computationally efficient because the permutations are not required.

We arrange the remaining paper as follows. In Section 2, we describe the gene-trait similarity regression model and the score tests for evaluating effects of interests (i.e., interaction effect, joint effect of gene and GxG, and conditional main effect test). We also describe how the similarity regression can be connected to a variance component model. We present the simulation study (Section 3) and a real data application on the warfarin study of Wysowski et al. (2007) (Section 4). Finally, we discuss the limitations and future research directions.

Method

The Gene-Trait Similarity Model

Denote Y_i as the trait value, X_i the $K \times 1$ covariant vector including the intercept term, and $1 \times l_m$ vector $G_{m,i}$ records genotypes at marker m for ith individual. Define S_{ij}^A to be the genetic similarity of gene A between subjects i and j ($i \neq j$). There are many ways to describe the genetic similarity between individuals. Here, we consider the weighted IBS sum across the M_A loci in gene A, i.e., for subjects i and j, the similarity level is $S_{ij}^A = \sum_{m=1}^{M_A} w_m s_{m,ij}$ where $s_{m,ij} = x/2$ if $G_{m,i}$ and $G_{m,j}$ have x alleles in common. The similarity level in gene B, S_{ij}^B, can be defined in the same manner. The trait similarity between individual i and j, denoted by Z_{ij}, is computed by

$$Z_{ij} = (Y_i - \mu_i)(Y_j - \mu_j),$$

where $\mu_i = E(Y_i | X_i, G_i) = X_i \gamma$, which is the conditional trait mean given covariate information but assuming no genotype effect and γ is the effect of the covariates. The gene-trait similarity regression for gene-gene interactions is given as

$$E(Z_{ij} | X, G) = \tau_A S_{ij}^A + \tau_B S_{ij}^B + \tau_{AB} S_{ij}^A S_{ij}^B. \tag{1}$$

In Model (1), we use the product of the similarity scores, $S_{ij}^{A}S_{ij}^{B}$, to model the GxG effect. Note that the regression model has zero intercept because the covariates have been incorporated when quantifying trait similarity (Tzeng et al., 2009).

The Interaction Test

The interaction test examines the hypothesis $H_{0,Int}$: $\tau_{AB} = 0$. Direct derivation of a test statistic under Model (1) can be burdensome because Model (1) is a regression whose observation unit is pairs of individuals. Consequently, the observations are correlated when two pairs share a common individual. The variance-covariance matrix for Z is a non-sparse matrix with dimension $\binom{n}{2}$ by $\binom{n}{2}$, which is computationally challenging to inverse so to obtain the test statistics. To bypass these issues, we derive the score test based on the equivalence between the similarity regression and a mixed effect model (Tzeng et al., 2009; Tzeng et al., 2011). To see this, consider the following working mixed model:

$$Y_i = X_i\gamma + g_{A,i} + g_{B,i} + g_{AB,i} + e_i, \tag{2}$$

where $e_i \sim N(0, \sigma)$, and $g_{A,i}$, $g_{B,i}$ and $g_{AB,i}$ are subject-specific genetic effects for gene A, gene B and interaction between genes A and B, respectively. Let $g_A^{\mathrm{T}} = \left[g_{A,1}, \cdots, g_{A,n}\right]$, $g_B^{\mathrm{T}} = \left[g_{B,1}, \cdots, g_{B,n}\right]$, and $g_{AB}^{\mathrm{T}} = \left[g_{AB,1}, \cdots, g_{AB,n}\right]$, and assume that

$g_A \sim MN(0, v_A S_A)$, $g_B \sim MN(0, v_B S_B)$, $g_{AB} \sim MN(0, v_{AB} S_{AB})$,

where matrix $S_A = \left\{S_{ij}^{A}\right\}_{n\times n}$, $S_B = \left\{S_{ij}^{B}\right\}_{n\times n}$, and $S_{AB} = \left\{S_{ij}^{A} \times S_{ij}^{B}\right\}_{n\times n}$. In other words, under the working model (2), the correlation of the Gene-A effect between individuals i and j

is governed by the genetic similarity between the two individuals in gene A (S_{ij}^{A}). Then the marginal trait covariance in Model (2) can be obtained by

$$cov\left(Y_i, Y_j | X, G\right) = cov_{g.}\{E(Y_i|\boldsymbol{X}, G, g_A, g_B, g_{AB}), E(Y_i|\boldsymbol{X}, G, g_A, g_B, g_{AB})\} \qquad (3)$$

$$= cov_{g.}\left\{X_i\gamma + g_{A,i} + g_{B,i} + g_{AB,i}, X_j\gamma + g_{A,j} + g_{B,j} + g_{AB,j}\right\}$$

$$= v_A S_{ij}^{A} + v_B S_{ij}^{B} + v_{AB} S_{ij}^{A} S_{ij}^{B}.$$

Comparing (3) with (1), we have $\tau_A = v_A$, $\tau_B = v_B$, and $\tau_{AB} = v_{AB}$. That is, the regression coefficients in the similarity regression model are variance components in the working linear mixed effect model.

We derive the score function of the REML log-likelihood function of Model (2) in Appendix A, and obtain the score statistic under $H_{0,Int}$ as

$$T_{Int} = \frac{1}{2}\boldsymbol{Y}' P_{Int} S_{AB} P_{Int} \boldsymbol{Y}\Big|_{\tau_A = \hat{\tau}_A, \tau_B = \hat{\tau}_B, \sigma = \hat{\sigma}},$$

where the trait value $\boldsymbol{Y}^{\mathrm{T}} = (Y_1, \ldots, Y_n)$, $\boldsymbol{X} = (X_1; X_2; \cdots; X_n)$ $P_{Int} = V_{Int}^{-1} - V_{Int}^{-1}\boldsymbol{X}(\boldsymbol{X}'^{-1} V_{Int}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}' V_{Int}^{-1}$, $V_{Int} = \tau_A S_A + \tau_B S_B + \sigma I$ and $(\hat{\tau}_A, \hat{\tau}_B, \hat{\sigma})$ are the maximum REML estimates obtained under $H_{0,Int}\colon \tau_{AB} = 0$. We describe an adaptive EM algorithm to obtain $(\hat{\tau}_A, \hat{\tau}_B, \hat{\sigma})$ in Appendix B.

Under the alternative hypothesis $\tau_{AB} \neq 0$, T_{Int} is a strictly increasing function of τ_{AB}. Therefore larger values of T_{Int} provides stronger evidence against $H_{0,Int}$. This suggests that the testing procedure should be one sided. As shown in Appendix A, the distribution of T_{Int} follows a weighted χ^2 distribution. That is, define $C_{Int} = \frac{1}{2} V_{Int}^{1/2} P_{Int} S_{AB} P_{Int} V_{Int}^{1/2}$, and then

T_{Int} has the same distribution as $\sum_{j=1}^{c} \lambda_{j,Int}\chi_1^2$, where $\lambda_{j,Int}$ is the ordered none zero eigenvalues of matrix C_{Int}. The p-values can be calculated using moment-matching approximations (Liu et al., 2009; Duchesne et al., 2010).

The Joint Test

Instead of evaluating the GxG effects between two genes, one may be interested in assessing the overall effect from the two genes, regardless the main effects or interaction effects. To do so, we construct a joint test to examine the null hypothesis of $H_{0,Joint}$: $\tau_A = \tau_B = \tau_{AB} = 0$ under the full model $E(Z_{ij}|X,G) = \tau_A S_{ij}^A + \tau_B S_{ij}^B + \tau_{AB} S_{ij}^A S_{ij}^B$. As shown in Appendix A, the test statistic is given as

$$T_{Joint} = \frac{1}{2}\boldsymbol{Y}' P_{Joint}(S_A + S_B + S_{AB}) P_{Joint}\boldsymbol{Y}\Big|_{\sigma=\breve{\sigma}},$$

where $P_{Joint} = \sigma^{-1}\{I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}$ and $\breve{\sigma} = \frac{\boldsymbol{Y}'\{I-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\boldsymbol{Y}}{(n-K)}$, where K is number of covariates. The distribution of T_{Joint} also follows a weighted χ^2 distribution, i.e., T_{Joint} has the same distribution as $\sum_{j=1}^{c} \lambda_{j,Joint}\chi_1^2$ with $\lambda_{j,Joint}$ the ordered nonzero eigenvalues of $C_{Joint} = \frac{1}{2} V_{Joint}^{1/2} P_{Joint}(S_A + S_B + S_{AB}) P_{Joint} V_{Joint}^{1/2}$.

The Conditional Main Effect Test

We also construct the score test for the main effect of a gene conditioning on the effect of the other gene. We consider the conditional main effect test only under the assumption of no gene-gene interactions. This is because in scenarios where the interaction effects do exist, the main effects are typically not well-defined, and its significance depends on the scale of the interacting variables. Consider the full model: $E(Z_{ij}) = \tau_A S_{A,ij} + \tau_B S_{B,ij}$. The effect of Gene A accounting for the effect of Gene B can be assessed by examining the hypothesis of $H_{0,A}$: $\tau_A = 0$. Similar to previous two tests, the score test statistic for gene A is given as:

$$T_A = \frac{1}{2}\boldsymbol{Y}'P_A S_A P_A \boldsymbol{Y}\Big|_{\tau_B=\tilde{\tau}_B,\sigma=\tilde{\sigma}},$$

where $P_A = V_A^{-1} - V_A^{-1}\boldsymbol{X}(\boldsymbol{X}'^{-1}V_A^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'V_A^{-1}$, $V_A = \tau_B S_B + \sigma I,$ and $(\tilde{\tau}_B, \tilde{\sigma})$ are the maximum REML estimates obtained under $H_{0,A}\colon \tau_A = 0$. We describe the adaptive EM algorithms to obtain $(\tilde{\tau}_B, \tilde{\sigma})$ in Appendix B. As shown in Appendix A, T_A has the same distribution as $\sum_{j=1}^{c} \lambda_{j,A}\chi_1^2$ with $\lambda_{j,A}$ the ordered nonzero eigenvalues of $C_A = \frac{1}{2}V_A^{1/2}P_A S_A P_A V_A^{1/2}$. The test statistic for assessing the conditional main effect of gene B, T_B, can be defined similarly.

Simulation study

Design for Simulation Study

In the simulation, we study the performance of the proposed method and benchmark it against 3 approaches: (1) the linear regression (referred to as LR); (2) the principal component method of Wang et al. (2009) (referred to as PCA), and (3) PLS: the partial least square method of Wang et al. (2009) (referred to as PLS). LR incorporates all SNPs from each of genes and the pairwise interactions between the SNPs from different genes. For the conditional main effect test, only the proposed test and the LR are performed because PCA and PLS are identical to LR when there is no interaction term.

To simulate genotype data with realistic LD patterns, we use genotype data of Gene RBJ (8 SNPs) and Gene GPRC5B (15 SNPs) downloaded from HapMap (http://hapmap.ncbi.nlm.nih.gov/). The LD structure of the two genes are given in Figure 1.

For each SNP in a given gene, we list in Table 2.1 its minor allele frequency (MAF) and LD, which is quantified by the average of R^2 between the target SNP and the remaining SNPs in the same gene. We considered two causal SNPs from each gene and simulate the trait values based on the model:

$$\begin{aligned} Y = \beta_A \times (SNP_1^A + SNP_2^A + SNP_1^A \times SNP_2^A) \\ +\beta_B \times (SNP_1^B + SNP_2^B + SNP_1^B \times SNP_2^B) \\ +\beta_{AB} \times (SNP_1^A \times SNP_1^B + SNP_2^A \times SNP_2^B) + e, \qquad (4) \end{aligned}$$

where SNP_1^A and SNP_2^A are the number of the minor alleles carried by a subject at the first and second causal loci in Gene A; SNP_1^B and SNP_2^B are defined similarity; e follows a normal distributed variable with mean 0 and variance 1. The trait value Y depended on three components: the gene effect from Gene A, the gene effect from Gene B and the interaction effect between Gene A and Gene B. There is no LD between *RBJ* and *GPRC5B* because the genotypes of a person for the two genes were sampled independently.

In Type I error rate evaluation of the joint test, we set $\beta_A = \beta_B = \beta_{AB} = 0$. For interaction test, we consider three situations of no GxG effects: (1) both genes have no genetic effect ($\beta_A = \beta_B = \beta_{AB} = 0$); (2) only gene A has a main effect to the trait ($\beta_A \neq 0, \beta_B = \beta_{AB} =$

0); (3) both Gene A and Gene B have main effect to the trait ($\beta_A \neq 0, \beta_B \neq 0, \beta_{AB} = 0$). For the conditional main effect ($H_0: \tau_B = 0$), we consider two scenarios of no gene B effect: (1) both gene have no genetic effect ($\beta_A = \beta_B = \beta_{AB} = 0$); (2) Gene A has a main effect to the trait ($\beta_A \neq 0, \beta_B = 0, \beta_{AB} = 0$). For each scenario, we ran the simulation for 1,000 replications and each replication consisted of 300 subjects to compute the type I error rates.

For power analysis, we considered three different scenarios based on the genetic architecture of the causal SNPs. Specifically, we selected three SNP pairs with certain LD and MAF as causal SNPs (Table 2.2). When generating trait values using the Model (4), we considered different β values so that power of different methods are between 20%~80%.

Simulation result

Table 2.3 shows the type I error rates for different methods at a significance level of $\alpha = 0.05$. The type I error rates of all approaches are around the nominal level in all different settings except in some scenarios for interaction test. The type I error rates of the proposed interaction tests were conservative when $\beta_A = \beta_B = 0$ (and hence $\tau_A = \tau_B = 0$). As one or both of τ_A and τ_B became further away from 0, e.g., 0.1, the type I error rates get better. As discussed in the Appendix B, the conservativeness of type I error rates was caused by the bias in the EM estimates of the variance components when their true values were0. The type I error for interaction test using PCA and PLS were around the nominal level while the results from LR were slightly inflated.

Because different scenarios have different causal SNPs, we used different β values so that the power of different methods were around 0.2~0.8. The results are shown in Table 2.4. For the interaction test, we see the proposed method was quite robust to the changes of LD and MAF, which either had a compatible power to the best approach or the highest power. LR always had lower power than the proposed method due to the large degrees of freedom. The performance of PCA and PLS were sensitive to the LD and MAF of the causal SNP. Specifically, in Scenario 1 where the casual SNPs had relatively high LD and large MAF, PCA and PLS performed the best, closely followed by the proposed method. In Scenario 2 where the causal SNPs had a relatively smaller LD and MAF, PLS still had the best power, closely followed by the proposed method. The power of PCA dropped a lot in this scenario and LR again had the least power. In Scenario 3, where the causal SNPs had low MAF, the proposed method has the best power, followed by LR, PLS and finally PCA. The possible reason for the low power of PCA and PLS is that the first component can only capture limit amount of information. The first component aimed to capture the max amount of SNP variations so it tends to upweight the SNPs with high LD or common MAF and capture little info about the causal SNPs. The situation was worsen in PCA, which had the least power, because without the supervision from the trait values, the first component put lower weights on the causal SNPs and higher weights on nuisance SNP.

For joint test, we see that proposed method always gain the best power, followed by PLS and PCA. PLS and PCA performed similarly but PLS always had a better power than PCA. The observation is consistent with Wang et al. (2009) since first component from PLS considers both LD and correlation between trait and gene information. All the three methods performed better than LR, as it may lose power due to large degree of freedoms. In joint test, the proposed method had the best performance while in interaction test, PLS/PCA may had better power gain in some scenarios. Possible reason is that the proposed approach needs to estimate nuisance under the H_0 in the interaction test. The corresponding EM estimates can be biased when the true values are 0 or close to 0. This may lead to too conservative results and cause power loss.

For conditional main effect test, proposed method always has a better power than linear model, but the difference between proposed method and linear model gets smaller as the LD and MAF of causal SNP is smaller. The finding is consistent with the finding in the interaction test: LR has a robust performance against LD and MAF.

Real Data Analysis

Warfarin is a widely used oral anticoagulant. In 2004, more than 30 million prescriptions contained this drug in United State (Wysowski et al., 2007). The optimal dose of warfarin is different from patient to patient, and an inappropriate dosage can lead to severe consequence such as bleeding, swelling of face, throat. Extensive researches have been conducted to develop method for predicting the appropriate dose.

We have conducted a genetic analysis using the data from the Warfarin study (The International Warfarin Pharmacogenetics Consortium, 2009). In this data set, 2 genes are involved, gene VKORC1 contains 7 SNPs and gene CYP2C9 is a tri-allelic locus. After quality control, this data set involves 301 individuals and records their stable warfarin dose. Also available are 4 detected covariates associated with warfarin therapy: age, sex, height and weight for each individuals in the study.

We apply the proposed method and the benchmark methods to detect the association between the warfarin dose and the genes. The results are summarized in Table 2.5. All methods identify significant association between trait and the two genes. But the proposed method exhibits the strongest evidence of association among all the approaches, which is consistent with the simulation results that the proposed method is more powerful than other methods. These results also suggest that there are no interactions between the two genes, and Gene VKORC1 has more significant impact on the warfarin dose.

Discussion

The focus of genetic association studies for complex diseases is being shifted from assessing genetic main effects to interaction effects. Mapping GxG would not only improve the power of detecting susceptible genes for the diseases, it would also facilitate the uncovering of the underlying mechanisms for complex diseases. Although the majority of GxG methods focus on SNP-SNP interactions, gene-based GxG methods may hold great promises in biological interpretation and power gain. In this paper, we propose a similarity regression model for studying GxG at gene level. We provide three tests to evaluate the effect of interest: interaction test, joint test and conditional main effect test. The simulation results showed that the proposed method had a consistent satisfactory performance under different genetic architectures. In contrast, PLS, PCA, and LR performed quite differently depending the LD and MAF of the causal SNPs.

In the proposed model, we used an adaptive EM algorithm to estimate the nuisance parameters τ_A and τ_B. Directly applying the traditional EM algorithm without any adjustment may lead to sizable biases and is time-consuming for converge when the true values of τ_A and τ_B were zero. The adaptive EM algorithm reduces the bias and computational time, and improves the type I rates although they are still conservative when true τ_A and τ_B equal to 0. How to improve the variance estimation when the true value is at the zero boundary would be an interesting topic for further research.

Our approach is proposed to find the GxG at the gene level. However, if a gene contains a large number of SNPs, direct summarizing all SNPs information into gene level would lower down the power since too many null SNPs are involved. A possible solution for that would be applying a LD analysis first, then splitting the large gene into several smaller blocks to do the GxG interaction detection.

Appendix A. Derivation of the score tests and their distributions

Consider the matrix presentation of model (2):

$$\mathbf{Y} = \mathbf{X}\gamma + g_A + g_B + g_{AB} + e.$$

The corresponding REML log-likelihood function, denoted as $L(\tau_A, \tau_B, \tau_{AB}, \sigma)$, is:

$$L(\theta) = -\frac{1}{2}[\log|V| + \log|\mathbf{X}'^{-1}V\mathbf{X}| + \mathbf{Y}'P\mathbf{Y}],$$

where $V = Var(Y) = \tau_A S_A + \tau_B S_B + \tau_{AB} S_{AB} + \sigma I$ is the marginal variance of $\mathbf{Y}$ and $P = V^{-1} - V^{-1}\mathbf{X}(\mathbf{X}'^{-1}V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}$ is the projection matrix for the model. The score functions of τ_A, τ_B, and τ_{AB} based on $L(\theta)$ are

$$U_{\tau_A}(\tau_A, \tau_B, \tau_{AB}, \sigma) = \frac{\partial L(\tau_A, \tau_B, \tau_{AB}, \sigma)}{\partial \tau_A} = \frac{1}{2}[\mathbf{Y}'PS_AP\mathbf{Y} - tr(PS_A)],$$

$$U_{\tau_B}(\tau_A, \tau_B, \tau_{AB}, \sigma) = \frac{\partial L(\tau_A, \tau_B, \tau_{AB}, \sigma)}{\partial \tau_B} = \frac{1}{2}[\mathbf{Y}'PS_BP\mathbf{Y} - tr(PS_B)], \text{and}$$

$$U_{\tau_{AB}}(\tau_A, \tau_B, \tau_{AB}, \sigma) = \frac{\partial L(\tau_A, \tau_B, \tau_{AB}, \sigma)}{\partial \tau_{AB}} = \frac{1}{2}[\mathbf{Y}'PS_{AB}P\mathbf{Y} - tr(PS_{AB})].$$

We construct the statistics based on the first terms of the score functions. Here after we define matrices P_h and V_h as P and V evaluated under $H_{0,h}$. Then for the interaction test $H_{0,Int}: \tau_{AB} = 0$, we set the test statistic as

$$T_{Int} = \frac{1}{2}\mathbf{Y}'P_{Int}S_{AB}P_{Int}\mathbf{Y}\Big|_{\tau_A=\hat{\tau}_A, \tau_B=\hat{\tau}_B, \sigma=\hat{\sigma}},$$

where the $P_{Int} = V_{Int}^{-1} - V_{Int}^{-1}\mathbf{X}(\mathbf{X}'^{-1}V_{Int}^{-1}\mathbf{X})^{-1}\mathbf{X}'V_{Int}^{-1}$, $V_{Int} = \tau_A S_A + \tau_B S_B + \sigma I$, and $(\hat{\tau}_A, \hat{\tau}_B, \hat{\sigma})$ are the maximum REML estimates obtained under $H_{0,Int}: \tau_{AB} = 0$.

For the conditional main effect test $H_{0,A}\colon \tau_A = 0$ under the constrain of no interaction (i.e., $\tau_{AB} = 0$), we set the test statistic as

$$T_A = \frac{1}{2}\mathbf{Y}' P_A S_A P_A \mathbf{Y}\Big|_{\tau_B=\tilde{\tau}_B, \sigma=\tilde{\sigma}},$$

where $P_A = V_A^{-1} - V_A^{-1}\mathbf{X}(\mathbf{X}'^{-1}V_A^{-1}\mathbf{X})^{-1}\mathbf{X}'V_A^{-1}$, $V_A = \tau_B S_B + \sigma I$, and $(\tilde{\tau}_B, \tilde{\sigma})$ are the maximum REML estimates obtained under $H_{0,A}\colon \tau_A = 0$. The test statistic T_B can be defined similarily for examining $H_{0,B}\colon \tau_B = 0$ under the constrain of $\tau_{AB} = 0$.

We describe the EM algorithms that we use to obtain $(\hat{\tau}_A, \hat{\tau}_B, \hat{\sigma})$ and $(\tilde{\tau}_B, \tilde{\sigma})$ in Appendix B.

For the joint test $H_{0,Joint}\colon \tau_A = \tau_B = \tau_{AB} = 0$, because τ's are non-negative variance components, $\tau_A = \tau_B = \tau_{AB} = 0$ if and only if $\tau_A + \tau_B + \tau_{AB} = 0$. This motivates us to construct the test statistic based on the sum of the three score functions. That is,

$$T_{Joint} = \frac{1}{2}\mathbf{Y}' P_{Joint}(S_A + S_B + S_{AB}) P_{Joint}\mathbf{Y}\Big|_{\sigma=\breve{\sigma}},$$

where the $P_{Joint} = V_{Joint}^{-1} - V_{Joint}^{-1}\mathbf{X}(\mathbf{X}'^{-1}V_{Joint}^{-1}\mathbf{X})^{-1}\mathbf{X}'V_{Joint}^{-1}$, $V_{Joint} = \sigma I$ and $\breve{\sigma} = \mathbf{Y}'\{I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{Y}/(n - K)$.

The distributions of the test statistics can be shown to follow a weighted chi-squared distribution via the fact that these statistics are quadratic form of $\mathbf{Y}$. To illustrate, consider $T_{Int} = \frac{1}{2}\mathbf{Y}' P_{Int} S_{AB} P_{Int}\mathbf{Y}$. Because P_{Int} is a projection matrix, $P_{Int}\mathbf{X}\gamma = 0$. Therefore,

$$T_{Int} = \frac{1}{2}\mathbf{Y}' P_{Int} S_{AB} P_{Int}\mathbf{Y} = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\gamma)' P_{Int} S_{AB} P_{Int}(\mathbf{Y} - \mathbf{X}\gamma) = \frac{1}{2}(\mathbf{Y} -$$

$$\mathbf{X}\gamma)' V_{Int}^{-1/2} V_{Int}^{1/2} P_{Int} S_{AB} P_{Int} V_{Int}^{1/2} V_{Int}^{-1/2}(\mathbf{Y} - \mathbf{X}\gamma) \sim \sum_{i=1}^{c} \lambda_i \chi_1^2$$

Define $\mathbf{z} = V_{Int}^{-1/2}(\mathbf{Y} - \mathbf{X}\gamma)$ and $C_{Int} = \frac{1}{2}V_{Int}^{1/2}P_{Int}S_{AB}P_{Int}V_{Int}^{1/2}$, we have $T_{Int} = \mathbf{z}'C_{Int}\mathbf{z} \sim \sum_{j=1}^{c} \lambda_{j,Int}\chi_1^2$, where $\lambda_{j,Int}$ is the ordered none zero eigenvalues of matrix C_{Int}. By the same mannar, one can obtain that $T_A \sim \sum_{j=1}^{c} \lambda_{j,A}\chi_1^2$ with $\lambda_{j,A}$ the ordered nonzero eigenvalues of $C_A = \frac{1}{2}V_A^{1/2}P_AS_AP_AV_A^{1/2}$, and $T_{Joint} \sim \sum_{j=1}^{c} \lambda_{j,Joint}\chi_1^2$ with $\lambda_{j,Joint}$ the ordered nonzero eigenvalues of $C_{Joint} = \frac{1}{2}V_{Joint}^{1/2}P_{Joint}(S_A + S_B + S_{AB})P_{Joint}V_{Joint}^{1/2}$.

Appendix B. The adaptive EM algorithm to obtain the maximum REML estimates

In interaction test and conditional one gene test, we use EM algorithm to estimate the nuisance parameter under the corresponding H_0. The regular EM algorithm has caused too conservative results since it provides non-negative estimates for the nuisance variance components τ_A, τ_B and σ. Thus, when τ_A and/or τ_B are 0 or close to 0, the EME can be biased as $E(\hat{\tau_A}) > \tau_A$ and $E(\widehat{\tau_B}) > \tau_B$. This motived us to use and adaptive procedure to address this problem. We first apply conditional main effect tests for $H_0: \tau_A = 0$ and $H_0: \tau_B = 0$. If we fail to reject the null hypothesis, the corresponding $\hat{\tau}$ is set to 0. If τ is significant different from 0, EM algorithm is used to estimate τ. By applying this adaptive EM, $E(\hat{\tau})$ would be closer to 0 when the τ is 0 or close to 0. When τ are relative large, the conditional main effect test would reject $H_0: \tau = 0$ and the estimate from t he adaptive EM is the same as the original EME.

For interaction test, when we do the type I error analysis, we found that the statistics T_{Int} does not fit the estimated weighted χ^2 distribution perfectly if the true value of τ_A and τ_B are 0 or close to 0. When the true value for τ_A and τ_B are relative large, the T_{Int} follows the estimated weighted χ^2 well. The main reason for this scenoraio is that the estimated $\widehat{\tau_A}$ and $\widehat{\tau_B}$ are always biased to the true value since EM algorithm could only give us positive estimates for $\hat{\tau}_A$ and $\hat{\tau}_B$. Thus, it is always true that $E(\hat{\tau}_A) > \tau_A$ and $E(\hat{\tau}_B) > \tau_B$, and the difference between $E(\hat{\tau})$ and τ would be huge when τ_A are 0 or close to 0.

We first describe the EM algorithm for $(\hat{\tau}_A, \hat{\tau}_B, \hat{\sigma})$, i.e., the maximum REML estimates under $H_{0,Int}: \tau_{AB} = 0$. Under $H_{0,Int}$, the LMM is

$$\mathbf{Y} = \mathbf{X}\gamma + g_A + g_B + e,$$

where $e \sim N(0, \sigma I)$, $g_A \sim N(0, \tau_A S_A)$ and $g_B \sim N(0, \tau_B S_B)$ as $\tau_A = \nu_A$ and $\tau_B = \nu_B$. Define $U = A^T \mathbf{Y}$ with the restriction that $A^T A = I_{n-K}$ and $AA^T = I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$. Then $U|g_A, g_B \quad \sim N(A'g_A + A'g_B, \sigma I_{n-K})$, which is independent of the fixed effect $\hat{\gamma} = (X^T X)^{-1} X^T Y$. Therefore the maximum REML estimates can be obtained by maximizing the marginal distribution of U, i.e., $f(U) = \int f(U|g_A, g_B) f(g_A) f(g_B) dg_A dg_B$. This motivated an expectation-maximization algorithm based on U (i.e., the observed data) and (g_A, g_B) (i.e., the missing data). The complete-data log likelihood is based on $f(U, g_A, g_B)$ is

$$\log f(U, g_A, g_B; \tau_A, \tau_B, \sigma)$$

$$= \log f(U|g_A, g_B; \tau_A, \tau_B, \sigma) + \log f\left(g_{A;\tau_A,\tau_B,\sigma}\right) + \log f(g_B; \tau_A, \tau_B, \sigma)$$

$$\begin{aligned} = -\frac{n-K}{2}\log\sigma - \frac{1}{2\sigma}(U - A'g_A - A'g_B)'(U - A'g_A - A'g_B) \\ -\frac{q_A}{2}\log\tau_A - \frac{1}{2}\log(|S_A|_+) - \frac{1}{2\tau_A} g_A^T S_A^- g_A \\ -\frac{q_B}{2}\log\tau_B - \frac{1}{2}\log(|S_B|_+) - \frac{1}{2\tau_B} g_B^T S_B^- g_B \end{aligned}$$

where q_A and q_B are the rank for matrix S_A and S_B respectively, $|S_A|_+$ is the pseudo-determinant, and S_A^- is the generalized inverse (as S_A and S_B may be singular).

In the expectation step, we compute

$$\begin{aligned}
&Q\left(\tau_A, \tau_B, \sigma; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}\right) \equiv E\left\{\log f(U, g_A, g_B; \tau_A, \tau_B, \sigma) | U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}\right\} \\
&\quad = -\frac{n-K}{2}\log\sigma - \frac{1}{2\sigma} E\left\{(U - A'g_A - A'g_B)'(U - A'g_A - A'g_B) | U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}\right\} \\
&\quad -\frac{q_A}{2}\log\tau_A - \frac{1}{2}\log(|S_A|_+) - \frac{1}{2\tau_A} E\left(g_A^T S_A^- g_A | U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}\right) \\
&\quad -\frac{q_B}{2}\log\tau_B - \frac{1}{2}\log(|S_B|_+) - \frac{1}{2\tau_B} E\left(g_B^T S_B^- g_B | U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}\right).
\end{aligned}$$

In the maximization step, we solve for $\partial Q / \partial \tau_A = 0$, $\partial Q / \partial \tau_B = 0$ and $\partial Q / \partial \sigma = 0$, and obtain

$$\hat{\tau}_A = \frac{1}{q_A} E\left(g_A^T S_A^- g_A | U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}\right) = \frac{1}{q_A}\left\{\tilde{\mathbf{g}}_A^{(t)\prime} S_A^- \tilde{\mathbf{g}}_A^{(t)} + tr(S_A^- \tilde{\mathbf{v}}_A^{(t)})\right\},$$

where $\tilde{\mathbf{g}}_A^{(t)} \equiv E(g_A | g_B, U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}) = \tau_A S_A P_{Int} Y |_{\hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}}$, and $\tilde{\mathbf{v}}_A^{(t)} \equiv Var(g_A | g_B, U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}) = \tau_A S_A - \tau_A^2 S_A P_{Int} S_A |_{\hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}}$. Similary, we have

$$\hat{\tau}_B = \frac{1}{q_B} E\left(g_B^T S_B^- g_B | U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}\right) = \frac{1}{q_B}\left\{\tilde{\mathbf{g}}_B^{(t)\prime} S_B^- \tilde{\mathbf{g}}_B^{(t)} + tr(S_B^- \tilde{\mathbf{v}}_B^{(t)})\right\}$$

Finally,

$$\begin{aligned}
&\hat{\sigma}^{(t)} = \frac{1}{n-K} E\left\{(U - A'g_A - A'g_B)'(U - A'g_A - A'g_B) | U; \hat{\tau}_A^{(t)}, \hat{\tau}_B^{(t)}, \hat{\sigma}^{(t)}\right\} \\
&= Y^{*T} AA'Y^* + tr[AA'(\hat{\tau}_A^{(t)} S_A - (\hat{\tau}_A^{(t)})^2 S_A P S_A + \hat{\tau}_B^{(t)} S_B - (\hat{\tau}_B^{(t)})^2 S_B P S_B \\
&\quad -2\hat{\tau}_A^{(t)} \hat{\tau}_B^{(t)} S_A P S_B)]
\end{aligned}$$

The EM algorithm for obtaining $(\tilde{\tau}_B, \tilde{\sigma})$ under $H_{0,A}: \tau_A = 0$ is similar to the above algorithm except that τ_A is set to be 0.

References

Barkoulas M, van Zon JS, Milloz J, van Oudenaarden A, F élix MA. Robustness and epistasis in the C. elegans vulval signaling network revealed by pathway dosage modulation. Dev Cell. 2013 Jan 14;24(1):64-75.

Bussey, H., Wittkowsky, A., hylek, E., and Walker, M. Genetic testing for warfarin dosing? *Pharmacotherapy*, 28:141-3, 2008.

Carlborg, O., & Haley, C. S. Epistasis: Too often neglected in complex trait studies? Nature Reviews Genetics 5, 618–62, 2004.

The International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *The new England journal of medicine*, 360:753-764, 2009.

Duchesne, P. and de Micheaux, P. L. Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54:858-862, 2010.

Evans, D. M., Marchini, J., Morris, A. P., and Cardon, L. R. (2006). Two-Stage Two-Locus Models in Genome-Wide Association. PLoS Genetics, 2(9), e157.

Howson JM, Cooper JD, Smyth DJ, Walker NM, Stevens H, She JX, Eisenbarth GS, Rewers M, Todd JA, Akolkar B, Concannon P, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Pociot F, Rich SS; Type 1 Diabetes Genetics Consortium. 2012. Evidence of gene-gene interaction and age-at-diagnosis effects in type 1 diabetes. Diabetes. 11: 3012-3017.

Jorgenson E, Witte JS.A gene-centric approach to genome-wide association studies. Nat Rev Genet 7(11):885–891, 2006.

Koh-Tan HH, McBride MW, McClure JD, Beattie E, Young B, Dominiczak AF and Graham D. 2013. Interaction between chromosome 2 and 3 regulates pulse pressure in the stroke-prone spontaneously hypertensive rat. Hypertension. 62, 33-40

Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. Genet. Epidemiol. 28: 157–170, 2005.

Marchini, J., Donnelly, P., and Cardon, L. R. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet, 37(4), 413–417.

Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity 56, 73–82, 2003.

Lin X, Hamilton-Williams EE, Rainbow DB, Hunter KM, Dai YD, Cheung J, Peterson LB, Wicker LS and Sherman LA. 2013. Genetic interactions among Idd3, Idd5.1, Idd5.2, and Idd5.3 protective loci in the nonobese diabetic mouse model of type 1 diabetes. J Immunol. 7, 3109-3120.

Liu, H., Tang, Y., and Zhang, H. H. A new chi-square approximation to the distribution of nonnegative definite quadratic forms in non-central normal variables. *Computational Statistics; Data Analysis*, 53(4):853 -856, 2009.

Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. Am J Hum Genet 75(3):353–362. 2004.

Phenix H, Perkins T, Kærn M. Identifiability and inference of pathway motifs by epistasis analysis. Chaos. 2013 Jun;23(2).

Pillai R, Waghulde H, Nie Y, Gopalakrishnan K, Kumarasamy S, Farms P, Garrett MR, Atanur SS, Maratou K, Aitman TJ and Joe B. 2013. Isolation and high-throughput sequencing of two-closely linked epistatic hypertension susceptibility loci with a panel of bicongenic strains. Physiol Genomics. June 11.

Sheriff A, Ott J. Applications of neural networks for gene finding. 2001. Adv. Genet. 42287–297.

Tzeng, J.-Y., Zhang, D., Chang, S.-M., Thomas, D. C., and Davidian, M. Gene-trati similarity regression for multimarker-based association analysis. *Biometrics*, 65:822-832, 2009.

Wang, T., Ho, G., Ye, K., Strickler, H., and Elston, R. C. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Genet. Epidemiol.*, 33(1):6-15, 2009.

Wysowski, D., Nourjah, P., and Swartz, L. Bleeding complications with warfarin use: a prevalent adverse effect resulting in regulatory action. *Arch Intern Med*, 167:1414-9, 2007.

Yee J, Kwon MS, Park T and Park M. A modified entropy-based approach for identifying gene-gene interactions in case-control study. PLoS One. Jul 18;8(7). 2013.

Zhang H, Bonney G. Use of classification trees for association studies. Genet. Epidemiol. 19: 323–332. 2000

Ziyab AH, Davies GA, Ewart S, Hopkin JM, Schauberger EM, Wills-Karp M, Holloway JW, Arshad SH, Zhang H and Karmaus W. 2013. Interactive effect of STAT6 and IL13 gene polymorphisms on eczema status: results from a longitudinal and a cross-sectional study. BMC Med Genet. Jul.

Table 2 1. LD and MAF information for Gene RBJ and GPRC5B.

Gene RBG			Gene GPRC5B		
SNP	LD[a]	MAF	SNP	LD	MAF
1	0.588	0.115	1	0.186	0.146
2	0.159	0.062	2	0.05	0.066
3	0.126	0.482	3	0.065	0.044
4	0.588	0.115	4	0.194	0.195
5	0.588	0.115	5	0.197	0.143
6	0.565	0.119	6	0.136	0.159
7	0.159	0.062	7	0.259	0.46
8	0.588	0.115	8	0.262	0.336
			9	0.23	0.482
			10	0.206	0.371
			11	0.285	0.394
			12	0.285	0.394
			13	0.138	0.155
			14	0	0.004
			15	0	0.005

[a] The LD is the average value between the corresponding SNP and the rest SNPs in the same gene.

Table 2.2 Causal SNPs used in power analysis under different scenarios

Scenario	SNP_1^A		SNP_2^A		SNP_1^B		SNP_2^B	
	LD	MAF	LD	MAF	LD	MAF	LD	MAF
1	0.58	0.12	0.13	0.48	0.26	0.46	0.23	0.48
2	0.57	0.12	0.16	0.06	0.14	0.16	0.23	0.48
3	0.58	0.12	0.16	0.06	0.07	0.04	0.14	0.16

Table 2.3 Type I error rate for different tests at significance level 0.05 for 3 tests. interaction test, joint test and Conditional main effect test

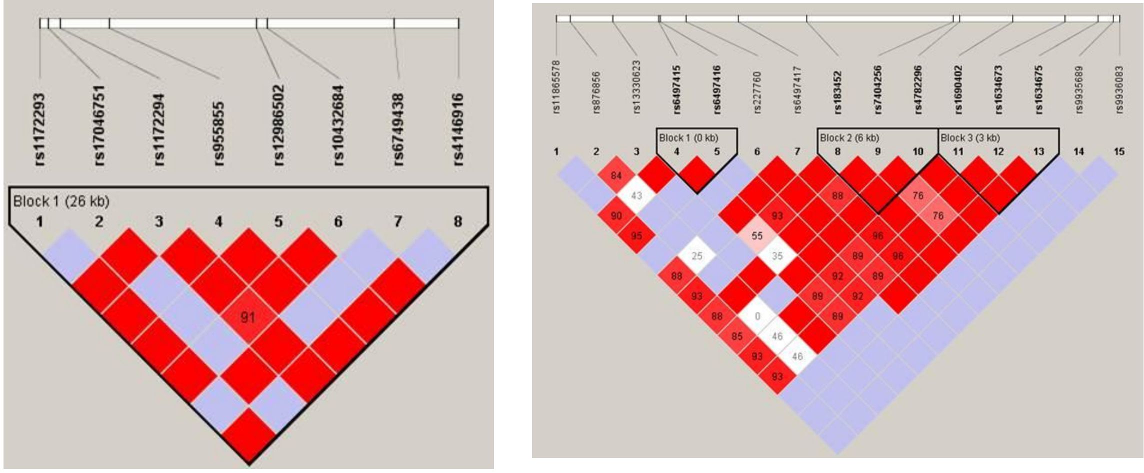
Tests	$(\beta_A, \beta_B, \beta_{AB})$					
	Interaction Test			Joint Test	Conditional main effect test (B|A)	
method	(0,0,0)	(0.1,0,0)	(0.1,0.1,0)	(0,0,0)	(0,0,0)	(0.3,0,0)
Proposed	0.038	0.04	0.046	0.049	0.048	0.052
LR	0.06	0.062	0.064	0.06	0.06	0.056
PCA	0.056	0.06	0.058	0.055		
PLS	0.046	0.048	0.049	0.054		

Table 2.4 Power analysis for different tests at significance level 0.05 For each test, 3 scenarios are used to test the proposed method and other method. In each scenario, β_A, β_B and β_{AB} are tuned so that the power are from 20%-80%

Test	Interaction test			Joint test			Conditional one gene test		
Scenario	S1	S2	S3	S1	S2	S3	S1	S2	S3
$(\beta_A, \beta_B, \beta_{AB})$	(0.1,0.1,0.3)	(0.1,0.1,0.5)	(0.1,0.1,1.5)	(0.05,0.05,0.05)	(0.1,0.1,0.1)	(0.3,0.3,0.3)	(0.07,0.07,0)	(0.1,0.1,0)	(0.3,0.3,0)
Proposed	0.690	0.460	0.815	0.670	0.875	0.580	0.65	0.685	0.495
LR	0.375	0.285	0.760	0.325	0.690	0.325	0.415	0.475	0.485
PCA	0.715	0.350	0.120	0.535	0.765	0.500			
PLS	0.745	0.490	0.470	0.545	0.785	0.505			

Table 2.5 *p*-values of four approaches in analysis Warfarin data

Methods	Joint test	Interaction test	Conditional main effect test	
			Gene 1	Gene 2
Proposed	5.6×10^{-20}	0.45	1.67×10^{-20}	1.02×10^{-7}
LR	1.92×10^{-16}	0.753	1.94×10^{-16}	2.65×10^{-8}
PCA	8.07×10^{-5}	0.53	-	-
PLS	7.43×10^{-5}	0.47	-	-



Gene RBJ

Gene GPRC5B

Figure 2.1 LD pattern of the two genes : Gene RBJ and Gene GPRC5B

Chapter 3

Pathway-guided Identification of Gene-Gene Interaction

Xin Wang[1,2], Daowen Zhang[2], Jung-Ying Tzeng[1,2,3*]

[1] Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA
[2] Department of Statistics, North Carolina State University, Raleigh, NC, USA
[3] Department of Statistics, National Cheng-Kung University, Tainan, Taiwan

*Correspondence: Jung-Ying Tzeng, Department of Statistics & Bioinformatics Research Center, Campus Box 7566, North Carolina State University, Raleigh, NC 27695-7566, USA

Running Title: Pathway-guided GxG Interactions

Abstract

Assessing gene-gene interactions (GxG) at gene level can examine epistasis at biologically functional units with amplified interaction signals from marker-marker pairs. Current gene-based GxG methods tend to be designed for studying interactions among two or a few genes. For complex traits, it is often to have a list of many candidate genes to explore GxG. In this work, we propose a pathway-guided approach based on penalized regression for detecting interactions among genes. Specifically, we apply the principal component (PC) analysis to summarize the multi-SNP genotypes and SNP-SNP interaction between a gene pair, and identify important main and interaction effects using L1 penalty, which incorporates adaptive weights based on biological guidance and trait supervision. Our approach aims to combine the advantages of biological guidance and data adaptiveness, and yields credible findings that have both biological and statistical supports and may have higher chances to shed insights in forming follow-up biological hypotheses for further cellular and molecular studies. The proposed approach can be used to explore the gene-gene interactions with a list of many candidate genes and is applicable even when sample size is smaller than the number of predictors studied. We evaluate the utility of the pathway-guided penalized GxG regression using simulation and real data analysis. The numerical studies suggest improved performance over the methods without using the pathway and trait guidance.

Introduction

Current focus of genetic association studies for complex diseases has been shifted from assessing the genetic main effect to interaction effect among genes [1]. Complex diseases, such as hypertension, cancer, diabetes, and psychiatric disorders are believed to have a polygenic basis and gene-gene interaction (GxG) may play significant roles in the disease etiology [2-6]. Understanding GxG may also help to uncover the missing heritability [7, 8] and to explain the inconsistent findings from main-effect analyses [9].

GxG can be defined from biological view and statistical view. Biologically, GxG refers to the physical interactions between biomolecules such as DNA, RNA or protein at the cellular level [10]. On the other hand, statistical GxG refers to deviation from additive main effects of genes. Although there were debates about the relationship between biological interaction and statistical interactions, evidences showed that statistical interactions and biological interactions can converge to the same scientific process [11]. For example, Bridge used statistical model to identify genes with interaction effects on *Drosophila* eye color [12], and the corresponding biological mechanism that depicts how these genes influence biological pathways was understood many years later [13].

Many methods have been proposed to detect GxG, such as logic regression [14], classification/regression tress (CART), multivariate adaptive regression splines (MARS) [15], multifactor dimensionality reduction (MDR) [16]. These methods have shown promising performances in detecting the interaction effects important to complex diseases or traits. [17]. However, most of these methods considered interactions among SNPs instead of interactions among genes. There are several advantages to assess GxG at gene level instead of SNP level. First，genes are the basic units in the biological mechanism and SNPs within a gene tend to work concordantly. Hence the gene-level results can be more biologically insightful, easier to interpret, more informative in revealing underlying mechanisms. Second, modeling the multi-SNPs information also incorporates the linkage disequilibrium (LD) among SNPs in the downstreaming analysis. Third, the polygenic nature of the complex diseases suggests moderate effect sizes for individual variants. Aggregating SNP effects at gene level can amplify the signals and make them more detectable. Finally, via appropriate dimension reduction to summarize the multi-SNP information, gene-level GxG methods are able to use less degrees of freedom, which further help to gain power improvement over the SNP-level analyses. Because of these reasons, several gene-level methods for GxG have been proposed, such as Turkey 1-df method [18], principle component analysis (PCA) and partial least square (PLS) based model [19], kernel-based regressions [20], nonparametric test based method [21]. These studies suggested that gene-level methods have higher power in detecting GxG than traditional SNP-SNP strategies, especially when the causal SNPs are not directly genotyped.

Most of the methods available for studying GxG interactions are for two or a few genes. However, for complex traits, it is often to have a list of many candidate genes to explore GxG. Even with a moderate size gene set, there can be a huge number of GxG terms even at gene level, e.g., a set of 10 genes would lead to 45 pairwise GxG interaction terms. Directly modeling all GxG would be inefficient due to computational challenge and lack of power. The arising solution is to reduce the search space of GxG by filtering out potentially unimportant genes [17]. In current practice, the GxG search space is reduced either in a trait-supervised fashion or using prior biological information.

To reduce the GxG search space supervised by the trait information, one would first apply main-effect association tests on each gene/SNP to remove unimportant ones and then model interactions among the remaining ones [22]. Two interaction mechanisms for Amyotrophic Lateral Sclerosis (ALS) have been identified by this method [23]. However, filtering out genes/SNPs through main-effect screening would have low power if the casual genes only have strong interaction effects but no main effects. To improve, several non-parametric methods are proposed to perform more effective filtering, e.g., the Relief [24] and Tuned Relief (TuRF) [25], which use the nearest neighbors method to find the important genes. The nearest neighbor of an individual is the one who has the highest genetic similarity at the focused genes/SNPs with the target individual. If the gene is important to the trait, the nearest neighbor pair tends to have similar traits. Reliefs sums up all the weighted trait differences to test whether one gene is important to the trait. These methods can successfully reduce the search space by eliminating unimportant genes/SNPs and retaining important ones that may be missed by main-effect screening [1].

Another way to reduce GxG search space is to use biological knowledge as a filter or prior knowledge [17], such as Biofilter [11]. Biofilter builds the list of important genes based on database such as KEGG, Protein interaction database (PID), Biocarter etc. Its underlying rationale is that if the interactions among a group of genes are supported by more biological evidence, the corresponding statistical finding for GxG is more credible. Biofilter uses an implication index, which is the number of databases supporting a focused GxG, to quantify the strength of biological support. If no database provides support to the focused GxG, it would be removed from the search space. Recent studies [26-28] showed that Biofilter can effectively reduce the GxG search space and result in biologically meaningful GxG findings.

Statistical analyses coupled with biological guidance can lead to credible findings that have both biological and statistical supports and may have higher chances to shed insight in forming follow-up biological hypotheses for further cellular and molecular studies. However, directly filtering out genes without incorporating trait information can be too arbitrary, especially when the prior knowledge is not trait-specific. In this paper, we propose a penalized method that incorporates biological guidance and trait supervision to detect GxG at gene level. Specifically, we apply the principal component (PC) analysis to summarize the multi-SNP genotypes and SNP-SNP interaction between a gene pair, and identify important main and interaction effects using L1 penalty, which incorporates adaptive weights based on association strength and trait-specific pathway supports. We demonstrate the utility of the pathway-guided penalized regression for GxG identification using simulation and real data analysis.

Method

For individual i, let Y_i be the trait value and $G_{m,i}$ be the multi-marker genotype vector of the l_m markers in gene m. Given the genotypes of genes s and t, $s \neq t$, the interaction design vector between the two genes is denoted by $H_{st,i} = G_{s,i} \otimes G_{t,i}$, where $\otimes$ is the Kronecker product. Also define genotype design matrix $G_m = \left[G_{m,1}, \cdots, G_{m,N}\right]^T$, interaction design vector $H_{st} = \left[H_{st,1}, \cdots, H_{st,N}\right]^T$ and trait vector $Y = [Y_1, \cdots, Y_N]^T$, where N is the sample size. Finally assume that there are M genes, and the total number of GxG among these genes is $q = M(M-1)/2$.

Obtaining gene-level genetic information

We first summarize the multi-SNP information at gene level for main-effect design matrix and interaction-effect design matrix by PC analysis. To fix the idea, we only use the first PC, but the model can be straightforwardly extended to multiple PCs. We use $X_{1,m}$ to denote the first PC of genotype design matrix G_m and use $X_{2,st}$ to denote the PC of the interaction design matrix H_{st}. Note that one can summarize the information of gene-gene interaction using $X_{1,s} \cdot X_{1,t}$. Doing so can bypass the need to compute and decompose the large matrix H_{st}. However, we found that $\mathrm{X}_{1,\mathrm{s}} \cdot \mathrm{X}_{1,\mathrm{t}}$ may not be able to capture much of the variability of H_{st} because $X_{1,s}$ and $X_{1,t}$ are obtained by maximizing the information captured in the main effect G_s and G_t, respectively. Alternatively, Wang et al. [2009] applied PLS to summarize the gene information at gene level, which aims to maximize both the SNP-SNP correlation and the SNP-trait correlation. Performance of GxG tests using leading components from PLS was shown to be superior than that using PCs from PCA (Wang et al., 2009). However, because PLS components were formed by maximizing their correlations with trait values, the corresponding GxG terms tend to stay significant even with no true interaction effects.

Variable selection guided by biological supports

We use the following model to assess the main and interaction effects of genes:

$$g(\mu) = \sum_{m=1}^{M} X_{1,m}\gamma_m + \sum_{\ell=1}^{q} X_{2,\ell}\beta_\ell = X_1\gamma + X_2\beta,$$

where $g(\cdot)$ is the link function, $\mu = E(Y|X)$ is the conditional mean trait value given covariates X_1 and X_2 with $X_1 = [X_{1,1}, \cdots, X_{1,M}]$, which is the PCs of M genes, and $X_2 = [X_{2,1}, \cdots, X_{2,q}]$, which is the PCs of $q = M(M-1)/2$ GxG terms. Parameter γ is the main effect vector with $\gamma = [\gamma_1, \cdots, \gamma_M]^T$ and β is the interaction effect vector with $\beta =$

$\left[\beta_1, \cdots, \beta_q\right]^T$. For quantitative traits, we set $g(\mu) = \mu$, i.e., the identify link function. For binary traits, we set $g(\mu) = \log(\mu/(1-\mu))$, i.e., the logit link function.

To detect important terms, we estimate γ and β by minimizing the following penalized log-likelihood

$$-\log L(\gamma, \beta; Y, X_1, X_2) + \lambda_1 \sum_{m=1}^{M} \omega_{1,m} |\gamma_m| + \lambda_2 \sum_{\ell=1}^{q} \omega_{2,\ell} |\beta_\ell|, \tag{1}$$

where $L(\gamma, \beta; Y, X_1, X_2)$ is the likelihood function of γ and β; $\omega_{1,m}$'s and $\omega_{2,\ell}$'s are the weights for main effects and interaction effects, respectively; λ_1 and λ_2 are the tuning parameters of main effects and interaction effects, respectively. The weights (either the weight for main effect ω_1 or interaction effect ω_2) are constructed based on 3 components: weights based on gene size (denoted by ω_{size}), weights based on pathway supports (denoted by ω_{path}) and weights based on effect size on the trait (denoted by ω_{effect}). That is, the overall weight is $\omega_m = \omega_{size,m} \cdot \omega_{path,m} \cdot \omega_{effect,m}$.

Weights for gene size $\boldsymbol{\omega_{size}}$. In gene-set association analysis, it has been noted that larger genes (i.e., gene with more SNPs) are more likely to be chosen as significant (Wang et al., 2010). Although here we summarize the gene information using the first PC, our results indicated the tendency of selecting large *M* genes if no penalty is imposed on large *M* genes (e.g., higher false positive rates (FPR) for larger genes when $\omega_{size} = 1$ in the Figure 3. 1). On the other hand, incorporating gene size in the penalty weights can make false positives (FPs) less concentrated in the category of pairs of large *M* genes. We note that while conventionally, gene size refers to the number of SNPs in a gene, in our work, gene size refers to the number of columns in the corresponding design matrix, e.g., G_m or H_ℓ. Specifically, we set $\omega_{size,m} = 1 + (s_m - \min\{s_i\})/(\max\{s_i\} - \min\{s_i\})$, where for main effect, s_m is the number of columns of G_m and for interaction effect, s_m is the number of columns of H_m. To coordinate ω_{size} with other weights (i.e., ω_{path} and ω_{effect}) and to avoid ω_{size} to dominate other weights, we consider the rescaled $s_m - \min\{s_i\}$ and divide it by $\max\{s_i\} - \min\{s_i\}$ so that ω_{size} is between 1 (no size weight) and 2 (maximize size weight). In other words, the maximum penalty from gene size is bounded at 2 times of the minimum penalty.

Weights for pathway supports $\boldsymbol{\omega_p}$. We use weight ω_{path} to incorporate the strength of pathway support. We focus only on biological evidence relevant to the trait of interest (e.g., via PubMed search) and quantify the support strength by the number of pathways that support the interaction among certain gene pairs. Define N_{path} as the total number of pathways related to the trait and n_ℓ is the number of sources supporting the ℓth gene-gene pair. We set $\omega_{path,\ell} = 1 - n_\ell/(2N_{path})$ so that a gene pair with greater pathway support receives less penalty. Because our focus is on GxG effects, we set $\omega_{path,m} = 1$ for main effect terms. The value of ω_{path} is between 0.5 and 1 to avoid the dominance of one weight over others.

The adaptive weight for effect size $\boldsymbol{\omega_{effect}}$. Weight ω_{effect} is the adaptive weight [29] that inversely weighs each effect term by an initial estimate of the effect size, i.e., $\omega_{effect,m} = 1/\ |\widetilde{\gamma_m}\ |$ for the main effect terms and $\omega_{effect,\ell} = 1/|\widetilde{\beta_\ell}\ |$ for interaction terms. As a result, important terms receive smaller penalty and tend to be retained in the selecting process while unimportant terms receive larger penalty and are more likely to be eliminated. We use the iterative L1 penalty method [30] to obtain the initial estimates $\widetilde{\gamma_m}$ and $\widetilde{\beta_\ell}$. Specifically, $\widetilde{\gamma_m}$ and $\widetilde{\beta_\ell}$ are obtained by minimizing:

$$-\log L(\beta,\gamma;Y,X_1,X_2)+\lambda_1^*\sum_{m=1}^{M}\frac{\left|\gamma_m^{(t)}\right|}{\left|\gamma_m^{(t-1)}\right|}+\lambda_2^*\sum_{\ell=1}^{q}\frac{\left|\beta_\ell^{(t)}\right|}{|\beta_\ell^{(t-1)}|}, \qquad (2).$$

where $\gamma_m^{(t)}$ and $\beta_\ell^{(t)}$ are the estimate for the *m*-th main effect and ℓ-th interaction effect in the *t*-th iterative. The difference between Equations (1) and (2) is that in Equation (2), the adaptive weights of the current iteration are the estimates from previous iteration. The iteration continues until γ_m and β_ℓ converge for all m and ℓ. Using a penalized estimator allows us to obtain the initial estimates even when the sample size is smaller than the total number of variables, and the iterative procedure yields more accurate estimates [31]. The estimates for some (potentially unimportant) variables can be 0 with the L1 penalty. When that occurs, we set $\widetilde{\gamma_m} = \min(\min_{\beta_k>0}\gamma_k\,, 10^{-3})$ and use similar treatment for $\widetilde{\beta_\ell}$ as well.

Computing tuning parameters.

For a given value of (λ_1, λ_2), we compute the L1-reguralized estimates of γ_m and β_ℓ, and calculate BIC$= -2\log L(\gamma, \beta; Y, X_1, X_2) + k\log N$ for the corresponding model, where k is the number of terms retained in the model. The (λ_1, λ_2) that gives the smallest BIC is used to obtain the final model. We use BIC to tune λ_1 and λ_2 because our goal is to select the true model structure and BIC has the consistency property in model selection [32-35].

Simulation

We use simulation to evaluate the performance of the proposed method and the impact of different choices of weights. We performed two sets of simulation. Simulation I was based on a well-controlled hypothetical data with *n*>>*p*. It aims to determine the optimal forms of the weights and understand the impact of different weight specifications. Simulation II was based on the Wellcome Trust Case-Control Consortium [36] for Crohn's disease with *n*<*p*. It aims to evaluate the utility of proposed approaches under realistic settings.

Simulation I

Designs of Simulation I

In Simulation I, we generated 11 genes with different sizes (Table 3.1). The genes are labeled as gene A to gene K, and the number of SNPs in each gene was randomly determined from Uniform(1, 100). The minor allele frequency (MAF) of a SNP was randomly determined from an Uniform(0.1, 0.5) distribution. The SNPs within each gene were sorted by their MAF and only the middle 50% were used as causal SNPs. In our simulation, genes with ≤30 SNPs were labeled as small (S), genes with ≥70 SNPs were labeled as large (L), and genes with 30 ~ 70 SNPs (exclusively) were labeled as medium-sized genes (M). We considered 6 categories of gene-gene pairs: SS, SM, SL, MM, ML and LL.

We generated trait value Y_i from Model (3) below, where we assumed that gene F has causal main effect and there exist causal interaction effects between genes A and B (i.e., 2 small genes), between genes C and D (i.e., a small gene and a large gene), and between genes I and J (i.e., 2 large genes):

$$Y_i = \left(\sum_{l \in \{casual\ SNPs\}} G_{F,li}\right)\phi_F + \sum_{st \in \{AB,CD,IJ\}}\left(\sum_{k,k' \in casual\ SNPs} G_{s,ki} \cdot G_{t,k'i}\right)\zeta_{\ st} + e_i,$$

(3)

where $G_{m,li}$ is the genotype of SNP l in gene m for subject i and e_i is generated from $Normal(0,1)$. Coefficients ϕ_F, ζ_{AB}, ζ_{CD} and ζ_{IJ} are effect size; in the simulation, we use a common value for these coefficients and the common value is determined so that the partial R^2 explained by interactions was around 30%. The partial R^2 of the interaction effect is defined as $R^2 = (R_{12}^2 - R_1^2)/(1 - R_1^2)$, where R_{12}^2 is the R-square value for Model (3) containing both main and interaction effects, and R_1^2 is the R-square value for Model (3) containing only main effects (i.e., $\zeta_{st} = 0$ for all s and t). The total number of relevant pathways was 20. In each replication, we simulate 1500 individuals and performed 200 replications per scenario.

To assess the impact of a weight type, we performed the analyses under 2 conditions: (a) setting the corresponding weight type as 1 (i.e., neutral weights) and (b) incorporating the proposed weight type. For example, to assess the impact of ω_{path}, we examine the performance of (a) using $\omega_{size} + \omega_{effect}$ (a) vs. the performance of (b) using $\omega_{size} + \omega_{path} + \omega_{effect}$. For each condition, we computed the true positive rate (TPR) of detecting the causal GxG gene pairs. We also computed FPR among the non-causal gene pairs. Finally, we calculated Statistics D [37], which is defined as $D = \log TPR - \log FPR$ and is commonly used as an omnibus index to integrate TPR and FPR. Higher D indicates better performance of the method has.

Results of Simulation I

Assessment of $\boldsymbol{\omega_{size}}$ (Figure 3.1). When evaluating ω_{size}, we set the number of pathways supporting each interaction pairs as 20, 10, and 0 for $A \times B$ (MS gene pair), $C \times D$ (ML gene pair) and $I \times J$ (LL gene pair), respectively. We considered $\omega_{size,m}$= 1 (i.e., no gene size weights) and $\omega_{size,m} = 1 + (s_m - \min\{s_i\})/(\max\{s_i\} - \min\{s_i\})$, where s_m is the number of columns in the design matrix. The results indicated that without penalizing the gene size, the FPRs for large genes were substantially larger than FPR for small genes, e.g., the higher FPRs of LL pairs relative to MM pairs and ML pairs. By setting ω_{size} as proposed, the FPRs became less clustered in the large gene pairs, i.e., the FPRs in LL, ML and MM pairs decreased, the FPRs for SL pairs remained similar, and the FPR for gene pairs not involving large genes (e.g., SS and SM) slightly increased. For TPR, we observed the TPR deceased as the gene size increases, which is because ζ_{AB}, ζ_{CD} and ζ_{IJ} were set to be the same and the number of pathway supports happen to decrease as gene size increases. By comparing the TPR with and without ω_{size}, we see that the TPR slightly increased for $A \times B$ (MS pair), and slightly decreased for $C \times D$ (ML pair) and for $I \times J$ (LL pair). This is because ω_{size} encouraged the model to select smaller terms, although the differences were small. According to the statistics D, adding size penalty always increase the overall performance.

Assessment of $\boldsymbol{\omega_{path}}$ (Figure 3.2). Our proposed weight for pathway support has a general form of $\omega_{path,\ell} = 1 - \frac{1}{2} \times \frac{n_\ell}{N_{path}}$ for GxG term ℓ, and is ranged between $\frac{1}{2}$ to 1. In other words, the maximum amount of penalty reduction from pathway support is set to be half. Note that ω_{path} actually encourage gene paris with pathway support to be selected more likely. When evaluating ω_{path}, we set the number of pathways supporting each interactions according to Table 3.2, where we considered three scenarios, i.e., all causal interactions with *little*, *moderate,* or *strong* pathway support. Figure 3.2 suggested that for the scenarios of *moderate* and *strong* support, incorporating w_{path} have little impact on FPR but can boost TPR. For the *little* support scenario, incorporating w_{path} (which relatively discourages the selection of gene pairs with little support) did not cause too much reduction in TPR. However, there is a slight increase on FPR in *little* and *moderate* support compared to using the null pathway weight. This is likely because under those two scenarios, majority of the pathway supports are assigned to the null GxG gene pairs (i.e., the last column of Table 3.2). Overall speaking, it is worth to incorporate the pathway weights --- the gain in the D statistic caused by ω_{path} in *moderate* and *strong* supports is substantially more than the loss in *little* supports, and the scenario of *little* support might occur less frequently in reality.

Assessment of $\boldsymbol{\omega_{effect}}$ (Figure 3.3). When evaluating ω_{effect}, we set the number of pathways supporting each interaction pairs as 20, 10, and 0 for $A \times B$, $C \times D$ and $I \times J$, respectively. We compared the performance of four different ways to obtain the adaptive weights: (1) using the effect estimates from the iterative L1 penalty (L1), (2) using $\omega_{effect} = 1$ (null weights), (3) using the effect estimates from linear regression (LR), and (4) using the effect estimates from penalized L2 regression (L2). The other two weights, i.e., w_{path} and w_{size}, were specified using the proposed form. Figure 3.3 suggests that a null weight can lead to high TPR and high FPR and result in a low D value. All three estimating methods yielded similar TPRs but different FPRs. The iterative L1 penalty method had the smallest FPR and was the best choice among the methods. In contrast, the linear regression had the worst performance, and it is infeasible when the number of variables exceeds the number of samples. The L2 penalty method had a FPR slightly smaller than the linear regression had.

Simulation II

Designs of Simulation II

In Simulation II, we used the data from Wellcome Trust Case-Control Consortium (WTCCC, 2007) for Crohn's disease to simulate genotypes. Wang et al. (2010) reported two important pathways for Crohn's disease: (1) the IL-12 and STAT4 pathway which contains12 genes, and (2) the T cell receptor pathway which contains 67 genes. Because 3 genes were involved in both pathways, there were in total 76 genes. We computed the number of pathway supports for each GxG gene pair, $n_\ell,$ by the number of pathways that contain the gene pair. For example, for the 3 gene pairs formed by the 3 genes that are involved in both pathways, the number of pathway supports is 2 (i.e., $n_\ell = 2$). There are 2271 pairs of genes with 1 pathway-support, and 576 pair of genes without pathway support. Different from Biofilter, we kept those gene pairs with 0 pathway supports in the model but with higher penalty; so the selection procedure is more likely to drop them unless the data support its importance.

We simulated 200 replicated datasets with 1500 subjects per replications. We assigned 2 genes as causal main-effect genes and another 10 gene pairs (different from the causal main-effect genes) with causal interaction effects. We sorted the SNPs within a causal gene by their MAF and use the middle 50% SNPs as casual. To generate phenotype, we set

$$g(\mu_i) = \alpha + \sum_{m=1}^{2}\left(\sum_{l\in\{casual\ SNPs\}} G_{m,li}\right)\phi_m + \sum_{s,t\in\{10\ causaal\ gene\ pairs\}}\left(\sum_{k,k'\in\{casual\ SNPs\}} G_{s,ki}\cdot G_{t,k'i}\right)\zeta_{st} \qquad (4),$$

where $G_{m,li} \in \{0,1,2\}$ is the genotype of the causal SNP l in gene m. For quantitative trait, we set $g(\mu_i) = \mu_i$ and generated Y_i from $N(\mu_i, 1)$ with $\alpha = 0$ and the values of ϕ's and ζ's such that the partial R^2 contributed from the interaction effects was around 30%. For binary trait we set $g(\mu_i) = log(\frac{\mu_i}{1-\mu_i})$ and generated Y_i from Bernoulli(μ_i). Parameter α was set to make the prevalence around 7%. Similar to quantitative traits, the values of ϕ's and ζ's were determined so that the partial R^2 from the interaction effects was around 30%. For binary traits, we used Nagelkerke R^2[38] which is defined as $\{1-\left(\frac{-2L_1}{-2L_{12}}\right)^{\frac{2}{N}}\}/\left\{1-(-2L_1)^{\frac{2}{N}}\right\}$, where L_{12} is the likelihood of the logistic regressions containing both main and interaction effects, and L_1 is that of the logistic regression containing only main effects. For each replication, we oversampled cases so to obtain a balanced case-control sample (i.e., 750 cases and 750 controls).

We considered 3 scenarios as listed in Table 3.3 by carefully selecting 10 interactive gene pairs to evaluate the performance of the proposed procedure. Its performance was benchmarked against the penalized regression with only gene-size weight. In the "*no* *support*” scenario, most of the causal gene pairs were with 0 pathway support. In “*Random*” scenario, we randomly selected 10 gene pairs as causals. In the “*strong support*” scenario, which was the opposite of the “*no support*” scenario, the 10 causal gene pairs were selected from those with strong pathway support. For each scenario, we computed the TPR across the 10 gene pairs, the FPR across the non-causal gene pairs, and the D statistics.

Results of Simulation II

For quantitative traits (Figure 3.4), we see the TPRs of the proposed methods, which incorporated pathway supports and adaptive effect weights, were much higher than the TPRs of the benchmark methods that did not used these weights. The TPRs of the proposed methods were 1.5~2 times higher for the benchmarks while the FPRs of all methods were similar. The proposed methods also have higher D values. These patterns held for all 3 scenarios (highly suppressed, random, and highly supported,). For binary traits (Figure 3.5), the results are similar to the quantitative traits. While FPRs were also retained around 0.002~0.003, the TPRs were smaller, e.g., about 70% of the TPRs for the quantitative traits. This is not unexpected because binary trait values contained less information than quantitative trait values.

Real Data Analysis

Crohn's disease, also known as Crohn syndrome and regional enteritis, is a type of inflammatory bowel disease that may affect any parts of the gastrointestinal tract from mouth to anus, causing a wide variety of symptoms. Crohn’s disease is a complex genetic disease and lots of studies have been done to find the genetic factors for Crohn’s disease [39, 40].

We applied our approach to the WTCCC genome-wide association dataset for Crohn disease (CD) (WTCCC, 2007). The data contains 2005 cases and 3004 controls, and each individual had 469,557 SNPs genotyped by Affymetirx. We focused our analysis on the 2 important pathways to Crohn’s disease [41], the IL-12 and STAT4 pathway and the T cell receptor pathway. As mentioned in the design of Simulation II, there were 76 genes from the 2 pathways with 3 genes involved in both pathways. We extracted the SNPs of the 76 genes and removed SNPs with MAF smaller than 1%. We performed the analysis using the proposed method (i.e., incorporate all weights) and the benchmark method (only incorporate gene-size weight in the penalty). The significant genes and gene pairs are listed in Table 3.4. For GxG effects, we also listed the number of supporting pathways. For the proposed method, many significant gene pairs identified contain gene *GRB2*. *GRB2* has been found significant with Crohn’s disease [42]; it encodes protein GRB2, which is an adaptor protein involved in the signal transduction and cell communication. In contrast, the benchmark methods found 2 more GxG with 0 pathway support and misses 4 GxG with pathway support. It is consistent with the simulation study since the proposed method may discourage the detection of GxG with no pathway support.

Discussion

In this work, we proposed a pathway-guided approach for detecting interactions among genes. We construct weighted L1 penalty to select the important gene effect and gene-gene interactions; the weights are based on number of pathways supporting of the effects as well as the estimated effect size. The proposed approach can be used to explore the gene-gene interactions with a list of many candidate genes and is applicable even when sample size is smaller than the number of predictors studied. The numerical studies suggest an improved performance over the methods without using the guidance from pathway support and effect strength.

Our approach aims to combine the advantages of biological guidance and data adaptiveness, and our study suggested that both ω_{path} and ω_{effect} were necessary to obtain a robust *D* gain for the proposed method across different scenarios. Using ω_{path} would increase the TPR for the causal GxG which have strong pathway support. However, it may also decrease the TPR for those GxG effects with no or little pathway support. For these scenarios, incorporating ω_{effect} can help to minimize the TPR reduction and even boost the TPR because it encourages pairs with effects to be selected in the model.

In real practice, pathway knowledge is often used to guide the search of biologically meaningful variables. This practice is based on the presumption that the pathway knowledge can reflect the underlying biological mechanisms. However, it is likely that the pathway structures depend on phenotypes and hence the “canonical” pathway information would only represent the status of healthy controls. From this point of view, treating the biological information as prior knowledge and performing data adaptive selection can provide robustness against vague information and minimize the false positive and false negative findings.

In this paper, we only used the pathway membership in the variable selection process. There exist other types of information, such as pathway structure, the regulation directions between genes, protein interaction, RNA networking or metabolite information, that can provide valuable guidance in the exploration of gene-gene interaction in the large search space. It is worth further study to appropriately formulate the biological knowledge from multiple resources into the statistical model and lead to efficient variable selections.

References

1 Cordell HJ: Detecting gene-gene interactions that underlie human disease. Nature Reviews Genetics 2009; **10**: 392-404.

2 Lin X, Hamilton-Williams EE, Rainbow DB *et al*: Genetic interactions among Idd3, Idd5.1, Idd5.2, and Idd5.3 protective loci in the nonobese diabetic mouse model of type 1 diabetes. J Immunol. 2013; **7**: 3109-3120.

3 Pillai R, Waghulde H, Nie Y *et al*: Isolation and high-throughput sequencing of two-closely linked epistatic hypertension susceptibility loci with a panel of bicongenic strains. Physiol Genomics. 2013; **45**(16):729-36.

4 Koh-Tan HH, McBride MW, McClure JD *et al*: Interaction between chromosome 2 and 3 regulates pulse pressure in the stroke-prone spontaneously hypertensive rat. Hypertension. 2013: **62**: 33-40

5 Howson JM, Cooper JD, Smyth DJ *et al*: Evidence of gene-gene interaction and age-at-diagnosis effects in type 1 diabetes. Diabetes 2012; **11**: 3012-3017.

6 Ziyab AH, Davies GA, Ewart S *et al*: Interactive effect of STAT6 and IL13 gene polymorphisms on eczema status: results from a longitudinal and a cross-sectional study. BMC Med Genet. 2013; **14**: 67.

7 Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet, 2005; **37**(4): 413–417.

8 Evans DM, Marchini J, Morris AP, Cardon LR: Two-Stage Two-Locus Models in Genome-Wide Association. PLoS Genetics, 2006; **2**(9), e157.

9 Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: A comprehensive review of genetic association studies. Genet Med. 2002; **4**(2): 45-61.

10 Cordell HJ: Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans. Human Molecular Genetics 2002; **11**: 2463-2468.

11 Bush WS, Dudek SM, Ritchie MD: Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput 2009: 368-379.

12 Bridges CB: Specific modifiers of eosin eye-color in Drosophila. Jour. Exper. Zool. 1919; **28**: 337-384.

13 Lloyd V, Ramaswami M, Kramer H: Not just pretty eyes: Drosophila eye-color mutations and lysosomal delivery. Trends Cell Biol. 1998; **8**: 257–259.

14 Kooperberg C, Ruczinski I, LeBlanc M, Hsu L: Sequence analysis using logic regression. Genet Epidemiol 2001; **21**(Suppl1): S626–S631.

15 Cook N, Zee R, Ridker P: Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. Stat. Med. 2004; **23**: 1439–1453.

16 Ritchie M, Hahn L, Moore J: Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol. 2003; **24**: 150–157.

17 Ritchie MD: Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. Ann Hum Genet. 2011 Jan; **75**(1):172-82.

18 Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S: Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 2006; **79**:1002-1016.

19 Wang T, Ho G, Ye K, Strickler H, Elston R: A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genet. Epidemiol. . 2010; **33**(1): 6-15.

20 Larson NB, Schaid DJ: A Kernel Regression Approach to Gene-Gene Interaction Detection for Case-Control Studies. Genet Epidemiol. 2013; **37**(7): 695-703.

21 Aschard H, Zaitlen N, Tamimi RM, Lindstr öm S, Kraft P: A nonparametric test to detect quantitative trait Loci where the phenotypic distribution differs by genotypes. Genet Epidemiol. 2013; **37**(4):323-33.

22 Wu J, Devlin B, Ringquist S, Trucco M, Roeder K: Screen and clean: a tool for identifying interactions in genome-wide association studies. Genet Epidemiol 2010 Apr; **34**(3): 275-85.

23 Sha Q, Zhang Z, Schymick JC, Traynor BJ, Zhang S: Genome-Wide Association Reveals Three SNPs Associated With Sporadic Amyotrophic Lateral Sclerosis Through a Two-Locus Analysis. BMC Med Genet. 2009; **10**: 86.

24 Robnik-Sikonja M, Kononenko I: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 2003; **53**: 23–69.

25 Moore JH, White BC: Tuning ReliefF for genome-wide genetic analysis. Lecture Notes in Computer Science. 2007; **4447**: 166–175.

26 Pendergrass SA, Verma SS, Holzinger ER *et al*: Next-generation analysis of cataracts: determining knowledge driven gene-gene interactions using Biofilter, and gene-environment interactions using the PhenX Toolkit. Pac Symp Biocomput. 2013:147-58.

27 Turner SD, Berg RL, Linneman JG *et al*: Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. PLoS One 2011; **6**(5): e19586.

28 Bush WS, McCauley JL, DeJager PL *et al*: International Multiple Sclerosis Genetics Consortium. A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility. Genes Immun. 2011; **12**(5):335-40.

29 Zou H: The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 2006; **101**: 1418-1429.

30 Bühlmann P, Meier L: Discussion of "One-step sparse estimates in nonconcave penalized likelihood models" (authors Zou H and Li R). Ann Stat 1919; **36**:1534–1541

31 Li Z, Sillanpää MJ: Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet. 2012 Aug; **125**(3): 419-35.

32 French B, Lumley T, Monks SA *et al*: Simple estimates of haplotype relative risks in case-control data. Genet Epidemiol. 2006; **30**: 485–494.

33 Guo W, Lin S: Generalized linear modeling with regularization for detecting common disease rare haplotype association. Genet Epidemiol. 2009; **33**: 308–316.

34 Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. Springer-Verlag; 2009.

35 Lake SL, Lyon H, Tantisira K *et al*: Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered. 2003; **55**: 56–65.

36 Wellcome Trust Case Control Consortium: Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; **447**: 661–678

37 Athanasiou T: Evidence Synthesis in Healthcare; A practical handbook for Clinicians. Springer. 2011.

38 Nagelkerke, NJD: A note on a general definition of the coefficient of determination. Biometrika 1991; **78**: 691-692.

39 Holmans, P *et al*: Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am. J. Hum. Genet. 2009; **85**: 13-24.

40 Abraham C, Cho J: IL-23 and autoimmunity: new insights into the pathogenesis of inflammatory bowel disease. Annu. Rev. Med. 2009; **60**: 97-110.

41 Wang K, Li M, Hakonarson H: Analyzing biological pathways in genome-wide association studies. Nature Reviews Genetics 2010; **11**: 843- 854.

42 Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM: Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011 Jul; **21**(7): 1109-21.

Table 3.1 Gene information for Simulation I. In Simulation I, 11 genes are generated, the number of SNPs in each gene ranges from 7 to 99.

Gene ID	A	B	C	D	E	F	G	H	I	J	K
Number of SNPs	42	7	48	77	46	20	31	99	84	72	14

Table 3.2 The biological supports under four scenarios. Each scenario contains 20 pathways. The differences between scenarios are the number of pathways supporting the causal GxG

Scenarios	Number of pathways supporting GxG of a gene pair				Total # of pathway supports[b]	% for causal gene pairs (1-a/b)
	Causal GxG gene pairs			Non-causal GxG gene pairs (52 pairs)[a]		
	$A \times B$	$C \times D$	$I \times J$			
Little support	2	1	0	92	95	3%
Moderate support	11	10	9	166	196	18%
Strong support	20	19	18	160	177	48%

Table 3.3 Different level of biological supports among the 10 gene pairs considered in Simulation II. *No Support:* the causal gene pairs do not have much pathway support; *Random:* the causal gene pairs that are randomly selected; *Strong support*: the causal gene pairs have strong pathway support.

# of supporting pathways Scenario	2 Pathways	1 Pathway	0 Pathway
1. No Support	0	2	8
2. Random	1	7	2
3. Strong Support	3	6	1

Table 3.4 List of significant main-effect genes and GxG gene pairs identified by the proposed method and the benchmark method. (“--“ means not found by the corresponding method.)

	Gene Names	# of pathways supporting the gene/gene pair	
		$\omega_{path} + \omega_{effect} + \omega_{size}$	ω_{size}
Main effect	GRB2	1	1
	IL12B	1	1
	PPP3CA	1	--
GxG effect	AKT3&GRB2	1	1
	CD247&IL12B	1	1
	CD4&FYN	1	1
	CHP&GRB2	1	1
	FYN&IKBKB	1	--
	GRB2&GSK3B	1	--
	GRB2&MAP3K14	1	--
	GRB2&NCK2	1	--
	ETV5&PPP3CA	0	0
	CHP&IL12B	--	0
	IL18R1&RASGRP1	--	0

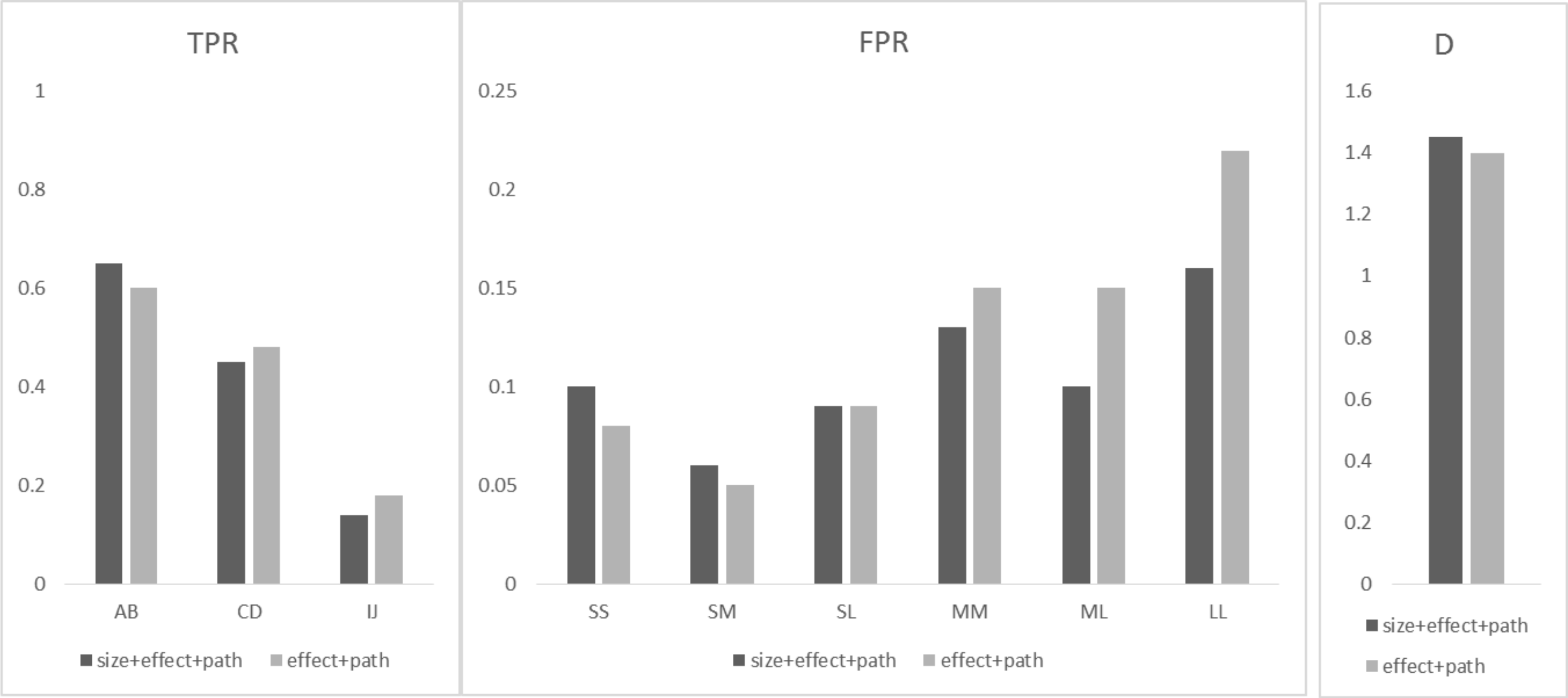


Figure 3.1 Assessment of size weight The True positive rate (TPR), False Positive Rate (FPR) and D statistic ($D = \log TPR - \log FPR$) for detecting GxG gene pairs in Simulation I. Here 2 types of ω_{size} are considered: (1) $\omega_{size,m} = 1$ (i.e., no size weights), represented by gray bar, and (2) $\omega_{size,m} = 1 + \frac{s_m - \min_m s_m}{\max_m s_m - \min_m s_m}$ where s_m is the gene size, represented by shaded bar. The x-axis represents the gene labels in the TPR plot and represents the gene sizes in the FPR plot, i.e., S/M/L for small/medium/large genes.

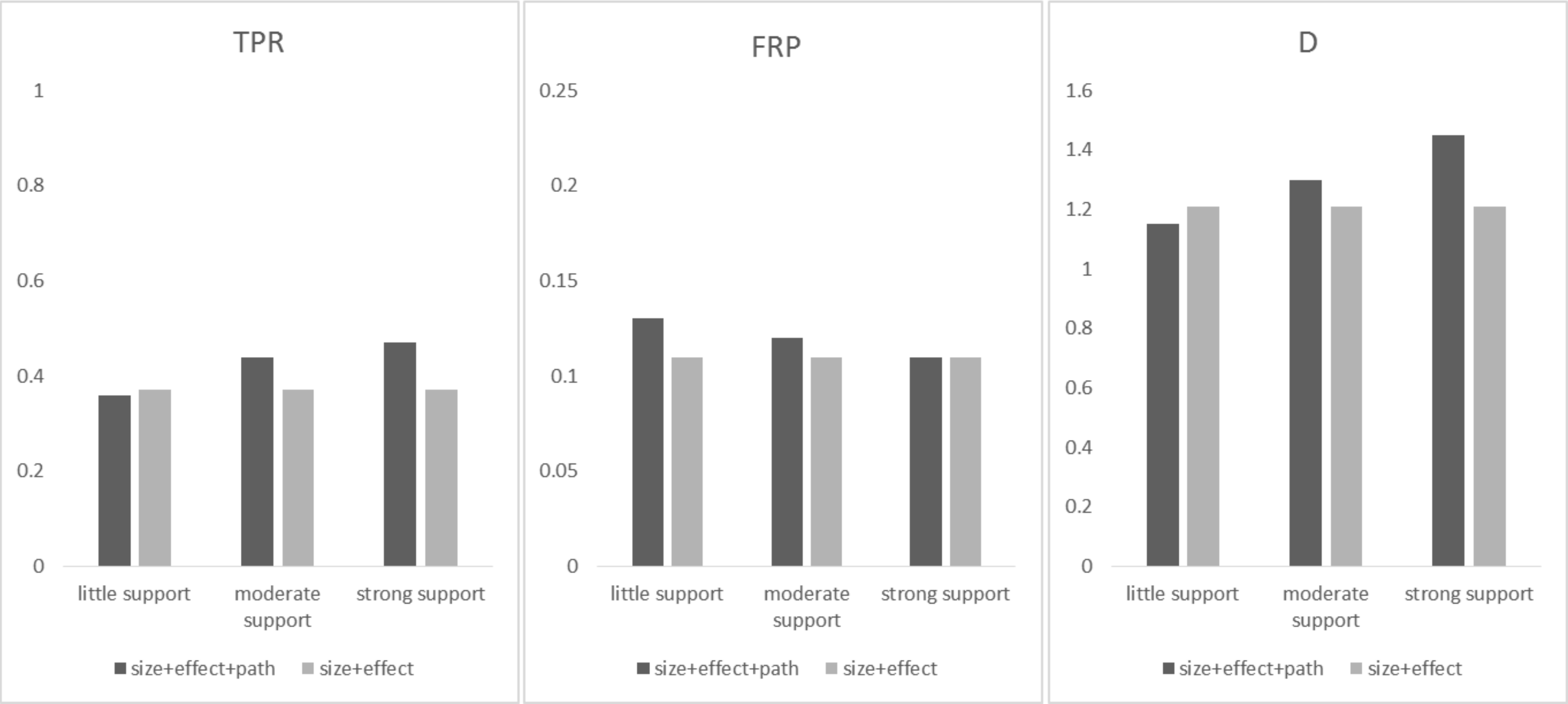


Figure 3.2 Assessment of pathway weight The True positive rate (TPR), False Positive Rate (FPR) and D statistic ($D = \log TPR - \log FPR$) for detecting GxG gene pairs in Simulation I under 3 different scenarios, i.e., causal gene pairs with *little*, *moderate* and *strong* pathway supports (as detailed in Table 2). In each scenario, shaded bar represents the result of incorporating pathway support, i.e., setting ω_{path} as proposed and gray bar represents the results of setting $\omega_{path} = 1$, i.e., no pathway support.

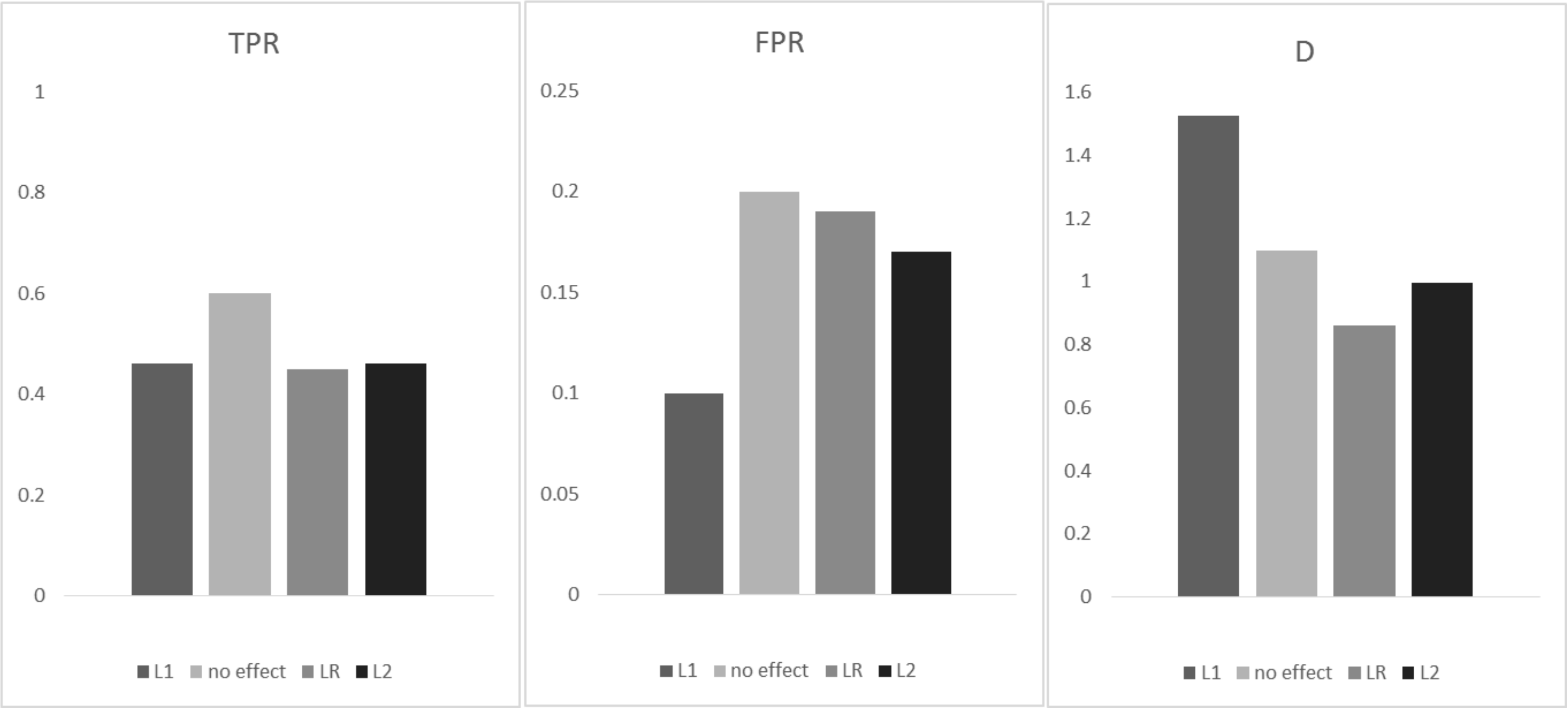


Figure 3.3 Assessment of effect weight The True positive rate (TPR), False Positive Rate (FPR) and D statistic ($D = \log TPR - \log FPR$) for detecting GxG gene pairs in Simulation I using ω_{effect} calculated by different methods: (1) Iterative L1 penalty regression (L1; shaded bar); (2) $\omega_{effect} = 1$ (No effect weight; light gray bar); (3) linear regression (LR; dark gray bar) and (4) L2 penalty regression (L2; black bar).

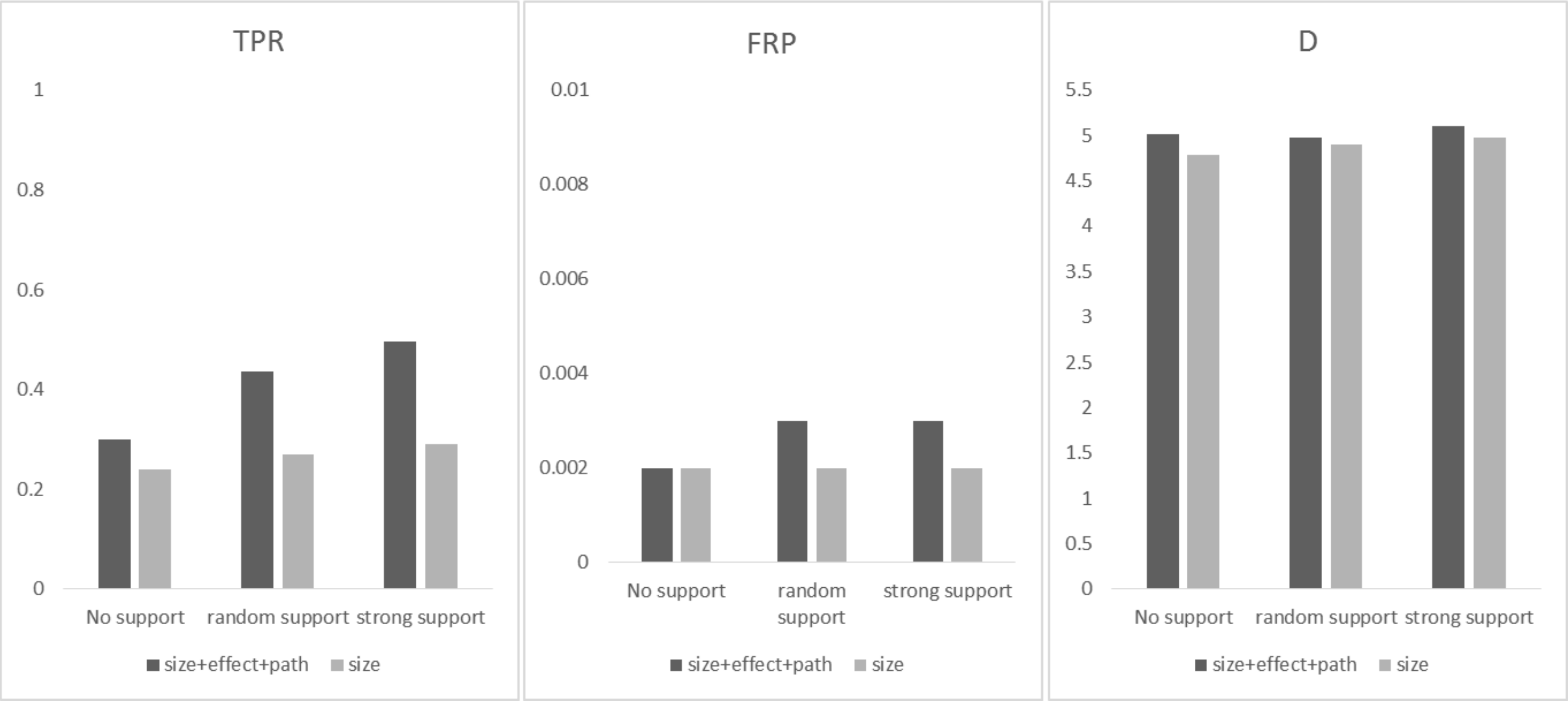


Figure 3.4 Simulation II: quantitative phenotypes Result of Simulation II based on Crohn's disease with quantitative phenotypes. True positive rate (TPR), false positive rate (FPR) and the D statistic ($D = \log TPR - \log FPR$) for detecting GxG gene pairs under 3 different scenarios as defined in Table 3, i.e., the causal gene pairs do not have much pathway support (*no support*), have strong pathway support (*strong support)*, and the causal gene pairs that are randomly selected (*random*). The shaded bar represents the results of using all 3 weights and the gray bar represents the results of only using ω_{size}.

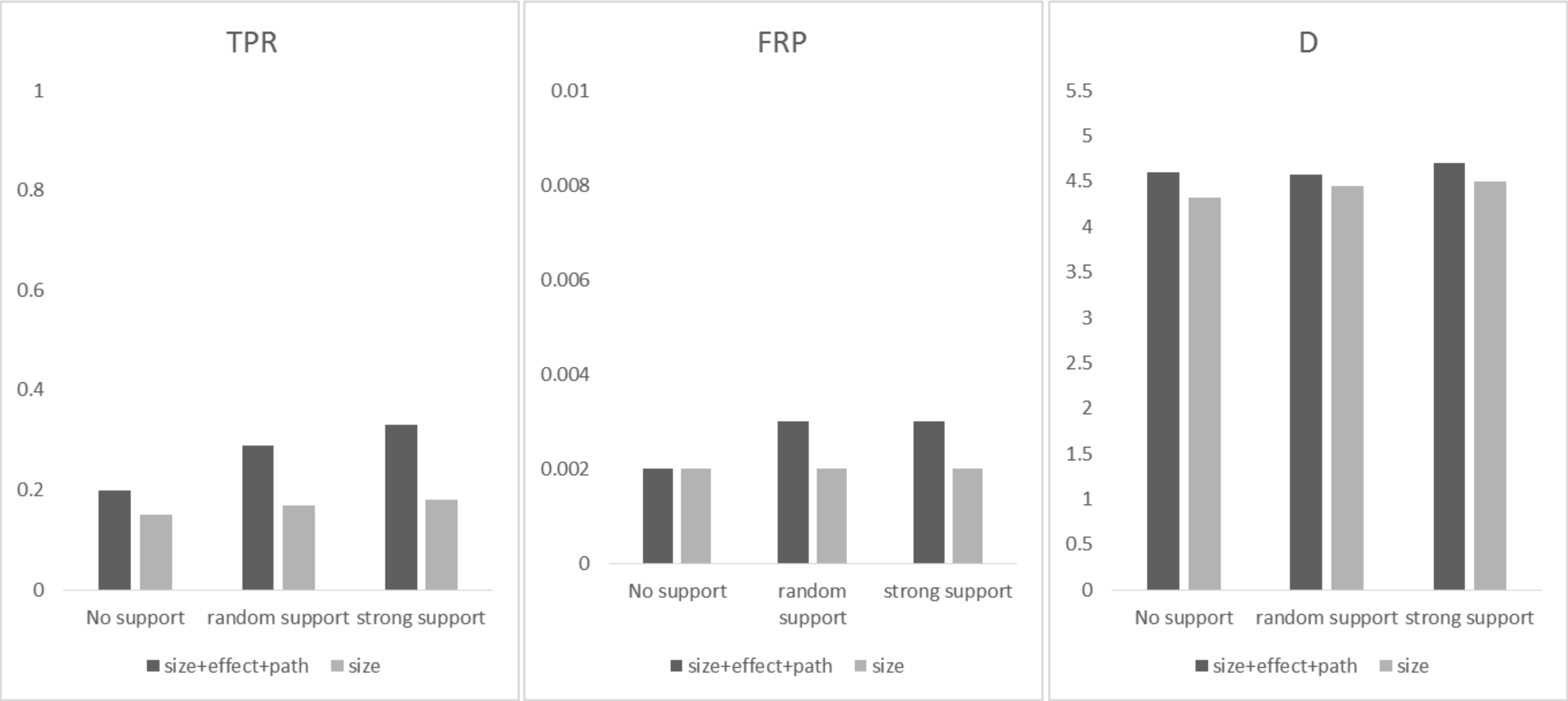


Figure 3.5 Simulation II: binary phenotype Results of Simulation II based on Crohn's disease with binary phenotypes. True positive rate (TPR), false positive rate (FPR) and the D statistic ($D = \log TPR - \log FPR$) for detecting GxG gene pairs under 3 different scenarios as defined in Table 3, i.e., the causal gene pairs do not have much pathway support (*no support*), have strong pathway support (*strong support)*, and the causal gene pairs that are randomly selected (*random*). The shaded bar represents the results of using all 3 weights and the gray bar represents the results of only using ω_{size}.

Chapter 4

Multiclass Probability Estimation via Kernel SVMs

Xin Wang[1, 2], Hao Helen Zhang[2, 3], Yichao Wu[2*]

[1] Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA

[2] Department of Statistics, North Carolina State University, Raleigh, NC, USA

[3] Department of mathematics, University of Arizona, Tucson, AZ, USA

*Correspondence: Yichao Wu, Department of Statistics, North Carolina State University, 4274 SAS Hall, 2311 Stinson Dr., NC State University, Raleigh, NC 27607, USA

Submitted to Biometrics.

Abstract

Multiclass classification and probability estimation have many important biological and medical applications. Support vector machines (SVMs) have shown great success in many real applications due to high classification accuracy. One limitation of SVMs is that they do not provide probability estimation for classification, though this uncertainty measure of prediction is often more informative and useful for decision making. In this paper, we propose a simple yet very effective framework to endow kernel SVMs with the feature of multiclass probability estimation. The new procedure enjoys desired theoretical, computational, and numerical properties. The produced multi-class probability estimators are shown to be consistent. The new estimator does not rely on any parametric assumption on data distribution, therefore it is robust and flexible. Computationally, only minimal programming effort is needed, since the procedure can be conveniently implemented using standard SVM softwares. Our numerical examples shows the very competitive performance of the SVM-based probability estimators, compared to traditional multi-class probability estimators including those produced from multiple logistic regression, linear discrimination analysis, or quadratic discrimination analysis.

Key Words and Phrases: multiclass classification, probability estimation, logistic regression, support vector machines.

Introduction

Multiclass classification and prediction are commonly encountered in biomedical studies. In cancer diagnosis, even for the same type of tumors, it is usually critical to divide them into several subgroups based on their histopathological type, grade, stage, and genetic information, as the knowledge of a specific subtype helps to tailor the treatment approaches and dose levels for increased efficacy and drug sensitivity, low toxicity, and the best outcome. For example, based on the combined clinical and pathological criteria, leukemia can be divided into a variety of subtypes and the four major subtypes are acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), acute myelogenous leukemia (AML), and chronic myelogenous leukemia (CML).

Recently, cancer classification based on microarray data has received much attention. One motivating example of this work is the classification of small round blue cell tumors (SRBCT) of childhood (Khan et al., 2009). The data consist of 2,308 gene expression measurements, which were obtained from glass-slide cDNA microarrays following the standard National Human Genome Research Institute protocol. There are four tumor subtypes: Burkitt lymphoma (BL), Ewing sarcoma (EWS), neuroblastoma (NB), or rhabdomyosarcoma (RMS). The training set contains 63 samples and the test set contains 25 samples. The goal is to use the subject's gene expression information to accurate classify the tumor into a specific subtype.

In multiclass classification problems, we are typically given a sample $\{(\boldsymbol{x}_i, y_i), i = 1,2, \cdots, n\}$ of identically and independently distributed observations from some unknown distribution $P(\boldsymbol{X}, Y)$, where $\boldsymbol{x}_i \in \mathcal{S} \subset IR^d$ and $y_i \in \{1,2,\cdots,K\}$ denote the input vector and output label, respectively. Here n is the sample size, d is the dimensionality of the input space, and K denotes the number of categories (or classes) for the response. Depending on whatthe ultimate goal is, classification can be generally divided into hard classification and soft classification. In hard classification, one is only interested in estimating a classification rule (or classifier) which shall be used to assign a label to a new input vector. Popular examples of hard classifiers include support vector machines, nearest neighbor classifiers, and classification trees. On the other hand, the goal of soft classification is to estimate the conditional probabilities of the response belonging to different subclasses, namely to estimate $p_k(\boldsymbol{x}) = P(Y = k|\boldsymbol{X} = \boldsymbol{x})$ for $k = 1,2,\cdots,K$. The estimated conditional class probabilities can be used to obtain a classification rule. For example, when equal costs are used for misclassifications, the argmax rule $\hat{k} = \arg \max_{\{k=1,\cdots,K\}} p_k(\boldsymbol{x})$ is for classification. Traditional probability estimation methods are based on either regression techniques such as multiple logistic regression, or the density estimation approach such as linear (or quadratic) discriminant analysis (LDA or QDA). See Agresti and Coull (1998) and references therein for an overview of these methods. The probability functions are usually more complex than the classification boundary, so in some sense soft classification aims to solve a more difficult problem than hard classification.

There is a recent surge of the literature on hard classification. One of the most well-known hard classification methods is support vector machine (SVM, Cortes and Vapnik, 1995; Vapnik, 1998). Lin (2002) proved that the binary SVM targets directly at the Bayes classification boundary without estimating the conditional class probabilities at all. From binary to multiclass problems, there are several extensions such as Weston and Watkins (1999); Lee et al., (2004) and Liu (2007). Zhang (2004) provides some interesting findings on the consistency of multi-category classification methods. Despite their great success in real problems for classification, support vector machines and their multiclass extensions can not provide probability estimates. For example, in cancer diagnosis using gene expression information, in addition to labeling a patient as “subtype A” or “subtype B”, the doctor often desires to have a reliable estimate for the probabilities of belonging to the substypes, since the probabilities provide valuable measure of uncertainty in classification and hence more informative for decision makings.

The SVMs have shown high classification accuracy for many applications in assorted scientific areas such as cancer diagnosis, handwritten digits recognition, junk email detection. Consequently, questions like “Is the SVM capable of estimating the conditional class probabilities?” have been posted for the possibility of taking advantage of the SVM’s impressive classification performance. In the context of binary classification problems, Wang et al. (2008) demonstrated that soft classification can be achieved by training a series of weighted support vector machines and then aggregating decision rules to form conditional class probabilities. Wu et al. (2010) generalized this method from the binary case to the multiclass case by training weighted multiclass classifiers. However, the number of weighted multi-category classifiers to be trained increases exponentially fast when the number of classes K gets larger or the weight grid becomes finer. The computational cost increases dramatically with K. In addition, when the overall classification problem changes by adding another class, the results for the original problem cannot be used any more and one needs to start over by training all weighted hard classifiers. In this work, we propose a very simple yet effective approach to generalize weighted binary SVMs to multiclass SVMs capable of estimating probabilities. The new framework still takes advantage of aggregating multiple hard classifiers to estimate conditional class probabilities, but the key idea is to decompose multiclass probability estimation problems into multiple binary problems. The procedure does not require solving a complex optimization problem, and any standard SVMs software can be used to implement the procedure. The computational cost increases quadratically with K. As suggested by our numerical examples, the new method is much faster and yet delivers very competitive performance.

The rest of the paper is organized as follows. Section 2 presents the main methodology and studies theoretical properties of the probability estimator. Section 3 gives an efficient computational algorithm for implementation. Section 4 contains numerous examples to illustrate performance of the new procedure, which is followed by the concluding section. The appendix collects proofs for theoretical results as well as the derivation of our algorithm.

Main Methodology

Background: Binary Classification and Probability Estimation

In binary classification problems, the class label y is typically coded as {+1, -1} for convenience. The large-margin classifier is constructed by solving the following regularization problem

$$\min_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^{n} L\left(y_i f(\boldsymbol{x}_i)\right) + \lambda J(f) \tag{1}$$

where the loss function $L(\cdot)$ is a function of functional margin $yf(x)$, $\mathcal{F}$ is some functional space, $J(f)$ is a penalty term for model complexity, and $\lambda > 0$ is the regularization parameter which balances the data fit measured by the loss function and the model complexity measured by the penalty. For example, the SVM uses the hinge loss $L(z) = (1 - z)_+ = \max\{0,1 - z\}$. If the function f is linear with the form $f(\boldsymbol{x}) = \boldsymbol{x}\beta_1 + \beta_0$, we call it a linear classifier. In kernel classification associated with a bivariate Mercer kernel $K(\cdot,\cdot)$, we end up with a nonlinear classifier f with the kernel representation $\beta_0 + \sum_{i=1}^{n} \theta_i\, K(\boldsymbol{x}_i, \boldsymbol{x})$ due to the representor theorem (Kimeldorf and Wahba, 1971), and in this case $\mathcal{F}$ is the reproducing kernel Hilbert space (RKHS Wahba, 1990) induced by $K(\cdot,\cdot)$, denoted as $\mathcal{H}_K$. When $J(f) =\| f \|_{\mathcal{H}_K}^2$, the optimization problem is expressed as

$$\min_{\beta_0,\theta} n^{-1}\sum_{i=1}^{n} L\left(y_i f(\boldsymbol{x}_i)\right) + \lambda\sum_{i=1}^{n}\sum_{j=1}^{n} \theta_i\, \theta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{2}$$

where $f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \theta_i\, K(\boldsymbol{x}_i, \boldsymbol{x})$.

Denote the conditional class probability $p_1(\boldsymbol{x}) = P(Y = +1|\boldsymbol{X} = \boldsymbol{x})$. Lin (2002) proved that the SVM solution $\hat{f}$ to (2) has the same sign as the Bayes classification rule $\text{sign}[p_1(\boldsymbol{x}) - \frac{1}{2}]$. In other words, the SVMs directly target on the Bayes rule without estimating $p_1(\boldsymbol{x})$.

Wang et al. (2008) suggested a novel approach to constructing the conditional class probabilities using the weighted SVMs for binary problems. The basic idea is to first assign samples from Class -1 (and Class $+1$) with weights π (and $1-\pi$)and then minimize the weighted hinge loss

$$\min_{f \in \mathcal{H}_K} n^{-1} \left[(1-\pi) \sum_{y_i=1} L\left(y_i f(\boldsymbol{x}_i)\right) + \pi \sum_{y_i=-1} L\left(y_i f(\boldsymbol{x}_i)\right)\right] + \lambda J(f), \tag{3}$$

where $0 \leq \pi \leq 1$. The sign of the minimizer to (2) is proved to be consistent for estimating $\text{sign}[p_1(\boldsymbol{x}) - \pi]$. Therefore, by using different π values, $0 = \pi_1 < \cdots < \pi_{m+1} = 1$, one can repeatedly solve (2) and obtain a series of classifiers, say, $\hat{f}_{\pi_1}, \cdots, \hat{f}_{\pi_{m+1}}$. For any $\boldsymbol{x}$, there exists a unique $j*$ such that $\hat{f}_{\pi_{j*}}(\boldsymbol{x})$ and $\hat{f}_{\pi_{j*+1}}(\boldsymbol{x})$ have opposite signs. It implies that π_{j*} and π_{j*+1} satisfy that $\text{sign}[p_1(\boldsymbol{x}) - \pi_{j*}] \neq \text{sign}[p_1(\boldsymbol{x}) - \pi_{j*+1}]$, which leads to a natural probability estimator $\hat{p}_1(\boldsymbol{x}) = \frac{1}{2}(\pi_{j*} + \pi_{j*+1})$. More technical details can be found in Wang et al. (2008). Their numerical examples show that their probability estimator performs competitively comparing to existing approaches such as Platt's method (Platt, 1990).

New Method for Multiclass Probability Estimation

From now on, we focus on multiclass problems where the number of classes $K > 2$. The response y is typically coded as $\{1,2,\cdots,K\}$. Wu et al. (2010) extended the estimation scheme of Wang et al. (2008) from the binary case to the multiclass case. Their framework is designed to separate all K classes altogether by solving a complicated weighted multiclass SVM problem, which assigns a weight π_k to points from class k for $k = 1, \cdots, K$. All the possible weights form the K-cube hyperplane $A_K = \{(\pi_, \cdots, \pi_K): 0 \leq \pi_k \leq 1, \sum_{i=1}^{K} \pi_k = 1\}$. To construct the multiclass probabilities from weighted classifiers, one needs to repeatedly fit weighted multiclass SVMs for all the points (or based on a fine grid) in A_K.The accuracy of the proposed probability estimates is very competitive. However, the computational cost of that procedure can be very high for a large K, since the number of grid points in A_K increases exponentially fast in K. In this paper, we propose an alternative, conceptually much simpler, and computationally much faster method to estimate multiclass probabilities using the weighted SVM technique.

Define the class probabilities $p_k(\boldsymbol{x}) = P(Y = k|\boldsymbol{X} = \boldsymbol{x})$, $k = 1, \cdots, K$. For each pair of two classes $(k, k^{'})$, define the pairwise conditional probability $q_{k|(k,k^{'})}(\boldsymbol{x})$ as

$$q_{k|(k,k^{'})}(\boldsymbol{x}) = \frac{P(Y = k|\boldsymbol{X} = \boldsymbol{x})}{P(Y = k|\boldsymbol{X} = \boldsymbol{x}) + P(Y = k^{'}|\boldsymbol{X} = \boldsymbol{x})}.$$

The pairwise conditional probability $q_{k|(k,k^{'})}(\boldsymbol{x})$ can be interpreted as the conditional probability of $\boldsymbol{x}$ belonging to Class k given the condition that it belongs to either Class k or Class $k^{'}$. The idea behind the new procedure isas follows. We decompose the multiclass classification problem into multiple binary classification problems, discriminating Class k from Class $k^{'}$ for $1 \leq k < k^{'} \leq K$, and fit the binary weighted SVM for each pair. There are totally $K(K-1)/2$ binary classification problems. The solution to each binary problem, Class k vs Class $k^{'}$, turns out to be a consistent estimator $q_k|(k, k^{'})$. We then integrate all the binary problem solutions to estimate the class probabilities p_k's. The following gives the detailed procedure:

1. For each pair of classes, Class k vs Class $k^{'}$, define a univariate function $R_{k,k^{'}}(y) = 1$ if $y = k$ and $= -1$ if $y = k^{'}$. Then we fit the weighted SVM by solving

$$\min_{f\in\mathcal{H}_K} n^{-1}\left[(1-\pi_m)\sum_{y_i=k} L\left(R_{k,k'}\,(y_i)f(\boldsymbol{x}_i)\right) + \pi_m \sum_{y_i=k'} L\left(R_{k,k'}\,(y_i)f(\boldsymbol{x}_i)\right)\right] + \lambda J(f) \qquad (4）$$

over a grid $0 < \pi_1 < \cdots < \pi_M < 1$. For each π_m, the solution is $\hat{f}_{k,k'\,,\pi_m}(\cdot)$.

2. Define

$\hat{q}_{k|(k,k'\,)}(\boldsymbol{x}) = [\arg\min_{\pi_m}\{\hat{f}_{k,k'\,,\pi_m}(\boldsymbol{x}) < 0\} + \arg\max_{\pi_m}\{\hat{f}_{k,k'\,,\pi_m}(\boldsymbol{x}) > 0\}]/2$. The final class probabilities are given as

$$\hat{p}_k(\boldsymbol{x}) = \frac{\hat{q}_{k|(k,k'\,)}(\boldsymbol{x})}{\sum_{l=1}^{K}\hat{q}_{l|(l,k'\,)}(\boldsymbol{x})}, \qquad (5)$$

Here we abuse the notation slightly by defining $\hat{q}_{k|(k,k)}(\boldsymbol{x}) = 1$ for any $\boldsymbol{x}$. In the above, we assume that a proper regularization parameter λ is identified for fitting each weighted SVM classifier. The selection of tuning parameter is discussed in next session.

In the following are theoretical justifications for the proposed class probability estimators.

Lemma 1. Assume $0 < \pi < 1$.Define

$$A(f) = E[(1-\pi)I(Y=k)L\left(R_{k,k'}(Y)f(\boldsymbol{X}) + \pi I(Y=k')L\left(R_{k,k'}(Y)f(\boldsymbol{X})\right)\right].$$

The minimizer of $A(f)$ is given by $f^*(\boldsymbol{x}) = q_{k|(k,k')}(\boldsymbol{x}) - \pi$.

Lemma 1 suggests that the minimizer of the binary weighted SVM (4) is equivalent to the Bayes rule for separating two classes. Following the same argument of Wang et al. (2008), we can show that $\hat{q}_{k|(k,k')}(\boldsymbol{x})$converges to $q_{k|(k,k')}(\boldsymbol{x})$ as the sample *n* goes to infinity. Using the relationship between $q_{k|(k,k')}(\boldsymbol{x})$ and $p_k(\boldsymbol{x})$, it is easy to see that $\hat{p}_k(\boldsymbol{x})$ provides a consistent estimator to $p_k(\boldsymbol{x})$. Although we focus on the hinge loss function of the SVM in this article, the new procedure can be extended to other supervised learning methods of which the loss is Fisher consistent.

From (5), it seems that the class probability estimators depends on the value of k'. It turns out, from the theoretical viewpoint, different k' all give the same quantity. So it does not matter which k'class is used as the baseline class. In practice, there could be slight difference in the estimators from different k''s due to numerical convergence and stability. In our implementation, we choose k' which leads to the maximal class probability estimator $\hat{p}_k(\boldsymbol{x})$.

Implementation

We describe the implement of the new approach in details. We start with the simple linear learning and then discuss nonlinear learning using the kernel trick.

Kernel learning

We select a uniform grid $0 = \pi_0 < \pi_1 < \cdots < \pi_M < \pi_{M+1} = 1$ with $\pi_M = m/(M + 1)$ for $m = 1, 2, \cdots, M$ and some integer $M > 0$. The kernel trick relies on a bivariate kernel function $K(\cdot,\cdot)$, which maps from $\mathcal{S} \times \mathcal{S}$ to IR^d

For each pair of two classes $(k, k^{'})$, we need to solve (4). Due to the representer theorem of Kimeldorf and Wahba (1971) (also see Wahba, 1990), the solution to (4) has the finite representation $f(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i\, K(\boldsymbol{x}_i, \boldsymbol{x}) + \beta_0$. This representation greatly facilitates the implementation of the weighted SVM, as the problem reduces to finding finite-dimensional coefficients θ_i, $i = 1,2, \cdots, n$ and β_0. Correspondingly, the roughness penalty becomes $J(f) = \sum_{i=1}^{n} \sum_{j=1}^{n} \theta_i\, a_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. By introducing the slack variables ξ_i, $i = 1,2, \cdots, n$, we can reformulate the optimization problem (4) as:

$$\min_{\beta,\beta_0,\xi_1,\cdots,\xi_n}\left[(1-\pi_m)\sum_{y_i=k}\xi_i+\pi_m\sum_{y_i=k^{'}}\xi_i\right]+\lambda\sum_{i=1}^{n}\sum_{j=1}^{n}\theta_i\,\theta_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \qquad (6)$$

$$\text{subject to } \xi_i\geq 0, i=1,2,\cdots,n$$

$$\xi_i\geq 1-R_{k,k^{'}}(y_i)(\sum_{j=1}^{n}\theta_j\,K(\boldsymbol{x}_j,\boldsymbol{x}_i)+\beta_0), i\in\{m\colon y_m=k \text{ or } k^{'}\}.$$

We can easily see that (6) is a linearly constrained quadratic programming (QP) problem, which can be solved by many numerical software packages. Denote the optimizer of (6) by $\hat{\theta}$'s and $\hat{\beta}_0$ and denote $\hat{f}^{\lambda}_{k,k^{'},\pi_m}(\boldsymbol{x})=\sum_{i=1}^{n}\hat{\theta}_i\,K(\boldsymbol{x}_i,\boldsymbol{x})+\hat{\beta}_0$.

Once we obtain the estimated functions $\hat{f}^{\lambda}_{k,k^{'},\pi_m}(\cdot)$ for all π_m, we can estimate the pairwise conditional probability $q_{k|(k,k^{'})}(\boldsymbol{x})$ by

$$\hat{q}^{\lambda}_{k|(k,k^{'})}(\boldsymbol{x})=[\min\{\pi_m\colon\hat{f}_{k,k^{'},\pi_m}(\boldsymbol{x})<0\}+\max\{\pi_m\colon\hat{f}_{k,k^{'},\pi_m}(\boldsymbol{x})>0\}]/2,$$

for any $\boldsymbol{x}\in\mathcal{S}$. The notation $\hat{p}^{\lambda}_{k|(k,k^{'})}(\boldsymbol{x})$ indicates that our estimates of the pairwise conditional probability depends on the regularization parameter λ, which need to be tuned.

Tuning parameter selection

As discussed above, the regularization parameter λ needs to be properly tuned to get desired results. For every λ, we estimate the pairwise conditional probability $q_{k|(k,k^{'})}(\cdot)$ by $\hat{q}^{\lambda}_{k|(k,k^{'})}(\cdot)$ as discussed above. The performance of $\hat{q}^{\lambda}_{k|(k,k^{'})}(\cdot)$ can be evaluated by the generalized Kullback-Leibler distance

$$GKL\left(q_{k|(k,k'\ )}, \hat{q}^{\lambda}_{k\big|k,k'\ )}\right)$$

$$= E\left[q_{k|(k,k'\ )}(\boldsymbol{X}) \log \frac{P_{k|(k,k'\ )}(\boldsymbol{X})}{\hat{q}^{\lambda}_{k|(k,k'\ )}(\boldsymbol{X})} + (1 - q_{k|(k,k'\ )}(\boldsymbol{X})) \log \frac{1 - q_{k|(k,k'\ )}(\boldsymbol{X})}{1 - \hat{q}^{\lambda}_{k|(k,k'\ )}(\boldsymbol{X})}\right]$$

$$= -E\left[q_{k|(k,k'\ )}(\boldsymbol{X}) \log\ \hat{q}^{\lambda}_{k|(k,k'\ )}(\boldsymbol{X}) + (1 - q_{k|(k,k'\ )}(\boldsymbol{X})) \log\left(1 - \hat{q}^{\lambda}_{k|(k,k'\ )}(\boldsymbol{X})\right)\right] + C$$

as in Wang et al. (2008), where the expectation is taken with respect to $\boldsymbol{X}$. The constant $C = E\left[q_{k|(k,k'\ )}(\boldsymbol{X}) \log\ q_{k|(k,k'\ )}(\boldsymbol{X}) + (1 - q_{k|(k,k'\ )}(\boldsymbol{X})) \log\left(1 - q_{k|(k,k'\ )}(\boldsymbol{X})\right)\right]$ does not depend on the estimate $\hat{q}^{m}bda_{k|(k,k')}(\cdot)$. Since the GKL depends on the unknown true pairwise conditional probabilities, it is not computable in practice for selecting the regularization parameter.

By noting that $E\left[(R_{k,k'}(Y) + 1)/2|\boldsymbol{X}, Y \in \{k, k'\}\right] = q_{k|(k,k')}(\boldsymbol{X})$ and removing the constant C in GKL, we obtain the empirical generalized Kullback-Leibler distance

$$EGKL(\hat{q}^{\lambda}_{k|(k,k')}) = -\frac{1}{2n_{k,k'}} \sum_{i: y_i = k \text{ or } k'} \left[(1 + R_{k,k'}(y_i)) \log\ \hat{q}^{\lambda}_{k|(k,k')}(\boldsymbol{x}_i) + (1 - R_{k,k'}(y_i)) \log\left(1 - \hat{q}^{\lambda}_{k|(k,k')}(\boldsymbol{x}_i)\right)\right],$$

where $n_{k,k'}$ denotes the number of observations with $y_i = k$ or k'. The EGKL tries to approximate GKL up to a constant C. Thus the EGKL gives a measure of the goodness-of-fit for estimating $q_{k|(k,k')}(\cdot)$ with $\hat{q}^{\lambda}_{k|(k,k')}(\cdot)$.

In our simulation studies, we simulate an independent tuning data set in the same way the corresponding training data have been generated. The EGKL of $\hat{q}^{\lambda}_{k|(k,k')}(\cdot)$ is evaluated over the tuning data set for a grid of a. We select the λ with the smallest $EGKL$ of the tuning set as the best regularization for estimating the pairwise conditional probability for pair $(k, k^{'})$.

Merging pairwise conditional probabilities

By abusing our notation slightly and defining $\hat{p}_{j|(j,j^{'})}(\boldsymbol{x}) = 1$ for any $\boldsymbol{x}$ when $j = j^{'}$, we merge these estimates $\hat{p}_{j|(j,j^{'})}(\boldsymbol{x})$ of the pairwise conditional probabilities to estimate the conditional probability by defining $\hat{p}_{j|j^{'}}(\boldsymbol{x})$ using (5). Here the notation $\hat{p}_{j|j^{'}}(\boldsymbol{x})$ means that we are estimating the conditional probability using class $j^{'}$ as the baseline class.

As discussed above, theoretically it does not matter which class is used as the baseline class while estimating the conditional probability in that all lead to a consistent estimate. However empirically it matters. In our finite-sample implementation, we define our final estimate $\hat{p}_j(\boldsymbol{x})$ as $\hat{p}_{j|j^{'}}(\boldsymbol{x})$ using the class $j^{'}$ with the biggest conditional probability as the baseline class. This is achieved as follows. For each $\boldsymbol{x}$, we compare $\hat{p}_{j|(j,j^{'})}(\boldsymbol{x})$ and $\hat{p}_{j^{'}|(j,j^{'})}(\boldsymbol{x})$ to see which one is bigger for each pair of class j and class $j^{'}$. After all pairs of classes have been compared, we denote $\hat{l}(\boldsymbol{x})$ to be the class with the maximum number of times having a larger estimated pairwise conditional probability. Then our final estimator for the conditional class probability is given by $\hat{p}_j(\boldsymbol{x}) = \hat{p}_{j|\hat{l}(x)}(\boldsymbol{x})$ for any $\boldsymbol{x}$ in the domain $\mathcal{X}$.

Result

In this section, we use several numerical simulation studies to compare our proposed method with some existing competitive methods. To be more precise, we compare with the cumulative logit model (CLM), baseline logit model (BML), kernel multi-category logistic regression (KMLR, Zhu and Hastie, 2005), classification tree (TREE, Breiman et al., 1984), random forest (RF, Breiman, 2001) and the method of Wu et al. (2010). Here the CLM assumes that $\log \frac{\sum_{m=1}^{k} p_k(x)}{1-\sum_{m=1}^{k} p_k(x)} = \beta_{k0} + \boldsymbol{x}^T\beta_k$ for $k = 1,2,\cdots,K-1$. On the other hand, the baseline logit model chooses one class (say class K) as the baseline class and assumes $\log \frac{p_k(x)}{p_K(x)} = \beta_{k0} + \boldsymbol{x}^T\beta_k$ for $k = 1,2,\cdots,K-1$. Whenever a kernel method is involved, we use the Gaussian kernel $R(\boldsymbol{x}_1,\boldsymbol{x}_2) = e^{-\|x_1-x_2\|_2^2/\sigma^2}$, where $\| \boldsymbol{x}_1 - \boldsymbol{x}_2 \|_2$ denotes the 2-norm of $\boldsymbol{x}_1 - \boldsymbol{x}_2$. Ten separate data sets are generated to tune the data width parameter σ among a grid of $\{1,2,3,4,5,6\}\sigma_M/4$, where $\sigma_M = \text{Median}\{\| \boldsymbol{x}_i - \boldsymbol{x}_j \|_2 : y_i \neq y_j\}$ is the median pairwise Euclidean distance. Whenever necessary, a default five-fold cross validation is used to select any tuning parameter that is involved. For the TREE based method, we use the R package “Tree” and its build-in cross validation function is used to prune trees with fold number set to be 10. Similarly we use the build-in tuning for RF provided in the R package.

In each simulation setting, data are generated with the true conditional class probabilities $p_k(\cdot)$ known. Therefore, we can use the following score to evaluate the performance of different methods:

- norm error： $\frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}}\sum_{k=1}^{K} |\, \hat{p}_k(\bar{\boldsymbol{x}}_i) - p_k(\bar{\boldsymbol{x}}_i)|$;
- norm error： $\frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}}\sum_{k=1}^{K}(\hat{p}_k(\bar{\boldsymbol{x}}_i) - p_k(\bar{\boldsymbol{x}}_i))^2$;
- EGKL loss： $-n^{-1}\sum_{i=1}^{n}(\frac{1}{2}(Y_i) + 1) log\hat{p}(X)(1 - \frac{1}{2})(Y_i) + 1)) log(1 - \hat{p}(X)))$.

Here $\{(\bar{\boldsymbol{x}}_i, \bar{y}_i), i = 1,2,\cdots,\bar{n}\}$ denotes the test set and $\bar{n}$ is the size of the test set.

The only parameter we tuned is the penalize parameter λ. λ was tuned among the grid of 10^i with i=-8, -7,...,7,8. We used each λ with the same train data set to compute the model and then model was used to predict the tune data set. The performance of the prediction was measured by the 3 different statistics (1-norm, 2-norm and EGKL). So finally we got 17 numbers for each statistics corresponding to 17 λ. The λ with the smallest value was selected to compute the test data set.

Our new method is implemented with a weight grid $\pi_i = i/20$ for $i = 0,1,\cdots,20$. For each example, the simulation is repeated for 100 times. We report the average errors over 100 replications and the corresponding standard deviations. Denote the size of training set by n. A test set of size $\bar{n} = 10n$ are generated to calculate the test error and the standard deviations.

Example 1 We generate a three-class linear learning example as follows: 1) Y is uniformly generated from {1,2,3}; 2) Given $Y = y$, we sample a two-dimension predictor $\boldsymbol{X}$ from a normal distribution $N(\mu(y), \Sigma)$, where $\mu(y) = (cos(2y\pi/3), sin(2y\pi/3))^T$ and $\Sigma = 0.7^2\boldsymbol{I}_2$, where $\boldsymbol{I}_2$ is a 2×2 identity matrix. The sample size is $n = 400$. The performance of different methods are summarized in Table 4.1, where the top half shows the result from training data and the bottom half shows the result from testing data.

Example 2 In this example, we consider a three-class non-linear example. The data are generated using the following two steps: 1) Generate independent X_1 and X_2 from Uniform $[-3,3]$ and $[-6,6]$, respectively; 2) Conditional on $\boldsymbol{X} = \boldsymbol{x} = (x_1, x_2)^T$, define $f_1(\boldsymbol{x}) = -x_1 + 0.1x_1^2 - 0.05x_2^2 + 0.1$, $f_2(\boldsymbol{x}) = -0.2x_1^2 + 0.1x_2^2 - 0.2$ and $f_3(\boldsymbol{x}) = x_1 + 0.1x_1^2 - 0.05x_2^2 + 0.1$. Let $p_k(\boldsymbol{x}) = P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = e^{f_k(\boldsymbol{x})}/(\sum_{m=1}^{3} e^{f_m(\boldsymbol{x})})$ for $k = 1,2,3$. Given $\boldsymbol{X} = \boldsymbol{x}$, the label Y takes the value k with probability $p_k(\boldsymbol{x})$ for $k = 1,2,3$. We set the training data size to be $n = 1000$. The result is listed in Table 4. 2.

For this example, since it is a nonlinear case, we try three different methods to achieve nonlinear multi-class soft classification. The first one is basis expansion, in which we expand the two dimension predictor $\boldsymbol{x} = (x_1, x_2)$ into a five dimension predictor $\widetilde{\boldsymbol{x}} = (x_1^2, x_2^2, x_1x_2, x_1, x_2)$ and then solve (3) with f being a linear function of $\widetilde{\boldsymbol{x}}$. The second one is to use the kernel trick with the Gaussian kernel $R(x_{\mathbf{1}}, x_{\mathbf{2}}) = e^{-||x_{\mathbf{1}}-x_{\mathbf{2}}||_2^2/\sigma^2}$ with σ is the width of the Gaussian, here we set $\sigma^2 = 2 * \sigma_M^2$, where σ_M^2 is the median of the Euclidean distances from each positive example to the nearest negative example. The third one is use use the kernel trickwith the Spline kernel Gu et al. (2000).

Example 3 In the previous examples, the BLM is the true model so it shows a as good performance as our method does. In this example, we design an experiment that does not fit the parametric method. The two-dimension predictor $\boldsymbol{X}$ is uniformly sampled from a disc $\{\boldsymbol{x}: x_1^2 + x_2^2 \leq 100\}$.Define $h_1(\boldsymbol{x}) = -5x_1\sqrt{3} + 5x_2$, $h_2(\boldsymbol{x}) = -5x_1\sqrt{3} - 5x_2$ and $h_3(\boldsymbol{x}) = 0$. Then we get a set of function by a transformation $f_k(\boldsymbol{x}) = \Phi^{-1}(T_2(h_k(\boldsymbol{x})))$, where $\Phi(\cdot)$ is the cumulative distribution function(cdf) of the standard normal distribution and $T_2(\cdot)$ is the cdf of t distribution with degree of freedom 2. Similar with example 2, we get set the probability $p_k(\boldsymbol{x}) = P(Y = k|fX = \boldsymbol{x}) = \exp\ f_k\,(\boldsymbol{x})/(\sum_{m=1}^{3} \exp\,(f_m(\boldsymbol{x})))$ for $k = 1,2,3$. The sample size is 400. The result is listed in Table 4.3.

Example 4 The previous examples are all 3-class problem. In this example, we simulate a data with 4 classes and nonlinear true classification boundaries. Data are sampled in two steps: 1). Generate the two dimension predictor $\boldsymbol{X}$ uniformly for the disc $\{\boldsymbol{x}: x_1^2 + x_2^2 \leq 16\}$, similarly in example 2, we define $f_1(\boldsymbol{x}) = -|x_1^2 + x_2^2|$, $f_2(\boldsymbol{x}) = -|x_1^2 + x_2^2 - 1.5^2|$, $f_3(\boldsymbol{x}) = -|x_1^2 + x_2^2 - 2.5^2|$ and $f_4(\boldsymbol{x}) = -|x_1^2 + x_2^2 - 3.5^2|$. Let $p_k(\boldsymbol{x}) = P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \exp\, f_k\,(\boldsymbol{x})/(\sum_{m=1}^{4} \exp\,(f_m(\boldsymbol{x})))$ for $k = 1,2,3,4$. Given $\boldsymbol{X} = \boldsymbol{x}$, the label Y takes the value k with probability $p_k(\boldsymbol{x})$ for $k = 1,2,3,4$. We study an example with sample size 400. In this example, we apply the basic expansion method as well. Results are reported in Table 4.4.

Example 5 The example 5 is to show our method's ability in dealing with classification problem with even larger number of classes. Here we generate a 5 class linear example with sample size 500. Similarly as in Example 1, the data is sampled in two steps: 1). Y is uniformly generated from 1 to 5; 2). Given $Y = y$, we sample a two-dimension predictor $\boldsymbol{X}$ from a normal distribution $N(\mu(y), \Sigma)$, where $\mu(y) = (\cos\,(2y\pi/5), \sin\,(2y\pi/5))^T$ and $\Sigma = \boldsymbol{I}_2$, $\boldsymbol{I}_2$ is a 2×2 identity matrix. The result is listed in Table 4.5.

Example 6 Previous examples are focusing on 2 dimension data. Here, we applied our method to a higher dimension data. For simplicity, we simulated the example 6 based on example 3. Here, we considered a ten dimension predictor $\boldsymbol{X}$ was uniformly sampled froma sphere $\{\boldsymbol{x}: x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2 + x_7^2 + x_8^2 + x_9^2 + x_10^2 \leq 100\}$. And similarly, $h_1(\boldsymbol{x}) = -x_1\sqrt{3} - x_2\sqrt{3} - x_3\sqrt{3} - x_3\sqrt{3} - x_5\sqrt{3} + x_6 + x_7 + x_8 + x_9 + x_10$, $h_2(\boldsymbol{x}) = -x_1\sqrt{3} - x_2\sqrt{3} - x_3\sqrt{3} - x_3\sqrt{3} - x_5\sqrt{3} - x_6 - x_7 - x_8 - x_9 - x_10$ and $h_3(\boldsymbol{x}) = 0$. Then we applied the same way to generate $f_k(\boldsymbol{x})$ function and the probability function. The sample size is 500. The result are shown in Table 4.6.

Real Example

In this section, we use two real data sets to demonstrate our new method. Two different scenarios are considered: one with the number of predictors much larger than the sample size ($p \gg n$) and the other with the number of predictors smaller than the sample size ($p < n$).

***p* ≫ *n*.**

Nowadays, mRNA expressions of thousands of genes can be monitored simultaneously though the DNA microarray technology at a reasonable cost. In this section, we apply our method on the children cancer data set in Khan et al. (2001). Using the cDNA gene expression profiles, this data set classified the small round blue cell tumors (SRBCTs) of childhood into 4 classes: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma(NHL), and the Ewing family of tumors (EWS). After filtering, 2308 gene profiles out of 6567 genes are given in the data set, available at http://research.nhgri.nih.gov/microarray/Supplement/. In this data set, the sample size of training set is 63 and test set is of size 20. The distribution of the four distinct tumorcategories in the training and test sets is given in Table 4.7. Here, Burkitt lymphoma (BL) is a subset of NHL.

We first standardize the data set by simply linear transformation on the training data. Specifically, we using the following formula to standardize the gene expression value $\tilde{x}_{gi}$ corresponding to the gth gen of subject i to get x_{gi}:

$$x_{gi} = \frac{\tilde{x}_{gi} - \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_{gj}}{sd(\tilde{x}_{g1}, \cdots, \tilde{x}_{gj})}$$

After the standardization, all genes are ranked by the marginal relevance in the class separation, using a criterion in Dudoit et al. (2002). Specifically, the relevance measure of the gene g is calculated to be the ratio of between classes sum of squares to within class sum of the squares as follows:

$$R(g) = \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} I\,(y_i = k)(\bar{x}_{g\cdot}^{(k)} - \bar{x}_{g\cdot})^2}{\sum_{i=1}^{n} \sum_{k=1}^{K} I\,(y_i = k)(x_{gi} - \bar{x}_{g\cdot}^{(k)})^2},$$

where n is the sample size of the training set, $\bar{x}_{g\cdot}^{(k)}$ is the average expression level of gene g for class k observations, and $\bar{x}_{g\cdot}$ is the overall mean expression level of gene g in the training set. Our goal is to exam the performance of the method under the situation that $p \gg n$, we select the top 100 genes and the bottom 100 genes as the covariates according the relevance measure R. The result is list in Table 4.8. Figure 4.6 shows the probability vectors $(Pr(EWS), Pr(BL), Pr(NB) and Pr(RMS))$ for the test data. For example, the purple bar represent the EWS samples, the ideal probability vector is (1, 0, 0, 0). Among the 20 samples, 19 of them are quite close the ideal vector. Only 1 NB sample is mis-predicted to be RMS.

$\boldsymbol{n > p.}$

Comparing with three other methods (Tree, Multi logistic and Random forest), we consider 3 examples to show our method’s performance, zip code, ecoli data and Yeast data. The zip code are normalized handwritten digits which are automatically scanned from envelopes by the U.S. Postal Service. Since the original scanned digits have binary different sizes and orientations; we desalinate and normalized the image, resulting in 16 x 16 gray-scale images (Le Cun et al., 1990). The whole sample size of the train set is 7291 and the one of the test set is 2007. As the data is quite huge, we first generate a sub data set by picking up 3 digits classes (digit 3, 6 and 9) to test the methods, then we use the whole set to see each method's ability to deal with large data set. The objective of ecoli data set is to predict the cellular localization sites of E.coli proteins (Horton and Nakai, 1996). In the original data, there are 8 different cellular sites, since some of classes have fewer observations than other, we combine 4 classes into 1 class so finally we use 4 classes to test the method. The Yeast data set is similar to the E.coli data, which is to determine the cellular localization of the yeast proteins (Horton and Nakai, 1996). There are 10 different sites, which include: CYT (cytosolic or cytoskeletal); NUC (nuclear); MIT (mitochondrial); ME3 (membrane protein, no N-terminal signal); ME2 (membrane protein, uncleaved signal); ME1 (membrane protein, cleaved signal); EXC (extracellular); VAC (vacuolar); POX (peroxisomal) and ERL (endoplasmic reticulum lumen). Combing 6 classes which have fewer observations into 1 class, we treat the Yeast data set as a 5 classification problem. More details are list in Table 4.9.

For each data set, the training set is split into 2 parts: one for train set (50%） and the other for the tune set (50%). The result is list in Table 4.10.

References

Agresti, A. and Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. The American Statistician, 52 119-126.

Breiman, L. (2001). Random forests. Machine Learning, 45 5-32.

Breiman, L., H. Friedman, J., A. Olshen, R. and J. Stone, C. (1984). Classification and regression trees. Wadsworth Publishing Company, Belmont, California, U.S.A.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine Learning, 20 273-297.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 7 673-679.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. Journal of Mathematical Analysis and Applications, 33 82-95.

Lee, Y., Lin, Y. and Wahba, G. (2004). Multi-category support vector machines, theory, and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association, 99 67-81.

Lin, Y. (2002). Support vector machines and the Bayes rule in classification. Data Mining and Knowledge Discovery, 6 259-275.

Liu, Y. (2007). Fisher consistency of multi-category support vector machines. In Eleventh International Conference on Artificial Intelligence and Statistics. 289-296.

Platt, J. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, chap. Advances in large margin classifiers. MIT Press.

Vapnik, V. (1998). Statistical Learning Theory. Wiley, New York.

Wahba, G. (1990). Spline models for observational data. In CBMS-NSF Regional Conference Series. SIAM Society for Industrial and Applied Mathematics.

Wang, J., Shen, X. and Liu, Y. (2008). Probability estimation for large margin classifiers. Biometrika, 95 149-167.

Weston, J. and Watkins, C. (1999). Proceedings of the 7th European Symposium on Artificial Neural Networks, chap. Support vector machines for multi-class pattern recognition. Bruges, Belgium, 219-224.

Wu, Y. and Liu, Y. (2007). Robust truncated-hinge-loss support vector machines. Journal of the American Statistical Association, 102 974-983.

Wu, Y., Zhang, H. H. and Liu, Y. (2010). Robust model-free multiclass probability estimation. Journal of the American Statistical Association. 105 424-436.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research, 5 1225-1251.

Zhu, J. and Hastie, T. (2005). Kernel logistic regression and the import vector machine. Journal of Computational and Graphical Statistics, 14 185-205.

Table 4.1 Simulation result from Example 1

		Proposed method		Wu et al. (2010)	KMLR	TREE	BLM(Oracle)
		GKL	EGKL				
Training	1-norm	10.36(1.53)	10.75(1.72)	22.05(4.42)	52.95(2.74)	26.72(3.47)	6.17(1.96)
	2-norm	0.61(0.29)	0.69(0.31)	5.15(2.19)	11.90(1.13)	5.61(1.21)	0.36(0.23)
	EGKL	2.31(0.55)	2.59(2.86)	8.81(2.98)	24.12(1.73)	Inf(NaN)	0.77(0.47)
Testing	1-norm	10.14(1.57)	10.70(1.53)	22.37(4.36)	53.64(2.58)	27.48(3.34)	6.19(1.95)
	2-norm	0.56(0.22)	0.64(0.24)	5.27(2.09)	12.38(1.12)	5.99(1.21)	0.36(0.23)
	EGKL	2.11(0.45)	2.17(2.20)	9.02(2.81)	24.83(1.66)	Inf(NaN)	0.78(0.48)

Table 4.2 Simulation result from Example 2

		Proposed method						Wu et al. (2010)	KMLR	TREE	BLM (Oracle)
		Basic expansion		Gaussian		Spline					
		GKL	EGKL	GKL	EGKL	GKL	EGKL				
Training	1-norm	21.47(3.65)	23.97(4.97)	26.45(4.27)	28.35(4.43)	30.19(6.18)	29.93(6.39)	41.62(2.91)	54.47(5.54)	56.48(11.95)	18.31(4.76)
	2-norm	3.25(0.99)	4.09(1.65)	4.63(1.29)	4.78(1.36)	5.45(2.10)	5.64(2.27)	9.13(1.27)	13.77(2.22)	19.69(5.57)	2.79(1.54)
	EGKL	7.06(1.77)	8.98(9.09)	8.63(2.11)	8.67(2.13)	10.13(3.36)	10.43(3.40)	16.21(1.82)	25.23(3.40)	Inf(NaN)	7.95(6.90)
Testing	1-norm	24.48(4.71)	24.09(6.24)	31.45(5.17)	30.79(5.37)	30.46(5.05)	31.06(5.11)	45.78(3.46)	59.32(4.78)	60.08(10.57)	19.38(5.21)
	2-norm	4.08(1.58)	5.01(2.57)	5.88(1.98)	5.91(2.06)	5.56(1.64)	5.43(1.64)	10.89(1.69)	16.38(2.25)	22.23(4.75)	3.19(1.90)
	EGKL	9.06(2.94)	10.90(10.55)	10.92(3.26)	10.83(3.43)	10.25(2.83)	9.69(10.25)	18.69(2.43)	28.92(3.24)	Inf(NaN)	9.11(9.23)

Table 4.3 Simulation result from Example 3

		Proposed method		Wu et al. (2010)	KMLR	TREE	BLM
		GKL	EGKL				
Training	1-norm	17.36(2.55)	18.93(3.29)	21.67(2.76)	59.99(2.16)	21.75(3.48)	30.89(1.42)
	2-norm	0.90(0.09)	0.93(0.11)	4.55(1.26)	14.96(0.98)	5.79(1.32)	6.84(0.44)
	EGKL	6.39(1.02)	7.14(7.19)	10.14(3.02)	29.32(1.47)	Inf(NaN)	12.53(0.56)
Testing	1-norm	18.29(2.58)	19.81(3.55)	21.89(2.56)	63.06(1.89)	24.44(3.29)	31.02(1.07)
	2-norm	0.30(0.03)	0.33(0.04)	4.70(1.28)	16.66(0.94)	7.69(1.35)	6.85(0.27)
	EGKL	6.97(1.41)	7.75(7.67)	10.34(2.85)	31.77(1.39)	Inf(NaN)	12.72(0.40)

Table 4.4 Simulation result from Example 4.

		Proposed method		RF	TREE	KMLR	BLM(Oracle)
		GKL	EGKL				
Training	1-norm	23.40(1.48)	23.40(1.48)	24.58(2.13)	36.95(5.64)	129.53(1.33)	10.86(1.62)
	2-norm	4.05(0.68)	4.04(0.68)	8.23(0.99)	13.31(2.57)	57.21(1.14)	1.53(0.44)
	EGKL	10.76(1.20)	10.73(10.73)	Inf(NaN)	Inf(NaN)	107.69(1.78)	4.26(1.19)
Testing	1-norm	25.82(1.55)	25.83(1.55)	29.81(1.81)	43.87(5.43)	131.98(8.19)	11.37(2.13)
	2-norm	5.31(0.79)	5.32(0.79)	8.57(1.00)	18.28(2.43)	86.12(12.35)	1.68(0.64)
	EGKL	12.80(1.43)	12.81(12.81)	Inf(NaN)	Inf(NaN)	247.83(91.00)	4.63(2.09)

Table 4.5 Simulation result from Example 5.

		Proposed method		RF	TREE	KMLR	BLM(Oracle)
		GKL	EGKL				
Training	1-norm	15.57(1.91)	15.55(2.19)	61.85(1.35)	38.79(3.16)	105.18(1.60)	6.93(1.65)
	2-norm	0.89(0.25)	1.02(0.30)	20.17(1.19)	7.36(1.24)	32.71(0.98)	0.25(0.12)
	EGKL	3.51(0.52)	3.74(3.79)	Inf(NaN)	Inf(NaN)	76.01(2.39)	0.57(0.24)
Testing	1-norm	15.28(1.59)	15.99(2.08)	41.46(1.76)	40.10(3.18)	105.55(1.46)	7.33(1.72)
	2-norm	0.96(0.22)	1.11(0.34)	8.95(0.84)	7.83(1.24)	32.87(0.90)	0.28(0.13)
	EGKL	3.58(0.41)	3.99(4.02)	Inf(NaN)	Inf(NaN)	76.41(2.22)	0.63(0.27)

Table 4.6 Simulation result from Example 6

		Proposed method		KMLR	TREE	BLM
		GKL	EGKL			
Training	1-norm	20.69(1.61)	22.15(1.94)	63.82(2.45)	22.38(3.86)	24.36(1.40)
	2-norm	0.83(0.05)	0.86(0.06)	12.84(0.86)	5.35(1.19)	4.12(0.43)
	EGKL	6.04(0.63)	6.49(6.49)	25.94(1.16)	Inf(NaN)	7.60(0.75)
Testing	1-norm	21.63(1.59)	22.61(1.84)	67.49(2.43)	22.46(2.59)	25.31(1.35)
	2-norm	0.28(0.02)	0.28(0.02)	13.61(0.83)	5.83(1.51)	4.55(0.51)
	EGKL	6.65(0.76)	6.88(6.88)	27.38(1.72)	Inf(NaN)	8.23(0.88)

Table 4.7 Class distribution of the mircoarray data

Data set	NB	RMS	BL	EWS	Total
Training	12	20	8	23	63
Test	6	5	3	6	20

Table 4.8 Classification error of the microarray data using the top 200 genes.

Proposed method	MSVM	L1 MSVM	Supnorm MSVM	Adaptive Supnorm MSVM
1	0	1	1	1

Table 4.9 Real data information. 3 Real data sets are used: the zip code data, Ecoli data set and Yeast data set. For the Zip code data set, a subset just contains the digits: 3, 6 and 9 and the whole data are used. For each data set, the number of class, the sample size of training set and the sample size for test set is listed.

		No. of attributions	No. of training set	No. of test set
Zip code	3 6 9	3	1966	513
	full	10	7291	2007
Ecoli		4	222	110
Yeast		5	989	495

Table 4.10 Error rate of real data using different methods.

		Proposed	RF	TREE	BLM
Zip code	3 6 9	1.17	1.37	8.18	3.31
	full	7.10	5.87	27.8	58.6
Ecoli		14.7	14.4	18.9	29.8
Yeast		37.3	36.1	41.2	39.0

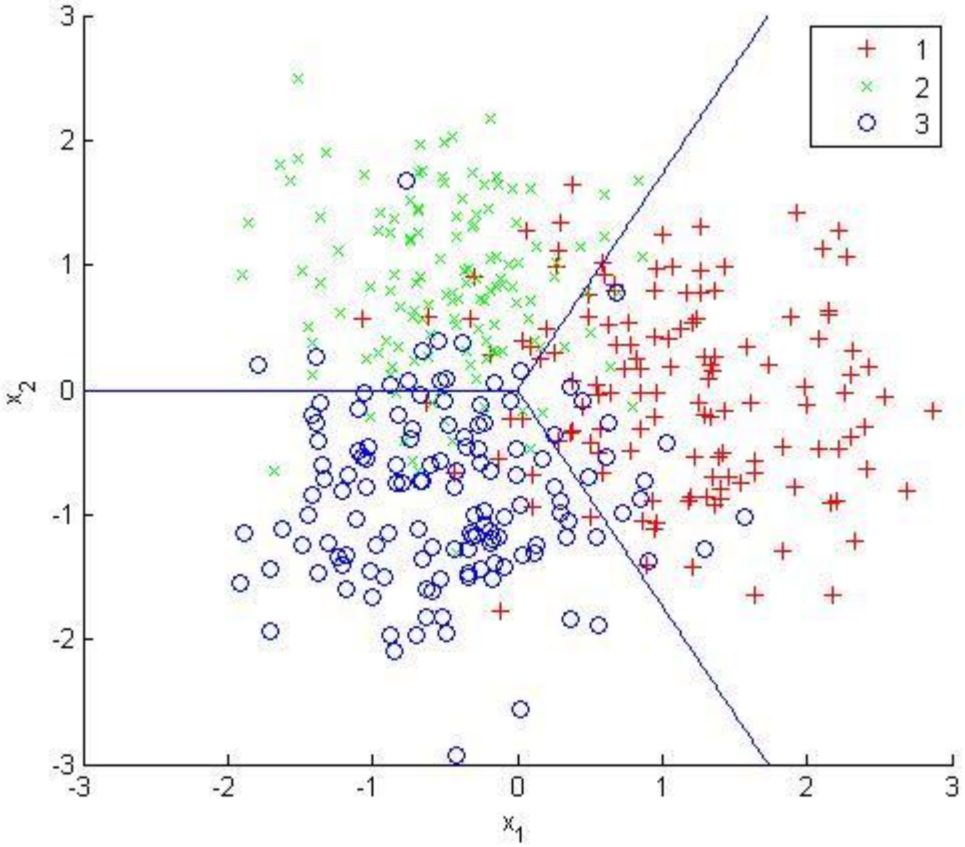


Figure 4.1 Data distribution of example 1. Data is shown as a two dimension data point $\{x_1, x_2\}$. Three classes with different colors (red, green and blue) are available. The blue lines are the theoretical boundaries of different classes.

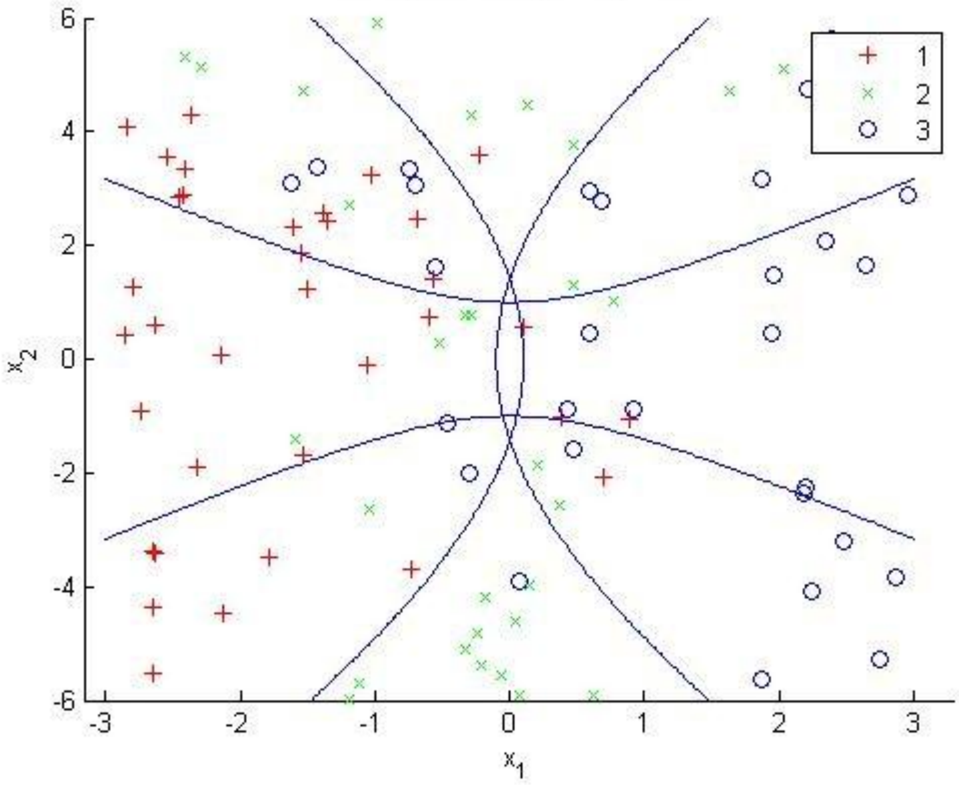


Figure 4.2 Data distribution of example 2.Data is shown as a two dimension data point $\{x_1, x_2\}$. Three classes with different colors (red, green and blue) are available. The blue curves are the theoretical boundaries of different classes.

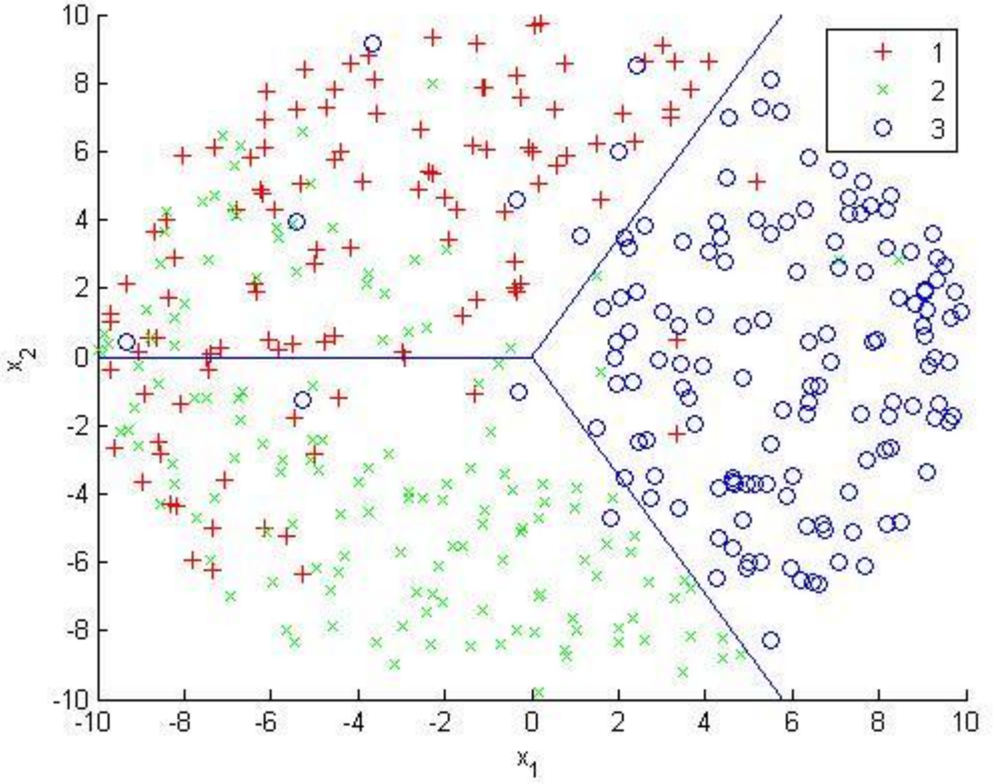


Figure 4.3 Data distribution of example 3 -Data is shown as a two dimension data point $\{x_1, x_2\}$. Three classes with different colors (red, green and blue) are available. The blue lines are the theoretical boundaries of different classes according to the underlying probability functions

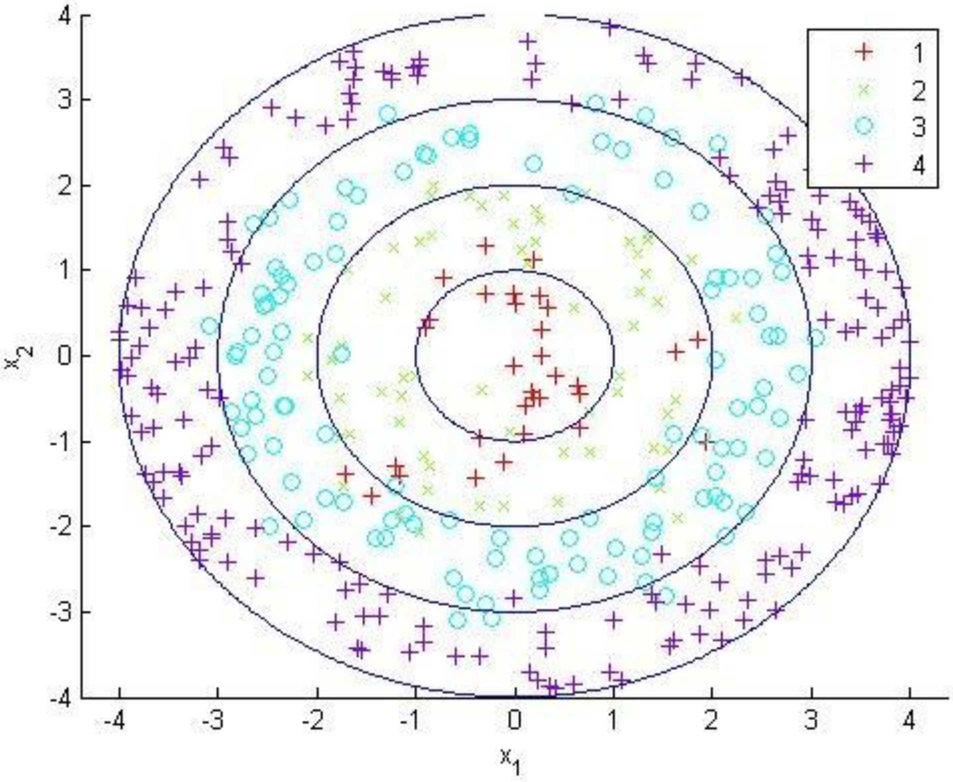


Figure 4.4 Data distribution of example 4 Data is shown as a two dimension data point $\{x_1, x_2\}$. Four classes with different colors (red, green, blue and purple) are available. The blue curves are the theoretical boundaries of different classes

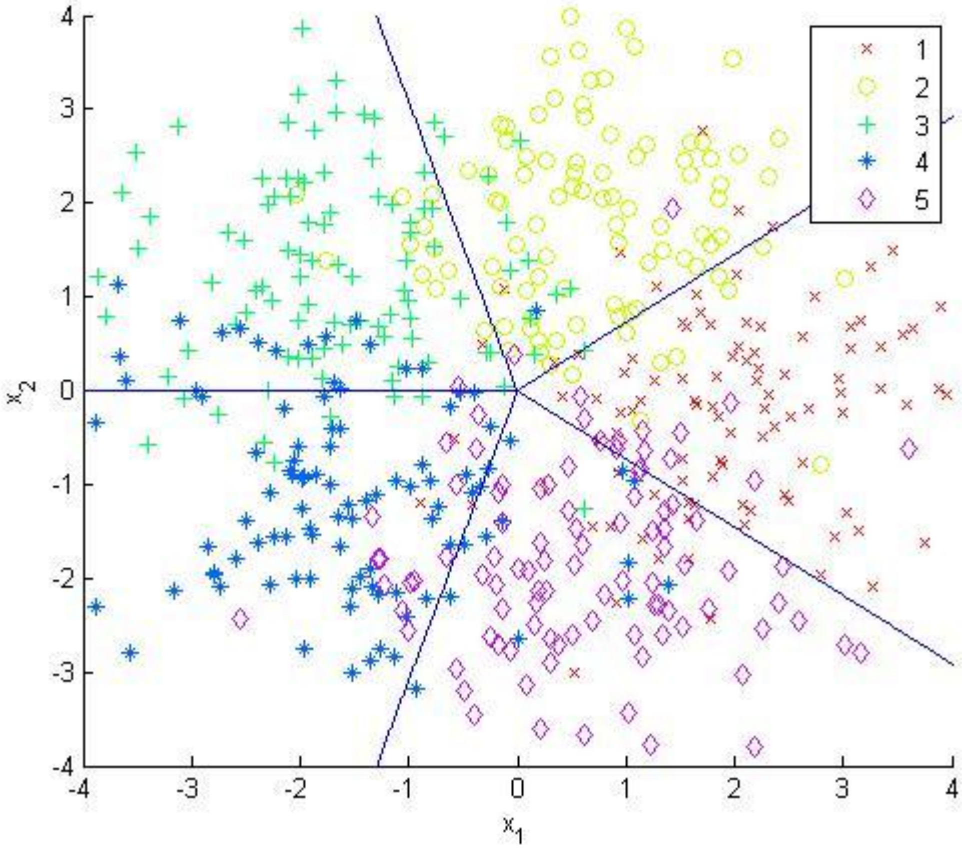


Figure 4.5 Data distribution of example 5 Data is shown as a two dimension data point $\{x_1, x_2\}$. Five classes with different colors (red, yellow, green blue and purple) are available. The blue lines are the theoretical boundaries of different classes.

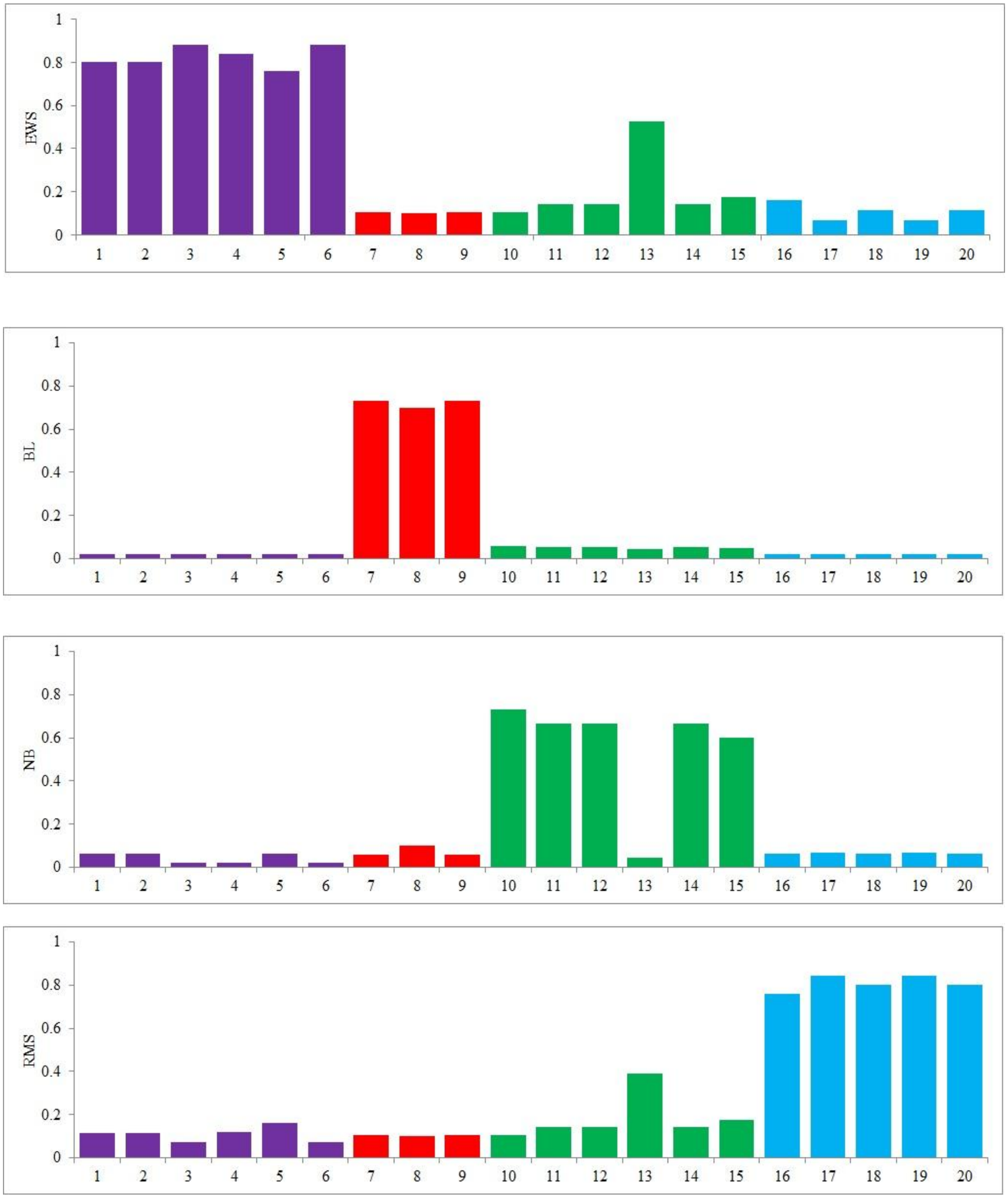


Figure 4.6 Probailibty estimates using real data There are 4 figures, corresponding to 4 cancer subtypes. For each sub figure, the probabilties of each individual belong to the corresponding subtypes are shown as histogram. For each indiviudal, the bar's color repsents its real subtype (Purple: EWS, Read: BL, Green: NB, Blue: RMS).