PROBABILISTIC MODELS FOR AGGREGATE ANALYSIS OF NON-GAUSSIAN DATA IN BIOMEDICINE

A Dissertation

by

MENG LU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Xiaoning Qian
Co-Chair of Committee,	Jianhua Z. Huang
Committee Members,	I-Hong Hou
	Nicholas G. Duffield
Head of Department,	Miroslav M. Begovic

December 2015

Major Subject: Computer Engineering

ABSTRACT

Aggregate association analysis is a popular way in genome-wide association studies (GWAS) that analyzes the association between the trait of interest and regions of functionally related genes, which has the advantage of capturing the missing heritability from the joint effects of correlated genetic variants while providing a better understanding of disease etiology from a systematic perspective. However, traditional methods lose their power for biomedical data with non-Gaussian data types. We proposed innovative statistical models to derive more accurate aggregated signals to enhance the power by taking account of the special data types. Based on general exponential family distribution assumptions, we developed supervised logistic PCA and supervised categorical PCA for pathway based GWAS and rare variant analysis. A general framework, sparse exponential family PCA (SePCA), is further developed for aggregate analyses for various types of biomedical data with good interpretation. We derived an efficient algorithm to find the optimal aggregated signals by solving its equivalent dual problem with closed-form updating rules. SePCA is extended for aggregate association analysis in hierarchical levels for better biological interpretation, from groups to individual variables. Both simulation studies and real world applications have demonstrated that our methods can achieve higher power in association analysis and population stratification by taking good care of the correlations among the non-Gaussian variables in biomedical data. Another analytic issue in aggregate analysis is that biomedical data often have special stratified data structures due to the experiment design to solve confounding issues. We extended SePCA to low-rank and full-rank matched models to take account of the stratified data structures. The simulation study has demonstrated their capability of recon-

structing more relevant PCs for the signals of interest compared to standard ePCA. A sparse low-rank matched PCA model outperforms the existing Bayesian methods in detecting differentially expressed genes for a benchmark spike-in gene study with technical replicates. In summary, our proposed statistical models for non-Gaussian biomedical data can derive more accurate and robust aggregated signals that help reveal underlying biological principles of human disease. Other than bioinformatics, these probabilistic models also have rich applications in data mining, computer vision, and social science areas.

DEDICATION

To my love Yan Liu and our families.

ACKNOWLEDGEMENTS

First, I would like to thank my PhD advisors, Prof. Xiaoning Qian and Jianhua Z. Huang, for their insightful advice and support during these past five years. This dissertation would not have been possible without their support and encouragement. I also would like to thank the members of my PhD committee, Professor I-Hong Hou and Nicholas G. Duffield, for providing guidance and support. I am grateful to our collaborators, Prof. Shuai Huang, David Hadley and Hye-Seung Lee for their support and valuable discussions about the research.

I thank the nice present members in our GSP lab: Navadon Khunlertgit, Hyundoo Jeong, Xingde Jiang, Zhengyu Guo, Priyadharshini S Venkat, Chung-Chi Tsai, Siamak Zamani, Shaogang Ren, and Yijie Wang. I also thank Ting Chen, an alumni of GSP lab, who has been so helpful when I first arrived at Texas A&M. I also would like to thank the dear friends I have made at University of South Florida: Bingxiong Lin, Fillipe Souza, Yingying Zhang, Xiang Zuo, and Mu Zhou. I will forever be grateful for their help and miss the fun time we have spent together.

I also thank my wonderful roommate, Jingjia Li, for her accompany and lots of fun she brought to me during the most difficult time I have been through. I also would like to thank my old roommates, Dongping Du and Ying Tan. I will never forget the many wonderful dinners and fun activities I have done with my roommates. Best wishes to all of them.

Finally, I thank my husband Yan Liu, our parents and my sister for their endless love, support, encouragement, and advice. Without their love and support, I would not be here, and could not accomplish anything in my life.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Motivation

One critical challenge in biomedicine is to understand the association between genotypes and phenotypes. For example, genome-wide association studies (GWAS) aim to detect the association between genetic variants and traits of interest like disease phenotypes. They have been very successful in identification of susceptibility loci through single-marker-based tests of the association of each individual single nucleotide polymorphisms (SNP) marker with diseases [54]. Limited by small sample size, however, the detected common variants at these loci have been found with individually only modest effects [10] that can only explain a small portion of the heritability of complex diseases. Several possible explanations for the missing heritability could be: 1) the sample size is very small which could cause the irreproducibility of the detected SNPs [33, 30] and it is even worse for those with weak effects; 2) the effect size may be underestimated due to incomplete disequilibrium between the causal genetic variants and the genotyped SNP markers; 3) the effects from rare variants are usually ignored due to the low minor allele frequency; 4) these suspected SNPs with weak effects may cooperatively work together with strong joint effects. The last explanation is considered from a systematic genetics prospective which prompts the development of new analytic approaches to unravel the relationships and interactions among groups of genetic variants underlying the complex diseases based on annotations and prior knowledge of functionalities of gene networks. Aggregate association analysis is a popular way to achieve these goals by analyzing the statistical significance of functional regions (e.g. pathways and networks) that are composed of multiple functionally related genes. It will not only

play a role in explaining the remaining heritability that GWAS fail to explore, but also provide a good understanding on which gene has significant association, what is the biological mechanism, and how they interact with each other from a systematic perspective. Moreover, aggregate association analysis could also enhance the power of rare variant analysis to address the issue in the third explanation for the missing heritability.

To estimate the statistical significance for the joint effects from multiple SNPs in a functional region, testing the combined signals representing the functional region is an appealing approach in aggregate association analysis due to its ability to take account of correlations among SNPs and the low degree of freedom of the test statistic. The major concern in this approach is how to summarize the optimal combined signals for a set of SNPs. Principal component analysis (PCA) is an attempting approach which has been applied for deriving combined signals [11, 4]. However, it may lose power in aggregating SNP data or other genomic data which have special data types or sometimes even has complex data structure determined by the experimental design.

1.2 Problem statement

This thesis proposes several probabilistic models to derive combined signals that are able to take account of the non-Gaussian data types or stratified data structures in biomedical data, with the aim to enhance the power of aggregate association analysis in the applications to pathway based analyses, rare variant analyses, and matched case-control studies. To take care of the special data types, we develop a new general dimension reduction method with the capability of variable selection, sparse exponential family PCA (SePCA), suitable for any data following exponential family distributions, including SNP data. PCA is usually suitable for dimension reduction

of continuous data by making Gaussian assumption of the data distribution, which is however not an appropriate assumption for recently emerging big omics data, such as SNP data containing categorical values. This motivates us to derive the aggregated signals by SePCA to handle mixed types of biomedical data more appropriately. As opposed to exponential family PCA (ePCA), it allows for variable selection during dimension reduction, which provides a better understanding of the results and helps focus on the informative variables, especially for high dimensional non-Gaussian data. SePCA has wide applications in data mining, image processing and social science other than bioinformatics.

SePCA method treats the dimension reduction as a maximum likelihood problem where the data likelihood is derived based on exponential family distributions in which the canonical parameters are further approximated by a latent variable decomposition. Instead of directly solving the primal problem which is a non-convex problem with orthogonal constraints and suffers from many local optima, we propose an efficient algorithm to find the optimal solutions by solving its equivalent dual problem via alternate updating with closed-form update rules. Our algorithm is more scalable to high-dimensional data due to its efficiency compared to the existing methods handling non-Gaussian data. In addition, SePCA can also be easily extended for aggregate association analysis by involving outcome information to build an integrate model, named as supervised SePCA, to study the joint effects from an optimal subset of predictors. The dual transformation can also be applied to solve this optimization problem due to the similar form of objective functions in general linear models and ePCA. Supervised SePCA allows for association analysis in hierarchical levels, from groups to individual variables, in an integrate framework that takes good care of the correlations among variables.

For the second issue arise in deriving aggregated signals, biomedical data often

have special data structures, which are determined in the design stage of studies to address confounding issues. Basically, they are stratified data matched on the confounding factors, which requires more complex approaches to perform the usual analyses such as dimension reduction and regression analysis. Directly applying standard PCA/ePCA on matched data could result in biased PCs whose variance are dominated by their confounding factors, which motivates us to derive exponential family matched PCA methods to adjust the biased PCs by eliminating the confounding effects. We propose low-rank and full rank models to derive unbiased PCs by explicitly modeling and estimating the confounding effects with the low-dimensional projections. Similar strategies in solving ePCA problem can also be applied here to find optimal solutions. The exponential family matched PCA methods are exempt from tedious steps to choose parameters that are usually concerns for the Bayesian methods. They are quite general to be applied in many areas that have demands in dimension reduction on stratified data.

1.3 Organization

The rest of the thesis is organized as follows.

In section 2, we review two categories of existing methods for aggregate association analysis for GWAS. The advantages and disadvantages for these methods are discussed, which suggests that testing combined signals is a promising approach for aggregate association analysis.

In section 3, we propose SePCA method and develop an efficient algorithm to solve a dual problem of the primal optimization problem. The results from both simulation experiments and real-world applications have demonstrated the superiority of our sparse exponential family PCA in reconstruction accuracy and computational efficiency over a previous sparse PCA and a previous sparse logistic PCA algorithm.

In section 4, we focus on studying the applications of two cases of SePCA: logistic PCA and categorical PCA in aggregate association analysis of SNP data represented by two different forms. In both studies, the generalized PCs are regarded as the combined signals, and are further refined in a supervised framework to achieve the highest power for association analysis. The superiority of logistic PCA and categorical PCA for aggregate association analysis are demonstrated by both the simulation study of pathway based GWAS and real world application in detecting significantly associated pathways for Crohn's disease.

In section 5, we investigate the performance of logistic PCA in aggregate association analysis of rare variants. The performance is assessed via both simulation study and GAW17 study which have demonstrated that LPCA could extract more accurate combined signals leading to higher power in detecting associated rare variants, in comparison of other collapsing methods.

In section 6, we propose exponential family matched PCA methods to derive aggregated signals for matched SNP data and gene expression data with technical replicates. Our low-rank and full-rank models could lead to more accurate reconstruction of PC loadings compared to standard PCA/LPCA when applied on simulated matched Gaussian/binary data sets. A sparse version of low-rank matched PCA is also proposed and applied for detecting differentially expressed genes for a benchmark microarray data with technical replicates. Our method could detect all the spike-in genes in the benchmark data, which however could not be achieved by a previous Bayesian method.

In section 7, we conclude the thesis by summarizing the major contributions of the thesis and listing the directions for future work.

2. LITERATURE REVIEW

2.1 Genome-wide association study

Genome-Wide Association Study (GWAS) has been one of main critical efforts to address genotype-phenotype association powered by next-generation sequencing profiling techniques. Whole genome single nucleotide polymorphisms (SNPs) in different individuals can be examined to see if any genetic variant is associated with a trait of interest. Traditional GWAS mainly focus on single-locus based analysis concerning individual SNPs, which has successfully detected many susceptible SNPs that however explains only a small portion of the heritability of diseases. Those missing heritability could come from rare variants or a set of SNPs with small individual effects but having strong joint effects or interactions with genetic or environmental factors [53].

Many alternative or complementary approaches have been proposed in recent decades to deal with these limitations of single-marker-based analysis. For example, genotype imputation is used to boost the number of SNPs tested for association in order to increase the power of GWA studies, the ability to fine-map causal variant and facilitate meta-analysis [55]. Multi-locus analysis methods are also proposed to simultaneously test multiple SNPs belonging to a functional region. They perform statistical tests such as multivariate regressions on their individual main effects as well as interaction effects among them to extract maximum information about linkage disequilibrium (LD) [81]. Although they are more informative, these methods will lose their power when LD in the region are weak or a single marker takes the main portion of the effect. As human genetic variations are structured into haplotypes, haplotype-based and its relevant methods have been developed to assess the

relationship between the trait of interest and multiple markers through an overall test of association across haplotypes [13, 68, 67]. However, it may lose power as extremely large number of degrees of freedom are involved in these haplotype-based methods. The ideal method for joint association tests of multiple SNPs is to generate the best combination across all the SNPs while reducing the number of degrees of freedom.

2.2 Aggregate analysis in GWAS

The aggregate analysis of SNPs in GWAS aims to analyze the association of the trait of interest with the joint effects of multiple SNPs from functional regions such as genes, pathways and molecular networks. Some popular gene-based association tests alternatively examine whether a genomic region is associated with the trait of interest using either a combined signal or a combined test statistic based on multiple SNP markers within a gene [44, 24, 81, 88, 40, 80, 46]. Similarly, pathway-based methods examine whether there is significant association between the trait of interest and a pathway composed of a group of related genes defined by some gene annotation database. These methods have been increasingly popular by analyzing cellular pathways which are often involved in disease susceptibility and disease progression [66]. By integrating prior biological knowledge, pathway-based analysis could enhance the detection rate of SNPs that are truly associated with disease but have only weak individual effects. Pathway-based analysis offers an appealing alternative to improve the power of standard GWAS and unravel the biological process underlying complex diseases.

There are two general categories of methods for pathway based GWAS: (1) calculate a summary statistic for each pathway based on the individual test statistics of SNPs and then test the summary statistic. (2) derive combined signals for each path-

way using SNP data and then test the association between these combined signals and the trait of interest. Determining an accurate aggregated statistic or a combined signal for a given set of SNPs is crucial in pathway-based analysis. In other words, the way to aggregate SNPs could determine the detection rate of truly associated SNPs with weak individual effects.

2.2.1 Testing a summary statistic

After mapping SNPs to those genes in prior curated pathways, it comes the primary part of pathway analysis, which is aggregating the effects of the individual SNPs or genes in a pathway. Given the individual statistical significance of each SNP or gene in a pathway, Fisher's combination is a simple way to combine p-values of all SNPs or genes into a summary statistic determining the gene-wise or pathway-wise significance. Denoting p-values for k individual SNPs or genes by p_i, the summary test statistic is $t = -2\sum_{i=1}^{k}\log(p_i)$ under the assumption that all the p-values p_i are independent and uniformly distributed under their null hypotheses. The statistical significance of a gene or pathway can then be represented by a p-value obtained by testing the summary statistic t, which follows a χ^2_{2k} distribution. With p-values of individual SNPs as the input, this method can directly work on the preliminary results obtained from GWAS with an advantage of largely reduced storage space and computational time. However, the independence assumption may be violated because of LD among SNPs or correlations among genes.

Set association is a simple alternative approach to summarize the individual statistical significance of SNPs by involving the correlation between a SNP and its neighbor gene. For each SNP, its test statistic is calculated by a product of χ^2_{Assoc} and χ^2_{HWE}. The first item is the standard Chi-square statistic obtained from a contingency table to compare the genotype frequencies between cases and controls. The

second item represents the Chi-square statistic for Hardy-Weinberg disequilibrium which measures the closeness between a SNP and a gene. The p-value of each gene or each pathway is evaluated by testing the sum of an optimal number of largest test statistics of SNPs mapping to it. In addition, three alternative test statistics are proposed by Luo to combine dependent p-values in consideration of their correlations by performing linear combination test, quadratic test and decorrelation test [51].

Several popular pathway-based methods for GWAS are extended from pathway based approaches for gene expression analysis of microarray data with continuous values in gene levels. The difference between GWAS and gene expression analysis is that the object in gene expression analysis is a gene instead of a SNP. To extend pathway-based approaches for gene expression analysis to GWAS, a step to calculate a test statistic or significance score for a gene is desirable for pathway-based GWAS. As we mentioned, Fisher's method is a feasible way to calculate p value for a gene based on the p-values of SNPs belonging to it. In addition to the violence of independent assumption, another problem facing it is the bias of gene size. Gene-wise significance has preference for larger genes with greater number of SNPs because they tend to have more associated SNPs by chance alone. The bias on gene and pathway size has been a non-ignorable problem that most researchers care about in pathway based GWAS. The most frequent approach to solve this limitation is selecting the SNP with most significant association in each gene to represent the gene. It could remove bias in some extent, however this is not optimal as the joint effect from multiple SNPs is ignored. Several alternative approaches are proposed including correcting for LD among SNPs [31], taking the most significant multiple testing-adjusted p-value as the significance score for a gene [90]. Wang uses the maximum test statistic of SNPs near a gene to represent the significance of the gene with adjustment of multiple testing.

Fisher's exact test for pathway based gene expression analysis examines if a pathway is enriched by significant genes based on a list of significant genes across all pathways. Genes with p-value less than a specific significance threshold (e.g. 0.05) are claimed as significant genes. Fisher's exact test could be extended for pathway-based GWAS by calculating the p-values for all genes using the approaches mentioned in the above paragraph. Denoting the total number of genes of interest by N, the number of genes significantly associated with the disease (p-value$\leq$ 0.05) by S and the number of genes in a given pathway by m, if there are k significantly associated genes in the pathway, the p-value of observing k-significant genes in the pathway is calculated by

$$p = 1 - \sum_{i=0}^{k} \frac{\binom{S}{i}\binom{N-S}{m-i}}{\binom{N}{m}}.$$

One limitation of this method is the ad hoc significance threshold for genes.

Gene Set Enrichment Analysis (GSEA) is another popular method for pathway-based gene expression analysis and could be further extended for pathway-based GWAS. To test the association significance for a pathway, given a rank list of genes based on their statistical significance, GSEA examines whether the members of a pathway are randomly distributed throughout the rank list or primarily found at the top or bottom [73]. Compared with Fisher's exact test, this method takes use of all genes across all pathways instead of certain number of significant genes contingent on an ad hoc threshold. Wang extends GSEA for pathway based GWAS using the maximum test statistic of SNPs near a gene to represent the statistical significance of the gene [80]. For each SNP, its test statistic could be chosen as the χ^2 statistic calculated by Cochran-Armitage trend test. This method follows the same procedure in GSEA for gene expression analysis except small modification on pathway size

adjustment. GSEA compares test statistics of genes in a given pathway with test statistics for genes in other pathways by calculating an enrichment score (ES) based on a weighted Kolmogorov-Smirnov running sum statistic. Given a list L of all N genes, $L = \{g_1, ..., g_N\}$, ranked by their statistical significance $\{r_1, ..., r_N\}$ from largest to smallest, the ES for a given pathway S is calculated by

$$ES(S) = \max_{1 \le i \le N} \Big\{ \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^q}{N_R} - \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{N - |S|} \Big\},$$

where $N_R = \sum_{g_j \in S} |r_j|^q$, q is a parameter that gives higher weights for genes with larger absolute statistic values, and $|S|$ denotes the number of genes in S. The significance of each pathway is obtained by testing their ES respectively based on an empirical phenotype-permutation test. q is set to 1 by the author and when q is set to 0, ES reduces to standard Kolmogorov-Smirnov statistic with no weight for each gene. Specifically, the null distribution of ES for each pathway is estimated by recalculating the ES based on permuted phenotypes. To account for pathway size effect and adjust for multiple hypothesis testing, ES are normalized for each pathway and the proportion of false positives are controlled by calculating false discovery rate (FDR) corresponding to each normalized ES. Another alternative approach to extend GSEA named as GSEA-SNP includes two main steps: (1) use several representative SNPs determined by an adaptive truncated product statistic to represent each gene in a pathway; (2) extend standard GSEA by testing if the set of representative SNPs from a particular pathway is significantly enriched with high ranks using a weighted Kolmogorov-Smirnov test [83]. This approach corrects for the gene size by selecting multiple SNPs as representatives for each gene and performs GSEA directly on a set of representative SNPs without calculating gene-wise statistical significance.

2.2.2 Testing combined signals

Unlike those approaches extended from gene set enrichment analysis where one examines whether significant genes are overrepresented in a pathway under study, another category of approaches for pathway based GWAS focus on testing the joint effect from multiple SNPs within genes and/or multiple genes within pathways. They derive combined signals to represent a set of SNPs or genes and then test the association between them and the trait of interest. These methods are getting increasing attention in recent several years in set association studies with a focus on analysis of the association between the trait and the dimension-reduced variables. Their power of identifying associated SNPs is claimed to be higher with no worry about correction criteria in multiple single-variate tests or high degrees of freedom in multivariate test, especially when the multiple SNPs in a gene or pathway have weak individual effects. To generate low-dimensional variables for SNP data in high dimension, existing statistical methods can be categorized into two broad categories.

2.2.2.1 Nonlinear models to generate combined signal

The first direction involves kernel-machine based approaches that combine all the SNPs in a pathway or gene together based on a user-specific kernel function which measures the similarity between individuals [40]. In this flexible framework, the comparison of multiple SNPs are reduced into a scalar by different kernel functions. Basically, these methods try to measure the similarity over multiple SNPs for all pairs of subjects and compare the pairwise genetic similarity with the pairwise trait similarity. Test statistics are thus derived based on these kernel functions or similarity scores typically leading to small degrees of freedom.

A simple kernel-based approach for case-control study proposed by Schaid et.al [69]. tests a Z_{global} statistic with only one degree of freedom developed using U statistics

based on similarity scores defined by a simple kernel. Consider a subject group l of n subjects where the genotype of a subject i at k-SNP is denoted by g_{ik}. The U statistic for a specific SNP k is defined as an average similarity scores across all pairs of subjects, represented by

$$\bar{U}_l^k = \frac{\sum_{i<j} h_l(g_{ik}, g_{jk})}{\binom{n}{2}},$$

where $h(g_{ik}, g_{jk})$ represents the similarity score between subject i and j at SNP k and l denotes the group they belong to. The overall similarity between these two subjects can be achieved by weighted summing up U^k over all SNPs, which corresponds exactly to a kernel function $h(\boldsymbol{g}_i, \boldsymbol{g}_j) = \sum_{k=1}^{K} w_k h(g_{ik}, g_{jk})$. Let $\bar{U}_1$ and $\bar{U}_2$ be the vectors of U-statistics for all markers for the case and control groups respectively. The test statistic is defined as $Z_{global} = \frac{w'(\bar{U}_1 - \bar{U}_2)}{\sqrt{w' V_0 w}}$ where V_0 denotes the variance-covariance matrix. Its distribution can be approximated by a normal distribution. However, with this method, the direction of the genotype score at each SNP affects the test power and the trait is limited to dichotomous phenotype.

Another approach for testing high-dimensional data is multivariate distance matrix regression (MDMR) [84], which evaluates the relationship between variation of genomic dissimilarity (distance) among a set of individuals and the variation of their trait values. An F statistic with a reduced degree of freedom is constructed to test the association by involving a matrix of genomic similarity among individuals. The distance matrix D can be calculated as $D = 11' - S$ where S is the similarity matrix which can be calculated by several ways [58]. Let $A = (a_{ij}) = (-\frac{1}{2} d_{ij}^2)$ where d_{ij} is the (i, j)-th element in D and X be the matrix of phenotype variables. Given $H = X(X'X)^{-1}X'$ and a centralized matrix of A represented by

$G = (I - \frac{1}{n}11')A(I - \frac{1}{n}11')$, the F statistic is given by

$$F = \frac{tr(HGH)}{tr[(I-H)G(I-H)]}.$$

This statistic can be used for testing both continuous and discrete traits, but it might lead to lower power for a set of independent SNPs.

The kernel-based association test (KBAT) [58] was claimed to be able to handle correlated SNPs without assumption on the direction of individual SNP effects. This test uses an analysis of variance paradigm to compare the variation between cases and controls with the variation within groups. This test is also developed based on the similarities between individuals. The similarity scores are then considered as the observations for the ANOVA model. Let $y^k_{l(ij)}$ denotes the similarity score between individuals i and j in group l at SNP k. It can be modeled using one-way ANOVA as follows:

$$y^k_{l(ij)} = \mu^k + \alpha^k_l + \epsilon^k_{l(ij)}, i < j = 1, ..., n_l; l = 1, 2.$$

Here μ denotes the general effect for pairs of individuals, α^k_l is the group specific treatment effect, and $\epsilon_{l(ij)}$ are the error components. The null hypothesis for testing the association is $H_0 : \alpha^k_1 = \alpha^k_2$. The authors also compared their method with Z_{global} and MDMR by performing simulation studies and have found KBAT have more power.

Other existing kernel machine based approaches [88, 40, 46] apply kernel functions into a linear or logistic regression framework for quantitative analysis and discriminant analysis. Given a data set of n subjects, y_i is a continuous trait following a normal distribution in regression framework or a dichotomous outcome following

Bernoulli distribution, $\boldsymbol{x}_i$ is a vector of p clinical covariates and $\boldsymbol{g}_i$ is a vector of genotypes of K SNPs in a pathway or gene. For continuous trait, the linear regression model on outcome y_i is:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + h(\boldsymbol{g}_i) + e_i,$$

For dichotomous trait, the logistic regression model on outcome y_i is:

$$\text{logit}P(y_i = 1) = \boldsymbol{x}_i^T \boldsymbol{\beta} + h(\boldsymbol{g}_i) + e_i,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, $h(\boldsymbol{g}_i)$ is an unknown centered smooth function, and the error e_i is assumed to be independent and follows a normal distribution $N(0, \sigma^2)$. The pathway effect is modeled by a nonparametric function $h(.)$ assumed belonging to the function space generated by a kernel function $K(.,.)$. By maximizing a penalized likelihood function based on the above model, $h(.)$ is generally estimated as $\sum_{i=1}^{n} \alpha_i K(., \boldsymbol{g}_i)$ by the Representer theorem (Kimeldorf and Wahba, 1970) where α_i are unknown parameters need to be estimated. A challenging thing here is the choice of kernel function. A simple example is the identical by state (IBS) kernel, defined as the count of matched alleles between two subjects which has the form

$$K(\boldsymbol{g}_i, \boldsymbol{g}_j) = \frac{\sum_{k=1}^{K} w_k \text{IBS}(g_{ik}, g_{jk})}{\sum_{k=1}^{K} w_k},$$

where w_k are weights for SNPs which can be determined by involving prior information. One can find the discussion of more choices of kernel functions from [69]. These kernel-based regression models are further shown to have a connection with linear mixed model or logistic mixed model by fitting which the pathway effect can be

easily estimated using existing statistical softwares [46]. Another advantage of kernel machine based approaches is the ability to model the nonlinear effects of SNPs while taking their interactions and correlations into consideration [18, 8]. These kernel based methods provide either non-parametric or semi-parametric models for pathway based association analysis.

2.2.2.2 Linear models to generate combined signal

In addition to nonlinear modeling of SNP effects, dimension reduction methods based on linear operators or linear combinations of SNPs including weighted summation, Fourier transformation (FT) and principle component analysis (PCA) have been also proposed for GWAS. By involving LD information from external databases, association tests by combining optimally weighted markers (ATOM) within a genomic region calculates a weighted summation of them with the optimal weights estimated by borrowing strength of LD [44]. Another dimension reduction approach [] involves Fourier transformation to transform the genotypes into low-dimensional frequency components with a larger weight for the component with lower frequency which contains more information. The global test statistic is calculated based on the weighted score statistics of the FT components and is proved following an asymptotic standard normal distribution as in Schaid's work [69]. The score statistic of a FT component x_k is defined as $U_k = \sum_{i=1}^{n} Y_i(x_{ik} - \bar{x}_k)$ where Y_i and $\bar{x}_k$ denote the trait for subject i and the sample mean of the FT component x_k. The weight for k-th FT component is assigned as $[1/(k+1)]^2$.

Rather than including external knowledge or ad hoc weights, principle component analysis (PCA) [24] provides simple optimal linear combinations of multiple SNPs, which could capture their variation as much as possible while reducing the dimensionality. The obtained principal components which are the optimal linear combinations

can be further used as the linear predictors in linear or logistic regression model to analyze the association of a set of SNPs with the trait of interest. Typically, people focus on the first principle component which contains the most variation of the SNP data and regard it as the combined signal in the joint effect analysis. By performing linear or logistic regressions on a combined signal of multiple SNPs with continuous or dichotomous trait as the outcome, the joint effect of multiple SNPs can be analyzed using aggregated information with low degrees of freedom to eliminate the issue of low power in multi-variate tests. However, noisy SNPs might deteriorate the accuracy of joint effect of multiple SNPs in a set when performing a linear combination of all of them. How to achieve useful and informative combined effect that results in small number of degrees of freedom of test statistic is a challenging topic for researchers studying SNP set association. Involving outcome information to select the informative SNPs entering PCA is a potential way that is similar to the idea of forward selection which is the basic strategy for variable selection in machine learning area. This idea has been employed to develop several models to microarray expression analysis and GWAS in bioinformatics area by several works [4, 11]. We will review those models in section 4 as they are closely related with our methods for pathway based analyses.

There also exists a Bayesian method, Bayesian hierarchical generalized linear model (BhGLM) [89], which performs generalized linear regression on the weighted summation of a group of genes or SNPs with flexible weights following some prior distributions. The relative contributions of individual genes or SNPs in a group are reflected by their respective weights which are estimated in a generalized linear model framework. Therefore, this model assigns more reasonable trait-guided weights for the genes or SNPs by learning the explicitly modeled weights in a generalized linear model framework. However, a number of parameters need to be defined to find

appropriate solutions, which will be tedious and annoying.

In summary, PCA based methods are promising linear models that can take good care of the correlations among SNPs to derive aggregated signals with low degree of freedom. Moreover, the contribution of each SNP or gene is explicitly represented by the weights in the principal loading vectors, which however is not available in non-linear models. To apply PCA based methods to non-Gaussian biomedical data, we make modifications and generalizations of standard PCA in this thesis to deal with the special data types and data structures existing in aggregate association analysis. More reviews of the relevant methodology ameliorating these issues will be included correspondingly in the following sections.

3. SPARSE EXPONENTIAL FAMILY PCA

3.1 Introduction

Dimension reduction methods are widely used for data analysis in many areas such as computer vision, data mining, and bioinformatics. In addition to the low dimensional projections, to reduce model complexity and enhance reproducibility of learning results, people often would like to know the physical meanings of the original variables and how they contribute to these projections. For example, in image analysis, it is of much interest to know which regions are crucial to represent or capture the essential information of the images, which could also help save the memory space during image collections. A knowledge of a group of variables expressing the maximum data variation will also be of much interest for next-generation sequencing data analysis since such a screening of variables could help reduce the profiling cost that are usually very high. To achieve these goals in diverse real-world applications, one faces two critical challenges: how to handle diverse data types arising from different applications and how to obtain meaningful interpretation of analysis results. Exponential family PCA (ePCA) methods [14, 70, 28] and sparse PCA (SPCA) methods [36, 91, 71, 20] are well-known to address these two issues separately. However, to the best of our knowledge, it seems that no one has proposed a method to address these two issues together.

In this section, we propose a sparse exponential family PCA (SePCA) method for dimension reduction with both the capability of addressing the interpretation issue and the generality of applications to any type of data following exponential family distributions. The rest of this section is organized as follows. Section 2 briefly reviews PCA in a probabilistic modeling framework, from which it could be

naturally extended to the exponential family PCA. We also introduce the SPCA problem and the algorithm solving it at the end of this section. Section 3 describes the formulation of SePCA, with an efficient alternative updating algorithm to solve it, and provides the computational complexity analysis. Section 4 illustrates the performance of SePCA compared with Zou's SPCA [91] and a previous sparse logistic PCA method [42] via experiments on both simulated data and real-world data.

3.2 Review of related work

In this section, we review some concepts and probabilistic models that form the foundations of SePCA. We introduce PCA from a probabilistic modeling perspective and naturally extend it to the exponential family. From this point of view, PCA is formulated as a maximum-likelihood estimation (MLE) problem which estimates the low-dimensional projections of a set of canonical parameters by assuming that the conditional probability of each data point given its canonical parameters follows a Gaussian distribution [75]. Similarly, the ePCA tailored to some other types of data could also be modeled as such a MLE problem by assuming that the conditional probability follows a corresponding distribution in the exponential family other than Gaussian. To give a flavor of SePCA, we also introduce SPCA as a simple case and discuss an efficient strategy to solve it at the end of this section.

3.2.1 Principal component analysis

Given a set of samples $\boldsymbol{x}_1, ..., \boldsymbol{x}_N \in R^D$, PCA projects the data into a principal-component subspace with a lower dimension $L(\leq D)$ and meanwhile attempts to preserve the maximum data variation. An alternative interpretation of PCA from a probabilistic perspective assumes that the data points are approximated by linear projections of low-dimensional latent variables plus a Gaussian noise. For each sample $\boldsymbol{x}_n (1 \leq n \leq N)$, given its corresponding vector of latent variables $\boldsymbol{z}_n$ that lies in

the principal-component subspace, the assumption is

$$\boldsymbol{x}_n = W^T \boldsymbol{z}_n + \boldsymbol{b} + \boldsymbol{\epsilon},$$

where W is a principal loading matrix whose rows span the principal-component subspace; $\boldsymbol{b}$ is a bias vector and $\boldsymbol{\epsilon}$ follows a Gaussian distribution $N(0, \sigma^2 I)$. Assuming a vector of canonical parameters $\boldsymbol{\theta}_n = W^T \boldsymbol{z}_n + \boldsymbol{b}$, the conditional probability of $\boldsymbol{x}_n$ given $\boldsymbol{\theta}_n$ is:

$$p(\boldsymbol{x}_n|\boldsymbol{\theta}_n) \sim N(\boldsymbol{x}_n|\boldsymbol{\theta}_n, \sigma^2 I)$$

and the conditional probability of $\boldsymbol{x}_n$ given $\boldsymbol{z}_n$ is:

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n) \sim N(\boldsymbol{x}_n|W^T \boldsymbol{z}_n + \boldsymbol{b}, \sigma^2 I).$$

PCA is then formulated as an optimization problem of maximizing the log-likelihood of the data set with respect to $\boldsymbol{z}_n, W$, and $\boldsymbol{b}$, where the objective function is:

$$\sum_n -||\boldsymbol{x}_n - (W^T \boldsymbol{z}_n + \boldsymbol{b})||^2 \ s.t.\ WW^T = I \tag{3.1}$$

up to a constant. Obviously, this problem is equivalent to minimize the sum of Euclidean distances from data points to their projections in the principal-component subspace, which is exactly one of the interpretations of PCA [60].

3.2.2 Exponential family PCA

From a probabilistic perspective, it is natural to generalize PCA to the exponential family. In the exponential family, a probabilistic latent variable model represent-

ing the conditional distribution of a data sample $\boldsymbol{x}_n$ has such a general form [14]:

$$p(\boldsymbol{x}_n|\boldsymbol{\theta}_n) = \exp(\boldsymbol{\theta}_n^T \boldsymbol{x}_n + \log q(\boldsymbol{x}_n) - A(\boldsymbol{\theta}_n)), \tag{3.2}$$

where $\boldsymbol{\theta}_n$ is the corresponding canonical parameters. $A(\boldsymbol{\theta}_n)$ is the log-normalization factor with a form of $\log \int \exp(\boldsymbol{\theta}_n^T \boldsymbol{x}_n) q(\boldsymbol{x}_n) d\boldsymbol{x}_n$, which ensures that the sum of the conditional probabilities over the domain of $\boldsymbol{x}_n$ equals 1. The probability functions for the members in the exponential family are mainly differentiated by the form of $A(.)$ function. Consequently, the data log-likelihood with respect to the canonical parameters may be of a quadratic form (for Gaussian) or not (for others). Take Gaussian for instance, $A(\boldsymbol{\theta}_n)$ takes a form of $\boldsymbol{\theta}_n^2/2$ to ensure a Gaussian distribution function. Then, its data log-likelihood function given $\boldsymbol{\theta}$ is equivalent to

$$\sum_n -||\boldsymbol{x}_n - \boldsymbol{\theta}_n||^2 \tag{3.3}$$

up to a constant. The canonical parameters $\boldsymbol{\theta}_n$ are further parameterized with a form of $W^T \boldsymbol{z}_n + \boldsymbol{b}$ using lower-dimensional latent variables $\boldsymbol{z}_n$, principal loading matrix W and a bias vector $\boldsymbol{b}$ for dimension reduction. After substituting $\boldsymbol{\theta}_n$ into (3.3), we arrive at (3.1), which is the objective function of PCA in the minimum reconstruction-error interpretation.

In general, ePCA could be achieved by maximizing the generalized likelihood based on a general form of the probability function shown by (3.2). After substituting $\boldsymbol{\theta}_n$ by $\boldsymbol{z}_n, W$ and $\boldsymbol{b}$, ePCA is then formulated as the following problem:

$$\min_{Z:Z^TZ=I} \min_{W,\boldsymbol{b}} \sum_n A(W^T \boldsymbol{z}_n + \boldsymbol{b}) - tr((ZW + \mathbf{1}\boldsymbol{b}^T)X^T), \tag{3.4}$$

where Z is the $N \times L$ principal component score matrix whose n-th row is $\boldsymbol{z}_n$. A

probabilistic graphical model to illustrate ePCA is illustrated in Figure 3.1.

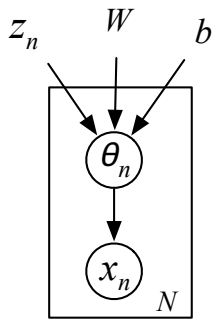

Figure 3.1: A probabilistic graphical model for ePCA.

One drawback that ePCA suffers in common with PCA is the interpretation issue, which motivates us to derive sparse PC loadings for ePCA, especially for high-dimensional data with many highly correlated variables.

3.2.3 Sparse PCA

Before discussing SePCA, we briefly review SPCA, which is a special case of SePCA with the Gaussian distribution assumption, and has been studied by numerous papers [91, 71, 20, 37]. Assuming that the data set has been centralized with zero-mean across samples, an intuitive formulation of the sparse PCA problem can be considered as:

$$\min_{Z} \min_{W:WW^T=I} ||X - ZW||^2 + \sum_{l=1}^{L} \lambda_l |W_l| \tag{3.5}$$

with the aim to achieve dimension reduction using only a set of significantly contributing original variables. The sparsity of each loading vector W_l is controlled by

the regularization coefficient λ_l, where W_l denotes the l-th row of W. A larger value of λ_l will cause fewer non-zero elements in W_l, which are explicitly interpreted as the variables contributing to the l-th principal component. A major difficulty in solving this problem is caused by the orthonormal constraints and l_1-norm penalty imposed simultaneously on the loading vectors.

Zou *et al.* [91] treat SPCA as a penalized regression problem and attempt to minimize the regression error when regressing the PCs on the original variables. They propose a "self-contained" regression approach to decouple the orthonormal and l_1 constraints, which then turns the problem into the following penalized regression problem

$$\min_{A:A^TA=I} \min_{W} ||X - XW^TA^T||^2 + \sum_{l=1}^{L} \lambda_l |W_l| \tag{3.6}$$

if the l_2-norm in the elastic net penalty is omitted. When A is fixed, the minimization of W is equivalent to solving a LASSO problem

$$\min_{W} ||XA - XW^T||^2 + \sum_{l=1}^{L} \lambda_l |W_l|.$$

Alternatively, we can also decouple the constraints by reformulating (3.5) as

$$\min_{Z:Z^TZ=I} \min_{W} ||X - ZW||^2 + \sum_{l=1}^{L} \lambda_l |W_l|, \tag{3.7}$$

which is equivalent to sPCA-rSVD [71] when $l = 1$.

Problems (3.6) and (3.7) are closely related with similar forms in both constraints and objective functions. Although their objective functions are not jointly convex and the constraints are non-convex, both of them can be solved using the same strategy by alternately minimizing variables using closed-form update rules. When A

or Z is fixed, the minimization of W is a LASSO problem solved by soft-thresholding operators. When W is fixed, A or Z can be updated according to Theorem 1 proposed in [91] and [56] given below:

Theorem 1. *Reduced-Rank Procrustes Rotation. Given two matrices $M_{N\times D}$ and $N_{N\times L}$, consider the constrained minimization problem*

$$\min_{A} ||M - NA^T||^2 \quad s.t. \quad A^T A = I_{L\times L}.$$

Suppose that the Singular Value Decomposition (SVD) of $M^T N$ is in the form of UDV^T, then $\hat{A} = UV^T$.

When A is initialized by the first L right eigenvectors of X, the above strategy solving the problem (3.6) can be illustrated as a two-stage procedure: first PCA is performed on X; and then sparse approximations are estimated for the loading vectors. This approach leads to efficient solutions demonstrated by the experimental results in [91]. Analogously, when Z is initialized by the first L left eigenvectors of X, the above strategy solving the problem (3.7) also acts as the same two-stage analysis and is expected to lead to efficient solutions. We will introduce such a similar strategy to help solve SePCA and investigate its efficiency in the following sections.

3.3 Model formulation and algorithm

3.3.1 Problem formulation

We formulate the SePCA problem by adding a regularization term on the generalized loading vectors to the objective function of the ePCA problem (3.4) as follows:

$$\min_{Z:Z^TZ=I} \min_{W,\boldsymbol{b}} \sum_n A(W^T\boldsymbol{z}_n + \boldsymbol{b}) - tr((ZW + \mathbf{1}\boldsymbol{b}^T)X^T) + \Omega(W,\boldsymbol{b}), \quad (3.8)$$

where $\Omega(W, \boldsymbol{b})$ is the regularization term that equals $\lambda_0||ZW + \mathbf{1}\boldsymbol{b}^T||^2 + \sum_{l=1}^{L} \lambda_l|W_l|$. The l_2-norm regularization term is involved here to ensure the stable reconstruction of principal components when $N < D$ and X is not a full rank matrix. It could also be interpreted as a Gaussian prior for canonical parameters to ensure the stability of the model. This penalized maximum likelihood estimator attempts to estimate the optimal loading vectors that are sparse and meanwhile maintain the minimum reconstruction error. Only the variables corresponding to the non-zero elements in a loading vector are selected to construct the corresponding principal component. The tuning parameter λ_l controls the sparsity of loading vectors and SePCA will reduce to ePCA when λ_l equals 0.

3.3.2 Reformulation of the objective function

Based on the observation that for general exponential family distributions, the objective function can be complex and is not jointly convex on Z and W, it is difficult and unsatisfactory to directly solve (3.8) by alternating updates based on gradient descent, which suffers from local optima. Instead, we transform this problem into an equivalent problem by conjugate dual that can be solved more effectively and efficiently.

The reformulation is achieved via replacing the term $A(W^T\boldsymbol{z}_n)$, which is not jointly convex in Z and W by introducing its convex conjugate. In mathematics, the convex conjugate for a function $h(\boldsymbol{\alpha})$ is defined as:

$$h^*(\boldsymbol{u}) = \sup_{\boldsymbol{\alpha} \in M} < \boldsymbol{u}, \boldsymbol{\alpha} > -h(\boldsymbol{\alpha}),$$

where $h^*(\boldsymbol{u})$ is always convex since the maximum of a linear function is convex. Let $A^*(.)$ denotes the convex conjugate of $A(.)$. The explicit form of $A(.)$ and $A^*(.)$ specific to a distribution in the exponential family are specified in [79].

Let Θ be a $N \times D$ matrix whose $n-$th row is $\boldsymbol{\theta}_n$. We first rewrite (3.8) by introducing linear constraints as follows:

$$\begin{aligned} \min_{Z:Z^TZ=I} \min_{W,\boldsymbol{b}} \min_{\Theta} \quad & \sum_n A(\boldsymbol{\theta}_n) + g(Z, W, \boldsymbol{b}) \\ s.t. \quad & \boldsymbol{\theta}_n = W^T \boldsymbol{z}_n + \boldsymbol{b} \quad \textit{for all } n, \end{aligned} \tag{3.9}$$

where $g(Z, W, \boldsymbol{b}) = -tr((ZW + \mathbf{1}\boldsymbol{b}^T)X^T) + \Omega(W, \boldsymbol{b})$. Since the complex form of function $A(.)$ introduces difficulty in directly solving (3.9), we propose Lemma 1 to first transform the minimization of $A(.)$ to its equivalent dual problem to replace the complex $A(.)$.

Lemma 1. *Let U be the $N \times D$ matrix whose $n-$th row is $\boldsymbol{u}_n$. The inner minimization of* (3.9) *with respect to Θ is equivalent to a dual problem:*

$$\max_U - \sum_n A^*(-\boldsymbol{u}_n) - < \boldsymbol{u}_n, W^T \boldsymbol{z}_n + \boldsymbol{b} > + g(Z, W, \boldsymbol{b}).$$

Proof. The Lagrangian of (3.9) is defined as:

$$\sum_n A(\boldsymbol{\theta}_n) + < \boldsymbol{u}_n, (\boldsymbol{\theta}_n - W^T \boldsymbol{z}_n - \boldsymbol{b}) > + g(Z, W, \boldsymbol{b}).$$

Then, the inner minimization of (3.9) on Θ is reformulated as the saddle point problem:

$$\min_{\Theta} \max_U \sum_n A(\boldsymbol{\theta}_n) + < \boldsymbol{u}_n, \boldsymbol{\theta}_n > - < \boldsymbol{u}_n, W^T \boldsymbol{z}_n + \boldsymbol{b} > + g(Z, W, \boldsymbol{b}). \tag{3.10}$$

Since the inner minimization of (3.9) on Θ is a convex problem with feasible linear constraints, it satisfies Slater's conditions for strong duality and the order of

minimization and maximization in (3.10) can be exchanged:

$$\begin{aligned}
&\max_{U} \min_{\Theta} \sum_{n} A(\boldsymbol{\theta}_n) + < \boldsymbol{u}_n, \boldsymbol{\theta}_n > - < \boldsymbol{u}_n, W^T \boldsymbol{z}_n + \boldsymbol{b} > + g(Z, W, \boldsymbol{b}) \\
&= \max_{U} - (\max_{\Theta} \sum_{n} -A(\boldsymbol{\theta}_n) - < \boldsymbol{u}_n, \boldsymbol{\theta}_n >) - < \boldsymbol{u}_n, W^T \boldsymbol{z}_n + \boldsymbol{b} > + g(Z, W, \boldsymbol{b}) \\
&= \max_{U} - \sum_{n} A^*(-\boldsymbol{u}_n) - < \boldsymbol{u}_n, W^T \boldsymbol{z}_n + \boldsymbol{b} > + g(Z, W, \boldsymbol{b}),
\end{aligned}$$

which completes the proof for Lemma 1. □

Then, based on Lemma 1, the original optimization problem (3.8) can be transformed to an equivalent dual problem illustrated by Theorem 2.

Theorem 2. *The optimization problem* (3.8) *is equivalent to*

$$\min_{Z: Z^T Z = I} \min_{W, \boldsymbol{b}} \max_{U} - \sum_{n} A^*(-\boldsymbol{u}_n) - tr((ZW + \mathbf{1}\boldsymbol{b}^T)(U + X)^T) + \Omega(W, \boldsymbol{b}). \quad (3.11)$$

Proof. It suffices to show that (3.9) is equivalent to (3.11). From Lemma 1, it is straightforward to prove that (3.9) is equivalent to its dual problem:

$$\begin{aligned}
&\min_{Z: Z^T Z = I} \min_{W, \boldsymbol{b}} \max_{U} - \sum_{n} A^*(-\boldsymbol{u}_n) - < \boldsymbol{u}_n, W^T \boldsymbol{z}_n + \boldsymbol{b} > + g(Z, W, \boldsymbol{b}) \\
&= \min_{Z: Z^T Z = I} \min_{W, \boldsymbol{b}} \max_{U} - \sum_{n} A^*(-\boldsymbol{u}_n) - tr((ZW + \mathbf{1}\boldsymbol{b}^T)(U + X)^T) + \Omega(W, \boldsymbol{b}),
\end{aligned}$$

which leads to (3.11) and completes the proof for Theorem 2. □

We will then focus on solving the equivalent dual problem (3.11) in the following subsection.

3.3.3 Closed-form update rules

Despite of the non-quadratic objective function and non-convex constraints, we can still find closed-form update rules to solve (3.11). The algorithm based on these

update rules will converge much faster than first-order iterative updating approaches. The solutions are achieved by alternately updating the unknown variables based on the closed-form solutions, which are given below.

Let $f(Z, W, \boldsymbol{b}, U)$ denote the objective function of this min-max problem (3.11). Obviously, $f(.,.,.,U)$ is concave in U. In each iteration, we can update U by solving the following optimization problem

$$\max_{U} -\sum_{n} A^*(-\boldsymbol{u}_n) - tr((ZW + \mathbf{1}\boldsymbol{b}^T)U^T).$$

The optimal $\boldsymbol{u}_n$ is obtained as the negative mean vector of a sample $\boldsymbol{x}_n$ given in Theorem 3. The mean vector is further shown to be equal to the first derivative of the log-normalization factor $A(.)$ according to the following property shown in Proposition 1 proposed by [79]. Therefore, a closed-form solution for $\boldsymbol{u}_n$ is $\hat{\boldsymbol{u}}_n = -\frac{\partial A(\boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n}|_{\boldsymbol{\theta}_n = W^T \boldsymbol{z}_n + \boldsymbol{b}}$. One can also verify this solution by setting the first derivative with respect to $\boldsymbol{u}_n$ equal to 0.

Theorem 3. *Let θ denote the canonical parameters of the exponential family distribution for random variables $x \in \mathcal{X}$. Consider the variational representation of the log-normalization factor $A(\theta)$ in terms of its dual $A^*(\mu)$ as $A(\theta) = \sup_{\mu \in \mathcal{M}}\{\langle \theta, \mu \rangle - A^*(\mu)\}$. Then, as proposed by [79], for all $\theta \in \Omega$, the supremum in this equation is attained uniquely at the vector μ^* specified by the moment matching conditions*

$$\mu^* = \int_{\mathcal{X}} xp(x|\theta)dx = E_\theta[X].$$

Similarly, consider an optimization problem: $\max_{u \in \mathcal{M}'} -\langle \theta, u \rangle - A^*(-u)$ where

$\mathcal{M}' = \{m : -m \in \mathcal{M}\}$. *The maximum is attained at the vector* u^* *specified by*

$$u^* = -\mu^* = -E_\theta[X].$$

Proposition 1. *The log-normalization factor* $A(\theta)$ *associated with any regular exponential family has the following property:*

It has the first derivative on its domain Ω *and the first derivative yields the cumulant of the random vector* X *as follows:*

$$\frac{\partial A(\theta)}{\partial \theta} = E_\theta[X] := \int_{\mathcal{X}} xp(x|\theta)dx.$$

For the outer minimization problem on Z, W and $\boldsymbol{b}$, the objective function $f(Z, W, \boldsymbol{b}, .)$ is quadratic as shown below:

$$\begin{aligned} & f(Z, W, \boldsymbol{b}, .)|_{Z^T Z = I} \\ = \; & -tr((ZW + \mathbf{1}\boldsymbol{b}^T)(U + X)^T) + \lambda_0 ||ZW + \mathbf{1}\boldsymbol{b}^T||^2 + \sum_{l=1}^{L} \lambda_l |W_l| + C_0 \\ = \; & \lambda_0 ||\frac{1}{2\lambda_0}(X + U) - ZW - \mathbf{1}\boldsymbol{b}^T||^2 + \sum_{l=1}^{L} \lambda_l |W_l| + C_1, \end{aligned}$$

where C_0 and C_1 are constant terms unrelated to Z, W and $\boldsymbol{b}$. The minimization problem on Z, W and $\boldsymbol{b}$ has a similar form as the SPCA problem (3.7) and thus can be solved by the same strategy mentioned in Section 3.2.3. Although this problem involves non-convex constraints, an efficient solution will be achieved owing to the elegant problem structure. Specifically, in the $(t+1)$-th iteration, given an optimal U^t, $\boldsymbol{b}^{t+1}$ is updated as

$$\boldsymbol{b}^{t+1} = \frac{1}{N}\Big(\frac{1}{2\lambda_0}(X + U^t) - Z^t W^t\Big)^T \mathbf{1}.$$

To update Z, the minimization problem with respect to Z is:

$$\begin{aligned} & \min_{Z:Z^TZ=I} ||\frac{1}{2\lambda_0}(X+U) - \mathbf{1}\boldsymbol{b}^T - ZW||^2 \\ = & \min_{Z:Z^TZ=I} ||\frac{1}{2\lambda_0}(X+U)^T - \boldsymbol{b}\mathbf{1}^T - W^TZ^T||^2. \end{aligned}$$

Denote Q as $\frac{1}{2\lambda_0}(X+U) - \mathbf{1}\boldsymbol{b}^T$. We first compute the SVD of $Q^tW^{tT} = R\Lambda V^T$ and then update Z^{t+1} by $R[,1:L]V^T$ according to Theorem 1.

To update W, the minimization problem with respect to W is a LASSO problem

$$\begin{aligned} & \min_W ||\frac{1}{2\lambda_0}(X+U) - ZW - \mathbf{1}\boldsymbol{b}^T||^2 + \sum_{l=1}^{L}\frac{\lambda_l}{\lambda_0}|W_l| \\ = & \min_W ||Q - ZW||^2 + \sum_{l=1}^{L}\frac{\lambda_l}{\lambda_0}|W_l|. \end{aligned}$$

Then the optimal W_l^{t+1} for $l = 1, ..., L$ is given by $\left(|Q^{tT}Z_l^t| - \frac{\lambda_l}{2\lambda_0}\right)_+ Sign(Q^{tT}Z_l^t)$, where Z_l denotes the l-th column of Z corresponding to the l-th PC.

In summary, the detailed procedure for solving SePCA is illustrated by Algorithm 1.

3.3.4 Computational complexity

In the initialization step, it takes $O(ND^2)$ computational operators to compute the SVD. Our algorithm contains two main steps: maximization of U and minimization of Z, W and $\boldsymbol{b}$. Computing U has the computational complexity of $O(NDL)$ in each iteration. In each iteration of optimizing Z, computing QW^T and the SVD of it has the complexity $O(NDL)$ and $O(NL^2)$ respectively. The estimation of W using the soft-thresholding operation have the complexity of $O(NDL)$ in each iteration. In total, the computational complexity is $O(ND^2) + rO(NDL)$ if it takes r iterations to converge. If $N << D$, the cost of SVD in the initialization step can be reduced

Algorithm 1 SePCA

1. Set $\boldsymbol{b}$ as $\frac{1}{N}X^T\mathbf{1}$. Compute SVD of $X - \mathbf{1}b = APB^T$ and set $Z = A[, 1 : L]$ and $W = B[1 : L,]$.
2. Update $\boldsymbol{u}_n$ by $-\frac{\partial A(\boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n}$ where $\boldsymbol{\theta}_n = W^T\boldsymbol{z}_n + \boldsymbol{b}$.
3. Update $\boldsymbol{b}$ by $\frac{1}{N}\Big(\frac{1}{2\lambda_0}(X + U) - ZW\Big)^T\mathbf{1}$.
4. Calculate $Q = \frac{1}{2\lambda_0}(X+U)-\mathbf{1}\boldsymbol{b}^T$. Compute the SVD of $QW^T = R\Lambda V^T$. Update $Z = R[, 1 : L]V^T$.
5. Given a fixed Z,
   for $l = 1, ..., L$, $W_l = \Big(|Q^TZ_l| - \frac{\lambda_l}{2\lambda_0}\Big)_+ Sign(Q^TZ_l)$.
6. Repeat 2-5 until convergence.
7. Normalize W_l as $\hat{W}_l = W_l/||W_l||_1$. And then calculate $\hat{Z}_l$ as $Z_l||W_l||_1$. Rank $\hat{W}_l$ and $\hat{Z}_l$ in the decreasing order of $||W_l||_1$.

to N^2D and the total computational complexity is $O(N^2D) + rO(NDL)$. It usually takes only a few iterations to converge according to our empirical experience.

3.3.5 Connections with ePCA and SPCA

Our SePCA is a generalization of both ePCA and SPCA and could reduce to them correspondingly in special cases. When λ_l is set as 0 and the bias term $\boldsymbol{b}$ is dropped, SePCA reduces to an optimization problem with the same objective function and constraints as ePCA problem proposed in [28] but with a different alternating order of optimization on Z, W and U. Fortunately, these two problems are shown equivalent irrespective of the optimization order in [28]. Without the l_1 norm regularization term on W, our algorithm updates W and Z by $\frac{1}{2\lambda_0}Z^T(X + U)$ and the first L left vectors of matrix $X + U$ respectively, which conform to the updates given by [28]. As for the U update step, we directly update U by a closed-form solution instead of the gradient ascent method used in [28]. Eventually, the gradient ascent approach will find the same solution since the objective function is

concave with respect to U; however it will take longer time than our algorithm.

In the cases that the data set X is assumed to be sampled from a Gaussian distribution, we have $A(\boldsymbol{\theta}_n) = \boldsymbol{\theta}_n^T\boldsymbol{\theta}_n/2$ and $A^*(\boldsymbol{u}_n) = \boldsymbol{u}_n^T\boldsymbol{u}_n/2$ correspondingly. Consequently, U is estimated as $-ZW$ when the data is centralized. After substituting the estimated U into the objective function in (3.11), we will arrive at the SPCA problem (3.7) with an elastic net regularization term.

3.3.6 Choice of tuning parameters

In Algorithm 1, the parameter λ_0 acts as a scaling factor for W, analogous to the role of tuning parameters for l_2 term in elastic net and Zou's sparse PCA. The default choice of λ_0 can be 1. For simplicity, we treat λ_l for different principal components equally and only need to determine one common parameter λ. The l_1 regularization parameter controls the model complexity. To compromise the goodness of fit and model complexity, we use Bayesian Information Criterion (BIC) to achieve the maximum likelihood with the most model generalization. λ is chosen by minimizing the following BIC criterion:

$$BIC = -2ln\hat{\ell}(U, Z, W, \boldsymbol{b}) + \log(ND) \times m(\lambda),$$

where $\hat{\ell}$ is the estimated log-likelihood and $m(\lambda)$ is the number of free parameters to be estimated: $m(\lambda) = ND + NK + D + |W(\lambda)|$ where ND is the total number of elements of U, NK is the total number of elements of Z, D is the length of the vector $\boldsymbol{b}$, and $|W(\lambda)|$ is the number of nonzero loadings in W when the penalty parameter is λ.

3.4 Experimental results

We have investigated the performance of our SePCA model and the efficiency of the algorithm via a simulated study and real-world applications. The simulation study aims to examine the accuracy and computational efficiency of SePCA on binary data and count data. These performances are further investigated by real-world applications in image clustering and population stratification that involve count data and binary data respectively. We show the superiority of SePCA by comparing it with Zou's SPCA [91]. For binary data cases, we also compare SePCA with a previous sparse logistic PCA method solved by coordinate descent Majorization-Minimization algorithm [41], which is denoted as SLPCA_MM in this paper. We denote SePCA under a certain exponential family distribution such as Bernoulli distribution or Poisson distribution as SePCA_Bern or SePCA_Pois, respectively.

3.4.1 Simulation study

3.4.1.1 Simulation design

In this set of simulation experiments, we studied the performance of SePCA under Bernoulli distribution for dimension reduction of binary data and Poisson distribution for count data. For each distribution, we simulated a matrix $X_{N\times D}$ of N independent D-dimensional samples by its corresponding $A(.)$ function of the canonical parameter matrix Θ. The performance of SePCA_Bern and SePCA_Pois in reconstruction of sparse PC loadings for binary data and count data are examined respectively. As we introduced, Θ is parameterized as $ZW + \mathbf{1}\boldsymbol{b}^T$ where Z is a $N \times L$ principal component score matrix and W is a $L \times D$ principal component loading matrix. L is the number of principal components, which is set to 2 in this simulation study. For simplicity, we assume that the bias vector $\boldsymbol{b}$ is 0. The principal component scores Z_l are generated randomly from a Gaussian distribution with zero mean and

a variance of σ_l^2. The loading matrix W is set as a sparse matrix, in which only $W[1, 1:20]$ and $W[2, 21:40]$ are set to 1 while the other elements are set to 0. Each loading vector W_l is normalized with a unit l_2 norm. To have a thorough evaluation, we studied simulated data with different sizes of samples and data variables as well as different variances of PC scores.

3.4.1.2 Binary data study

In this study, each element X_{nd} in the data set X is sampled from a corresponding Bernoulli distribution with a success probability that equals $\frac{exp(\Theta_{nd})}{1+exp(\Theta_{nd})}$.

We considered 8 different settings of (N, D) where N has two choices: 100, 200 and D has four choices: 50, 100, 200, 500 respectively. Since the variance σ_l^2 measures the signal level of the l-th PC, we set up PC variance relative to a suitably defined baseline noise level as $\sigma_l^2 = SNR_l \times \sigma_0^2$ (baseline noise level), where SNR_l is the signal to noise ratio for the l-th PC. The baseline noise level is defined as the variance of PC scores from binary data under the unstructured model. To compute this, we generated the independent $N \times D$ binary data from a Bernoulli distribution with the success probability 0.5 and apply SePCA_Bern on it. We calculate the baseline noise level as the average of sample variances of the obtained L PC scores. The details about calculating the baseline noise level can be found in Section 6.1 of [42]. Finally, we simulated 100 binary data sets for each of these 8 settings under SNR= (3, 2) and (5, 3) respectively.

To evaluate the performance of SePCA_Bern, we computed for each simulated data set the maximum angle between the estimated PC loadings and true PC loadings to evaluate the reconstruction accuracy, the percentage of non-zero elements estimated in the loading matrix among the true 40 non-zero elements (true positives) and other zero elements (false positives) to examine the accuracy in spar-

sity. By specifying $A(\theta)$ as $\log(1 + \exp(\theta))$ for Bernoulli distribution, we applied SePCA_Bern to the simulated data under each configuration and compared its performance with SPCA and SLPCA_MM. The results shown in Table 3.1 suggest that our SePCA_Bern outperforms the others for all configurations in reconstructing more accurate principal component loading matrix with smaller angles between estimated PC loadings and true PC loadings. A visualization of the angles for the three methods across all 100 replicates under the first setting with N=100, D=50 and SNR= (3,2) are shown by Figure 3.2. One can observe that SePCA_Bern could always achieve the smallest degree among the three methods. By observing true positive rates and false positive rates from these three methods in each case, we see that almost all of them take use of all the non-zero elements to predict the PC scores. However, SLPCA_MM and SPCA use more additional true zero elements which indeed have no contribution to the true PCs. These above observations demonstrate that SePCA_Bern could detect the best sparse structure of PC loadings and moreover achieve the highest reconstruction accuracy using those contributing variables. The superiority of SePCA_Bern over SPCA and SLPCA_MM is believed to benefit from an explicit modeling of binary data and from a direct optimization of the exact objective function respectively. In addition, the computational time of SePCA_Bern is much less than SPCA especially for higher dimensional data since SPCA requires an expensive computation of Gram matrix $X^T X$ which is avoided in SePCA_Bern.

3.4.1.3 Count data study

In this study, each element X_{nd} in the data set X is sampled from a corresponding Poisson distribution with mean that equals $\exp(\Theta_{nd})$. We implemented SePCA_Pois on the this count data set with $A(\theta)$ set as $\exp(\theta)$. We also compared it with SPCA and SePCA_MM based on the criterion of running time, maximum angle of

Table 3.1: Comparison of the performance of SPCA, SLPCA_MM and SePCA_Bern on simulated binary data. The average (standard deviation) of the running time, maximum angle of PC loadings, true positive rate and false positive rate over 100 simulations are presented for these three methods.

N	D	Method	Time(sec.)	Angle(degree)	True Pos.(%)	False Pos.(%)
SNR=(3,2)						
100	50	SPCA	0.68 (0.61)	11.18 (4.03)	100 (0)	7.20 (2.76)
		SLPCA_MM	0.50 (0.42)	8.35 (2.31)	100 (0)	7.48 (4.35)
		SePCA_Bern	0.09 (0.08)	5.37 (1.08)	100 (0)	0.92 (1.12)
	100	SPCA	1.98 (0.86)	16.67 (5.99)	100 (0)	5.82 (3.48)
		SLPCA_MM	1.29 (0.95)	8.83 (1.96)	100 (0)	10.0 (2.85)
		SePCA_Bern	0.16 (0.06)	7.15 (2.17)	100 (0)	0.35 (0.48)
	200	SPCA	5.95 (9.36)	26.61 (5.78)	99.95 (0.35)	15.1 (2.08)
		SLPCA_MM	2.22 (2.22)	8.09 (2.51)	100 (0)	0.83 (1.11)
		SePCA_Bern	1.3 (1.72)	7.69 (2.35)	100 (0)	0.12 (0.21)
	500	SPCA	17.91 (15.84)	37.69 (5.72)	96.33 (1.23)	6.39 (8.02)
		SLPCA_MM	2.00 (2.09)	7.15 (2.25)	100 (0)	0.75 (0.29)
		SePCA_Bern	1.03 (1.23)	6.96 (3.22)	100 (0)	0.09 (0.09)
200	50	SPCA	0.46 (0.58)	5.95 (1.35)	100 (0)	11.4 (2.57)
		SLPCA_MM	0.85 (0.64)	6.35 (1.45)	100 (0)	8.13 (4.67)
		SePCA_Bern	0.26 (0.01)	3.39 (0.66)	100 (0)	0.92 (1.12)
	100	SPCA	1.72 (1.63)	10.7 (1.57)	100 (0)	26.1 (1.99)
		SLPCA_MM	3.66 (2.60)	7.50 (1.10)	100 (0)	7.58 (3.86)
		SePCA_Bern	0.50 (0.02)	5.58 (1.04)	100 (0)	0.34 (0.54)
	200	SPCA	32.1 (63.6)	20.4 (2.10)	100 (0)	29.7 (1.53)
		SLPCA_MM	1.14 (0.13)	10.9 (1.78)	100 (0)	0.73 (0.40)
		SePCA_Bern	0.89 (0.05)	10.7 (1.02)	100 (0)	0.11 (0.18)
	500	SPCA	81.8(115.3)	29.5 (3.08)	100 (0)	20.7 (1.07)
		SLPCA_MM	2.42 (0.26)	8.32 (1.33)	100 (0)	0.78 (0.28)
		SePCA_Bern	2.08 (0.15)	7.70 (1.07)	100 (0)	0.11 (0.11)
SNR=(5,3)						
100	50	SPCA	0.63 (0.61)	11.4 (4.19)	100 (0)	6.65 (2.61)
		SLPCA_MM	0.58 (0.42)	7.58 (2.22)	100 (0)	8.05 (4.40)
		SePCA_Bern	0.09 (0.02)	4.45 (1.02)	100 (0)	0.90 (1.12)
	100	SPCA	1.29 (0.99)	18.0 (5.17)	100 (0)	14.4 (2.95)
		SLPCA_MM	1.45 (0.96)	7.50 (1.72)	100 (0)	8.93 (3.47)
		SePCA_Bern	0.16 (0.02)	5.77 (1.34)	100 (0)	0.33 (0.48)
	200	SPCA	2.75 (1.87)	27.1(5.50)	99.90 (0.49)	14.3 (2.26)
		SLPCA_MM	0.82 (0.24)	6.62 (1.62)	100 (0)	0.81 (0.97)
		SePCA_Bern	0.34 (0.01)	6.12 (1.86)	100 (0)	0.09 (0.18)
	500	SPCA	7.89 (4.73)	41.5 (5.03)	95.40 (3.58)	9.38 (0.96)
		SLPCA_MM	1.47 (0.18)	5.92 (1.61)	100 (0)	0.75 (0.29)
		SePCA_Bern	0.76 (0.08)	5.80 (1.81)	100 (0)	0.08 (0.09)
200	50	SPCA	0.31 (0.36)	5.61 (1.15)	100 (0)	10.9 (2.65)
		SLPCA_MM	0.97 (0.62)	5.48 (1.33)	100 (0)	7.85 (4.09)
		SePCA_Bern	0.28 (0.07)	2.70 (0.47)	100 (0)	0.93 (1.17)
	100	SPCA	2.84 (5.07)	9.54 (1.55)	100 (0)	24.8 (2.18)
		SLPCA_MM	4.08 (2.61)	6.30 (0.91)	100 (0)	7.91 (3.68)
		SePCA_Bern	0.54 (0.04)	4.32 (0.71)	100 (0)	0.39 (0.55)
	200	SPCA	9.67 (9.92)	17.0 (1.73)	100 (0)	28.5 (1.81)
		SLPCA_MM	1.23 (0.07)	8.30 (1.31)	100 (0)	0.74 (0.44)
		SePCA_Bern	0.91 (0.04)	7.44 (1.36)	100 (0)	0.12 (0.19)
	500	SPCA	76.3 (164.1)	29.6 (3.40)	100 (0)	20.9 (1.03)
		SLPCA_MM	2.83 (0.28)	6.74 (0.93)	100 (0)	0.76 (0.26)
		SePCA_Bern	2.03 (0.05)	5.71 (0.98)	100 (0)	0.10 (0.10)

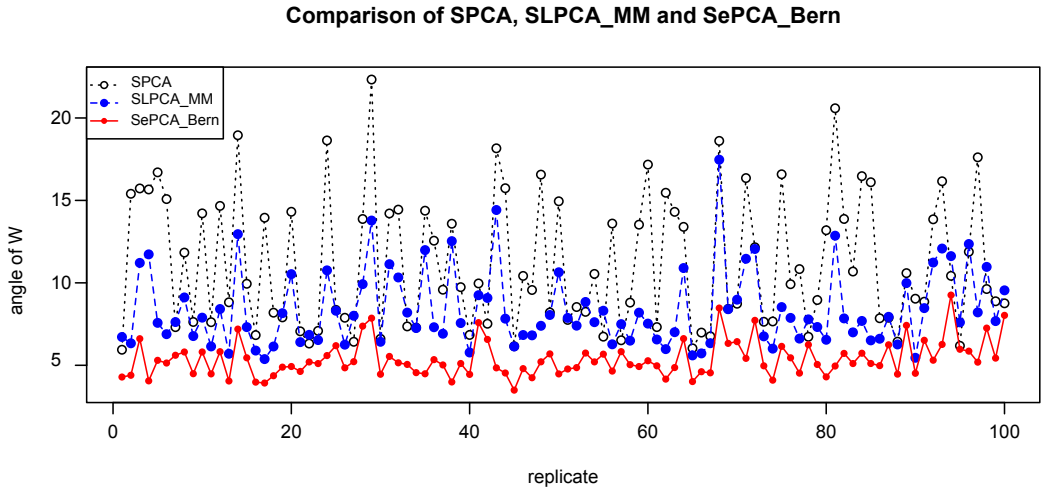

Figure 3.2: Plot of the maximum angle of PC loadings for SPCA, SLPCA_MM and SePCA_Bern across 100 replicates at N= 100, D=50 and SNR= c(3,2).

PC loadings, true positives and false positives. In order to apply SePCA_MM to analyze this count data set, we set a cut threshold at the half of the maximum value to dichotomize this data set. We show the results of these three methods for 100 replicates of data set simulated under one case with N=100, $D = 50$ and $SNR = (3, 2)$ in Table 3.2. As shown in this table, SePCA_Pois could estimate the most accurate PC loadings with the smallest difference angle while remaining the most approximate sparse structure. SePCA_MM has pretty bad performance due to the information loss in the dichotomization. Both the experimental results on binary data study and count data study suggest that it is crucial to assume the most appropriate distribution for a given data set according to its special data type to achieve the best dimension reduction with sparse loading vectors. The superiority of computational efficiency of our SePCA algorithm was also verified by both studies.

Table 3.2: Comparison of the performance of SPCA, SLPCA_MM and SePCA_Pois on simulated count data at N=100, D = 50 and SNR = (3,2). The average (standard deviation) of the running time, maximum angle of PC, true positive rate and false positive rate over 100 simulations are presented for these three methods.

N	D	Method	Time(sec.)	Angle(degree)	True Pos.(%)	False Pos.(%)
SNR=(3,2)						
100	50	SPCA	0.0616 (0.02)	6.14 (9.51)	99.5 (5)	0 (0)
		SLPCA_MM	0.0076 (0.00347)	74 (33.3)	59 (18.6)	0 (0)
		SePCA_Pois	0.0133 (0.00656)	1.31 (1.5)	100 (0)	1.2 (2.7)

3.4.2 Face image clustering

As image data is usually high-dimensional and involves redundant information caused by locally related pixels, it is desirable to reduce the dimension and redundancy via penalized latent factor analysis with variable selection pursuing better performance and interpretation. We applied SePCA on a data set from the Yale image database [25] to compare its clustering performance with other methods based on k-means clustering.

There are 11 different images of each of 15 distinct subjects in the Yale database. The images of each subject vary in different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised and wink. We randomly select a data set containing 44 images from 4 subjects: 1, 4, 6 and 8, corresponding to 4 clusters respectively. We just use a center region of 128×128(= 16,384) pixels from the original images by removing the redundant white background pixels. Then this data set is represented by a 44×16,384 matrix with each row corresponding to one image. Our goal is to cluster these images to 4 clusters corresponding to the four selected subjects. Due to the high dimensionality of images, we perform the clustering in a lower L-dimensional space constructed by the generalized principal components obtained by performing SePCA_Pois on the

images where the pixel intensities are considered as count data following Poisson distributions. The irrelevant or redundant pixels are taken care by the sparse learning of PC loading vectors.

We have studied the clustering performance at $L = 1, 2, 3$ respectively. The clustering accuracy is calculated as the proportion of correctly assigned labels based on a best match with true labels [9, 12]. The results summarized in Table 3.3 have shown that SePCA_Pois achieved higher or competitive clustering accuracy than SPCA at a larger sparsity rate in all the three cases. In each case, the sparsity rate is summarized by the "Total(%)" column in Table 3.3, which is reflected by the number of nonzero elements in all the PC loadings followed by its percentage. The number of non-zeros in each PC loading as well as its percentage are shown in the following columns. These observations suggest that an appropriate assumption of the data distribution gives rise to more appealing dimension reduction results than standard sparse PCA using fewer variables. The visualization of 2-PC projections for SePCA_Pois and SPCA are shown in Figure 3.3, which demonstrated that the projections obtained from SePCA_Pois have more obvious clustering boundary and smaller within-cluster distance.

3.4.3 Population stratification

We applied our algorithm on Single Nucleotide Polymorphism (SNP) data from the International HapMap Project (HapMap3) [16] to analyze the subpopulation structure. This dataset contains 1,301 samples from 11 populations of European ancestry, Asian ancestry, and African ancestry. Those samples from the same population tend to have a common pattern of genetic variation expressed by SNPs, which can be detected by clustering. We treat SNP data as binary data with 0 representing the most prevalent homogeneous base pair (wild-type) and 1 representing the other

Table 3.3: Comparison of the clustering performance after SPCA and SePCA_Pois on Yale data. The clustering accuracy and the number of non-zero variables (percentage) in all the loading vectors as well as that in each loading vector are presented for these three methods.

Method	Acc.	Total(%)	PC1(%)	PC2(%)	PC3(%)
L=1					
SPCA	0.61	7064(43.1)	7064(43.1)	-	-
SePCA_Pois	0.61	6023(36.8)	6023(36.8)	-	-
L=2					
SPCA	0.64	7693(47.0)	7076(43.2)	4916(30.0)	-
SePCA_Pois	0.70	6998 (42.7)	5666(34.6)	5320(32.5)	-
L=3					
SPCA	0.84	7906(48.3)	6286(38.3)	5994(36.6)	4924(30.1)
SePCA_Pois	0.84	7640(46.6)	6133(37.4)	5414(33.0)	3762(23.0)

genotypes (mutant with minor alleles). Considering millions of SNPs are genotyped in the data set, we first reduce the number of SNPs by quality control via removing SNPs and samples with minor allele frequency less than 0.05, excluding regions with strong linkage disequilibrium (LD) such as Major histocompatibility complex (MHC) and LD pruning to make sure pairs of variants in a window size of 50 kb have correlation value r^2 less than 0.2. After randomly sampling 5,000 SNPs and removing the SNPs with missing values, we have a smaller data set with 1,184 samples and 748 SNPs for clustering analysis. Since these 11 populations could be clearly grouped into three clusters corresponding to the three categories of ancestry, we only present our results on identifying the four populations with the African ancestry: African ancestry in Southwest USA (ASW); Luhya in Webuye, Kenya (LWK); Maasai in Kinyawa, Kenya (MKK) and Yoruba in Ibadan, Nigeria (YRI) with 83, 90, 171 and 166 samples respectively. Thus, we have 511 samples in this study in total. We applied SPCA, SLPCA_MM and SePCA_Bern for clustering and found that the last two methods have competitive clustering accuracy and sparsity in all

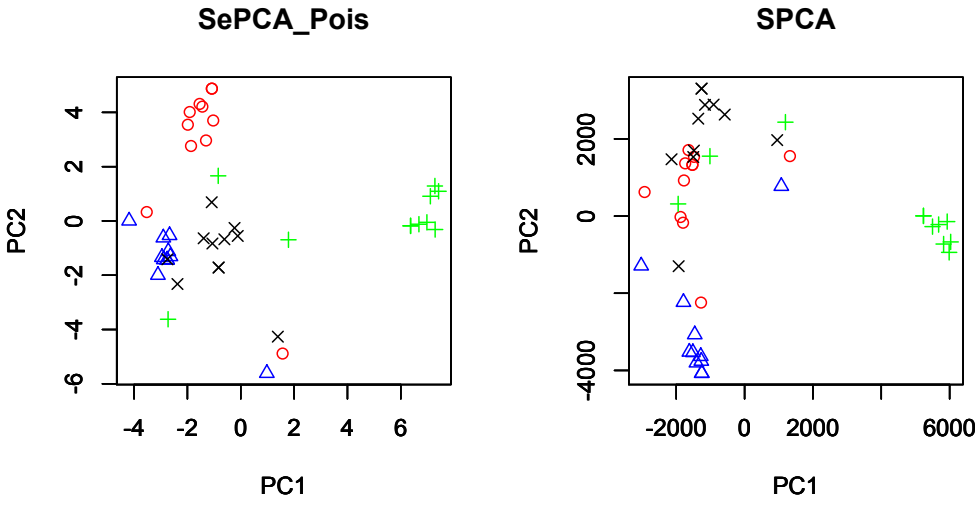

Figure 3.3: Visualization of the distribution of 44 Yale images from 4 subjects in 2-PC space.

three cases demonstrated by Table 3.4. It suggests that SePCA_Bern has no obvious improvement compared with SLPCA_MM when a fairly large portion of variables contribute to the projections. However, both of them have considerable improvement over SPCA in clustering accuracy at the competitive sparsity rate owing to their explicit modelings of binary data.

To examine the computational efficiency of SLPCA_MM and SePCA_Bern, we recorded their running time for different number of PCs. As shown in Figure 3.4, our method has the running time quite insensitive to the number of PCs and outperforms SLPCA_MM with increasing gain in computational efficiency as the number of PCs increases. The dramatic increasing trend in the running time of SLPCA_MM is believed as a result of the one by one coordinate-descent optimization for PCs. One may notice that the CD update in SLPCA_MM will cost slightly less than SVD in our algorithm when quite a few PCs are computed for a small data set with a relatively

Table 3.4: Comparison of the clustering performance after SPCA, SLPCA_MM and SePCA_Bern on HapMap data. The clustering accuracy and the number of non-zero variables (percentage) in all the loading vectors as well as that in each loading vector are presented for these three methods.

Methods	Acc.	Total(%)	PC1(%)	PC2(%)	PC3(%)
L=1					
SPCA	0.81	502(67.1)	502(67.1)	-	-
SLPCA_MM	0.85	567(75.8)	567(75.8)	-	-
SePCA_Bern	0.85	553(73.9)	553(73.9)	-	-
L=2					
SPCA	0.81	633(84.6)	496(66.3)	494(66.0)	-
SLPCA_MM	0.85	659(88.1)	568(75.9)	420(56.1)	-
SePCA_Bern	0.85	650(86.9)	553(73.9)	402(53.7)	-
L=3					
SPCA	0.66	688(92.0)	489(65.4)	483(64.6)	489(65.4)
SLPCA_MM	0.68	708(94.6)	568(75.9)	440(58.8)	420(56.1)
SePCA_Bern	0.68	700(93.6)	554(74.1)	414(55.3)	400(53.5)

low dimension. Our algorithm can be further speed up by doing a truncated SVD when the required number of PCs is less than the full rank of the data matrix.

3.5 Supervised sparse exponential family PCA (SSePCA)

In the last section, we introduced an extended SePCA with the ability of variable selection to pursue better interpretation for the generalized PCs based on the outcome of interest. SePCA is an unsupervised model for dimension reduction, which can not guarantee the resulting PCs and selected variables are meaningful for a supervised study, in which the label information are available. For example, the supervised study could be a study that analyzes the association between the outcome (i.e. labels) and the joint effects from all the predictors. In this section, we extend SePCA to a supervised model named as supervised SePCA (SSePCA) to learn supervised generalized PCs for association studies using only a set of dominant variables. As opposed to SePCA, the variable selection in SSePCA is guided by the

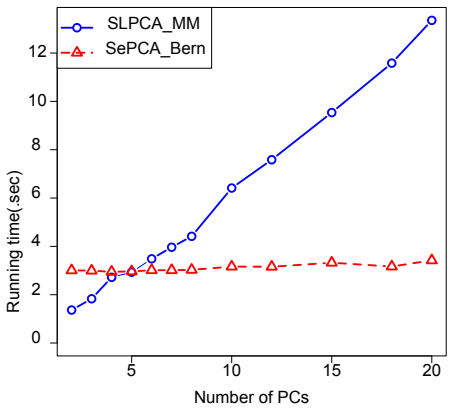

Figure 3.4: Comparison of the running time versus the number of PCs for SLPCA_MM and SePCA_Bern on HapMap data.

outcome. Thus, its power for association analysis could be improved by eliminating the effects from irrelevant variables under the guidance of label information in an integrate framework.

Given a design matrix X composed of n samples in d-dimensional space and its corresponding outcome $\boldsymbol{y} = (y_1, ..., y_n)$, we aim to analyze the association between the joint effect from the D predictors and the outcome $\boldsymbol{y}$. The data type of the predictors and the outcome could be any type of data following exponential family distributions. We propose a SSePCA model to enforce that the estimated principal components extract the maximal variation from some key variables to best predict/differentiate the outcome. The joint effects are also simultaneously estimated along with the PCs in an integrative framework. A brief graphical model illustrating SSePCA is shown in Figure 3.5. On the left side of this graphical model, $\boldsymbol{z}$ is the latent variable for a sample $\boldsymbol{x}$ lying in an orthogonal principal-component subspace and W is the corresponding principal component loading matrix. On the right side, the latent

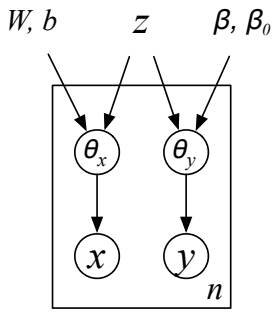

Figure 3.5: A probabilistic graphical model for SSePCA.

variable $\boldsymbol{z}$ is simultaneously involved in a generalized linear model as predictors to predict coefficients $\boldsymbol{\beta}$ that reflecting the joint effects. The association analysis and data aggregation are simultaneously integrated in this model with the aim of estimating the aggregated signal that best summarizes the data which meanwhile influence the outcome.

This model generalizes PCA for data sample $\boldsymbol{x}$ of a certain data type through an exponential family distribution $p(\boldsymbol{x}|\boldsymbol{\theta}_x)$ where $\boldsymbol{\theta}_x$ denotes the corresponding canonical parameters which are assumed to factorize in the form $W\boldsymbol{z} + b$. Analogously, this model also generalizes linear models for various types of the outcome $\boldsymbol{y}$ through an exponential family distribution $p(\boldsymbol{y}|\boldsymbol{\theta}_y)$ where $\boldsymbol{\theta}_y$ is the linear predictor that equals $\boldsymbol{z}^T\boldsymbol{\beta} + \boldsymbol{\beta}_0$.

$p(\boldsymbol{x}|\boldsymbol{\theta}_x)$ and $p(\boldsymbol{y}|\boldsymbol{\theta}_y)$ share a common general form of the exponential family distribution of an observation a given a canonical parameter vector θ, which can be written as:

$$p(a|\theta) = \exp(\theta^T a + \log q(a) - A(\theta)),$$

where $A(\theta) = \log \int \exp(\theta^T a) q(a) da$. In generalized linear models, a corresponds to the outcome and θ corresponds to the linear predictor. The derivative of $A(\theta)$ corresponds to the inverse link function h^{-1}, which provides the relationship between the mean of the distribution function and the linear predictor. The outcome is related to the linear predictor via the link function. In the case of exponential family PCA, the particular form of $A(\theta)$ function is determined by the distribution that the data sample a is assumed to follow.

Consequently, the data likelihood of SSePCA model can be written as:

$$\ell(X, \boldsymbol{y}|\theta_x, \theta_y) = \Pi_{i=1}^n p(\boldsymbol{x}_i|\theta_{xi}) p(y_i|\theta_{yi}) \tag{3.12}$$

Under the linear decomposition of θ_x and θ_y, we can further substitute Z, W and $\boldsymbol{\beta}$ into (3.12) and estimate the unknown parameters by maximizing the following log-likelihood function:

$$\begin{aligned} & \max_{Z:Z^TZ=I} \max_{W,\boldsymbol{\beta}} \sum_{i=1}^n \log p(\boldsymbol{x}_i|\boldsymbol{z}_i, W) + \sum_{i=1}^n \log p(y_i|\boldsymbol{z}_i, \boldsymbol{\beta}) \\ = & \min_{Z:Z^TZ=I} \min_{W,\boldsymbol{\beta}} \sum_{i=1}^n A_1(Z_{i:}, W) - tr(ZWX^T) + \sum_{i=1}^n A_2(Z_{i:}, \boldsymbol{\beta}) - tr(Z\boldsymbol{\beta}y^T) \end{aligned}$$

The current model can force the weights stored in the PC loading vectors to be directly associated with disease outcome when studying complex disease. However, the PC loading vectors take combinatorial contributions from all the predictors and therefore it is difficult for biological interpretation and hard to identify dominant contributing variables. To address this problem, we impose a sparsity penalty on the loading vectors similarly as what we did for SePCA to make the model interpretable

and simpler with lower degree of freedom.

$$\ell = \min_{Z:Z^TZ=I} \min_{W,\boldsymbol{\beta}} \sum_{i=1}^{n} A_1(Z_{i:}, W) - tr(ZWX^T) + \Omega(W) + \sum_{i=1}^{n} A_2(Z_{i:}, \boldsymbol{\beta}) - tr(Z\boldsymbol{\beta}y^T) \tag{3.13}$$

We reformulate our original optimization problem by applying conjugate duality as what we did for SePCA. By applying the same strategy on the generalized linear model part of (3.13), we achieve:

$$\begin{aligned}
&\min_{Z:Z^TZ=I} \min_{W,\boldsymbol{\beta}} \quad \sum_i \quad A_1(Z, W_{i:}) - W_{i:}[X^TZ]_{:i} + ||ZW + \mathbf{1}b^T||^2 + \sum_{l=1}^{L} \lambda_l |W_l|_1 \\
&\qquad\qquad + \quad A_2(Z, \boldsymbol{\beta}) - \boldsymbol{\beta}^T Z^T \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{\beta} \\
= &\min_{Z:Z^TZ=I} \min_{W,\boldsymbol{\beta}} \max_{U,\boldsymbol{v}} \quad \sum_i \quad -A_1^*(U_{i:}) + tr(Z^T(U - X)W^T) + ||ZW + \mathbf{1}b^T||^2 + \sum_{l=1}^{L} \lambda_l |W_l|_1 \\
&\qquad\qquad - \quad A_2^*(\boldsymbol{v}) + \boldsymbol{\beta}^T Z^T(\boldsymbol{v} - \boldsymbol{y}) + \boldsymbol{\beta}^T \boldsymbol{\beta}.
\end{aligned}$$

The solution can be achieved by alternatively updating unknown parameters with closed-form solutions. In each iteration, the optimal U_i is achieved when the moment matching condition holds in which U_i equals to $E(\boldsymbol{x}_i|\boldsymbol{\theta}_i)$ with $\boldsymbol{\theta_i} = W\boldsymbol{z}_i + b$. This mean variable is calculated as the derivative of $A_1(\theta_i)$. Given an estimated U_i and $\boldsymbol{z}_i$, we update b and W by solving a least square problem and a Lasso problem respectively. The optimal v_i, $\boldsymbol{\beta}$ and β_0 could be estimated based on the same logic. Given all the other estimated parameters, Z can be estimated by doing an SVD of a matrix involving both information from data matrix X and the outcome $\boldsymbol{y}$.

The detailed procedure for SSePCA is given in Algorithm 2.

To investigate the performance of SSePCA, we apply it in a simulation study for pathway based aggregate analysis of SNP data which will be illustrated in section 4. The corresponding simulation results are included in section 4.4.

Algorithm 2 Supervised sparse exponential family PCA

1. Set $\boldsymbol{b}$ as $\frac{1}{N}X^T\mathbf{1}$ and β_0 as $\frac{1}{N}\boldsymbol{y}^T\mathbf{1}$; Compute SVD of $[\eta(X-\mathbf{1}b),(1-\eta)(\boldsymbol{y}-\beta_0)] = APB^T$ and set $Z = A[,1:L]$, $W = \eta Z^T(X-\mathbf{1}b)$ and $\boldsymbol{\beta} = (1-\eta)Z^T(\boldsymbol{y}-\beta_0)$.

2. Update U by $\frac{\partial A(\Theta_x)}{\partial \Theta_x}$ where $\Theta_x = ZW + \mathbf{1}b$. Update $\boldsymbol{v}$ by $\frac{\partial A(\theta_y)}{\partial \theta_y}$where $\theta_y = Z\boldsymbol{\beta} + \beta_0$.

3. Update $\boldsymbol{b}$ by $\frac{1}{N}(\frac{\eta}{2}(X-U) - ZW)^T\mathbf{1}$ and β_0 by $\frac{1}{N}(\frac{1-\eta}{2}(\boldsymbol{y}-\boldsymbol{v}) - Z\boldsymbol{\beta})^T\mathbf{1}$.

4. Calculate $P_1 = \frac{\eta}{2}(X-U) - \mathbf{1}b$ and $P_2 = \frac{1-\eta}{2}(\boldsymbol{y}-\boldsymbol{v}) - \beta_0$. Denote $P = [P_1, P_2]$ and $Q = [W, \boldsymbol{\beta}]$. Compute SVD of $PQ^T = MDN^T$. Update Z by $M[,1:L]N^T$.

5. Given a fixed $Z = [\boldsymbol{z}_1, ..., \boldsymbol{z}_L]$,
   for $j = 1, ..., L$, $\boldsymbol{w}_j = \left(|z_j^T P_1| - \frac{\gamma_j}{2}\right)_+ Sign(z_j^T P_1)$.
   $\boldsymbol{\beta} = Z^T P_2$.

6. Repeat 2-5 until convergence.

8. Normalize $\boldsymbol{w}_j$ and scale $\boldsymbol{z}_j = \boldsymbol{z}_j||\boldsymbol{w}_j||_1$, $\beta_j = \beta_j/||\boldsymbol{w}_j||_1$. Rank $\boldsymbol{z}_j$, $\boldsymbol{w}_j$ and β_j by the decreasing order of $||\boldsymbol{w}_j||_1$.

3.6 Conclusion

We proposed a sparse model of exponential family PCA, SePCA, to enable variable selection in low-dimensional analysis of exponential family data for better systematic interpretation in real-world applications. In comparison with SPCA, it is a more suitable method for sparse learning of exponential family data. Our experimental results have empirically demonstrated that SePCA could achieve more accurate and sparse principal component loadings compared with Zou's SPCA via explicit modelings for those non-Gaussian data. By optimizing the exact log-likelihood function when analyzing binary data, SePCA_Bern outperforms an existing logistic PCA method SLPCA_MM in either clustering accuracy or sparsity of PC loadings for high dimensional binary data. Moreover, SePCA_Bern achieves much higher computational efficiency compared with SPCA and SLPCA_MM by avoiding calculation of Gram matrix or CD updates for each principle component. The elegant problem

structure of the dual form of SePCA model and the closed-form update rules lead to higher computational efficiency.

Our model is flexible and highly extensible. With integration of additional label information into the current framework, it can be adjusted as a supervised-learning model to solve classification or regression problems involving high-dimensional exponential family data. The dual-transformation strategy is still applicable for this problem due to the similar form of models shared by ePCA and GLMs. Closed-form update rules are still available for this problem. Our model also provides a way for hierarchical analysis of the latent variables based on several dominant variables. It can simultaneously estimate the principal component effects as well as the individual effects of the dominant variables. Moreover, one could also apply other regularization terms on PC loadings to achieve smoothness or perform graph-regularized learning.

4. PATHWAY BASED AGGREGATE ANALYSIS OF SNP DATA*

4.1 Introduction

As it is commonly conjectured that complex diseases arise due to disruptions by complex interplay between multiple genetic factors (that may interact with environmental exposures as well), pathway based methods enable more efficient and reproducible association analyses than conventional GWAS focusing on studying individual effects of SNPs. As in other existing pathway-based association analysis [11, 76], we take functional pathways defined a priori as functional units determining disease outcome, for example, by manual curation as in many publicly accessible pathway databases [2, 38]. As discussed in the literature review of pathway based GWAS, testing the combined signals representing a functional region is an appealing approach in pathway based aggregate association analysis due to its ability to take account of correlations among SNPs and the low degree of freedom of the test statistic. The major concern in this approach is how to summarize the optimal combined signals for a set of SNPs. Principal component analysis (PCA) is an attempting approach which has been applied for deriving combined signals [11, 74]. However, it may lose power in aggregating SNP data since PCA is usually suitable for dimension reduction of continuous data by making Gaussian assumption of the data distribution, which is however not an appropriate assumption for SNP data containing categorical values. The SePCA method illustrated in section 3 provides ways to derive more accurate PCs for SNP data as combined signals.

In this section, we study the applications of two cases of SePCA: logistic PCA [48]*

*Reprinted with permission from "Supervised logistic principal component analysis for pathway based genome-wide association studies" by Meng Lu, Jianhua Z. Huang, and Xiaoning Qian, 2012. Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, page: 52-59. Copyright ©2012 by ACM, Inc. http://doi.acm.org/10.1145/2382936.2382943.

and categorical PCA [50] in aggregate association analysis of SNP data represented by two different forms. Logistic PCA assumes Bernoulli distributions for SNP data when they are represented by a dominant model where the most prevalent homozygous pair is represented by 0 and the other two pairs are represented by 1. Categorical PCA assumes multinomial distributions for SNP data in which each SNP takes one of the three possible categorical values: AA, Aa, and aa with "a" representing the minor allele. This categorical representation model makes no bias assumption of the SNP effects compared with the dominant model. In both studies, the generalized PCs are regarded as the combined signals, and are further refined in a supervised framework to achieve the highest power for association analysis. The superiority of logistic PCA and categorical PCA for aggregate association analysis are demonstrated by both the simulation study of pathway based GWAS and real world application in detecting significantly associated pathways for Crohn's disease.

4.2 Supervised logistic PCA

In this section, we take LPCA, which is specifically designed to model and compute the optimal rank reduced representation of given SNP data [42]. As the underlying data model of LPCA fits SNP data, aggregated variables from LPCA may capture more information in the original SNP data, which will directly affect the significance analysis of corresponding pathways.

However, without integrating disease outcome information into LPCA, the derived aggregated variables only have the optimal representation of the original data (as in PCA), but may not have any discriminating power regarding disease. Furthermore, since typically only a limited number of SNPs in each pathway contribute to disruptions that trigger disease, there may be redundant information if we consider all SNPs in LPCA. In order to infer the most associated aggregated variables

for pathways, we adopt a supervised selection procedure to search for subsets of SNPs that are most associated with disease outcome. By combining this supervised SNP selection procedure together with LPCA, we expect that the derived aggregated variables by SLPCA accurately capture the association of pathways with disease and consequently we will obtain more accurate and reproducible results.

4.2.1 Review of logistic PCA

To effectively analyze high-dimensional and categorical SNP data, we first review the LPCA model [42], which generalizes traditional PCA. The SNP genotype data can be represented by a high-dimensional vector of categorical values $\{0, 1, 2\}$ for homozygous or heterozygous alleles. We focus on their binary representation with 0 representing the most prevalent homogeneous base pair (wild-type) and 1 for the other genotypes (mutant with minor alleles). This method of encoding SNPs corresponds to testing for a dominant/recessive genetic effect on outcome. Our task is to derive summary statistics for given functional pathways that aggregate weak effects from individual SNPs.

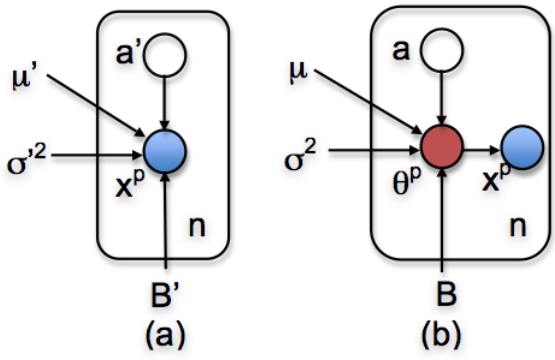

Figure 4.1: The probabilistic graphical models of $\mathbf{x}^p$ with n observations for a pathway p in (a) PCA; and (b) LPCA ([48] © 2012 by ACM).

For a given pathway p, the corresponding genotype information can be represented by a d-dimensional vector $\mathbf{x}^p$, where d is the number of SNPs within the pathway. Traditional PCA assumes that the underlying random variable $\mathbf{x}^p$ is a linear transformation of a K-dimensional latent variable $\mathbf{a}'$ ($K \leq d$) with additive Gaussian noise $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma'^2\mathbf{I})$: $\mathbf{x}^p = \boldsymbol{\mu}' + \mathbf{B}'\mathbf{a}' + \boldsymbol{\epsilon}$ with the graphical model illustrated in Fig. 4.1(a), where parameters $\boldsymbol{\mu}'$ is the mean vector and $\mathbf{B}'$ is a $d \times K$ linear transformation matrix whose columns correspond to principal components. For a given $d \times n$ data matrix $\mathbf{X}^p$ with n samples, we can compute the optimal reduced rank representation $\hat{\mathbf{X}}^p = \boldsymbol{\mu}'\mathbf{1}^T + \mathbf{B}'\mathbf{A}'$ with the minimum mean squared distance to $\mathbf{X}^p$ [42], in which $\mathbf{A}'$ is a $K \times n$ matrix with corresponding principal scores. Unlike traditional PCA, we assume in LPCA that each individual SNP x follows a Bernoulli distribution: $Pr(x|\theta) = \sigma(\theta)^x \sigma(-\theta)^{(1-x)}$, where $\sigma(\theta) = [1 + \exp(-\theta)]^{-1}$ is the logistic function, and θ is the canonical parameter of the Bernoulli distribution in the exponential families. For a given pathway p, we write the log-likelihood of canonical parameters Θ, a $d \times n$ matrix, for the given SNP data $\mathbf{X}^p$:

$$L(\Theta) = Pr(\mathbf{X}^p|\Theta) = \prod_{i=1}^{d}\prod_{j=1}^{n} \sigma(\theta_{ij})^{x_{ij}} \sigma(-\theta_{ij})^{(1-x_{ij})}. \tag{4.1}$$

Assuming that Θ has a reduced rank representation $\Theta = \boldsymbol{\mu}\mathbf{1}^T + \mathbf{B}\mathbf{A}$, where $\boldsymbol{\mu}$, $\mathbf{B}$, $\mathbf{A}$ are correspondingly the mean, principal components (loading vectors), and principal scores similarly as in traditional PCA. We note that the difference is that in LPCA, we have introduced a new intermediate random variable $\boldsymbol{\theta}^p$ to model canonical parameters of SNP data distribution and this new random variable itself is associated with the K-dimensional latent variable $\mathbf{a}$ as shown in Fig. 4.1(b). Substituting $\boldsymbol{\mu}$, $\mathbf{B}$, and $\mathbf{A}$ into (4.1), we can compute the log-likelihood:

$$
\begin{aligned}
l(\boldsymbol{\mu}, A, B) &= \sum_{i=1}^{d}\sum_{j=1}^{n}\Big\{x_{ij}\log\sigma\Big(\mu_i + \sum_{k=1}^{K} a_k^j b_k^i\Big) \\
&+ (1-x_{ij})\log\sigma\Big(-\mu_i - \sum_{k=1}^{K} a_k^j b_k^i\Big)\Big\}. \qquad (4.2)
\end{aligned}
$$

We can rewrite the principal score matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K]^T$ by row vectors $\mathbf{a}_k$'s, which are the aggregated variables of interest. These corresponding principal scores will be used to compute test statistics to detect significant pathways that are associated with disease. Magnitudes in principal components (loading vectors) $\mathbf{b}_k$ reflect the actual contributions of corresponding SNPs to summary statistics.

To estimate $\boldsymbol{\mu}$, $\mathbf{B}$, especially principal scores $\mathbf{A}$ in LPCA, we need to optimize the log-likelihood function (4.2) which is highly non-convex with orthonormal constraints for $\mathbf{b}_k$'s. We construct tight lower-bound functions and iteratively solve the optimization problem as done in the Majorization-Minimization (MM) algorithm [42]. The algorithmic detail is discussed in the original paper [42].

With derived principal scores $\mathbf{A}$ from LPCA as aggregated variables for each pathway, we can estimate their statistical significance by analyzing association of these aggregated variables with disease outcome. Specifically, With derived principal scores $\mathbf{A}$ as predictors and disease outcome as the response variable, we learn a logistic regression model for each pathway p by assuming the following model:

$$
\log(\frac{\pi_j}{1-\pi_j}) = \beta_0 + \sum_{k=1}^{K}\beta_k a_k^j \qquad (4.3)
$$

where π_j is the posterior probability of the jth subject having disease given known aggregated variables $\mathbf{A}$ and statistical significance of β_k's indicates whether the pathway is significantly associated with disease outcome. We take the first principal score

with $K = 1$ to use one summary statistic for each pathway, which has been done in previous studies [11] and has provided satisfactory performances in our experiments.

4.2.2 Methodology

In LPCA, the derived aggregated variables only depend on the data distribution but are not directly related to disease outcome. As we only take the first principal scores for each pathway, a simple LPCA may not give the principal components and corresponding scores that are most associated with outcome. In order to make sure that the resulting principal components and summary statistics contain the most influential information regarding disease, we propose SLPCA method by adopting the similar idea in SPCA [4, 5, 11] to derive summary statistics that are associated with outcome as much as possible.

The key idea of SLPCA for association study is to estimate pathway-level significance by deriving the most associated summary statistics from a group of SNPs with maximum combined effect in pathways. However, simply selecting several top SNPs based on their individual significance as in conventional GWAS has limited power as the most disease-associated group of SNPs may contain SNPs with relatively mild individual effects but significant combined effects due to the interactions among these SNPs. To solve this problem while avoiding the time consuming tasks as in traditional forward or backward feature selection [65] to select such most associated SNP groups from a large number of potential candidates, we adopt a similar heuristic procedure in a recent pathway based analysis using SPCA [11]. Specifically, in each pathway, we first rank its mapped SNPs based on their statistical association with outcome and group SNPs as candidate units by gradually increasing the size as in forward selection. We implement LPCA to derive multiple potential summary statistics for the formed candidate groups respectively. The final statistical signifi-

cance of each pathway is the best value derived from the candidate group with the most discriminating power in (4.3).

In order to have a fair comparison of our proposed SLPCA with SPCA for pathway based analysis, we follow the same SNP selection procedure in [11]. For each pathway, we sequentially form 20 candidate groups by selecting 20 thresholds at each increment of 5 percentiles of total SNPs based on SNP association measure, which is computed as the coefficient p-value by fitting a logistic regression model with genotype data and outcome. For each of these 20 candidate groups, we implement LPCA to derive the first principal scores and thereafter compute t-statistic $(= \hat{\beta}_1/s.e.(\hat{\beta}_1))$ as the test statistics based on (4.3). For each pathway, the final test statistic is based on the maximum absolute value of t-statistics among all the candidate groups, which is denoted as M statistic:

$$M = \{t_\ell : |t_\ell| = max_{1\leq\ell\leq 20}|t_\ell|\}, t_\ell = \hat{\beta}_1^\ell/s.e.(\hat{\beta}_1^\ell), \tag{4.4}$$

where ℓ indexes candidate groups of SNPs. This new test statistic can not be approximated well by t-distribution. Hence, we estimate the final pathway association significance by computing nominal p-values based on permutation test to generate a null distribution of M statistics with random disease outcome.

In summary, our SLPCA takes the following steps to measure the significant association of pathways with disease:

(1) *Generate candidate SNP groups for each pathway*

For each individual SNP assigned to a given pathway, its significance (p-value) can be computed by fitting a logistic regression model. Given all SNPs belonging to a pathway, we generate 20 incremental candidate groups by setting 20 thresholds at each increment of 5 percentiles of p-values for those SNPs. Hence, for each pathway,

20 groups of SNPs $\{S_1, \ldots, S_{20}\}$ are formed by sequentially grouping SNPs with p-values less than each corresponding threshold.

(2) *LPCA on candidate SNP groups*

LPCA can be implemented to compute the first PC scores for 20 candidate groups respectively in each pathway.

(3) *Calculate M statistics for candidate groups*

For each candidate group $S_\ell (1 \leq \ell \leq 20)$, we fit the logistic regression model (4.3) using the corresponding first PC scores and estimate t-statistic $t_\ell = \hat{\beta}_1^\ell / s.e.(\hat{\beta}_1^\ell)$. Let $M = \{t_\ell : |t_\ell| = max_{1 \leq \ell \leq 20} |t_\ell|\}$.

(4) *Estimate the null distribution of M statistics*

For each pathway, we perform permutation test by generating random disease status for each sample from a Bernoulli distribution with the success probability set to the disease prevalence. Based on randomly generated outcomes, the empirical null distribution of M statistics can be estimated by repeating steps (1) to (3) and pooling together corresponding M values from all pathways as a random sample from the null distribution of M.

(5) *Calculate p-value for each pathway*

Given a null distribution of M statistic and M values for all pathways based on true disease status, an empirical p-value for each pathway can be calculated to estimate the pathway significance. This provides a self-contained test which compares pathways to the non-associated genomic background.

With such an implementation, our SLPCA has the potential to aggregate weak signals from individual SNPs with the explicit modeling of categorical SNP data considering outcome.

4.2.3 Simulation study

To demonstrate the advantages of explicit modeling of categorical SNP data in our SLPCA model, we first carry out a simulation study following the experiment design in [11], to which we compare the performance of SLPCA for the detection of causal pathways. First, to account for the pathway size effect in association study, 50 gene sets are randomly selected as testing pathways from those Gene Ontology (GO) categories from Biological Process Ontology in GO database [2]. Based on the Ensembl database (Release 67) [23], each selected pathway contains SNPs within 5KB upstream or downstream from its corresponding genes. After limiting SNPs to those on the Perlegen GV4 chip and filtering them for quality control to guarantee minor allele frequency (MAF) > 0.05, we have 3,584 SNPs in total across the selected 50 pathways with their size ranging from 25 to 195. Further, to construct causal pathways, five pathways with SNPs across different chromosomes are randomly selected from 50 pathways. These five causal pathways have 105,116,171,177, and 195 SNPs, respectively.

In order to generate samples of SNP genotype data with realistic allele frequencies and LD patterns, we use HAP-SAMPLE simulation tool (command-line version) [87]. HAP-SAMPLE simulates genotype data for case-control studies by resampling from HapMap Phase I/II public database (Release 21a) [16]. In this simulation, we use the Caucasian cohort (CEU) population database as the source dataset from which samples of SNP data are generated. We further generate case-control status based on sampled genotype data using the following disease model:

$$\log(f/(1-f)) = \beta_0 + \beta_1 g_1 + \beta_2 g_2 + \cdots + \beta_D g_D \tag{4.5}$$

where D is the total number of causal SNPs associated with outcome status; g_i

represents the genotype of causal SNP i ($1 \leq i \leq D$); and $f = \Pr(disease|g_1, \ldots, g_D)$ is the probability of disease given genotypes $\{g_1, \ldots, g_D\}$. To guarantee independent effects from different causal SNPs in our experiments, we randomly select D causal SNPs on different chromosomes in each selected causal pathway, where D=3, 4, and 5. The coefficients $\beta_i (1 \leq i \leq D)$ also are independently generated from a Gaussian distribution $N(\mu, \sigma^2)$. As in [11], we have tested four different scenarios with $\mu = \log(1.1)$ and σ^2= 0.15, 0.2, 0.25, and 0.3 respectively for each D, resulting in $4 \times 3 = 12$ different settings. We note that the coefficient β_i reflects the effect of a causal SNP i in affecting disease outcome. A larger absolute value of β_i indicates that SNP i has more significant association with disease. We assume disease prevalence to be 5% and the estimation of β_0 follows the solution in [43]. Given genotype data and values of $\{\beta_0, \beta_1, \ldots, \beta_5\}$, f is computed using the disease model (4.5) with which the absolute risk (AR) is computed in HAP-SAMPLE to generate case-control status for five causal pathways. In the remaining 45 null pathways, the case-control status are randomly generated from a Bernoulli distribution with the probability of disease equal to 5%. We generate 500 case and control samples respectively for each scenario of our simulation experiments and have replicated each scenario 100 times, resulting in 500 (5×100) causal pathways and 4500 (45×100) null pathways in total.

As in [11], statistical power is used as the criterion to evaluate the performance of our SLPCA, which is compared with SPCA in [11]. Specifically, it is computed as the proportion of detected causal pathways that are significantly associated with case-control outcome using two methods. We have independently implemented SPCA for comparison as we do not have the access to the original code by the authors in [11].

We first compute the power at the significance level 0.05 for all the scenarios using both methods. Based on the results shown in Table 4.1, it is clear that our SLPCA consistently performs better than SPCA due to the explicit modeling of

Table 4.1: Comparison of power obtained by SLPCA and SPCA at significance level 0.05 ([48] © 2012 by ACM).

Scenario	SLPCA	SPCA
D=3, σ^2=0.15	0.68	0.61
D=3, σ^2=0.2	0.74	0.68
D=3, σ^2=0.25	0.76	0.72
D=3, σ^2=0.3	0.78	0.76
D=4, σ^2=0.15	0.78	0.71
D=4, σ^2=0.2	0.79	0.76
D=4, σ^2=0.25	0.84	0.82
D=4, σ^2=0.3	0.85	0.83
D=5, σ^2=0.15	0.82	0.79
D=5, σ^2=0.2	0.86	0.84
D=5, σ^2=0.25	0.92	0.90
D=5, σ^2=0.3	0.93	0.92

categorical SNP data. We further investigate the performances of both methods for a more thorough analysis with different false positive rates at ten different levels $\{0.001, 0.002, 0.004, 0.006, 0.008, 0.01, 0.02, 0.03, 0.04, 0.05\}$. We plot the Receiver Operating Characteristic (ROC) curves for each scenario obtained by two methods in Figure 4.2. From the figure, SLPCA again has consistently higher power than SPCA at different significance levels under different scenarios. In addition, we notice that SPCA and SLPCA both have higher power with increasing variances of coefficients β_i's in the disease model (4.5). As β_i indicates the relative effect of the ith causal SNP and the absolute value of β_i tends to be large if it has a large variance, the ith SNP has larger effect on outcome with increasing variance. In other words, with a larger variance, power increases as causal SNPs tend to have higher significance and causal pathways are easier to be detected. Our results have demonstrated this tendency. More importantly, the performance improvement of SLPCA over SPCA is more obvious when we have smaller variances for β_i's, which indicates that causal

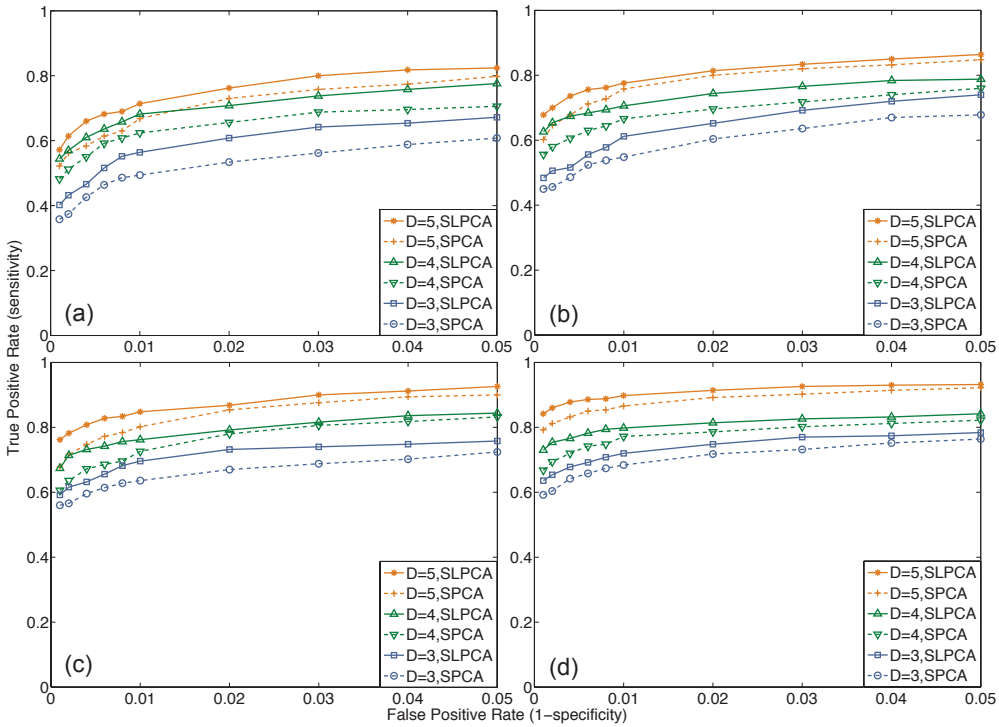

Figure 4.2: ROC curves for SPCA and SLPCA with D=3,4,5: (a) σ^2=0.15; (b) σ^2=0.2; (c) σ^2=0.25; (d) σ^2=0.3 ([48] © 2012 by ACM).

SNPs have less effect on disease outcome and hence causal pathways are more difficult to detect. Due to the better modeling of SNP data to capture more influential information, SLPCA has demonstrated its advantages over SPCA in these difficult cases.

We also find that SLPCA and SPCA both have larger power with more causal SNPs in the disease model (4.5) with the same variances of coefficients β_i's. It is natural that the enrichment of causal SNPs in a pathway contributes to its significance. Moreover, in the case with fewer causal SNPs (D=3) and smaller coefficient variances (σ^2=0.15) in causal pathways, it is more difficult to detect causal pathways and the power is relatively low. These cases resemble the situations of individual SNPs with weak effects, which pose a challenging problem for both methods. However, we find

that the superiority of SLPCA over SPCA is prominent especially for these difficult cases with fewer causal SNPs in the disease model, again due to the explicit modeling of SNP data by LPCA.

Overall, it is clear that our SLPCA can achieve higher power to identify SNPs that are significantly associated with disease status, especially for SNPs with relatively weak effects. By explicitly modeling the distribution of categorical SNP data with LPCA instead of using traditional PCA with the underlying assumption of Gaussian distribution, which may not perfectly capture the characteristics of SNP data, SLPCA has demonstrated its advantages over traditional SPCA for pathway based association analysis.

4.2.4 Analysis for Crohn's disease

Using the case-control data from Wellcome Trust Case Control Consortium (WTCCC) (http://www.wtccc.org.uk), we further apply our SLPCA model to identify pathways that are significantly associated with Crohn's Disease (CD), which is a inflammatory bowel disease conjectured to be affected by multiple genetic factors as well as environmental exposures. WTCCC provides 2,005 case samples and 3,004 control samples consisting of 1,504 individuals from the 1958 British Birth Cohort and 1,500 individuals from the UK blood services. These samples are genotyped by Affymetrix GeneChip 500K. After quality control, there are 1,748 case samples and 2,938 controls in total with 469,557 SNPs in each sample [82].

Table 4.2: Representative pathways identified by SLPCA in WTCCC Crohn's disease data set ([48] © 2012 by ACM).

Pathway	No. of genes	No. of SNPs
Protein oligomerization	40	706
Positive regulation of cytokine secretion	10	117
Interleukin 1 secretion	10	116
Positive regulation of DNA binding	26	300

Table 4.2 continued

Pathway	No. of genes	No. of SNPs
Regulation of transcription factor activity	40	375
Positive regulation of binding	28	318
Positive regulation of transcription factor activity	24	294
Regulation of binding	58	535
Regulation of DNA binding	47	406
Activation of NF-κB transcription factor	18	261
Detection of biotic stimulus	10	75
Response to bacterium	30	177
Detection of chemical stimulus	18	497
Defense response to bacterium	24	146
Regulation of cytokine production	25	244
KEGG NOD like receptor signaling pathway	62	499
Positive regulation of I-κ kinase NF-κB cascade	80	442
Regulation of secretion	40	405
Positive regulation of cytokine production	15	131
Positive regulation of protein secretion	12	126
Regulation of cytokine secretion	16	135
Detection of external stimulus	23	178
Positive regulation of secretion	20	168
Peptide metabolic process	10	474
Cytokine secretion	18	190
Regulation of protein secretion	22	223
Protein secretion	32	364
Regulation of I-κ kinase NF-κB cascade	86	502
Transcription initiation	35	279
Cytokine production	72	593
I-κ kinase NF-κB cascade	107	704
Regulation of signal transduction	213	2440
Positive regulation of T cell proliferation	13	130
Regulation of T cell proliferation	16	140
Regulation of T cell activation	28	225
T cell proliferation	19	200
KEGG JAK-STAT signaling pathway	155	1205
KEGG antigen processing and presentation	89	318
Glycerophospholipid biosynthetic process	30	158
Glycerophospholipid metabolic process	43	297
Phospholipid biosynthetic process	39	334
ST phosphoinositide 3 kinase pathway	33	402
Phospholipid metabolic process	71	568

To carry out pathway-based genome-wide analysis, we again follow the experiment in [11] to obtain the pathway information from Molecular Signature Database (MSigDB: http://www.broadinstitute.org/gsea/msigdb), in which we collect two categories of pathways—C2-CP and C5-BP, corresponding to annotated canonical pathways (CP) from online pathway databases such as KEGG, BioCarta and Reactome pathway databases and GO biological processes (BP) respectively. To increase the specificity by avoiding overly broad pathways, we further filter out those pathways with more than 250 genes, resulting in 866 CPs and 751 BPs with 8,354 uniques genes in total. We map SNPs in CD data to these pathways based on the *Homo sapiens* Variation (dbSNP 130) and *Homo sapiens* genes (GRCh37.p7) datasets in the Ensembl database (Ensembl 67) using BiomaRt (http://www.biomart.org/). Similarly as in our simulation study, SNPs located within 5KB upstream or downstream of corresponding genes are assigned to different pathways respectively. With the WTCCC CD data and mapped SNPs, we implement SLPCA to each pathway and nominal p-values from permutation tests are estimated to identify significant pathways at significance level 0.05.

From our results, the identified pathways are typically involved in the following cellular functions: (1) Regulation of protein secretion and transcription factor activity regulation; (2) Detection of stimulus and response to bacterium; (3) NOD like receptor signaling and regulation of the Nuclear Factor-κB (NF-κB); (4) Regulation of cytokine production and secretion; (5) Regulation of signal transduction; (6) T-cell proliferation or activation related pathways; and (7) Interleukin-1 secretion. Most of these pathways are related to the immune system, which reacts abnormally in people with CD. Some of representative significant pathways are given in Table 4.2. Among these pathways, NOD2 appears in multiple pathways, especially in pathways involving NF-κB regulation and other signaling pathways. It is indeed the first identified

gene associated with CD in previous analysis [62]. It plays an important role in the immune response by recognizing specific pattern of intracellular bacteria and stimulating immune reaction through activating the NF-κB protein. T-cell proliferation or activation related pathways contain genes IL12B, IL18, IL21 in common. Among these genes, IL12B has been verified as an associated gene in [6]. JAK-STAT signaling pathway contains IL23R, a previously identified associated gene for CD [59]. We also have identified antigen processing and presentation pathway containing HLA-DQA2, HLA-DOB, and HLA-DRA which have been found highly associated with CD by Ballard *et al.* [6] We also find a group of significant pathways related with lipid metabolism including glycerophospholipid metabolic process, phospholipid metabolic process and glycerophospholipid biosynthesis pathways.

We further investigate frequent genes that appear in these significant pathways associated with CD. The top ten most frequent genes enriched in top 40 pathways are NOD2, PYCARD, BCL10, NLRP3, CARD8, PYDC1, RELA, UBE2N, CRTAM and NOD1. In addition to NOD2 as the first identified CD susceptibility gene, RELA is another important member of NF-κB family, which plays critical roles in regulating the cellular response to infection. Incorrect regulation of NF-κB is thought to be related with inflammatory or autoimmune diseases, which may trigger CD [29]. Other members of the NOD-like receptor (NLR) family: NOD1, NLRP3, NLRC4, NLRC12 and NLRP2 also emerge from these significant pathways. These NLR proteins have been proven to be related in cytokine processing and NF-κB activation and are widely accepted as critical to regulate the innate immune response [78]. In previous association analysis [63], the protein encoded by BCL10 has been shown to induce apoptosis and to activate NF-κB. Moreover, another gene MALT1 which synergizes in the activation of NF-κB with this protein also occurs frequently in identified pathways. Tumor Necrosis Factor (TNF) as pro-inflammatory cytokines, is also

considered as an inducer of NF-κB activity and most TNF receptor members activate NF-κB pathways through their interaction with TNF Receptor-Associated Factors (TRAFs) [19]. Our results confirm that TNF, TRAF6 and genes interactive with TRAF6 such as UBE2N and UBE2V1 are included in significant pathways associated with CD. In addition, stimulation of Toll-Like Receptors (TLRs) identified as specific pattern recognition molecules will also lead to activation of NF-κB [39]. We find several TLR genes in these identified pathways, including TLR6, which are believed to be key regulators of both innate and adaptive immune responses [29].

In summary, our SLPCA method shows the potential to detect association effects for CD and our results have a large overlap with the results reported in [11]. These findings agree well with the recent literature of multiple GWA studies [6, 77, 80]. But due to the difficulty of obtaining the exact same testing dataset as in [11] and the fact that there still lacks a complete understanding of the etiology of CD, it is difficult to give a conclusive evaluation. Further validation of the proposed method will be the focus of our future research when a better-understood benchmark data is available.

4.2.5 Conclusion

SLPCA captures more information of original data compared with SPCA by explicit modeling of SNP data distribution and guarantees the association of aggregated variables corresponding to pathways with disease outcome by a heuristic selection of SNPs in pathway. Both simulation data and Crohn's disease data with real LD structure have testified our SLPCA method has favorable results compared to SPCA. In our future research for better "supervised" learning to improve on heuristic SNP selection, we will consider to add sparsity penalty in SLPCA as in sparse PCA [91] to automatically generate subsets of SNPs with aggregated variables that

are significantly associated with outcome. We will further explore new statistical learning models to integrate outcome information directly into the LPCA procedure to achieve adaptive instead of heuristic summary statistics. In addition, the power of current pathway-based methods for GWAS is limited by the curated pathway definition. Network-based analysis for GWAS provides a promising framework considering interaction among genes to better understand the underlying mechanisms of disease development.

4.3 Supervised categorical PCA

We have previously developed logistic PCA (LPCA) methods [48, 42] for gene- and pathway-based analysis of SNP data by explicitly modeling the categorical nature of SNP data. For LPCA, we first transform the genotype data from the domain $\{0, 1, 2\}$ to binary data $\{0, 1\}$, which is assumed to follow a Bernoulli distribution. We have obtained promising results compared with traditional PCA-based SNP analysis that inherently assumes continuous normally distributed SNP data. However, due to the data transformation, LPCA also has an inherent assumption that the risk effect takes either recessive or dominant model. The important information in the original SNP data, especially when we have more general underlying risk effect models, may be lost due to the transformation.

In this section, we develop a more general PCA denoted as categorical PCA (CPCA) that does not make any specific model assumptions of the effect of genetic mutants on the given trait. We first derive an optimization algorithm for CPCA suitable for categorical data analysis. Similar as conventional PCA, CPCA finds the optimal linear combinations that best explain the observed data but may not derive the principal components that are the most associated with a trait of interest. In order to derive the best principal components capturing the maximum combined

effect from multiple SNPs with respect to a given trait of interest, we then apply it in a similar supervised framework as SPCA method.

By our supervised CPCA (SCPCA) [50], the resulting principal components have the most discriminating power and can be further taken as aggregated predictors for the disease outcome. It ensures that the principal components obtained by CPCA are not deteriorated by noisy SNPs that are irrelevant with the trait. With a more general data model and direct integration of trait information for identifying the most influential SNPs in a functional region, our preliminary results on both simulated genotype data and the Wellcome Trust Case Control Consortium (WTCCC) Crohn's Disease (CD) data [82] have demonstrated the advantages of our supervised CPCA over traditional SPCA and supervised LPCA for gene-based and pathway-based aggregated association analysis.

4.3.1 Methodology

It is desirable to develop variants of PCA based on respective modelings for different types of data such as integer, categorical, binary, and nonnegative data. PCA has been extended to the exponential family in previous work [14, 79, 27] by assuming data follows a general form of exponential family distributions:

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}_i) = \exp\Big(\boldsymbol{\theta}_i^T\boldsymbol{x}_i + \log p_0(\boldsymbol{x}_i) - G(\boldsymbol{\theta}_i)\Big).$$

Here, $\boldsymbol{x}_i \in R^d$ is the ith data point and $\boldsymbol{\theta}_i \in R^d$ is the "natural parameter" of the corresponding distribution. $G(\boldsymbol{\theta}_i)$ is a function of the form $\log \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} p_0(\boldsymbol{x}_i) \exp(\boldsymbol{\theta}_i^T \boldsymbol{x}_i)$ to ensure that the sum of $p(\boldsymbol{x}_i|\boldsymbol{\theta}_i)$ over the domain of $\boldsymbol{x}_i$ equals to 1 and p_0 is a function depending only on $\boldsymbol{x}_i$. Different members in the exponential family have their respective G functions specified in [14], which results in different distributions and different generalization of PCA. To generalize PCA based on the distributions of

exponential family, it starts from an important assumption of $\boldsymbol{\theta}_i$ where it is assumed to be a linear combination of bases $W = [\boldsymbol{w}_1, ..., \boldsymbol{w}_l]$ with the minimum reconstruction loss represented as $\boldsymbol{\theta}_i = \sum_{q=1}^{l} z_{iq}\boldsymbol{w}_q + \boldsymbol{\mu}$. The bases and their corresponding weights $\boldsymbol{z}_i = \{z_{iq}\}$ are called as principal component loading vectors and principal component scores respectively. Given the distribution for data points $\boldsymbol{x}_i$ and the representation of $\boldsymbol{\theta}_i$, the conditional log-likelihood function of the n data points with respect to their principal components can be written as:

$$\begin{aligned} \ell &= \sum_{i=1}^{n} \Big(\boldsymbol{\theta}_i^T \boldsymbol{x}_i - G(\boldsymbol{\theta}_i) \Big) \\ &= \sum_{i=1}^{n} \Big((\boldsymbol{z}_i^T W^T + \boldsymbol{\mu}^T)\boldsymbol{x}_i - G(W\boldsymbol{z}_i + \boldsymbol{\mu}) \Big), \end{aligned} \tag{4.6}$$

where p_0 can be considered as a constant term and ignored here. The principal components resulted from a generalized PCA can then be estimated by maximizing (4.6). In a special case of the data following a normal distribution, it turns to be the tranditional PCA derived by maximizing this log-likelihood with $G(\boldsymbol{\theta}_i)$ having a form of $\boldsymbol{\theta}_i^T\boldsymbol{\theta}_i/2$ where the corresponding parameters are $\boldsymbol{z}_i$, W, and $\boldsymbol{\mu}$. As mentioned earlier, the SNP data in GWAS only has three different genotypes $\{00, 10/01, 11\}$. We focus on the derivation of exponential family PCA for categorical data denoted as CPCA in the equivalent categorical domain $\{0, 1, 2\}$ instead of taking numerical values. For categorical SNP data which follows a multinomial distribution, each observation $\boldsymbol{x}_i$ is expressed as a set of observation vectors $\boldsymbol{x}_i^0, \boldsymbol{x}_i^1, \boldsymbol{x}_i^2$ with only 1 and 0 elements. A 1 or 0 in $\boldsymbol{x}_i^k, k \in \{0, 1, 2\}$ denotes the corresponding outcome equals to k or not. Each observation vector $\boldsymbol{x}_i^k$ corresponds to a natural parameter vector $\boldsymbol{\theta}_i^k$ determining the success probabilities of the outcomes belonging to category k. Each $\boldsymbol{\theta}_i^k$ is projected to a low-dimensional space spanned by its respective basis $W^k = [\boldsymbol{w}_1^k, ..., \boldsymbol{w}_n^k]$, sharing the common principal component scores $\boldsymbol{z}_i$. It can then

be represented as $\boldsymbol{\theta}_i^k = W^k \boldsymbol{z}_i + \boldsymbol{\mu}^k$. For multinomial distributions, the corresponding G function for $\boldsymbol{\theta}_i^k$'s is $\sum_{j=1}^{d} \log \sum_{k=1}^{c} \exp(\theta_{ij}^k)$, where θ_{ij} is the j-th element of $\boldsymbol{\theta}_i$. By substituting this G function into (4.6) and replacing $\boldsymbol{\theta}_i^k$ by the actual parameters $\boldsymbol{z}_i$, W^k, and $\boldsymbol{\mu}^k$, the log-likelihood function to be maximized for CPCA is rewritten as:

$$\begin{aligned} \ell &= \sum_{i=1}^{n} \{ \sum_{k=1}^{c} \boldsymbol{\theta}_i^{kT} \boldsymbol{x}_i^k - G(\{\boldsymbol{\theta}_i^k\}) \} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{d} \{ \sum_{k=1}^{c} (Z_{i:} W_{:j}^{kT} + \mu_j^k) X_{ij}^k - \log \sum_{k=1}^{c} \exp(Z_{i:} W_{:j}^{kT} + \mu_j^k) \} \end{aligned} \tag{4.7}$$

where $X^k = [\boldsymbol{x}_1^{kT}; ...; \boldsymbol{x}_n^{kT}]$, $Z = [\boldsymbol{z}_1^T; ...; \boldsymbol{z}_n^T]$ and $W^{kT} = [\boldsymbol{w}_1^{kT}; ...; \boldsymbol{w}_l^{kT}]$. $Z_{i:}$, $W_{:j}^{kT}$ and μ_j^k represent the i-th row of Z, the j-th column of W^{kT} and the j-th element of $\boldsymbol{\mu}^k$ respectively.

The principal component scores Z and principal component loading matrix W could be estimated by maximizing this log-likelihood function with the constraint that Z has orthonormal columns. We implement Newton's method for gradient ascent search for the local maximum as the objective function is not jointly concave with respect to Z, W, and $\boldsymbol{\mu}$. Given the objective function (4.7) with respect to $Z_{i:}$,$W_{:j}^{kT}$ and μ_j^k, we update $Z_{i:}$,$W_{:j}^{kT}$ and μ_j^k by computing their respective first-partial derivative and Hessian matrix for each iteration in Newton's method. Specifically,

$$Z'_{i:} = Z_{i:} - H(Z_{i:})^{-1} g(Z_{i:}), \tag{4.8}$$

where $Z'_{i:}$ represents the updated principal component scores in each iteration; $g(Z_{i:})$ and $H(Z_{i:})$ denote the first derivative and Hessian matrix of the objective function

ℓ with respective to $Z_{i:}$. By basic calculus, $g(Z_{i:})$ is computed as:

$$\begin{aligned} g(Z_{i:}) = \frac{\partial l}{\partial Z_{i:}} &= \sum_{j=1}^{d}\sum_{k=1}^{c}(W_{j:}^{k}X_{ij}^{k} - W_{j:}^{k}P_{ij}^{k}) \\ &= \sum_{k=1}^{c}(X_{i:}^{k} - P_{i:}^{k})W^{k}, \end{aligned}$$

where $P_{ij}^{k} = \frac{\exp(Z_{i:}W_{:j}^{kT})}{\sum_{k=1}^{c}\exp(Z_{i:}W_{:j}^{kT})}$. Similarly, $H(Z_{i:})$ is computed as:

$$H(Z_{i:}) = \sum_{j=1}^{d}\sum_{k=1}^{c}(P_{ij}^{k2} - P_{ij}^{k})W_{:j}^{kT}W_{j:}^{k}$$

In each iteration, we also alternatively update $W_{:j}^{kT}$ and μ_{j}^{k} based on the following equations:

$$W_{:j}^{\prime kT} = W_{:j}^{kT} - H(W_{:j}^{kT})^{-1}g(W_{:j}^{kT}), \tag{4.9}$$

$$\mu_{j}^{\prime k} = \mu_{j}^{k} - H(\mu_{j}^{k})^{-1}g(\mu_{j}^{k}), \tag{4.10}$$

and we have:

$$\begin{aligned} g(W_{:j}^{kT}) = \frac{\partial l}{\partial W_{:j}^{kT}} &= \sum_{i=1}^{n}(Z_{i:}^{T}X_{ij}^{k} - Z_{i:}^{T}P_{ij}^{k}) \\ &= Z^{T}(X_{:j}^{k} - P_{:j}^{k}). \end{aligned}$$

$$H(W_{:j}^{kT}) = \sum_{i=1}^{n}(P_{ij}^{k2} - P_{ij}^{k})Z_{i:}^{T}Z_{i:}$$

$$g(\mu_j^k) = \sum_{i=1}^{n}(X_{ij}^k - P_{ij}^k)$$

$$H(\mu_j^k) = \sum_{i=1}^{n}(P_{ij}^{k2} - P_{ij}^k)$$

The optimal solution of the corresponding parameters Z,W^{kT}, and $\boldsymbol{\mu}$ can be estimated by the following Algorithm 3. As any non-convex optimization problem, our algorithm is not guaranteed to converge to a global maximum. To overcome the problem of being trapped by local optima, we randomly start the algorithm with different initialization values several times and find the best solution with the maximum likelihood value. The time complexity for this whole procedure is $O(ks^3)$ where $s = min(n, d)$ and k is number of iterations it takes to converge. Specifically, the calculation of the first derivatives and Hessian matrices takes $O(dnl)$ and $O(dnl^2)$ respectively. The update of W, Z and $\boldsymbol{\mu}$ takes $O(ql^3)$ where $q = max(n, d)$. The whole time complexity is mainly determined by QR decomposition procedure which takes $O(s^3)$ in each iteration. Our CPCA is in the same magnitude of time complexity as LPCA. Although PCA has a lower time complexity $O(nd^2 + d^3)$ if $n > d$, one should be aware that our algorithms are designed for more general risk effect models and may achieve better performance with reasonable sacrifice on running time.

Similarly, we also apply CPCA in a similar supervised framework as SLPCA to guaranttee the derived PCs related with the disease. Our supervised CPCA takes the following steps to perform aggregated association analysis of a trait for a SNP set S:

(1) *Generate candidate SNP subsets for a SNP set S*

For each individual SNP in S, its statistical significance reflected by the corresponding p-value can be computed by fitting a logistic regression model. Given all SNPs

Algorithm 3 Categorical PCA (CPCA)

1. Initialize with $\boldsymbol{\mu} = (\mu_1^k, ..., \mu_d^k)^T$, $Z = [Z_{1:}; ...; Z_{n:}]$ and $W^k = [W_{1:}^k; ...; W_{d:}^k]$ by random values. Compute the transpose of W^k: $W^{kT} = (W_{:1}^{kT}, ..., W_{:d}^{kT})$.
2. Compute $g(Z_{i:})$, $H(Z_{i:})$, $g(W_{:j}^{kT})$, $H(W_{:j}^{kT})$, $g(\mu_j^k)$, and $H(\mu_j^k)$ respectively.
3. Update Z by $Z = [Z_{1:}; ...; Z_{n:}]$ where each $Z_{i:}$ is updated based on (4.8) respectively. Compute the QR decomposition $Z = QR$ and replace Z by Q for orthonormality constraints.
4. Update W^{kT} by $W^{kT} = [W_{:1}^{kT}, ..., W_{:d}^{kT}]$ where $W_{:j}^{kT}$'s are updated by (4.9) respectively.
5. Update $\boldsymbol{\mu}^k$ by $\boldsymbol{\mu}^k = [\mu_1^k, ..., \mu_d^k]^T$ based on (4.10) respectively.
6. Repeat steps 2 through 5 until convergence.

in S, we generate 20 incremental candidate subsets by setting 20 thresholds at each increment of 5 percentiles of p-values for those SNPs. Hence, 20 subsets of SNPs $\{S_1, \ldots, S_{20}\}$ are formed by sequentially grouping SNPs with p-values less than each corresponding threshold.

(2) *CPCA on candidate SNP subsets*

CPCA can be implemented to compute the first PC scores for 20 candidate subsets respectively.

(3) *Calculate M statistic for a SNP set S*

For each candidate subset $S_\ell (1 \leq \ell \leq 20)$, we fit the logistic regression model (4.3) using the corresponding first PC scores and estimate t-statistic $t_\ell = \hat{\beta}_1^\ell / s.e.(\hat{\beta}_1^\ell)$. Let $M = \{t_\ell : |t_\ell| = max_{1 \leq \ell \leq 20} |t_\ell|\}$.

(4) *Estimate the null distribution of M statistic*

We perform a permutation test by generating random trait status for each sample from a Bernoulli distribution with the success probability set to the disease prevalence. Based on randomly generated outcomes, the empirical null distribution of M statistic can be estimated by repeating steps (1) to (3) and pooled together as a random sample from the null distribution of M.

(5) *Calculate p-value for a SNP set S*

Given a null distribution of M statistic and the M value based on true trait, an empirical p-value for S can be calculated to estimate its significance.

4.3.2 Simulation experiment

To simulate SNP genotype data with real allele frequencies and linkage disequilibrium (LD) structure patterns, we use the HAPGEN2 [72] simulation tool to generate case and control samples based on a reference set, for which we choose Caucasian cohort (CEU) population on human chromosome 22 from 1000 Genomes project [15]. HAPGEN2 simulates genotype data by resampling this reference set of population haplotypes and an estimate of the fine-scale recombination rate across the region, so that the simulated data has the same LD patterns as the reference data [72]. Unlike other simulation tools simulating a single "disease SNP" on the same haplotype, such as HAPSAMPLE [87], HAPGEN [72], and GWAsimulator [43], HAPGEN2 can simulate multiple SNPs associated with the disease outcome on the same chromosome, which is often the case for many complex diseases [72]. First, we map a total of 6,129 SNPs genotyped with Affymetrix array 6.0 in the chosen reference set to their neighboring genes: SNPs within 5KB upstream or downstream from a gene are assigned to that gene based on the Ensembl database (Release 67). We randomly select 50 genes with their constituent SNPs as genotyped SNPs for our simulation. These selected genes have 11 to 175 constituent SNPs. Among them, five genes are randomly selected as causal genes for the simulated disease outcome. They contain 56,168,30,12, and 99 SNPs respectively, within which three SNPs for each causal gene are randomly selected as their corresponding disease SNPs respectively. The other 45 genes are considered as null genes with no risk effect on the outcome.

HAPGEN2 models the probability $\pi_i = P(Y_i = 1|G_i)$ that subject i has dis-

ease given SNP genotype $G_i \in \{0, 1, 2\}$, for which π_i could take three values: f_0, $f_1(= f_0 \times rr_1)$, or $f_2(= f_0 \times rr_2)$ corresponding to the genotype with different number of minor alleles ($G_i = 0$, 1, or 2). In this general disease model, f_0, f_1 and f_2 are the corresponding penetrance of the disease and rr_1, rr_2 are the relative risk for heterozygous ($G_i = 1$) or homozygous (($G_i = 2$)) pairs, respectivly. Under a null hypothesis SNP G_i has no effect on disease, $rr_1 = rr_2 = 1$. To test the power of our supervised CPCA method for detecting causal genes, we studied three different settings for risk effect sizes for disease SNPs in those causal genes. In order to model more general risk effect from different SNPs, we set the homozygote risk for a disease SNP slightly bigger than its corresponding heterozygote risk to avoid any proportional relationship assumptions between its genotype and risk effect size. For example, if we assume a commonly adopted additive model, the relative homozygote risk for a disease SNP is inherently assumed to be equal to the square of its relative heterozygote risk, which may not capture the actual genotype-phenotype relationships in real data. Therefore, we set the relative heterozygote risk and homozygote risk for all disease SNPs at three different levels at $(rr_1, rr_2) = (1.2, 1.3)$, $(1.3, 1.4)$, and $(1.5, 1.6)$. In our simulation study, 500 case and control samples are generated respectively in 100 replicates for each causal gene under different risk levels. The same number of cases and controls are also randomly generated in 100 replicates for 45 null genes. In summary, we simulate 500 (5×100) causal genes and 4500 (45×100) null genes for each scenario in total.

The performance of our supervised CPCA (SCPCA) method on this set of simulated data is evaluated by comparing with the results obtained by SPCA and supervised LPCA (SLPCA) based on two criteria: statistical power and receiver operating characteristic (ROC) curves. The statistical power is computed as the proportion of detected causal genes that are significantly associated with the case-control outcome,

for which we have the ground truth as we simulate the outcome based on selected "causal" SNPs. Table 4.3 provides the statistical power at the significance level 0.05 from different methods, which shows that our method has achieved consistently higher power than the other two methods. Due to explicit modeling of categorical data, our SCPCA performs better than SPCA, which inherently assumes that the data follows a normal distribution. We note that the performance of SLPCA is slightly worse than SPCA in this set of simulation experiments because it loses information when transforming the original categorical genotypes $\{0, 1, 2\}$ into a binary representation $\{0, 1\}$ by assuming an inappropriate dominant/recessive model. To further validate the superiority of our SCPCA method, we plot the ROC curves by these three methods for all three risk effect sizes as shown in Figure 4.3. The ROC curves by SCPCA are always on top of those from SPCA and SLPCA for all scenarios, which demonstrates that its statistical power is consistently higher than the others at different significant levels. In addition, both Table 4.3 and Figure 4.3 have illustrated that our SCPCA has achieved more significant performance improvement over the other two methods when the risk effect is small. This demonstrates that SCPCA can perform better due to its explicit modeling of categorical SNP data with more general model assumptions, especially when we have difficult cases where the causal genes are more difficult to detect with smaller risk effect from their constituent disease SNPs. As we expect, based on the results from this simulation experiment, SCPCA is clearly superior to SPCA and SLPCA.

4.3.3 Analysis for Crohn's disease

We further apply our SCPCA method for a pathway-based association analysis of Crohn's Disease (CD) based on the GWAS case-control data from Wellcome Trust Case Control Consortium (WTCCC) [82]. In this CD dataset, there are 2,005 case

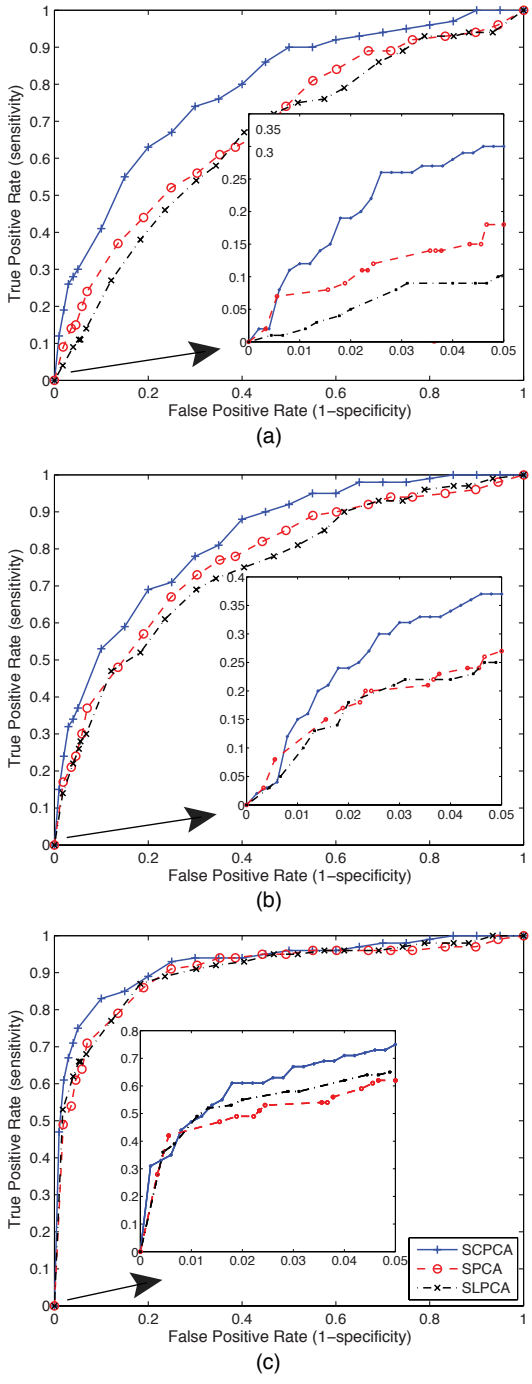

Figure 4.3: ROC curves for SCPCA, SPCA, SLPCA at risk level (relative heterozygote risk, relative homozygote risk)= (a) (1.2,1.3); (b) (1.3,1.4); and (c) (1.5,1.6) in gene-based association analysis on simulation data [50].

Table 4.3: Comparison of statistical power obtained by SCPCA, SPCA and SLPCA at significance level 0.05 for three risk levels: (relative heterozygote risk, relative homozygote risk)=(1.2,1.3); (1.3,1.4); (1.5,1.6) in gene-based association analysis on simulation data [50].

Power	**Method**		
Risk level	SCPCA	SPCA	SLPCA
(1.2,1.3)	0.30	0.24	0.14
(1.3,1.4)	0.37	0.37	0.30
(1.5,1.6)	0.75	0.71	0.68

samples and 3,004 control samples consisting of 1,504 individuals from the 1958 British Birth Cohort and 1,500 individuals from the UK blood services. After quality control, there are 1,748 cases and 2,938 controls in total with 469,557 SNPs in each sample [82].

To analyze the joint effect from multiple SNPs in functional regions that may be associated with Crohn's disease, we first map all the SNPs in the CD dataset into their corresponding pathways and thus implement SCPCA on each pathway to identify those pathways that are statistically significantly associated with the disease outcome. Specifically, we first download the pathway information from Molecular Signature Database (MSigDB: http://www.broadinstitute.org/gsea/msigdb) and collect two categories of pathways as the prior biology knowledge: C2-CP and C5-BP, corresponding to annotated canonical pathways (CP) from online pathway databases such as KEGG, BioCarta and Reactome pathway databases and GO biological processes (BP), respectively. We further filter out those pathways with more than 250 genes to increase the specificity by avoiding overly broad pathways, which has been similarly done in literature [48, 11]. The resulting 866 CP and 751 BP pathways are taken as candidate functional regions for our aggregated association analysis of Crohn's disease. With the same procedure as in [48], we map SNPs in the prepro-

cessed CD data to these pathways based on the *Homo sapiens* Variation (dbSNP 130) and *Homo sapiens* genes (GRCh37.p7) datasets in the Ensembl database (Ensembl 67) using BiomaRt (http://www.biomart.org/). SNPs are first assigned to their neighboring genes and then mapped to their corresponding pathways according to the previously described pathway information. With the WTCCC CD data and mapped SNPs in all pathways, we implement SCPCA to each pathway and calculate nominal p-values from permutation tests. To correct for the multiple-testing issue, we estimate the adjusted p-value for each pathway based on the Benjamini-Hochberg method. Significant pathways are identified at false discovery rate level 0.05.

We list 30 representative significantly associated pathways in Table 4.4. Those significant pathways are mostly involved in the following cellular functions: (1) initialization, activation and regulation of transcription factor activity; (2) lipid metabolism or lipid biosynthetic process; (3) regulation of protein kinase activity and protein transport; (4) regulation of cytokine secretion; (5) cellular catabolic process; (6) interleukin production; (7) response to inflammatory and virus; (8) epidermis and muscle development. Many of these pathways are related to the development of human immune system. Their alteration could cause potential malfunctioning of immune system that leads to CD.

To be more specific, those pathways with functions in regulation of cytokine secretion and initialization, activation and regulation of transcription factor are closely related with innate immunity and also have been claimed as statistically significant pathways associated with CD in previous SPCA and SLPCA based analysis [48, 11]. Among these pathways, their common gene NOD2 is the first identified gene associated with CD in previous analysis [62]. It plays an important role in immune response by stimulating immune activity through activating NF-κB. Another common group of causal pathways in these three methods includes gene categories related

to response to bacteria and inflammatory. The overly aggressive immune response to bacteria causes inflammatory and is more likely a factor causing CD [64]. Our results also have some other overlap with the previous reported results based on SLPCA [48] in those pathways related with lipid metabolism and interleukin secretion and production including genes: APOA1, IL18, NOD2, CARD8, PYCARD, NLRC4, NLRP12, NLRP3, PYDC1, NLRP2, TLR8 and others. These findings agree well with the recent literature of multiple GWA studies [80, 6, 77]. Substantial alternation of lipid metabolism has been shown in patients with acute CD associated with metabolic disturbances [32]. In addition, our SCPCA found a set of statistically significant pathways related with regulation of protein kinase activity. The mitogen activated protein kinases have been shown with a role in inflammatory bowel disease such as CD by acting as instigative controllers of many signaling pathways regulating the innate and adaptive immune system [7]. We also identified several pathways related with cellular catabolic process and muscle development. Abnormal cellular metabolic process could cause increased energy expenditure, which are typically shown in patients with CD and could further alter muscle mass and function with persist nutritional deficiencies [85]. However, given the fact that there still lacks a complete understanding of the etiology of CD, it is difficult to provide a conclusive evaluation, which will be studied in our future research.

4.3.4 Conclusion

We have derived CPCA for aggregated association analysis of categorical SNP data, which is further extended to SCPCA in a supervised framework. Our SCPCA captures more relevant information from SNP data based on a better data modeling and aggregates genotypic information from multiple SNPs into a combined signal that is the most associated with the trait by a heuristic selection procedure. By explicitly

Table 4.4: Top 30 representative pathways identified by SCPCA in WTCCC Crohn's Disease data set. This table lists the top 30 statistically significant pathways as well as the number of enriched genes and SNPs for each pathway. Overlapped pathways with those detected by SPCA or SLPCA are also indicated. In the table: the pathways marked as "Yes" have similar functions as the statistically significant pathways detected by SPCA or SLPCA [50].

Pathway	No. of genes	No. of SNPs	Overlap
Peptide metabolic process	10	474	Yes
Lipid biosynthetic process	97	1100	Yes
Transcription initiation	35	279	Yes
Phospholipid biosynthetic process	39	334	Yes
Glycerophospholipid biosynthetic process	30	158	Yes
Lipoprotein metabolic process	33	175	Yes
Membrane lipid biosynthetic process	49	619	Yes
Neuropeptide signaling pathway	14	95	
Steroid hormone receptor signaling pathway	20	177	
Cytokinesis	19	215	Yes
Activation of NF-κB transcription factor	18	261	Yes
Positive regulation of transcription from RNA polymerase II promoter	65	1511	
Cellular carbohydrate metabolic process	122	1610	
Positive regulation of cytokine secretion	10	117	Yes
Positive regulation of transcription factor activity	24	294	Yes
Epidermis development	70	581	
Regulation of DNA binding	47	406	Yes
Positive regulation of binding	28	318	Yes
Interleukin 1 secretion	10	116	Yes
Muscle development	92	1979	
Cellular catabolic process	209	2076	
Cellular lipid catabolic process	34	236	
Intracellular protein transport	139	1545	
Regulation of protein kinase activity	151	1393	Yes
Glycoprotein metabolic process	88	1466	
Regulation of transcription factor activity	40	375	Yes
Interleukin 8 production	11	74	Yes
Inflammatory response	124	1146	Yes
Positive regulation of cell proliferation	142	1991	Yes
Protein targeting	104	1202	
Response to virus	49	313	Yes

modeling SNP data as categorical data instead of continuous data with inherent assumptions on numerical effects related with genotypes, our SCPCA has shown higher power compared with SPCA and SLPCA in the gene-based simulation study as well as pathway-based Crohn's disease analysis. On the other hand, SCPCA will lose power if SNP data is indeed under the additive model assumption for introduced risk that affect the trait of interest. When the underlying model is dominant/recessive model or unknown, SLPCA or SCPCA is preferred as they make no assumptions on the numerical effects related with genotypes by assuming SNP data is either binary or categorical.

4.4 Supervised sparse ePCA

Both SLPCA and SCPCA illustrated in the previous sections take advantage of an ad-hoc supervised framework to address the issue that the aggregated signals, *i.e.* PCs from either LPCA or CPCA are not guaranteed to be directly related to the outcome of interest. The ad-hoc procedure performs LPCA or CPCA on a subset of SNPs determined by their individual statistical significance, which may fail for the cases when a group of SNPs have strong joint effects but weak individual effects. A flexible way to pursue a good subset of SNPs with their joint effects best related to the outcome is to perform the sparse ePCA and association analysis in an integrated framework. Our SSePCA method proposed in the previous section can be applied to achieve this goal, whose performance will be evaluated by the following simulation study on a synthetic benchmark data set.

4.4.1 Simulation study

To simulate SNP genotype data with real allele frequencies and linkage disequilibrium (LD) structure patterns, we still use the HAPGEN2 [72] simulation tool to generate case and control samples as what we did for the simulation study of

SCPCA. We still use the same 5 causal genes and the same 45 null genes in the simulation study of SCPCA. Three SNPs for each causal gene are randomly selected as their corresponding disease SNPs respectively. HAPGEN2 models the probability $\pi_i = P(Y_i = 1|G_i)$ that subject i has disease given SNP genotype $G_i \in \{0, 1, 2\}$, for which π_i could take three values: f_0, $f_1(= f_0 \times rr_1)$, or $f_2(= f_0 \times rr_2)$ corresponding to the genotype with different number of minor alleles (G_i = 0, 1, or 2). In this general disease model, f_0, f_1 and f_2 are the corresponding penetrance of the disease and rr_1, rr_2 are the relative risk for heterozygous ($G_i = 1$) or homozygous (($G_i = 2$)) pairs, respectively. Under a null hypothesis, SNP G_i has no effect on disease, thus $rr_1 = rr_2 = 1$. For simplicity, we investigate the performance of SSePCA on the SNP genotypes represented by a dominant model. We named SSePCA applied for binary data types as supervised sparse logistic PCA (SSLPCA). To test the power of our SSLPCA method for detecting causal genes, we studied two different settings for risk effect sizes for disease SNPs in those causal genes. We set the relative heterozygote risk and homozygote risk for all disease SNPs at two different levels at (rr_1, rr_2)=$(1.5, 1.6)$ and $(1.7, 1.8)$. Also, 500 case and control samples are generated respectively in 100 replicates for each causal gene under different risk levels. The same number of cases and controls are also randomly generated in 100 replicates for 45 null genes. In summary, we simulate 500 (5×100) causal genes and 4500 (45×100) null genes for each scenario in total.

The performance of our SSLPCA method on this set of simulated data is evaluated by comparing with the results obtained by SLPCA and a Bayesian method BhGLM based on receiver operating characteristic (ROC) curves. We plot the ROC curves by these three methods for both two risk effect sizes as shown in Figure 4.4. The ROC curves by SSLPCA are always on top of those from SLPCA and BhGLM for both scenarios except at a very low false positive rate, which demonstrates that its

statistical power is mostly higher than the others at different significant levels. This indicates that SSLPCA can achieve higher power in aggregate association analysis of the joint effects of SNPs in the causal genes. As we expected, based on the results from this simulation experiment, SSLPCA is superior to SLPCA due to the flexible variable selection by an integrative framework rather than an ad-hoc way. In addition, both SSLPCA and SLPCA has clearly better performance than the Bayesian method BhGLM, which suggests that the PCA based methods could extract higher joint effects by taking better care of the relations among SNPs.

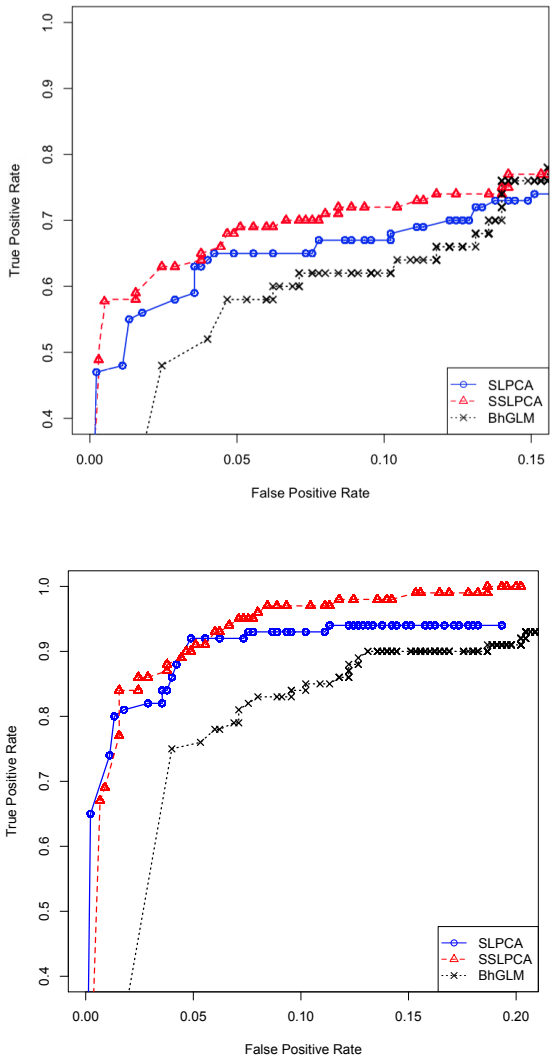

Figure 4.4: ROC curves for SSLPCA, SLPCA, and BhGLM from left to right at risk levels (relative heterozygote risk, relative homozygote risk)=(1.5,1.6); (1.7,1.8) in gene-based association analysis on simulation data.

5. RARE VARIANT ANALYSIS*

5.1 Introduction

Another interesting topic or challenge in association analysis of complex diseases is rare variant analysis, as some complex diseases have been shown to be related with rare variants and their interaction with environmental or other factors. Rare variant is the genetic variant with minor allele frequency (MAF) less than 0.05 or 0.01. It is complicated to study rare variants because they are not well represented in genome wide association arrays and there are a relatively small number of rare variants contained in SNP genotyping panels which are typically designed with a focus on common SNPs for GWAS [22]. Moreover, the characteristics of low minor allele frequency (MAF) and weak individual effects in rare variants will cause conventional statistical methods lose their power due to their small variation. Aggregate analysis is an effective way to detect rare variants with disease association by analyzing the collective effect of multiple rare variants through accumulating their individual effects. The idea behind aggregate analysis is that multiple rare variants are individually rare but accumulatively common enough to be analyzed as common variants in GWAS. Several collapsing methods have been proposed to generate an enriched signal or aggregated statistic for a set of rare variants [43, 57, 52, 61]. However, their aggregated signals or statistics are derived based on a simple summation of individual effects which ignores the correlations between rare variants.

In this section, we propose a novel framework of MAF-based logistic principal component analysis (MLPCA) [49]* to derive aggregated statistics by explicitly mod-

*Reprinted with permission from "Logistic principal component analysis for rare variants in gene-environment interaction analysis" by Meng Lu *et al*, 2014. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol:11, issue:6, page:1020-1028. Copyright ©2014 by IEEE. DOI: 10.1109/TCBB.2014.2322371.

eling the correlations between rare variant SNPs which are represented by categorical values. The derived aggregated statistics by MLPCA can then be tested as a surrogate variable in regression models to detect the gene-environment interaction from rare variants. In addition, MLPCA searches for the optimal linear combination from the best subset of rare variants that has the maximum association with the given trait. We compared the power of our MLPCA-based methods with four existing collapsing methods in gene-environment interaction association analysis using both our simulation data set and Genetic Analysis Workshop 17 (GAW17) data [1]. Our experimental results have demonstrated that MLPCA on two forms of genotype data representations achieves higher statistical power than those existing methods and can be further improved by introducing the appropriate sparsity penalty. The performance improvement by our MLPCA-based methods result from the derived aggregated statistics by explicitly modeling categorical SNP data and searching for the maximum associated subset of SNPs for collapsing, which helps better capture the combined effect from individual rare variants and the interaction with environmental factors.

5.2 Review of collapsing methods

Conventional statistical methods such as logistic or linear regression models have been successfully used to identify associated common variants in GWAS. However, these methods lose their power if the causal variants are rare with very low MAF (i.e. MAF$<$0.01 or $<$0.05) unless the effect size is large [22]. The power of single-variable tests on rare variants has been demonstrated to be very sensitive to the effect size [3]. One solution to enhance the detection rate may be to derive summary statistics for the corresponding genes to which rare variant SNPs belong. Fixed threshold collapsing (T1 or T5), weighted-sum (WS) and variable-threshold (VT)

are common collapsing methods proposed to generate such aggregated statistics for genes consisting of multiple rare variants [57, 52, 61]. These aggregated statistics can be used to evaluate the statistical significance of the association between a gene G and a given phenotype y. For example, if y is continuous, a linear regression model can be fitted:

$$y_i = \beta_0 + \beta_1 a_{iG} + \epsilon_i, \tag{5.1}$$

where y_i and a_{iG} denote the phenotype and the derived aggregated statistics for subject i respectively; and ϵ_i is a noise term. The regression coefficient β_1 reflects the effect size of aggregated statistics for rare variants on the phenotype y and β_0 is an intercept term. In the following, we briefly review four commonly used collapsing methods T1, T5, WS, and VT for rare variant association studies:

5.2.1 Fixed threshold collapsing (T1 or T5)

The simplest way to aggregate information from rare variant SNPs is to add up the number of minor alleles for all the rare variant SNPs that belong to corresponding genes with MAF smaller than a certain threshold (for example, 0.01 or 0.05). The total number of minor alleles of these rare variant SNPs in each gene is considered as the aggregated statistics to represent the genetic information in rare variants in each gene. Based on different threshold values, we have different collapsing methods. For example, T1 and T5 denote the fixed threshold collapsing method with MAF thresholds at 0.01 and 0.05, respectively. Mathematically, for a given gene G, the corresponding aggregated statistics a_{iG} for the ith subject can be computed as:

$$a_{iG} = \sum_{k \in S} x_k^i, \tag{5.2}$$

where $S = \{j \in G | x_j$ with MAF<0.01 (T1) or 0.05 (T5)$\}$ is a set containing the rare variants with MAF below the corresponding thresholds. Here, x_j represents the jth rare variant belonging to G, and x_k^i denotes the genotype of the kth rare variant for subject i.

5.2.2 Weighted summation collapsing (WS)

Fixed threshold methods may miss causal variants that are more common than the arbitrary cutoff for MAF threshold as these cutoff values might not be optimal thresholds which vary with different diseases. The weighted sum method (WS) [52] aggregates the effects of all SNPs in a given gene by summing the number of minor alleles for all the variants, weighted according to the corresponding MAF of each SNP. For a gene G, the aggregated statistics by WS for the ith subject is represented as:

$$a_{iG} = \sum_{j \in G} w_j x_j^i, w_j = 1/[p_j(1-p_j)]^{1/2}, \tag{5.3}$$

where p_j is the MAF of the jth variant, and w_j is the corresponding weight for the jth variant, which is in fact the inverse of the standard deviation if we assume each variant follows a Bernoulli distribution with the mean value p_j. The weight is larger for SNP with lower MAF, smaller otherwise.

5.2.3 Variable threshold collapsing (VT)

For both fixed threshold and WS methods, it is probable that associations shown in the derived summary statistics may be diluted by including SNPs which are not associated with the outcome. In order to remove possible noise from SNPs which do not influence the outcome and further improve the power, the variable threshold method (VT) [61] aims to search for an "optimal" MAF threshold with the highest

power for association studies instead of using arbitrary fixed thresholds. First, each possible MAF cutoff value T_t among all variants in a gene is chosen to compute the corresponding summary statistics:

$$a_{iG}^t = \sum_{k \in S} x_k^i, S = \{j \in G | x_j \text{ with MAF} < T_t\}. \tag{5.4}$$

For each possible summary statistic a_{iG}^t, z scores of regression coefficients β_1 fitted in the regression models (5.1) can be computed as the statistical significance of association with disease for corresponding genes. Among all the summary statistics computed with different cutoff values T_t, the one with the maximum absolute value of z scores is selected as the VT summary statistics. VT method determines aggregated statistics corresponding to the "optimal" MAF threshold by involving trait information and thus considers the effects of SNPs irrespective of their MAFs.

5.3 Aggregate statistics based on logistic PCA

In this section, we derive new summary statistics for rare variant association studies based on a binary genotype data representation using logistic principal component analysis (LPCA) [42, 48]. Instead of simply summing up the number of minor alleles as in the existing collapsing methods for association analysis, we derive the summary statistics by taking into account of the distribution of genotype data in a given data set. Simple summation does not necessarily lead to an optimal combination of rare variants for association analysis as it ignores the relations among SNPs. Principal component analysis (PCA) [35] provides an alternative that takes account of linkage disequilibrium among SNPs by approximating the original data using an optimal linear combination of genotypic information from SNPs, with the least loss of information. However, conventional PCA is designed for continuous data based on the Gaussian distribution assumption, which may not perform well for categorical

SNP data. For appropriate analysis of rare variant SNP data, we derive the new summary statistics by extending PCA to LPCA for binary data based on reduced rank representations [42].

For a given subset of d SNPs from n subjects, the corresponding genotype data can be represented by an $n \times d$ matrix $\mathbf{X}= \{x_j^i\}$, in which x_j^i denotes the genotype for the jth SNP of subject i. For each SNP, we may have two possible representations of x_j. One is the common representation with the state space $\{00, 10, 11\}$ corresponding to homozygous or heterozygous alleles. We note that this representation can be considered as the padded binary data and we do not impose any genetic effect model assumptions. The other possible representation is simply encoding genotype data into a binary representation with 0 representing the most prevalent homogeneous base pair (wild-type) and 1 for the other genotypes (mutant with minor alleles), which corresponds to testing for a dominant/recessive genetic effect on the outcome. With this dominant/recessive genetic effect model assumption, we also can test for corresponding genotype-phenotype associations. However, the recent studies have shown that the genetic effect model assumption may affect the detection power of associations [45, 33]. Therefore, between two possible representations, we expect that we may detect more general associations between genotype and the given trait based on the first form of SNP data representation.

As the given SNP data $\mathbf{X}$ can be expressed by binary values for both representations, the existing methods in the literature for analyzing binary data [42, 14, 70, 21] can be implemented to derive summary statistics for SNP data. To distinguish the implementations for two different data representations, we name the method with the binary representation $x_j \in \{0, 1\}$ the logistic principal component analysis (LPCA) while the method for the second padded binary representation $x_j \in \{00, 01, 11\}$ the padded logistic principal component analysis (PLPCA). For both LPCA and

PLPCA, under the assumption that the binary data follows a Bernoulli distribution, we can similarly summarize genotype data (as shown elsewhere [42]) to derive an improved aggregated statistics to enhance the detection accuracy of functional rare variants by further considering interaction effects from SNPs in a gene. Here, we propose to perform both methods on an "optimal" subset of SNPs in corresponding genes with respect to a given trait y so that we can evaluate gene association effects by analyzing the derived aggregated statistics. To identify the subset of SNPs, we adopt a similar MAF threshold selection procedure as in VT to determine a flexible MAF cutoff value for our novel MAF-based aggregated statistics.

We note that LPCA and PLPCA have the same optimization model and algorithm with PLPCA dealing with padded binary data since each SNP $x_j \in \{00, 01, 11\}$ is represented as two-dimensional binary vector instead of single binary value as in LPCA. Thus, LPCA can be extended for PLPCA in a straightforward way. For simplicity, we take the first PC score $\mathbf{a}_1$ as the aggregated statistics and estimate its significant association with outcome. When we take aggregated statistics from more than one principal component, we may be able to further improve power due to including more information in the model. We will investigate this issue more thoroughly in our future work.

With the derived aggregated statistics, we now can select the best subset of SNPs which gives the most significant association with the trait y by changing the MAF-based cutoff value T' as in VT. For a gene G and a possible MAF threshold T', the aggregated statistics for the ith subject is $a_{iG} = \mathbf{a}_1^i$ with $\mathbf{X} = \{x_j | x_j \text{ with MAF} < T', j \in G\}$.

In summary, based on the binary SNP data, our MLPCA collapsing method takes the following steps:

(1) *Generate candidate subsets for each gene*

For each unique MAF value of variants in a given gene, a corresponding subset is generated by collecting those variants with MAF less than the specific value.

(2) *Derive candidate aggregate statistics*

The first PC score obtained from LPCA implemented on each candidate subset is regarded as its corresponding aggregated statistics.

(3) *Fit simple linear regression models*

Each candidate aggregated statistics is substituted into the regression model. Each t-statistic of the estimated coefficient $\hat{\beta}_1$ can be calculated by $\hat{\beta}_1/s.e.(\hat{\beta}_1)$.

(4) *Determine the final aggregated statistics*

Return the aggregated statistics with the maximum absolute value of t-statistics.

For the implementation with the padded data representation, we can extend the previous procedure in a straightforward manner and call it MAF-based PLPCA (MPLPCA). As there are often only a small number of contributing SNPs with effects to given traits, to derive better aggregated statistics to avoid including nonfunctional SNPs that are not associated with the outcome, we can further apply a sparse LPCA by adding an L1 regularization term $P_\lambda = -\sum_{l=1}^{k}(\lambda||\mathbf{b}_l||_1)$ to the log-likelihood to search for sparse PC loading vectors in $\mathbf{B}$ that maximize the regularized log-likelihood function. A sparse loading vector with only a few number of non-zero elements implies the selection of SNPs that contribute to the derived aggregated statistics. The L1 regularization term is a sparsity inducing penalty with parameter λ controlling the sparsity. By maximizing the regularized log-likelihood function, the larger λ is, the more zero elements the loading vectors have, which means fewer SNPs contributing to aggregated statistics are selected. We implement the same MAF-based procedure for sparse LPCA and denote it as MSLPCA.

5.4 Pooled association test for gene-environment interaction

The above aggregation methods could be further extend to learn the effect of the gene-environment interaction factors. In order to do that, we test the association significance based on the following regression model:

$$y_i = \beta_0 + \beta_1 a_{iG} + \beta_2 z_i + \beta_{12} a_{iG} * z_i + \epsilon_i, \tag{5.5}$$

where z_i denotes the environmental exposure status of subject i; ϵ_i is a noise term; and β_0 is an intercept term. The regression coefficients β_1 and β_2 reflect the effect size of aggregated statistics for rare variants and environmental factor z on the outcome y while β_{12} represents the effect size of gene-environment interaction summarized by aggregated statistics. We note that in corresponding MAF-based methods (VT as well as LPCA-based methods) for interaction analysis, the association significance for all the aggregated statistics for corresponding genes are tested based on the regression coefficients β_{12} for interaction effects, instead of β_1 fitted by model (5.1) for main effect analysis. Although multiple hypothesis tests are performed to compare all the t-statistic values corresponding to different MAF-based cutoff values, there is no multiple hypothesis testing problem introduced since in the end we only consider one hypothesis test which tests the t-statistic with maximum absolute value corresponding to a specific MAF-based cutoff value under a null distribution based on permutation tests.

We test the prediction accuracy using all different collapsing methods based on this interaction model. Specifically, to measure association significance of rare variants in a given gene G, association tests are performed by fitting the model (5.5) involving interaction effects with its corresponding aggregated statistics a_{iG} as one predictor. For all the above collapsing methods, p-values for gene-environment inter-

action are estimated by permutation tests for t statistic scores of β_{12} since t statistics can no longer be approximated well by a t-distribution due to potential SNP subset selection. We have performed 1,000 permutations for the phenotype y and the null distribution of t statistics is estimated empirically by fitting the same model (5.5) based on the permuted traits. Nominal p-values for the interaction between the aggregated statistics and environmental factor are used to denote the statistical significance of its association with given traits.

5.5 Experimental results

We evaluate our MAF-based LPCA methods and compare the performance with the four existing collapsing methods (T1, T5, WS, VT) for gene-environment interaction association analysis for rare variants on a simulated rare variant SNP data set as well as GAW17 data.

Statistical power (true positive rate) and ROC (receiver operating characteristic) curves are taken as the criteria to evaluate and compare the performances of different methods. Statistical power is estimated by the proportion of successfully detected genes whose aggregated statistics and interaction with the environmental factor are tested to have statistically significant associations with a given trait. With N replicates of data, statistical power can be computed based on the known functional or "causal" genes as:

$$\sum_{G \in \mathbf{C}} |R_G|/(N * |\mathbf{C}|), R_G = \{r | p_G^r < \alpha, 1 \leq r \leq N\}, \tag{5.6}$$

where p_G^r denotes the association p-value for gene G in the r-th replicate; $\mathbf{C}$ represents a set of causal genes; and α denotes a given significance level.

A ROC curve is a plot of statistical power against false positive rate with the latter defined as the proportion of falsely detected null genes that do not affect the

given trait. False positive rate is calculated by:

$$\sum_{G \in \mathbf{C}'} |R_G|/(N * |\mathbf{C}'|), R_G = \{r | p_G^r < \alpha, 1 \leq r \leq N\}, \tag{5.7}$$

in which $\mathbf{C}'$ represents a set of all known null genes that have no effects on a given trait.

5.5.1 Simulation study

In this set of simulation experiments, we have simulated a Bernoulli random variable E with prevalence of 30% as the environmental factor. Similarly, SNP data has been randomly generated for 80 genes consisting of various number of SNPs ranging from 1 to 80 respectively for 700 samples. All SNPs were independently generated with their MAFs less than 0.05 as this paper focus on gene-environment interaction analysis for rare variants. To simulate the common genotypes in the state space {00,01,11} for a SNP in 700 samples, both alleles of each SNP have been sampled from a Bernoulli distribution independently with a randomly selected success rate less than 0.05. This representation could then be easily transformed to another genotype space {0,1,2} representing the number of minor alleles that each SNP has. A quantitative disease risk score Y also has been simulated based on the linear regression model which integrates individual or main effects from both simulated SNP data and environmental factor as well as their interactive effects. To be specific, we have randomly selected one gene as the only causal gene significantly associated with disease Y. This gene has 16 SNPs and half of them are selected as functional SNPs. The disease trait Y has been simulated for 200 replicates of 700 samples based on these eight functional SNPs and a randomly simulated environmental factor E sampled from a Bernoulli distribution with success rate 0.3 by the following linear

regression model:

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i * SNP_i + \alpha * E + \sum_{i=1}^{8} \gamma_i * SNP_i * E + \epsilon \quad (5.8)$$

with β_0 set as 0 for simplicity; ϵ denoting a white noise following a standard normal distribution; and α, β_i and γ_i representing the effect of environmental factor E, the main effect from SNPs, and the genotype-environment interaction effect of the ith causal SNP respectively, sampled from the same normal distribution $N(\log(1.1), 0.25)$ independently.

To make thorough and comprehensive comparison of our MAF-based LPCA methods with the four prevalent collapsing methods, we have tested their performances in both gene-environment interaction association analysis and main effect association analysis to identify genes significantly associated with Y with and without gene-environment interaction effect respectively. The pooled association analysis for each gene has been performed on 700 samples in all 200 replicates for both tests by fitting both models (5.5) and (5.1) using the derived aggregated statistics from different methods. The main effects and gene-environment interaction effects are detected based on their nominal p-values calculated by permutation tests as introduced in the method section.

Our MAF-based LPCA (MLPCA) method has shown consistently higher statistical power than the four existing collapsing methods at four different significance levels for both main effect and interaction effect analyses as given in Tables 5.1 and 5.2. Moreover, our method's power could be further improved either in the padded version (MPLPCA) as expected or in the sparse version (MSLPCA) by introducing an appropriate penalty parameter $\lambda = 1$. All the programs for the work in this paper are written in R and run on a Mac OS X with a 2.5 GHz Intel Core i5 processor.

The run time for each collapsing method in one replicate for both main effect and interaction effect analyses is included respectively in Tables 5.1 and 5.2. Due to the heuristic search procedure for the "optimal" MAF threshold in both VT and our MLPCA-based methods, it takes them more time for association tests compared to T1, T5, and WS. Our MLPCA-based methods are more time consuming as we need to optimize for maximum likelihood estimates of corresponding principal components. However, as shown in both tables, the run time is in the same magnitude to the run time of VT. To verify the consistent superiority of our methods, ROC curves and area under the curve (AUC) of all these methods for the main effect and interaction effect analyses are also plotted and calculated respectively as shown in Figure 5.1(c) and Figure 5.1(d), which demonstrate that our methods have achieved consistently higher power than those four collapsing methods for both analyses. We further verify this conclusion by re-running the same simulation experiments but with a smaller sample size 300. The respective ROC curves and AUCs for main effect and interaction effect analyses are plotted and calculated as given in Figure 5.1(a) and Figure 5.1(b). Based on the ROC curves for different sample sizes for both main effect and interaction effect analyses, we observe that: (1) the power of each method has been improved for a larger sample size as more accurate estimates of association significance have been obtained by running regression analyses with more samples; (2) our methods keep the superiority for different sample sizes; (3) the improvement of our methods, especially MSLPCA and MPLPCA, is larger with the increasing sample size compared to the other four collapsing methods.

5.5.2 Analysis of GAW17 data

To make further evaluation and comparison, we have also applied all the above methods to analyze gene-environment interaction effect for rare variants in GAW17

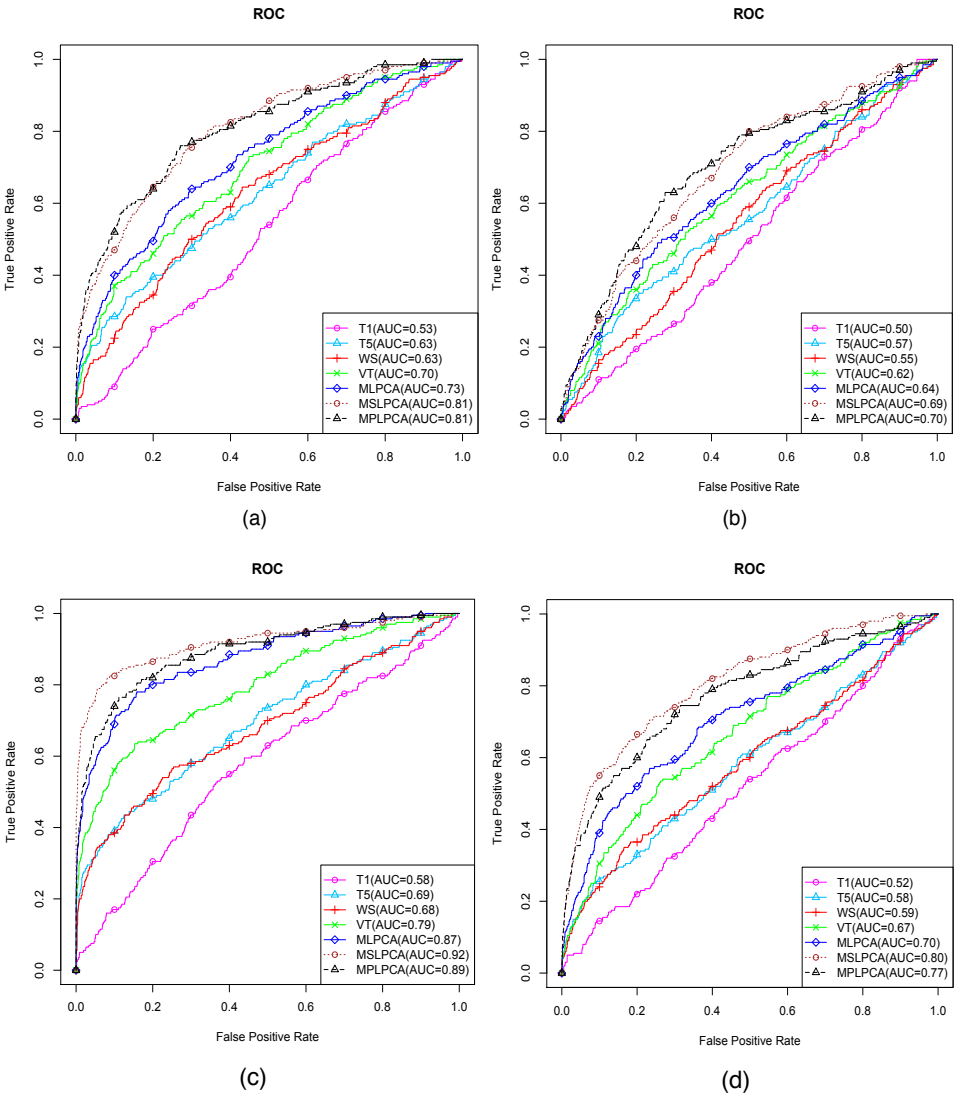

Figure 5.1: ROC curves for T1, T5, WS, VT, MLPCA, MSLPCA, MPLPCA in (a) main effect analysis with 300 samples; (b) gene-environment interaction analysis with 300 samples; (c) main effect analysis with 700 samples; and (d) gene-environment interaction analysis with 700 samples for simulation data ([49] © 2014 by IEEE).

Table 5.1: Performance comparison for T1, T5, WS, VT, MLPCA, MSLPCA, and MPLPCA at four significance levels for main effect analysis on simulation data ([49] © 2014 by IEEE).

α	T1	T5	WS	VT	MLPCA	MSLPCA	MPLPCA
0.005	0.030	0.170	0.165	0.240	0.345	0.555	0.350
0.001	0.050	0.215	0.200	0.305	0.390	0.615	0.405
0.05	0.090	0.320	0.315	0.445	0.575	0.755	0.650
0.1	0.170	0.390	0.385	0.560	0.690	0.825	0.740
time (min)	3.68	3.87	4.20	50.57	71.08	51.71	77.59

Table 5.2: Performance comparison for T1, T5, WS, VT, MLPCA, MSLPCA, and MPLPCA at four significance levels for gene-environment interaction analysis on simulation data ([49] © 2014 by IEEE).

α	T1	T5	WS	VT	MLPCA	MSLPCA	MPLPCA
0.005	0.010	0.045	0.045	0.055	0.075	0.125	0.095
0.01	0.030	0.075	0.060	0.075	0.115	0.185	0.180
0.05	0.055	0.170	0.170	0.185	0.230	0.375	0.420
0.1	0.145	0.255	0.240	0.305	0.390	0.490	0.550
time (min)	4.19	4.36	4.77	59.39	87.12	63.57	90.36

data which is a hybrid of simulated and real data based on the 1000 Genomes Project with a realistic pattern of number and frequency of SNPs including linkage disequilibrium structure. GAW17 provides three quantitive risk factors: Q1, Q2, Q4 and one binary common disease trait with 200 replicates in two samples of 697 individuals with 24,487 autosomal markers assigned to 3,205 genes [1]. Both rare variants and common variants are included in this simulation data set in a wide range of effect size with MAFs ranging from 0.07% to 25.8%. Here we choose the sample of unrelated individuals regardless of their pedigrees as we would like to guarantee the association test of various collapsing methods would not be disturbed by other factors such as

the correlation between individuals. Since genotype-smoking interaction effects on Q1 are included for variants in only one gene *KDR* on chromosome 4, we would like to design our experiments focusing on chromosome 4 with total 944 variants in 81 genes for efficiency. By performing gene-environment interaction effect test for each gene by the model (5.5) based on the derived aggregated statistics by different methods, we have detected variants with genotype-smoking interaction effects significantly associated with Q1 based on nominal p-values calculated by permutation tests as before.

In this genotype-smoking interaction association analysis, statistical power is the proportion of gene *KDR* detected interacting with smoking to influence the outcome Q1 in 200 replicates. The false positive rate is the proportion of falsely detected null genes that do not have interaction effects in 200 replicates. Based on the given statistical power values of different collapsing methods under four different significance levels in Table 5.3 and ROC curves and AUC shown in Figure 5.2, it is clear that our MLPCA method performs consistently better than all four existing collapsing methods due to the optimal linear combination of rare variants through the explicit modeling of categorical SNP data. Under the same framework, we have further investigated the power of MAF-based sparse LPCA (MSLPCA) with $\lambda = 0.1$. As shown in Table 5.3 and Figure 5.2, MSLPCA enhances the power of MLPCA when λ is large enough, since unimportant SNPs are filtered out by L1 norm penalization. In addition, the padded version of our method (MPLPCA) also improves the power by using a two-dimensional binary representation of genotype data without any genetic effect model assumption. The run time for each collapsing method is included in Table 5.3.

Table 5.3: Performance comparison for T1, T5, WS, VT, MLPCA, MSLPCA, and MPLPCA at four significance levels for genotype-smoking interaction analysis on GAW17 data ([49] © 2014 by IEEE).

α	T1	T5	WS	VT	MLPCA	MSLPCA	MPLPCA
0.005	0.045	0.070	0.085	0.090	0.100	0.070	0.095
0.01	0.060	0.085	0.100	0.135	0.135	0.105	0.165
0.05	0.180	0.170	0.175	0.25	0.270	0.295	0.300
0.1	0.285	0.260	0.230	0.315	0.340	0.390	0.410
time (min)	3.10	3.43	3.70	21.19	48.76	31.73	61.73

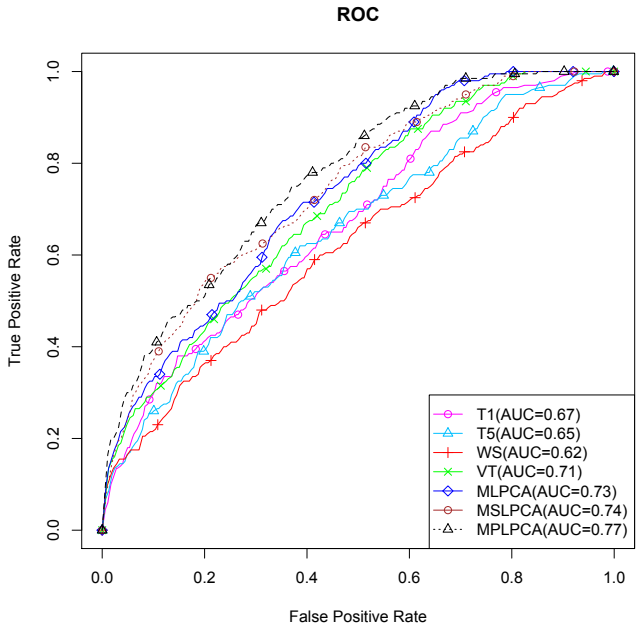

Figure 5.2: ROC curves for T1, T5, WS, VT, MLPCA, MSLPCA, MPLPCA in gene-environment interaction analysis for GAW17 data ([49] © 2014 by IEEE).

5.6 Conclusion

Different from conventional collapsing methods for rare variant analysis based on the simple summation of the total number of minor alleles for selected or weighted SNPs, we propose a set of novel MLPCA-based methods [49] to aggregate rare variants by an optimal linear combination of the best SNP subset by explicitly modeling categorical SNP data. Our new aggregated statistics captures the combined effect from individual rare variants belonging to the corresponding gene. Moreover, our method could be further improved by adding an L1-norm regularization term to perform additional unsupervised selection of causal SNPs based on the sparsity assumption. The experimental results have demonstrated our MLPCA-based methods have higher power compared to four commonly used collapsing methods and its power can be further enhanced by its sparse and padded versions.

6. EXPONENTIAL FAMILY MATCHED PCA

6.1 Introduction

In addition to the aforementioned challenges in aggregate analyses in the previous sections, it is often the case that genotypic variability may be due to potential confounding factors, including race, age, and other demographic characteristics of the population. We specifically focus on developing aggregate analysis methods for stratified data that contains specific structure information of the samples, due to potential confounding factors. This kind of data structure arises from specific experiment design which tries to adjust the effects from confounding factors in association analysis to identify critical factors that are specifically associated with the outcome of interest.

Typically, the confounding issues can be addressed by matching in the design stage of a study, in which samples are matched on one or more attributes (i.e. age, gender, smoking status, etc) in a stratum. When the sample size is small, based on the matched data, the confounding could be controlled more efficiently by balancing the distribution across strata to pursue more stable estimates of the odds ratio. Consequently, new algorithms are required for the subsequent analysis of stratified data. Ignoring matching in the analysis stage could cause estimation bias in the downstream analysis. For example, the combined signals representing a set of SNPs could be dominated by the confounding factors, which will further jeopardize the power of aggregate association analysis.

In this section, we focus on studying the dimension reduction for matched data when the disease or label information is not available or badly collected. We propose a novel dimension reduction method for matched data, namely matched PCA, as

well as its low-rank version to eliminate the bias effects from confounding factors with the aim of achieving PCs expressing the maximum variance from signals of interest. The results from simulation studies have verified the superiority of our two matched PCA models over standard PCA in reconstruction accuracy of PCs for matched data. The spike-in data experiment with microarray gene expression data also demonstrates the advantage of a sparse version of our method in detecting differentially expressed genes compared with a Bayesian method.

6.2 Matched PCA

The covariance of principle components resulting from standard PCA might be dominated by the confounding factors when matching is not taken care in the analysis. To eliminate the confounding effects, we propose two matched PCA models by involving an explicit modeling of the confounding effects to the standard PCA model, which can thus lead to more accurate estimation of principle components from the true signals of interest. Let X denote a $N \times D$ data matrix with N matched samples and U denote a $N \times G$ indicator matrix specifying the strata, to which each sample belongs to. Each strata comprises a set of matched samples. Given X and U, we developed two models to estimate the adjusted PC scores Z and the corresponding PC loading matrix W.

6.2.1 Full-rank model

We first propose a full-rank matched PCA model by explicitly modeling the confounding effects across the strata.

$$X = UV + ZW^T + \mathbf{1}\boldsymbol{\mu}^T + \boldsymbol{\epsilon}$$
$$s.t. \quad Z^T Z = I,$$

where Z is the $N \times L$ (L is the latent dimension) principal component score matrix; W is the $D \times L$ principal loading matrix; and V models the strata effects with the elements in each row reflecting the contribution of each variables to a certain stratum. $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$ are the bias and noise term respectively.

To estimate these unknown parameter matrix, we formulate a corresponding mean squared error (MSE) problem:

$$\min_{V,Z,W,\boldsymbol{\mu}} ||X - UV - ZW^T - \mathbf{1}\boldsymbol{\mu}^T||^2$$
$$s.t. \quad Z^T Z = I.$$

The optimization problem can be solved by Algorithm 4.

Algorithm 4 Full-rank matched PCA

1. Denote a random row-permuted matrix of X by X_p. Set $\boldsymbol{\mu}$ as $\frac{1}{N} X_p^T \mathbf{1}$. Compute SVD of $X_p - \mathbf{1}\boldsymbol{\mu}^T = APB^T$ and initialize Z as $A[, 1:L]$ and W as $B[, 1:L]$.
2. Update V by $(U^T U)^{-1} U^T (X - \mathbf{1}\boldsymbol{\mu}^T - ZW^T)$.
3. Update $\boldsymbol{\mu}$ by $\frac{1}{N}(X - UV - ZW^T)^T \mathbf{1}$.
4. Let $M = X - \mathbf{1}\boldsymbol{\mu}^T - UV$. Do SVD of $MW = PDQ^T$ and set Z by PQ^T.
5. Update W by $M^T Z (Z^T Z)^{-1}$.
6. Repeat 2-5 until convergence.
7. Normalize j-th PC loading vector $\boldsymbol{w}_j$ and scale $\boldsymbol{z}_j = \boldsymbol{z}_j ||\boldsymbol{w}_j||_1$. Rank $\boldsymbol{z}_j$, $\boldsymbol{w}_j$ by the decreasing order of $||\boldsymbol{w}_j||_1$.

6.2.2 Low-rank model

With a larger number of strata and involved variables, we will face the overfitting problem due to the large number of unknown parameters in the model. To address

this issue, we further improve the model by introducing a low-rank approximation of the group effect matrix.

The low-rank model is formulated as follows:

$$X = UZ_0W_0^T + ZW^T + \mathbf{1}\boldsymbol{\mu}^T + \boldsymbol{\epsilon} \tag{6.1}$$

$$s.t. \quad Z_0^T Z_0 = I, Z^T Z = I, \tag{6.2}$$

where the group effect matrix V in the full-rank model is further parameterized as a product of a $G \times K$ matrix Z_0 and a $K \times D$ matrix W_0 for a low-rank approximation.

To estimate these unknown parameters, we solve the following MSE problem:

$$\min_{Z_0,W_0} \min_{Z,W,\boldsymbol{\mu}} ||X - UZ_0W_0^T - ZW^T - \mathbf{1}\boldsymbol{\mu}^T||^2$$
$$s.t. \quad Z_0^T Z_0 = I, Z^T Z = I$$

The optimization problem can be solved by Algorithm 5.

6.2.3 Simulation study of matched Gaussian data

The simulation model used to generate the data is:

$$X = UV + ZW^T, \tag{6.3}$$

where each i-th row of V is generated from an independently identical Gaussian distribution $N(\mu_i, v_{var})$; each column j of Z is generated from an independently identical Gaussian distribution $N(0, w_{var})$; and each element in W is sampled from a uniform distribution.

We studied a data matrix X of 500 samples with the dimension set to 300. The values of the elements in X are then generated based on model (6.3). We just

Algorithm 5 Low-rank matched PCA

1. Denote a random row-permuted matrix of X by X_p. Set $\boldsymbol{\mu}$ as $\frac{1}{N}X_p^T\mathbf{1}$. Compute SVD of $X_p - \mathbf{1}\boldsymbol{\mu}^T = APB^T$ and initialize Z as $A[,1:L]$ and W as $B[,1:L]$. Do a SVD of a random matrix $X_0 = A_0PB_0^T$ and initialize Z_0 by $A_0[,1:K]$ and W_0 by $B_0[,1:K]$.
2. Update $\boldsymbol{\mu}$ by $\frac{1}{N}(X - UZ_0W_0^T - ZW^T)^T\mathbf{1}$.
3. Let $S = (U^TU)^{-1}U^T(X - \mathbf{1}\boldsymbol{\mu}^T - ZW^T)$. Do SVD of $SW_0 = P_0DQ_0^T$ and set Z_0 by $P_0Q_0^T$.
4. Update W_0 by $S^TZ_0(Z_0^TZ_0)^{-1}$.
5. Let $M = X - \mathbf{1}\boldsymbol{\mu}^T - UZ_0W_0^T$. Do SVD of $MW = PDQ^T$ and set Z by PQ^T.
6. Update W by $M^TZ(Z^TZ)^{-1}$.
7. Repeat 2-6 until convergence.
8. Normalize j-th PC loading vector $\boldsymbol{w}_j$ and scale $\boldsymbol{z}_j = \boldsymbol{z}_j||\boldsymbol{w}_j||_1$. Rank $\boldsymbol{z}_j$, $\boldsymbol{w}_j$ by the decreasing order of $||\boldsymbol{w}_j||_1$.

simulated one principal component for simplicity. μ_i is selected from a uniform distribution on the interval [0,1] and w_{var} is set as 5. We thoroughly investigated the performance of our two models in reconstruction of PCs for different numbers of strata and different effect sizes from the confounding factors by setting different values for G and v_{var}. We set four values of G: 25, 50, 100, and 125 which correspond to four different strata sizes: 20, 10, 5, and 4 respectively. These four different cases are studied under two choices of v_{var}: 0.01 and 0.05 respectively.

We investigated the performance of our full-rank matched PCA and low-rank matched PCA models in reconstruction of the PCs, by comparing with standard PCA. For simplicity, we just consider one PC for simple. The red line, black line, black dots, and blue line correspond to standard PCA, full-rank matched PCA, low-rank matched PCA, and benchmark respectively. The benchmark result is estimated from full-rank matched PCA using the true V. The performance is evaluated based on the angle between the estimated PC loading vector and the ground truth. We

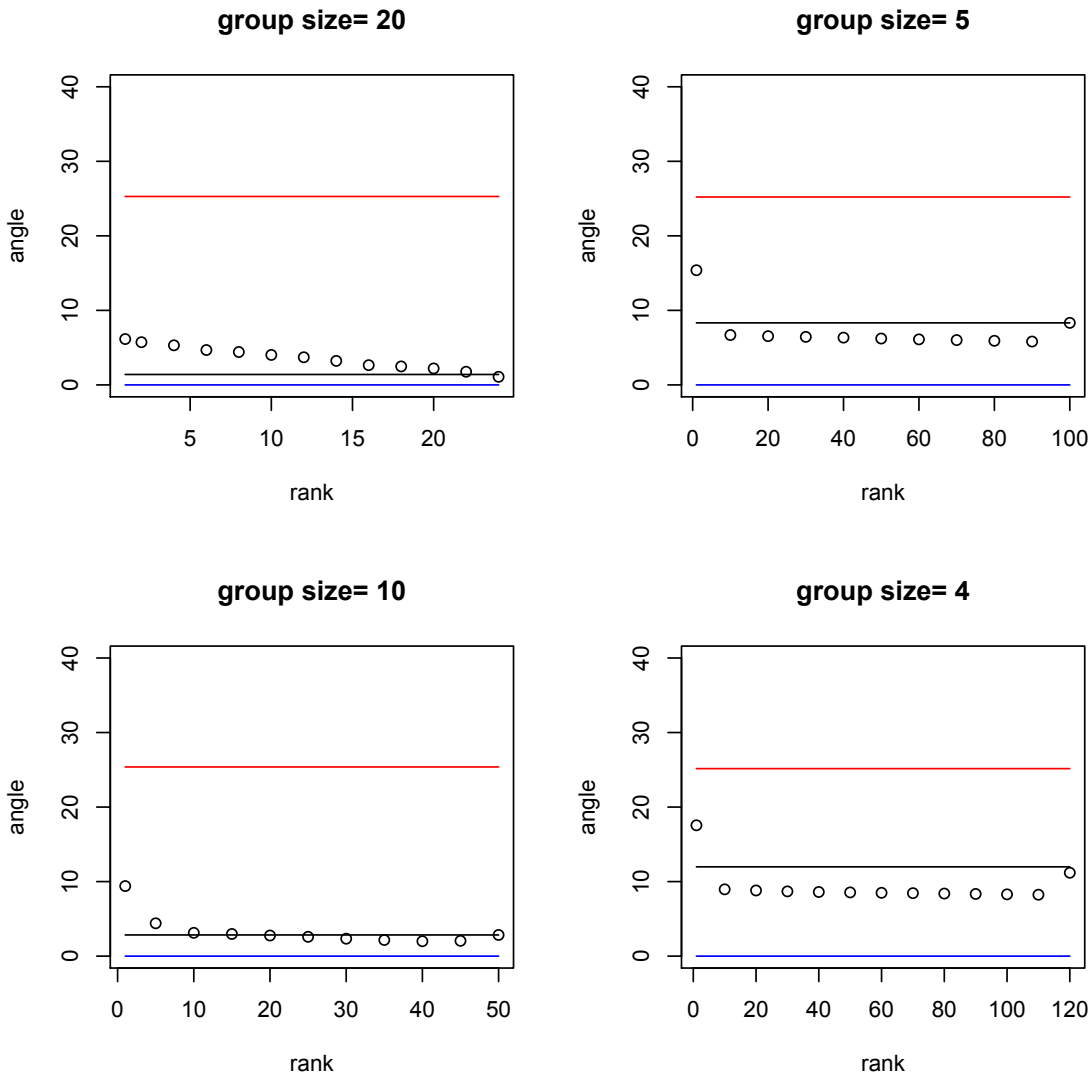

Figure 6.1: Comparison of angles between the estimated PC loading and ground truth versus different ranks under var_v= 0.01 and D=300 for matched Gaussian data. We compared standard PCA(red line), full-rank matched PCA(black line), low-rank matched PCA(black dots), and benchmark(blue line).

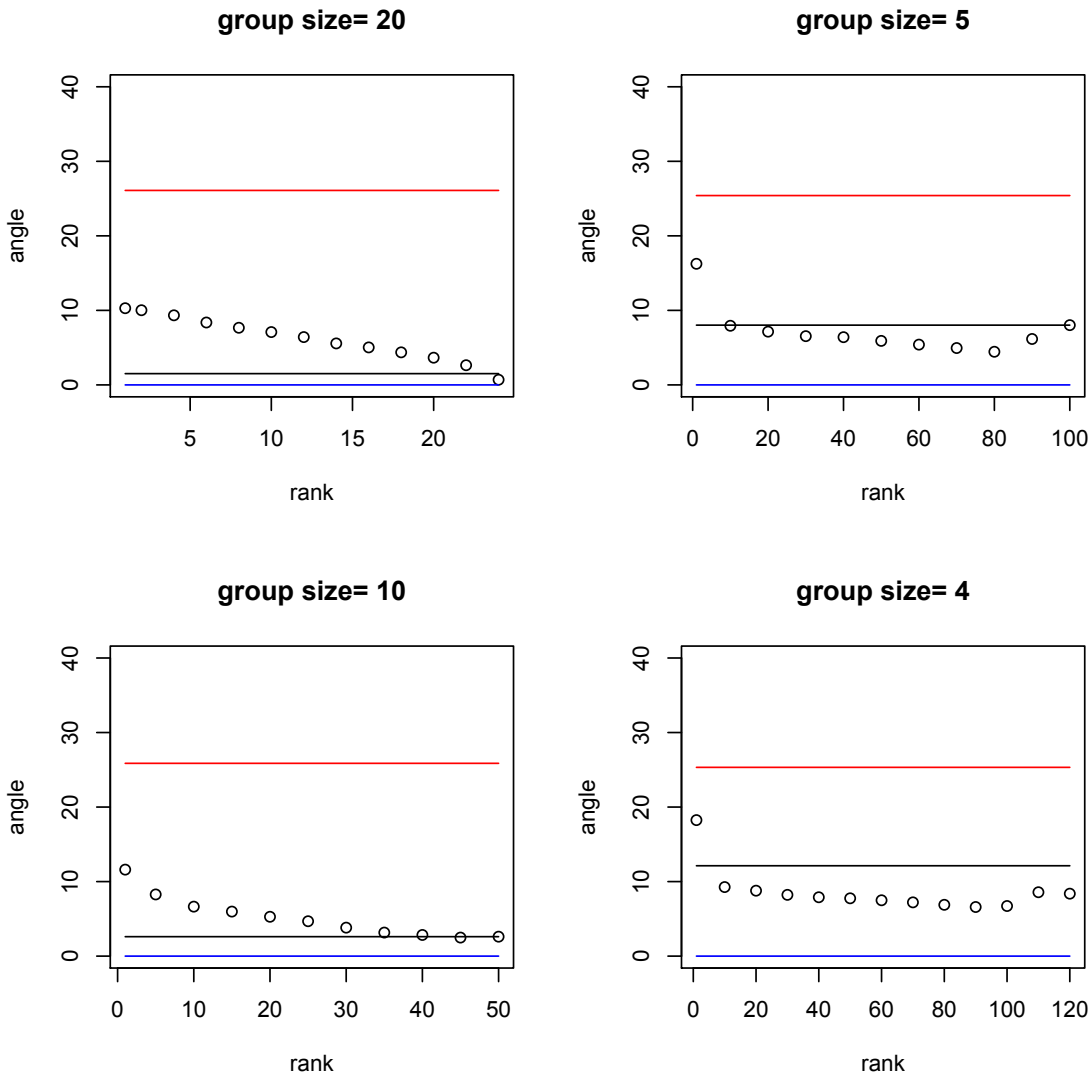

Figure 6.2: Comparison of angles between the estimated PC loading and ground truth versus different ranks under var_v= 0.05 and D=300 for matched Gaussian data. We compared standard PCA(red line), full-rank matched PCA(black line), low-rank matched PCA(black dots), and benchmark(blue line).

run the simulation study 10 times and plot the average angles from all the replicates versus rank for all the methods. Figure 6.1 shows that both of our methods achieved more accurate estimation of the true PC loading vector which is reflected by their smaller angles compared to the standard PCA. Moreover, the low-rank model can further improve the accuracy by cutting down the number of parameters to be estimated in V to avoid the overfitting problem that may jeopardize the power of our full-rank model. As shown in the cases with quite small group sizes, the full-rank model suffers from the overfitting issue that can be addressed well by our low-rank model. In addition, in the worse situations with smaller group sizes where a larger number of parameters need to be estimated in V, the improvement of accuracy in PC reconstruction from our low-rank model gets larger compared with the full-rank model. The results shown in Figure 6.2 with a larger var_v also agree well with the conclusions we made from Figure 6.1. In addition, we found that the improvement of reconstruction accuracy from our low-rank model gets more significant compared to the full-rank model when the variances of variables in V are smaller. Also, it is observed that a smaller var_v leads to a lower-rank Z_0 required for the low-rank model to achieve the same performance as the full-rank model.

6.2.4 Application to spike-in data

Microarray techniques are widely used for high-throughput measurements of gene expression levels, which consist of multiple quite complicated steps. However, there exist two critical issues in the microarray experiments: (1) microarray experiments are associated with low precision probe-level measurements, especially for weakly expressed genes; (2) each step in the experiment could induce variability to the measurements. Typically, a simple way to handle these issues is performing at least three replicate measures for each sample in order to take care of the technical noise [47]. In

microarray data analysis, one of the basic tasks is to detect differentially expressed genes. To achieve this goal, one need summarize the probe-level measurements to estimate gene expression levels for the downstream analysis. Alternatively, the summarization of expression levels can also be completed during the procedure of the subsequent analysis to meet its specific medical goals.

Our matched PCA allows to learn both the expression levels and differentially expressed genes simultaneously by involving an explicit term of the sample group structure to the standard PCA model. The differentially expressed genes can be identified via inducing a sparsity penalty on the PC loading vectors for variable selection. By introducing a sparsity penalty term to our low-rank matched PCA model (6.1), we formulated the corresponding new matched PCA problem shown as below:

$$\min_{Z,Z_0} \min_{W,W_0,\mu} ||X - \mathbf{1}\boldsymbol{\mu}^T - UZ_0W_0^T - ZW^T||_F^2 + \sum_l \lambda |W|_l$$
$$s.t. \quad Z_0^T Z_0 = I, Z^T Z = I. \tag{6.4}$$

We then apply this modified matched PCA method to perform gene expression analysis of spike-in data [17], which is a benchmark data set designed for assessing gene expression measures. In this data set, several replicates are measured for each sample to pursue robust and reproducible analysis. Our goal is to detect differentially expressed genes from this spike-in data containing technical replicates.

6.2.4.1 Description of spike-in data

In the spike-in data provided by Affymetrix (http://www.affymetrix.com/analysis /download_center2.affx), human cRNA fragments matching 14 probe-sets on the HGU95A GeneChip were added to the hybridization mixture of the arrays at con-

centrations ranging from 0 to 1024 pM [17]. All arrays use the same hybridization mixture obtained from a common tissue source. The cRNAs were spiked-in at 14 different concentration on each of 14 array groups based on a Latin square design. Each array group corresponds to a row of the Latin square, while each probe set corresponds to a column of the Latin square. The concentrations of 14 probe sets in the first row of the Latin square are: 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 pM. Each following array group has the concentration rotated by one column cyclicly. The design is described in detail by Irizarry *et al.* [34]. Table 6.1 shows the Affymetrix Latin square design.

Table 6.1: Affymetrix spike-in Latin square design. All probe set IDs end in _at, which is removed to save space. The design consists of 14 probe sets spiked-in in 14 array groups. Each row is an array group containing 14 probe sets.

	37777	684	1597	38734	39058	36311	36889	1024	36202	36085	40322	33818	1091	1708
A	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024
B	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024	0
C	0.5	1	2	4	8	16	32	64	128	256	512	1024	0	0.25
D	1	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5
E	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1
F	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2
G	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4
H	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8
I	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16
J	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32
K	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64
L	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128
M	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256
N	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512

There are three replicates in each array group except group C, for which there are only two. The two array groups denoted with M and N, for which there are 12

replicates. In total, the spike-in data contains 59 arrays of 12,626 probe sets, in which 14 of them are spiked-in. Given the array group information, we set a corresponding U matrix and apply our model (6.4) to identify spike-in genes.

6.2.4.2 Results

By applying our model (6.4) with λ set to 1 and the number of PCs to 1, we estimated the weights for all the genes stored in $W[,1]$ and found 28 genes with non-zero weights. We ranked these genes by their absolute values of the weights and have detected 18 spike-in genes with their weights significantly distinct from the others, as shown in Figure 6.3, where there seems an obvious gap between the top 18 genes and the others. These top 18 gene IDs include: 37777_at, 407_at, 1708_at, 36311_at, 1024_at, 684_at, 36889_at, 36202_at, 546_at, 39058_at, 38734_at, 1091_at, 40322_at, 36085_at, 1597_at, 33818_at, 1552_i_at, and 35127_at. The top 16 of them have also been reported as spiked-in probe sets by other researchers [17, 86] as opposed to the 14 originally described by Affymetrix. They claimed that 33818_at follows the pattern of the 12th column of Latin square design and should be included as the transcript in column 12, since no spike-in gene given by Affymetrix has the pattern consistent with the 12th column of the Latin square design. The spike-in gene 407_at which is designed in the 12th column however actually has the same concentration pattern as gene 37777_at in column 1. Another detected gene in the top 16 but not in the original 14 is 546_at, which should be considered with same concentration as 36202_at in column 9, because it is designed against the same target: Unigene ID Hs. 75209 and it shows the same pattern as 36202_at in the data [17]. Besides the top 16 genes, we also detected two extra interesting genes: 1552_i_at and 35127_at, which requires further effort to confirm their concentration patterns.

For comparison, we also applied another existing Bayesian method [47], which

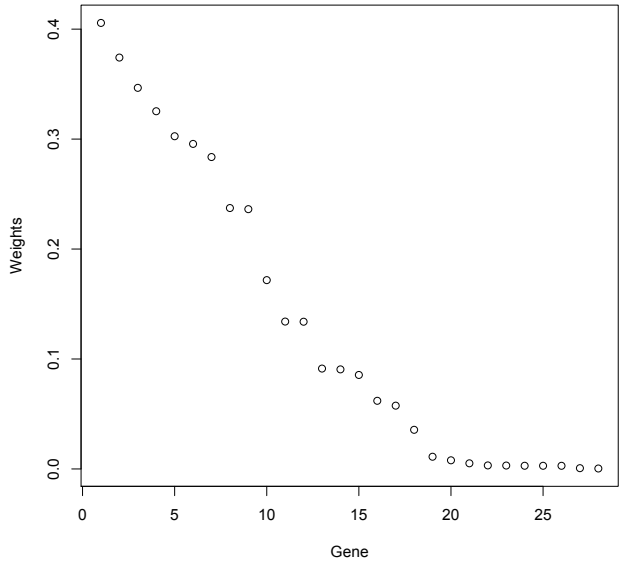

Figure 6.3: A plot of weights for the 28 ranked non-zero weighting genes in spike-in data.

takes account of the technical variance from microarrays in a probabilistic model to infer the differentially expressed genes from replicated experiments. The probability of positive log-ratio (PPLR) of expression levels between conditions are calculated in this model to score and rank all the genes. This method expects better results by involving both point estimates of expression levels and their variance estimates in the probabilistic model. Since this method is designed for case-control studies, we repeat the experiment 10 times by randomly assigning the label information to the array sets and then check their top-ranked genes. We run a Web-tool developed to implement this method, RepExplore [26], and found that among the top-ranked 16 genes from each of the 10 runs, this Bayesian method could only detect at most 15 spike-in genes of the 16 reported by both our method and previous researchers. This suggests that our low-rank matched PCA is superior in taking care of the variance

of technical measures in gene expression analysis.

6.3 Exponential family matched PCA

6.3.1 Methodology

Both low-rank and full-rank models of matched PCA can be easily extended to exponential family distributions to adjust confounding effects for matched data that follow exponential family distributions. Compared to matched PCA, exponential family matched PCA approximates the canonical parameters of the data rather than the original data by a combination of strata projections and lower dimensional projections of the signals of interest, similar as SePCA in previous sections. The optimization problem for exponential family matched PCA can be formulated for the full-rank model as:

$$\min_{Z:Z^TZ=I} \min_{V,W,\boldsymbol{b}} \sum_n A(V^T\boldsymbol{u}_n + W\boldsymbol{z}_n + \boldsymbol{\mu}) - tr\Big((UV + ZW^T + \mathbf{1}\boldsymbol{\mu}^T)X^T\Big),$$

and for the corresponding low-rank model as:

$$\min_{Z:Z^TZ=I, Z_0:Z_0^TZ_0=I} \min_{W_0,W,\boldsymbol{b}} \sum_n A(W_0Z_0^T\boldsymbol{u}_n + W\boldsymbol{z}_n + \boldsymbol{\mu}) - tr\Big((UZ_0W_0^T + ZW^T + \mathbf{1}\boldsymbol{\mu}^T)X^T\Big).$$

The developed optimization algorithms introduced in previous sections can be applied to solve these problems similarly.

6.3.2 Simulation study of matched binary data

The simulation model used to generate the data is:

$$\Theta = UV + ZW^T, \tag{6.5}$$

where each i-th row of V is generated from an independently identical Gaussian distribution $N(\mu_i, v_{var})$; each column j of Z is generated from an independently identical Gaussian distribution $N(0, w_{var})$; each element in W is sampled from a uniform distribution.

We studied a data matrix X of 500 samples with the dimension set as 300. Each element X_{ij} in the matched binary data set is sampled from a Bernoulli distribution with the success probability as Θ_{ij} generated based on model (6.5). We just simulated one principal component for simplicity. μ_i is selected from a uniform distribution on the interval [0, 5] and w_{var} is set as 100. We also investigated the performance of our two models in reconstruction of PCs for different numbers of strata and different effect sizes from the confounding factors by setting different values for G and v_{var}. We set four values of G: 25, 50, 100, and 125 which correspond to four different strata sizes: 20, 10, 5, and 4 respectively. These four different cases are studied under two choices of v_{var}: 0.1 and 1 respectively.

We investigated the performance of our full-rank matched LPCA and low-rank matched LPCA models in reconstruction of the PCs, by comparing with LPCA. The red line, black line, black dots, and blue line correspond to LPCA, full-rank matched LPCA, low-rank matched LPCA, and benchmark respectively. The benchmark result is estimated from full-rank matched LPCA using the true V. The performance is evaluated based on the angle between the estimated PC loading vector and the ground truth. We run the simulation study 10 times and plot the average angles from all the replicates versus rank for all the methods. Figure 6.4 shows that both of our methods achieved more accurate estimation of the true PC loading vector compared with standard PCA. Moreover, the low-rank model can further improve the accuracy using only rank 1 especially for quite smaller group sizes where larger number of parameters need to be estimated in V. The results shown in Figure 6.5

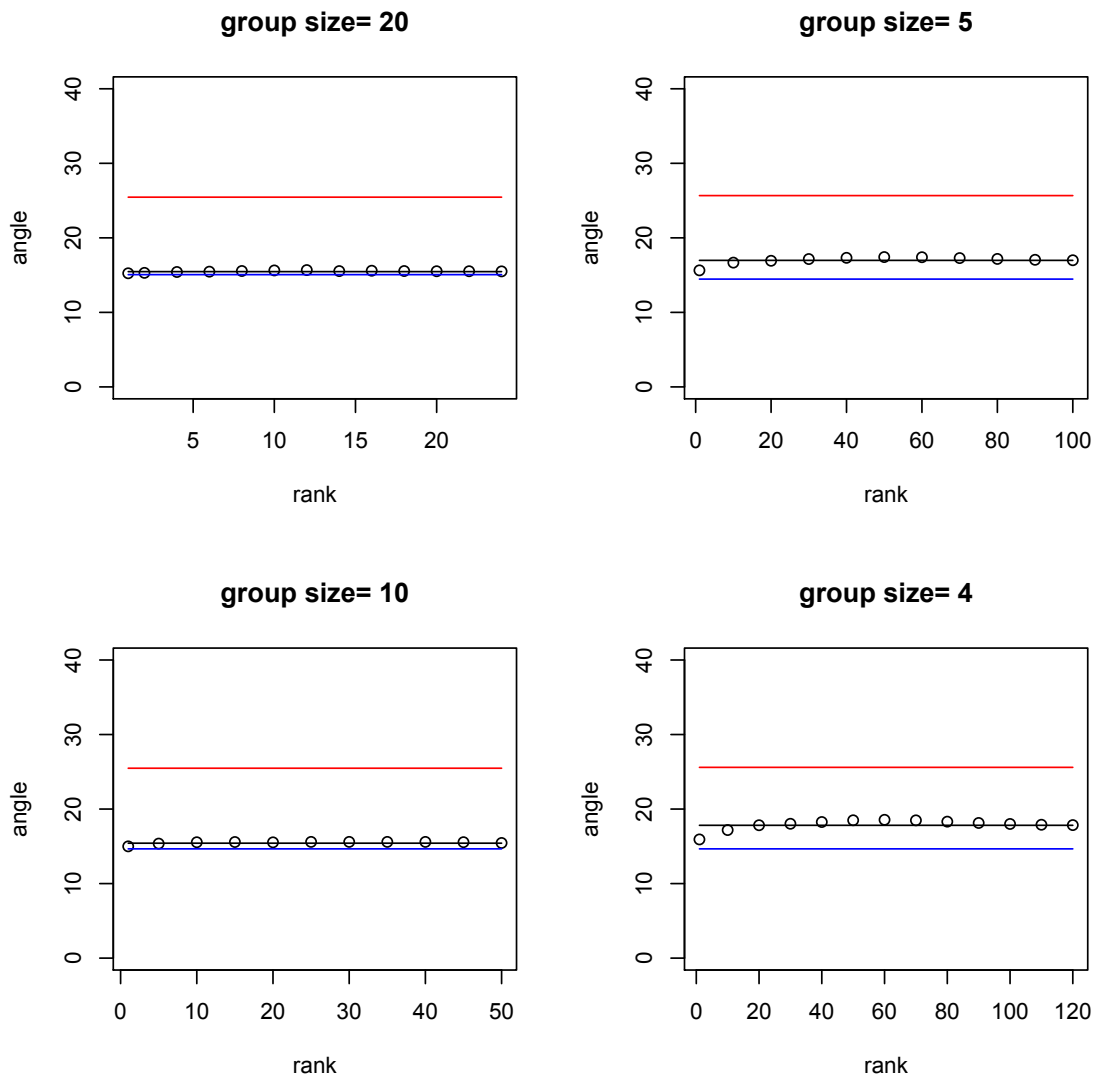

Figure 6.4: Comparison of angles between estimated PC loading and ground truth versus different ranks under var_v= 0.1 and D=300 for matched binary data. We compared LPCA(red line), full-rank matched LPCA(black line), low-rank matched LPCA(black dots), and benchmark(blue line).

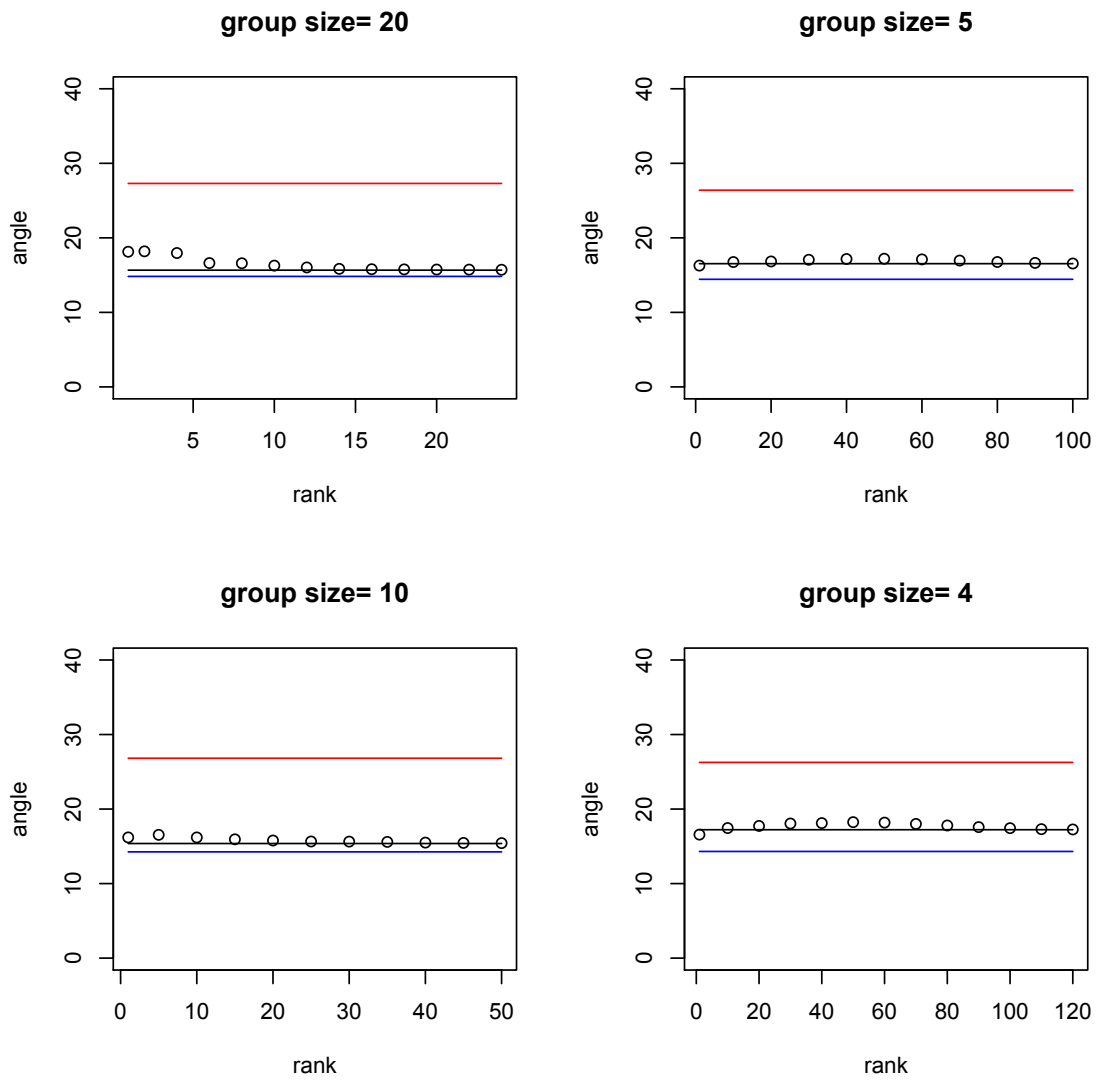

Figure 6.5: Comparison of angles between estimated PC loading and ground truth versus different ranks under var_v= 1 and D=300 for matched binary data. We compared LPCA(red line), full-rank matched LPCA(black line), low-rank matched LPCA(black dots), and benchmark(blue line).

with a larger var_v also agree well with the conclusions we made from Figure 6.4. Also, it is observed that a smaller var_v leads to a lower-rank Z_0 required for the low-rank model to achieve the same performance as the full-rank model.

Overall, both simulation studies on matched Gaussian data and matched binary data suggest that the low-rank models can improve the accuracy in reconstruction of PCs compared with the full-rank model for the cases with small group sizes and achieve more obvious improvement especially for the cases with smaller group sizes.

7. CONCLUSIONS AND FUTURE WORK

In this thesis, we derive general aggregate analysis methods for biomedical data which usually have special data types and structural information. Specifically, we derive accurate aggregated signals for SNP data based on explicit probabilistic models, which results in enhanced power of aggregate association analysis in pathway based GWAS, rare variant analysis, and matched data analysis, demonstrated by both simulation studies and real-world applications.

To take care of the special data types of SNP data in deriving aggregated signals, we developed a general dimension reduction method capable of variable selection, named as SePCA, suitable for any type of data following exponential family distributions including SNP data. To the best of our knowledge, as a dimension reduction method, it is the first general method with both the capability of interpreting PCs and the generality of applications to any type of data following exponential family distributions, which helps focus on analyzing informative variables for high dimensional non-Gaussian data such as next-generation sequencing data and genetic mutation data in genomics. SePCA has higher power in deriving more accurate PCs using pretty sparse variables for non-Gaussian data analysis compared to the sparse PCA, demonstrated by the simulation study and real applications in image clustering, and population stratification on binary data or integer data. Another contribution of our SePCA algorithm is that it is quite scalable to high dimensional data due to the efficient closed-form update rules for deriving optimal solutions. In the pathway based Crohn's disease analysis, two specific members of SePCA: SLPCA and SCPCA, could derive more accurate aggregated signals that help discover more validated immune system related pathways in explaining the underlying mechanisms of

disease development. In rare variant analysis, SLPCA can capture more effects from rare variants and enhance the power of detecting statistically significant genes in the GAW17 simulation study, compared to the other baseline methods.

We further developed supervised SePCA to involve the outcome information to SePCA and formulate an integrated framework for aggregate association analysis. Both supervised dimension reduction and variable selection are simultaneously achieved in this integrated framework. The contribution of this algorithm is that it is capable of extracting the aggregated signals that are most related to the outcome for non-Gaussian data using sparse variables. The learned sparse variables play significant roles in determining the outcome and provide a better interpretation of the results. Benefit from the supervised learning of sparse variables and aggregated signals from an integrative framework, the supervised SePCA could achieve higher power demonstrated by a simulation study for aggregate association analysis of SNP data, compared with a traditional supervised method which does variable selection in an ad-hoc way and a Bayesian method that imposes some prior distributions for the weights of the variables.

In addition to the special non-Gaussian data types, biomedical data often has special stratified data structures in some experiment designs that involve matching to solve confounding issues. Both our low-rank and full-rank exponential family matched PCA models can adjust the bias effects from the stratified data structures and reconstruct more relevant PCs for the signals of interest, as demonstrated by the simulation study in comparison with standard ePCA. The low-rank model outperforms full-rank model by ameliorating the overfitting issues arise in biomedical data with high dimensionality and small stratum size of samples. A sparse low-rank matched PCA model can further provide good interpretation of the results. It outperforms a Bayesian method in detecting the variables that significantly contribute

to the variance of signals of interest, which is verified by the application of detecting differentially expressed genes for a benchmark spike-in gene study with technical replicates. The results indicate that exponential family matched PCA methods are able to exempt the resulting PCs from the bias effects from confounding or strata factors, which are appealing tools for dimension reduction of those data with stratified structures.

In summary, our proposed statistical models for non-Gaussian biomedical data can derive more accurate and robust aggregated signals to help tackle the typical challenges from weak individual genetic effects, complex data types and stratified data structures in biomedical applications. The results have suggested that our methods can help reveal underlying biological principles of human disease. Other than bioinformatics, these probabilistic models also have rich applications in data mining, computer vision, and social science areas.

In the future work, we will explore the following two research topics: (1) investigate and pursue more favorable approaches to determine the choice of tuning parameters for regularization terms and the ratio for supervised learning trade-off; (2) perform matched case-control study by involving label information to develop supervised exponential family matched PCA models that are robust to technical noise, systematic noise and confounding factors, with the expectation to provide mechanistic insights in understanding complex diseases. In addition, it is also desirable to extend the applications of our supervised SePCA and exponential family matched PCA methods to other real-world data mining applications to further validate the performance of our proposed methods.

REFERENCES

[1] Laura Almasy et al. Genetic analysis workshop 17 mini-exome simulation. *BMC Proceedings*, 5(Suppl 9):S2, 2011.

[2] Michael Ashburner et al. Gene Ontology: Tool for the unification of biology. *Nat Genet*, 25(1):25–29, May 2000.

[3] Jennifer Asimit and Eleftheria Zeggini. Rare variant association analysis methods for complex traits. *Annual Review of Genetics*, 44:293–308, December 2010.

[4] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, March 2006.

[5] Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, 2(4):e108, April 2004.

[6] David Ballard, Clara Abraham, Judy Cho, and Hongyu Zhao. Pathway analysis comparison using Crohn's disease genome wide association studies. *BMC Medical Genomics*, 3:25, 2010.

[7] O. J. Broom, B. Widjaya, J. Troelse, J. Olsen, and O. H. Nielsen. Mitogen activated protein kinases: a role in inflammatory bowel disease? *Clin Exp Immunol*, 158(3):272–280, 2009.

[8] Michael P.S Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares, and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1):262–267, January 2000.

[9] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, December 2005.

[10] Paul Chaffee, Alan E. Hubbard, and Mark L. van der Laan. *Permutation-based pathway testing using the super learner algorithm.* University of California, Berkely, Berkely, 2009.

[11] Xi Chen, Lily Wang, Bo Hu, Mingsheng Guo, John Barnard, and Xiaofeng Zhu. Pathway-based analysis for genome-wide association studies using supervised principal components. *Genetic Epidemiology*, 34:716–724, 2010.

[12] Xinlei Chen and Deng Cai. Large scale spectral clustering with landmark-based representation. In *Twenty-Fifth Conference on Artificial Intelligence (AAAI'11)*, 2011.

[13] Andrew G. Clark. The role of haplotypes in candidate gene studies. *Genet. Epidemiol.*, 27:321–333, 2004.

[14] Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal component analysis to the exponential family. *In Advances in Neural Information Processing Systems*, 2001.

[15] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[16] The Internationa HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

[17] Leslie M. Cope, Rafael A. Irizarry, Harris A. Jaffee, Zhijin Wu, and Terence P. Speed. A benchmark for affymetrix genechip expression measures. *Bioinforamtics*, 20(3):323–331, 2004.

[18] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, 2000.

[19] Bryant G. Darnay, Jian Ni, Paul A. Moore, and Bharat B. Aggarwal. Activation of NF-κB by RANK requires tumor necrosis factor receptor-associated factor (TRAF) 6 and NF-κB-inducing kinase. *The Journal of Biological Chemistry*, 274(12):7724–7731, March 1999.

[20] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, July 2007.

[21] Jan de Leeuw. Principal component analysis of binary data by iterated singular value decomposition. *Comput Statist Data Anal*, 50:21–39, 2006.

[22] Carmen Dering, Claudia Hemmelmann, Elizabeth Pugh, and Andreas Ziegler. Statistical analysis of rare sequence variants: An overview of collapsing methods. *Genetic Epidemiology*, 35:S12–S17, 2011.

[23] Paul Flicek et al. Ensembl 2012. *Nucleic Acids Research*, 40(D1):D84–D90, 2012.

[24] James W. Gauderman, Cassandra Murcray, Frank Gilliland, and David V. Conti. Testing association between disease and multiple snps in a candidate gene. *Genet. Epidemiol.*, 31(5):383395, July 2007.

[25] Athinodoros S. Georghiades and Peter N. Belhumeur. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, June 2001.

[26] Enrico Glaab and Reinhard Schneider. Repexplore: addressing technical replicate variance in proteomics and metabolomics data analysis. *Bioinformatics*, 31(13):2235–2237, 2015.

[27] Geoff Gordon. Generalized linear models. *Proceedings of Advances in Neural Information Processing Systems*, 15, 2002.

[28] Yuhong Guo and Dale Schuurmans. Efficient global optimization for exponential family PCA and low-rank matrix factorization. *In Allerton Conf. on Commun., Control, and Computing*, 2008.

[29] MS Hayden, AP West, and S Ghosh. NF-κB and the immune response. *Oncogene*, 25(51):6758–6780, October 2006.

[30] Joel N. Hirschhorn, Kirk Lohmueller, Edward Byrne, and Kurt Hirschhorn. A comprehensive review of genetic association studies. *Genet.Med*, 4:45–61, 2002.

[31] Peter Holmans. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv Genet.*, 72:141–179, 2010.

[32] Vladimir Hrabovsky, Zdenek Zadak, Vladimir Blaha, Radomir Hyspler, Alena Ticha, and Tomas Karlik. Lipid metabolism in active crohn's disease: preresults. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub*, 150(2):363–366, NOV 2006.

[33] John P. A. Ioannidis, Peter Castaldi, and Evangelos Evangelou. A compendium of genome-wide associations for cancer: Critical synopsis and reappraisal. *Journal of the National Cancer Institute*, 102:846–858, 2010.

[34] Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003.

[35] Ian T Jolliffe. *Principal component analysis, 2nd ed.* Springer, New York, 2002.

[36] Ian T. Jolliffe, Nickolay T. Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[37] Michel Journee, Yurii Nesterov, Peter Richtarik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. 2008.

[38] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[39] Taro Kawai and Shizuo Akira. TLR signaling. *Cell Death and Differentiation*, 13:816–825, 2006.

[40] Lydia Coulter Kwee, Dawei Liu, Xihong Lin, Debashis Ghosh, and Michael P. Epstein. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397, February 2008.

[41] Seokho Lee and Jianhua Z. Huang. A coordinate descent MM algorithm for fast computation of sparse logistic PCA. *Journal of Computational Statistics & Data Analysis*, 62:26–38, June 2013.

[42] Seokho Lee, Jianhua Z. Huang, and Jianhua Hu. Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics*, 4(3):1579–1601, 2010.

[43] Chun Li and Mingyao Li. GWAsimulator: A rapid whole-genome simulation program. *Bioinformatics*, 24(1):140–142, 2008.

[44] Mingyao Li, Kai Wang, Struan F. A. Grant, Hakon Hakonarson, and Chun Li. Atom: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics*, 25(4):497–503, December 2008.

[45] Huiyi Lin, Wenquan Wang, Yung-Hsin Liu, Seng-Jaw Soong, Timothy P York, Leann Myers, and Jennifer J Hu. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer. *J Hum Genet*, 53:802–811, 2008.

[46] Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multi-dimensional genetic pathway data: Least squares kernel machines and linear mixed models. *Biometrics*, 63:1079–1088, July 2007.

[47] Xuejun Liu, Marta Milo, Neil D Lawrence, and Magnus Rattray. Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, 22(17):2107–2113, 2006.

[48] Meng Lu, Jianhua Z. Huang, and Xiaoning Qian. Supervised logistic principal component analysis for pathway based genome-wide association studies. In *ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM BCB)*, 2012.

[49] Meng Lu, Hye-Seung Lee, David Hadley, Jianhua Z. Huang, and Xiaoning Qian. Logistic principal component analysis for rare variants in gene-environment interaction analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(6), 2014.

[50] Meng Lu, Hye-Seung Lee, David Hadley, Jianhua Z. Huang, and Xiaoning Qian. Supervised categorical principal component analysis for genome-wide association analyses. In *BMC Genomics*, volume 15, 2014.

[51] Li Luo et al. Genome-wide gene and pathway analysis. *Eur. J. Hum. Genet.*, 18:1045–1053, 2010.

[52] Bo Eskerod Madsen and Sharon R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, 2009.

[53] Brendan Maher. Personal genomes: the case of the missing heritability. *Nature*, 456(7218):18–21, Nov 2008.

[54] Teri A. Manolio. Genomewide association studies and assessment of the risk of disease. *N Engl J Med.*, 363(2):166–176, 2010.

[55] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11:499–511, July 2010.

[56] Kanti V. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, New York, 1979.

[57] Andrew P. Morris and Eleftheria Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193, February 2010.

[58] Indranil Mukhopadhyay, Eleanor Feingold, Daniel E Weeks, and Anbupalam Thalamuthu. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol.*, 34(3):213–221, November 2010.

[59] Markus F. Neurath. IL-23: A master regulator in Crohn disease. *Nature Medicine*, 13(1):26–28, January 2007.

[60] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Pholosophical Magazine and Journal of Science, Sixth Series*, 2:559–572, 1901.

[61] Alkes L. Price et al. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*, 86:832–838, 2010.

[62] Graham Radford-Smith and Nirmala Pandeya. Associations between NOD2/CARD15 genotype and phenotype in Crohn's disease–Are we there yet? *World Journal of Gastroenterology*, 12(44):7097–7103, November 2006.

[63] Jurgen Ruland et al. Bcl10 is a positive regulator of antigen receptor-induced activation of NF-κB and neural tube closure. *Cell*, 104(1):33–42, January 2001.

[64] Balfour Startor Ryan. Bacteria in crohn's disease: mechanisms of inflammation and therapeutic implications. *J Clin Gastroenterol*, 41(6):637, 2007.

[65] Yvan Saeys, Inaki Iñza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[66] Eric E. Schadt, Stephen H. Friend, and David A. Shaywitz. A network view of disease and compound screening. *Nat Rev Drug Discov*, 8(4):286–295, April 2009.

[67] Daniel J. Schaid. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, 70:425–434, 2002.

[68] Daniel J. Schaid. Genetic epidemiology and haplotypes. *Genet. Epidemiol.*, 37:317–320, 2004.

[69] Daniel J. Schaid et al. Nonparametric tests of association of multiple genes with human disease. *The American Journal of Human Genetics*, 76:780–793, December 2005.

[70] Andrew I. Schein, Lawrence K. Saul, and Lyle H. Ungar. A generalized linear model for principal component analysis of binary data. *In Proceedings of the 9 th International Workshop on Artificial Intelligence and Statistics*, page 546431, 2003.

[71] Haipeng Shen and Jianhua Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 101:1015–1034, 2008.

[72] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(26):2304–2305, June 2011.

[73] Aravind Subramanian et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acac Sci*, 102(43):15545–15550, October 2005.

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

[75] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999.

[76] John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225, September 2005.

[77] Ali Torkamani, Eric J. Topol, and Nicholas J. Schork. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265–272, November 2008.

[78] Roland N. Wagner, Martina Proell, Thomas A. Kufer, and Robert Schwarzenbacher. Evaluation of nod-like receptor (NLR) effector domain interactions.

PLoS ONE, 4(4):e4931, April 2009.

[79] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[80] Kai Wang, Mingyao Li, and Maja Bucan. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum.Genet.*, 81:1278–1283, December 2007.

[81] Tao Wang and Robert C. Elston. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Bioinformatics*, 80(2):353–360, February 2007.

[82] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[83] Lingjie Weng et al. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, 12:99, 2011.

[84] Jennifer Wessel and Nicholas J. Schork. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet.*, 79(5):792806, November 2006.

[85] Jean Baptiste Wiroth et al. Muscle performance in patients with crohn's disease in clinical remission. *Inflamm Bowel Dis*, 2005(11):296–303, 2005.

[86] R. Wolfinger and M. T. Chu. Who are those strangers in the latin square? critical assessment of microarray data analysis. *CAMDA02*, 2002.

[87] Fred A. Wright et al. Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, 23(19):2581–2588, September 2007.

[88] Michael C. Wu et al. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(2):929–942, June 2010.

[89] Nengjun Yi, Nianjun Liu, Degui Zhi, and Jun Li. Hierarchical generalized linear models for multiple groups of rare and common variants: Jointly estimating group and individual-variant effects. *PLOS Genetics*, 7(12):e1002382, 2011.

[90] Kai Yu et al. Pathway analysis by adaptive combination of p-values. *Genet Epidemiol*, 33(8):700–709, December 2009.

[91] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.