STATISTICAL LEARNING OF PROTEOMICS DATA AND GLOBAL TESTING FOR DATA WITH CORRELATIONS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Donglai Chen

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Jun Xie, Chair

Department of Statistics

Dr. Hyonho Chun

Department of Mathematics and Statistics, Boston University

Dr. Lingsong Zhang

Department of Statistics

Dr. Anindya Bhadra

Department of Statistics

Approved by:

Dr. Jun Xie

Graduate Chair, Department of Statistics

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to Professor Jun Xie. She spent much time and efforts on deciding research directions, and guided me on literature review, proving theorems, thesis writing. She also gave advice on taking courses, writing emails, presentation and job searching. I am thankful for my committee members, Lingsong Zhang, Hyonho Chun, Anindya Bhadra, for helpful advice and comments.

I am supported by and NIH Grant R21 GM101504 and NSF Grant Award IOS-1127027, PRF fellowship. Part of my work is collaborative work with Daniel Szymanski's group. I appreciate their explanation of biological background. Their biological research problems make room for new data mining methods. Yaowu Liu is an expert on hypothesis testing. He provided insights on powers of global testing. I am grateful to his valuable comments and suggestions. I am indebted to Douglas Crabill for helps on large scale computing.

I appreciate my supervisors during consulting, research assistantship and internship. I would like to thank Bruce Craig, Ce-Ce Furtner and Arman Sabbaghi at statistical consulting service. Dr. Craig is helpful on design of experiments and random effect models. Dr. Sabbaghi is helpful on propensity score matching. I thank Upatising Benjavan, Kenneth Musselman at Regenstrief. I learned analysis of patients' health record data, writing reports from them. I am grateful to Robert Kill and Erica Romohr at Oriental Trading. I learned analysis of customers' order data and presenting to nontechnical people from them.

I am grateful to Hong Qu and Beihai Jiang for introducing me to bioinformatics research and the analysis of survival data. I remember the time when I studied, played and had dinners with my friends and colleagues at Purdue including Yaowu Liu, Min Ren, Zhou Shen, Xinlin Tao, Feng Wu, Botao Hao, Yuying Song, Hui Sun,

Yunfan Li. I miss my friends I met in Shenzhen middle school including Chenrui Fan, Hongfei Li and Ruinan Du. They are working in different cities, but we keep contact.

Last but not the least, I love my parents who always stand by me and support me to pursue a PhD at Purdue.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABBREVIATIONS

SEC	Size Exclusion Chromatography
IEX	Ion Exchange Chromatography
BIC	Bayesian Information Criterion
XIC	Extracted Ion Chromatogram
RSS	Residual Sum of Squares
TAIR	The Arabidopsis Information Resource
SNP	Single Nucleotide Polymorphism
GWAS	Genome-wide Association Study
PCA	Principal component analysis
SVD	Singular Value Decomposition
SVA	Surrogate variable analysis
LEAPP	Latent effect adjustment after primary projection
LC-MS/MS	Liquid chromatography tandem mass spectrometry
EIGENSTRAT	Literally it means eigenvector stratification. The method that uses PCA to detect and correct for population stratification.
GMinP	MinP under correlation
HC	Higher criticism
HC-corr	Higher criticism under correlation
GHC	Generalized higher criticism
MinP	Minimum p-value test
SKAT	Sequence kernel association test

ABSTRACT

Chen, Donglai Ph.D., Purdue University, May 2019. Statistical Learning of Proteomics Data and Global Testing for Data with Correlations. Major Professor: Jun Xie.

This dissertation consists of two parts. The first part is a collaborative project with Dr. Szymanski's group in Agronomy at Purdue, to predict protein complex assemblies and interactions. Proteins in the leaf cytosol of Arabidopsis were fractionated using Size Exclusion Chromatography (SEC) and mixed-bed Ion Exchange Chromatography (IEX). Protein mass spectrometry data were obtained for the two platforms of separation and two replicates of each. We combine the four data sets and conduct a series of statistical learning, including 1) data filtering, 2) a two-round hierarchical clustering to integrate multiple data types, 3) validation of clustering based on known protein complexes, 4) mining dendrogram trees for prediction of protein complexes. Our method is developed for integrative analysis of different data types and it eliminates the difficulty of choosing an appropriate cluster number in clustering analysis. It provides a statistical learning tool to globally analyze the oligomerization state of a system of protein complexes.

The second part examines global hypothesis testing under sparse alternatives and arbitrarily strong dependence. Global tests are used to aggregate information and reduce the burden of multiple testing. A common situation in modern data analysis is that variables with nonzero effects are sparse. The minimum p-value and higher criticism tests are particularly effective and more powerful than the F test under sparse alternatives. This is the common setting in genome-wide association study (GWAS) data. However, arbitrarily strong dependence among variables poses a great challenge towards the p-value calculation of these optimal tests. We develop a latent variable adjusted method to correct minimum p-value test. After adjustment, test

statistics become weakly dependent and the corresponding null distributions are valid. We show that if the latent variable is not related to the response variable, power can be improved. Simulation studies show that our method is more powerful than other methods in highly sparse signal and correlated marginal tests setting. We also show its application in a real dataset.

1. STATISTICAL LEARNING OF PROTEOMICS DATA

1.1 Introduction

The data analysis presented in this chapter is based on a collaborative project with Dr. Szymanski's group in Agronomy at Purdue to predict protein complex assemblies and interactions. A protein complex is a group of peptide chains. Proteins seldom act alone. Information from the prediction of protein complexes will give implication about protein functions. In biology, traditional methods study protein complexes one at a time, which is labor intensive and slow. High throughput techniques, such as the yeast two-hybrid system (Fields & Sternglanz, 1994; Jansen et al., 2003) and tandem affinity purification (Rigaut et al., 1999) have been developed to examine hundreds of protein complexes simultaneously. The high throughput techniques produce large data sets and demand advanced data analysis methods.

In this collaborative project, we develop new protein complex prediction methods based on gel-free protein separation and quantitative mass spectrometry. More specifically, size exclusion chromatography (SEC; Mori & Barth, 2013) and ion exchange chromatography (IEX; Jungbauer & Hahn, 2009) are used to characterize thousands of proteins and their complexes in native states. SEC separates molecules by their sizes, and IEX separates molecules by charges. The two technologies increase the number of detected proteins in complex samples, resulting in a large-scale protein data set with composition information of complexes. On the other hand, analysis of this high throughput data is challenging, requiring meaningful data representation and integration of multiple data types, i.e., SEC and IEX data. We have developed a series of methods for statistical learning of these proteomic data, including an integrative analysis of different data types and a data mining approach that eliminates the

difficulty of choosing cluster number in clustering analysis. It provides a statistical learning tool to analyze the composition of protein complexes globally.

The focus of this chapter is on data analysis and statistical learning for the specific proteomic data. Section 1.2 describes the experiment workflow and data representation. In Section 1.3 we explain data cleaning and preprocessing and Gaussian curves and their uses in representing biologically reproducible data. In Section 1.4 we describe data mining procedures through cluster analysis, propose two-round clustering to combine clustering results of two different data types, and develop a procedure that deals with multiple peak proteins and assign them to multiple clusters. Section 1.5 describes cluster validation methods to decide the number of clusters and evaluate cluster results. We develop a statistical learning approach for mining dendrogram trees that relaxes the need of deciding the number of clusters in Section 1.6.

1.2 Experiment workflow and data format

We use Arabidopsis, a model plant, to develop a data analysis system for automatic prediction of protein complexes. The data analysis methods can be applied to rice, cotton, soy, and other plants, and have been applied to predict cytosol or membrane-associated protein complexes as well as chloroplast protein complexes. Here we use the analysis of cytosolic proteins as an example. Cytosol is the fluid in a cell. It plays essential roles in metabolism, signaling, protein translation, and recycling. Proteins in cytosol seldom function alone. Understanding protein complexes in the cytosol can help us understand their functions.

Figure 1.1 shows the experiment workflow (McBride et al., 2018). Arabidopsis plants are grown, and their intact leaves are collected. Cytosols are extracted by grinding leaves and centrifugation. Soluble protein samples are generated. Samples are analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS; Ferrer & Thurman, 2003), which identifies and quantifies proteins. The LC-MS/MS

technique is based on the principle that proteins forming a complex should have the same molecular property, e.g., with similar molecular sizes or charges.

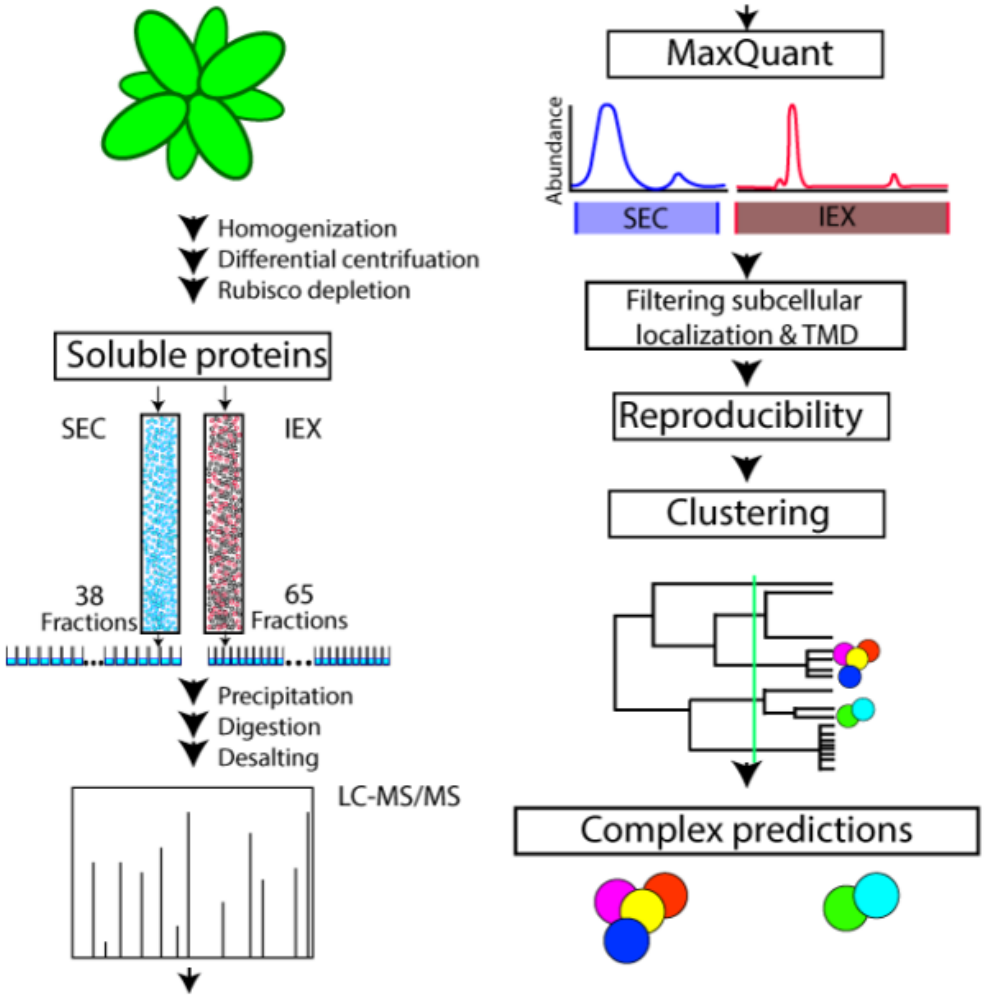


Figure 1.1. Experiment workflow. Proteins are separated by sizes and charges. Then MS/MS identifies peptides and creates protein profiles. Only reproducible protein profiles are used for cluster analysis. Cluster analysis is conducted to make complex predictions (McBride et al., 2018).

LC-MS/MS is the combination of liquid chromatography (LC) and tandem mass spectrometry (MS/MS). It separates soluble proteins, identifies peptides of these proteins and measures peptides intensities (Ferrer & Thurman, 2003). In the LC step, size exclusion chromatography (SEC) and ion exchange chromatography (IEX) are

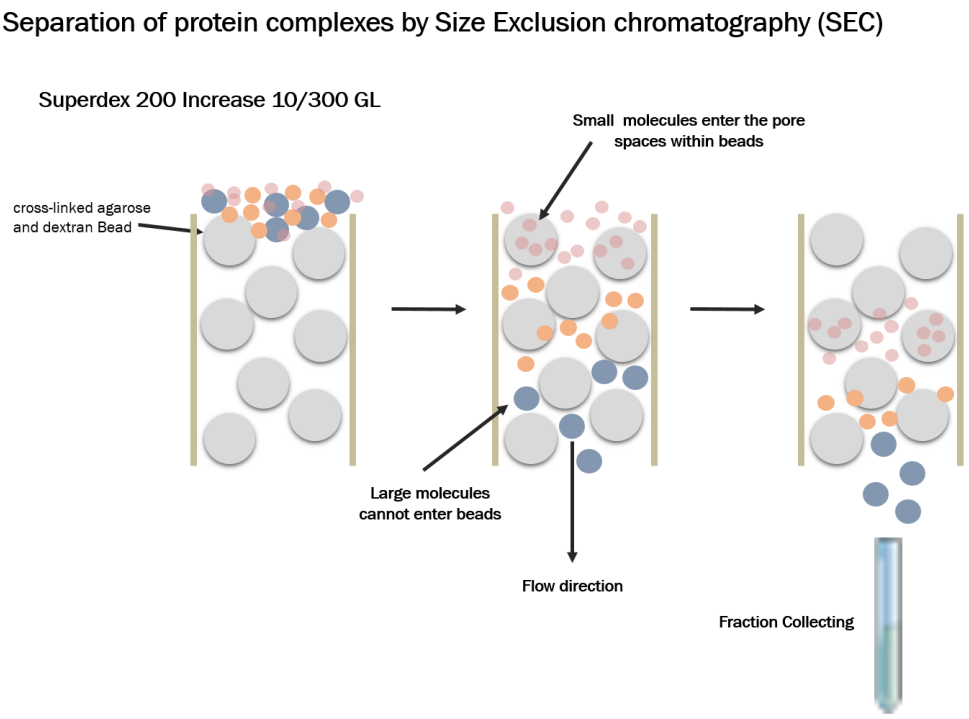


Figure 1.2. Schematic of SEC (Figure reprint with permission from Y. Lee, 2016)

two methods used to separate soluble proteins. Both SEC and IEX are methods of chromatography (Coskun, 2016). Chromatography is a type of technique to separate a mixture that is soluble in liquid or gas, which is called a mobile phase. In the process, a sample passes through a tube that holds solid material. The solid material causes a difference of speed of the mobile phase that moves in the structure, which in turn causes the mobile phase to separate. The action of one component exiting the structure and being collected is called elution. Co-elution means two or more proteins elute together. The time one component passing through the structure is called retention time. During the chromatography process, separated samples, called fractions, elute in time order. We assume protein complexes remain stable and proteins in the same complex coelute. Then results of chromatography provide useful information to predict protein complex. The data that we observe is presented a table for the number of molecules at different fractions.

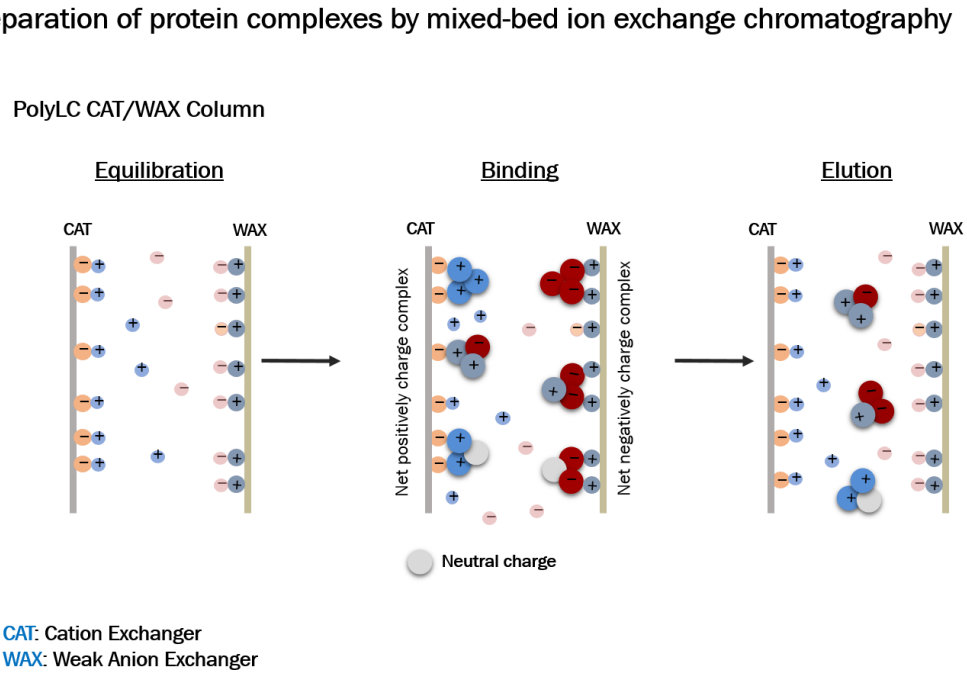


Figure 1.3. Schematic of IEX (Figure reprint with permission from Y. Lee, 2016)

Figure 1.2 (Y. Lee, 2016) shows the SEC process. SEC separates molecules by their sizes. Porous beads are spherical polymers with sponge-like structure. Porous polymer beads are filled in a long and hollow tube. Beads have different sizes of pores, and they can trap different sizes of molecules. Molecules of different sizes have different speeds to travel through porous materials. Larger molecules travel faster, and small molecules travel slower. Proteins of different sizes are thus separated by different speeds.

Figure 1.3 (Y. Lee, 2016) shows the IEX process. IEX separates charged molecules. In the beginning, proteins are injected into a tube. The tube is set to have specific pH so that proteins have charges. Proteins with positive charges bind with cation exchanger (ions with negative charges), and proteins with negative charges bind with anion exchanger (ions with positive charges). By adding salt, pH changes and ions in the salt bind with exchangers and in turn proteins that bind with exchangers elute, and weakly charged proteins elute faster. Again, separated protein samples are

collected. IEX has higher resolution than SEC in terms of the number of fractions for molecule separation.

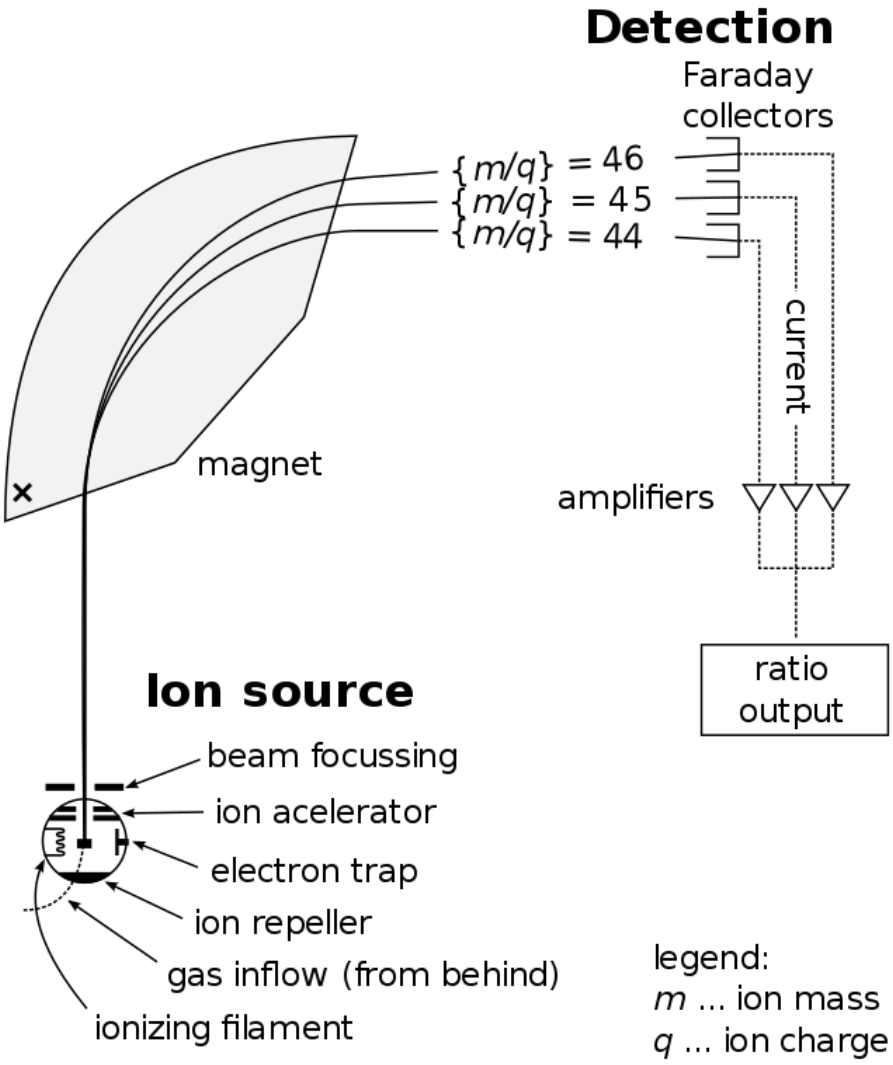


Figure 1.4. Schematic of mass spectrometry (Figure reprint from Wikipedia; Revez et al., 2001)

Separated proteins are digested into peptides using an enzyme called trypsin because mass spectrometry (MS) is better at identifying peptides than proteins. Separated samples, or fractions, are then analyzed by MS/MS (Gross, 2017). Figure 1.4 shows the process of mass spectrometry (Revez et al., 2001). Samples enter the MS

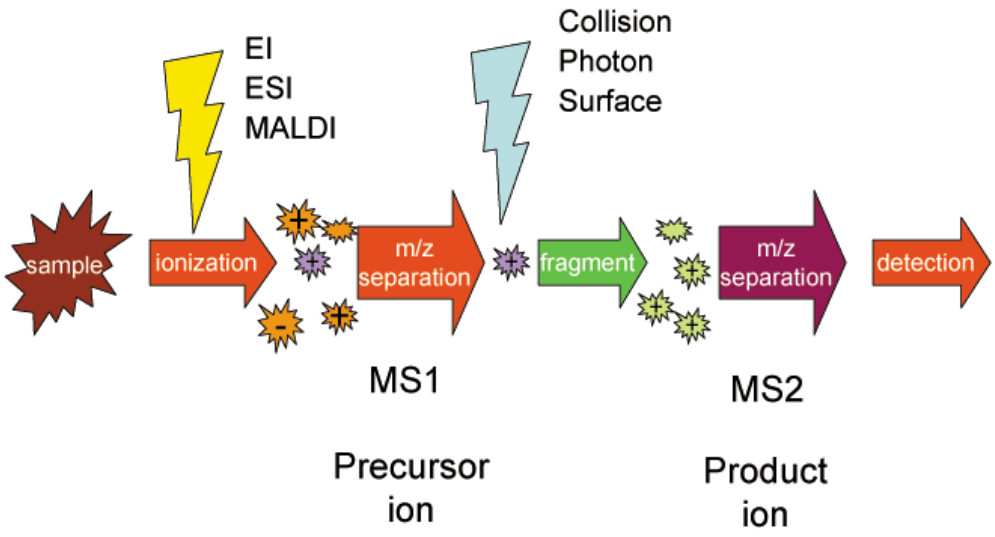


Figure 1.5. Schematic of MS/MS (Figure reprint from Wikipedia; Murray, 2006)

machine in the gas phase and are charged into ions, which is called ionization. MS manipulates the motion of ions by changing electric or magnetic fields to measure their mass to charge ratio, denoted as m/z, where m refers to the atomic mass unit and z refers to the number of charges per ion. Particles are accelerated in electric fields and deflected (change direction) in magnetic fields. Particles with less m/z have a larger change of direction due to magnetic fields. Ions are amplified. That means when ions are received, additional electrons are released to increase signal. Intensities of amplified ions are measured. The intensity is proportional to the number of ions detected. The intensities may not represent the actual number of ions due to variability in counting of ions, amplification, ionization, so intensities are usually labeled as arbitrary units. Two stages of MS work together to effectively identify peptide sequences and their intensities. Figure 1.5 shows the MS/MS procedure (Murray, 2006). At the first stage, peptides that enter the MS/MS machine are called precursor ions and their intensities are measured. At the second stage, peptides collide with molecules in the gas and break into pieces called product ions. This process is called fragmentation. Their m/z's are measured, and peptide sequences are identified.

There are two types of quantification of peptides, namely spectral counts and peak intensities. Figure 1.6 (Y. Lee, 2016) shows MS scans and intensities at each scan, where the term scan is referred to detection of ions. The number of scans per second is called frequency. The upper left of Figure 1.6 is the mass spectrum at one time. The upper right of Figure 1.6 shows MS scans at a certain frequency. Spectral count means the number of times that the specific peptides that belong to certain proteins are detected, which refers to the number of red vertical lines in the upper right panel of Figure 1.6. Extracted ion chromatograms (XIC; Koulman et al., 2009) refers to the lower part of Figure 1.6, which is the plot of time and intensity. The total intensities of peptides are the areas under the curves of intensities over time for peptides with specific m/z's. In this project, XIC is used because it has a larger dynamic range than spectral counts. That means it can better distinguish ions of low abundance and high abundance.

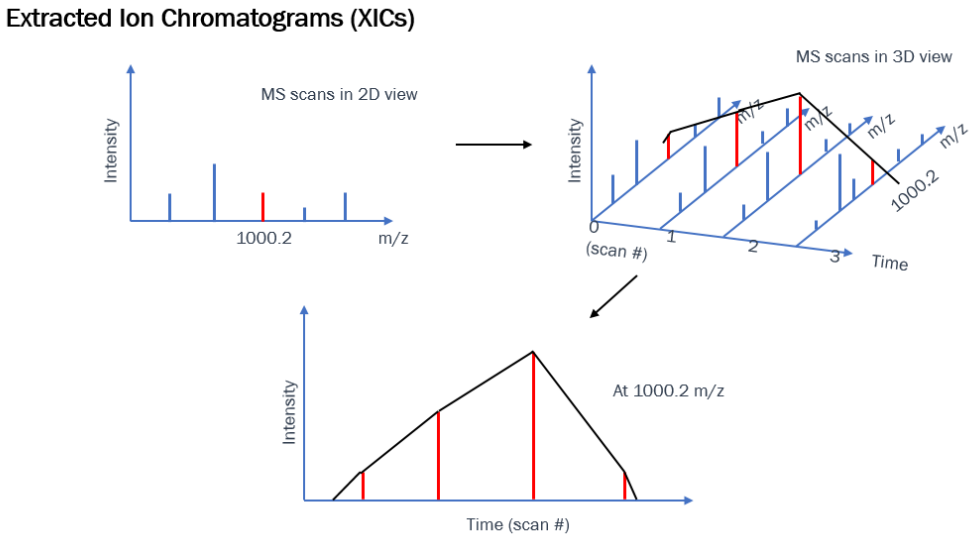


Figure 1.6. XIC (Figure reprint with permission from Y. Lee, 2016)

Data generated from MS is analyzed by MaxQuant software (Cox & Mann, 2008, 2011; Cox et al., 2014). It identifies peptides and matches peptides to proteins by searching amino acid sequences in the Arabidopsis Information Resource (Lamesch et

al., 2011) protein sequence database. False discovery rate (FDR), the percentage of wrong identification of proteins, is set to be 1%. Peptide profiles are the intensities of peptides at each fraction. Protein profiles are defined as the sum of their component peptide profiles.

After LC-MS/MS and the MaxQuant process, we obtain a data table of all proteins of the organism under study, where the rows represent proteins, the columns represent fractions as explained next, and the entry is the intensity of the corresponding protein at the corresponding fraction. Each column in the SEC data represents one apparent mass in kDa (kilodalton is a unit of mass. One dalton is 1/12 of the mass of carbon-12). Apparent masses are in decreasing order. Each column in the IEX data corresponds to a charge amount that represents elution time of proteins. It does not have a specific unit. There are 38 fractions in the SEC data and 65 fractions in the IEX data. The number of fractions is decided by the resolution of each separation technique. IEX has higher resolution than SEC and therefore has more fractions. Values in the data table are the proteins' relative intensity or abundance. After the protein database search through MaxQuant, 898 cytosolic proteins are identified by SEC and 1771 cytosolic proteins are identified by IEX. Two biological replicates of SEC and IEX are collected in our study.

1.3 Data processing through Gaussian curve fitting

Figure 1.7 shows the SEC profiles of a small group of proteins. The three dimensions represent protein IDs, fractions, and standardized abundance, respectively. For any protein, its standardized abundance across fractions is obtained as the observed abundance dividing the maximum abundance among all fractions. The standardized abundance is between 0 and 1. The standardization of protein profiles will be further discussed in Section 1.3.4. As Figure 1.7 shows, we observe bell-shaped curves in protein profiles due to the resolution of SEC and IEX. In other words, we observe a peak and non-zero intensities across several adjacent fractions. Information of peaks,

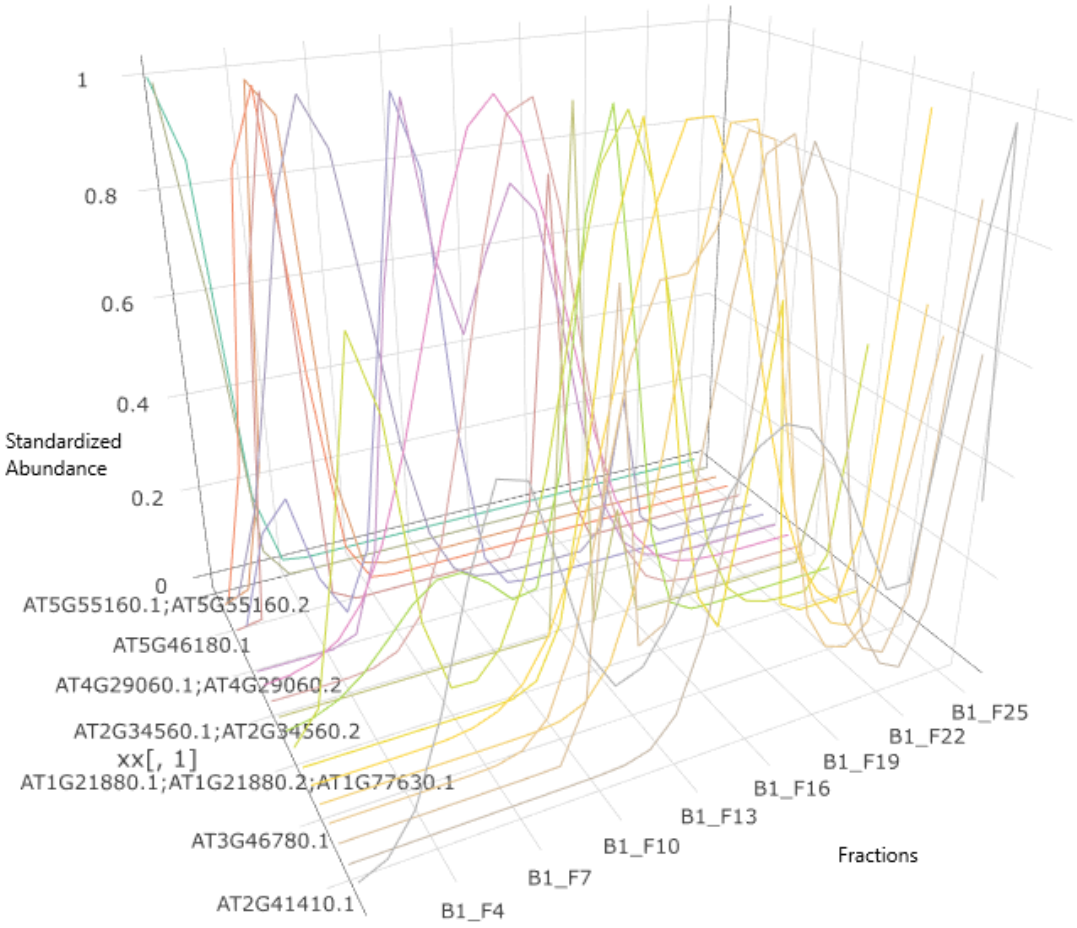


Figure 1.7. 3D plot of protein SEC profiles

like peak width, height, and location, characterizes protein profiles. Peak location is the most informative feature among the three parameters because it represents mass in the SEC data and charge in the IEX data. A peak in a protein profile indicates that the protein may belong to one protein complex with specific mass or charge that corresponds to the peak fraction. Typically, we expect a protein profile contains one peak, corresponding to the specific molecule mass or charge characteristic. However, a small number of proteins contains multiple peaks. A multiple-peak protein profile indicates that the protein is detected at multiple fractions. We infer that the proteins may belong to multiple complexes. In Section 1.4.3, we introduce a method to handle multiple peak proteins in protein complex prediction.

1.3.1 Existing methods to find peaks in protein profiles

Peakfinder (Yoder, 2011) is a simple algorithm that identifies peaks as local maxima in a protein profile. It returns a peak location if the peak is higher than its surrounding neighbor observations and the peak is at least 1/4 of the range of the data (maximum height minus minimum height). This algorithm performs better than other methods that use derivatives for local maximum, especially when data is very noisy. It is also very efficient and can identify peaks in millions of observations in seconds. However, it does not restrict the distance between peaks and does not smooth the profiles.

Another approach of finding peaks uses Matlab Curve Fitting Toolbox to deconvolve protein profiles into component Gaussian curves (Kristensen, Gsponer, & Foster, 2012). Fitting a Gaussian curve can be done by minimizing the squared error. A more challenging task is to determine the number of Gaussian peaks. The Matlab toolbox used leave-one-out cross-validation to decide the number of Gaussian peaks. At each iteration, the method drops one point from the profile and fits Gaussian curves. Then it calculates the squared error of the dropped point and the fitted value from the Gaussian curve and calculates means of those squared errors across itera-

tions. The fitted curve with the smallest sum of squares of errors gives the number of Gaussian peaks. However, the algorithm does not restrict the Gaussian curve to have a minimum distance between peaks.

1.3.2 Constrained Gaussian peak fitting for multiple peak protein profiles

The Gaussian curve fitting algorithm of our analysis is based on Kristensen et al. (2012) but adds constraints on Gaussian peaks, requiring two peaks to be separated by at least four fractions. The reason for adding constraints is that due to the resolution of chromatography, peaks that are within four fractions are not separable and hence cannot be used to indicate distinct peak locations.

A Gaussian curve function can be expressed in the following form,

$$f(x) = a \exp(-(x-b)^2/c^2) \tag{1.1}$$

where $f(x)$ represents the protein profile, or intensity, at fraction x, a is the peak height, b is the peak location, and c is the peak width. We generalize the single peak function to multiple peaks.

$$f(x) = \sum_{i=1}^{m} a_i \exp(-(x-b_i)^2/c_i^2) \tag{1.2}$$

where a_i b_i, c_i are the height, location, and width of the ith peak, m is the number of peaks with a value from 1 to 4, and $|b_i - b_j| > 4$ for any $i \neq j$. The differences between b_i's must be at least 4 due to the resolution issue mentioned above.

It is not necessary to fit Gaussian curves if most fractions have zero intensities. We fit Gaussian curves if a protein profile has at least two non-zero intensities in adjacent fractions. We fit one Gaussian peak with two to five non-zero fractions, up to two Gaussian peaks with six to eight non-zero fractions, up to three Gaussian peaks with nine to eleven non-zero fractions, and up to four Gaussian peaks with twelve or more non-zero fractions. Since the nonlinear curve fitting in Matlab is sensitive to starting values, a carefully chosen initial value set can improve curve fitting results and avoid local optimal solutions. We choose the initial value of the first peak location at the

maximum of the profile and then remove this point plus the neighboring 3 points and find the maximum of the rest points. Repeating this step, we define all starting points for the curve fitting. We use the Bayesian Information Criterion (BIC; Schwarz, 1978) to decide the number of Gaussian peaks. The definition of BIC is:

$$BIC = n\log(RSS/n) + k\log(n), \tag{1.3}$$

where n is the number of observations, k is the number of parameters, RSS is the residual sum of squares of the fitted curve. The number of peaks with the lowest BIC is selected. Among selected peaks, we removed peaks with heights less than 1/5 of the highest peak. Those small peaks are considered as noises and will not be used in the following analysis.

The constrained Gaussian fitting is more robust than the standard peak finding methods in the analysis of noisy data. Besides, the use of the BIC penalty prevents over-fitting and makes the algorithm efficient for a large number of proteins.

1.3.3 Use of Gaussian peak fitting for reproducible proteins

Note that we have two biological replicates for each separation approach, SEC or IEX. Peaks that only appear in one biological replicate are not reproducible, implying that the corresponding protein would not have a reliable complex conformation. We will not consider these proteins in our complex prediction. We use Gaussian curve fitting to identify peaks and determine reproducible protein profiles. Figure 1.8 shows an example of Gaussian peak fitting and reproducible peaks. We can see that there are two peaks in the first biological replicate (bio 1) at fractions 13 and 20 and these two peaks are within two fractions from the corresponding peaks in the second biological replicate (bio 2). Both two peaks are reproducible peaks and will be used for cluster analysis.

We use Gaussian curve fitting to remove inconsistent protein profiles. Reproducible peaks are defined as the peaks whose locations from biological replicates are close to each other. More specifically, we consider a peak is reproducible if its lo-

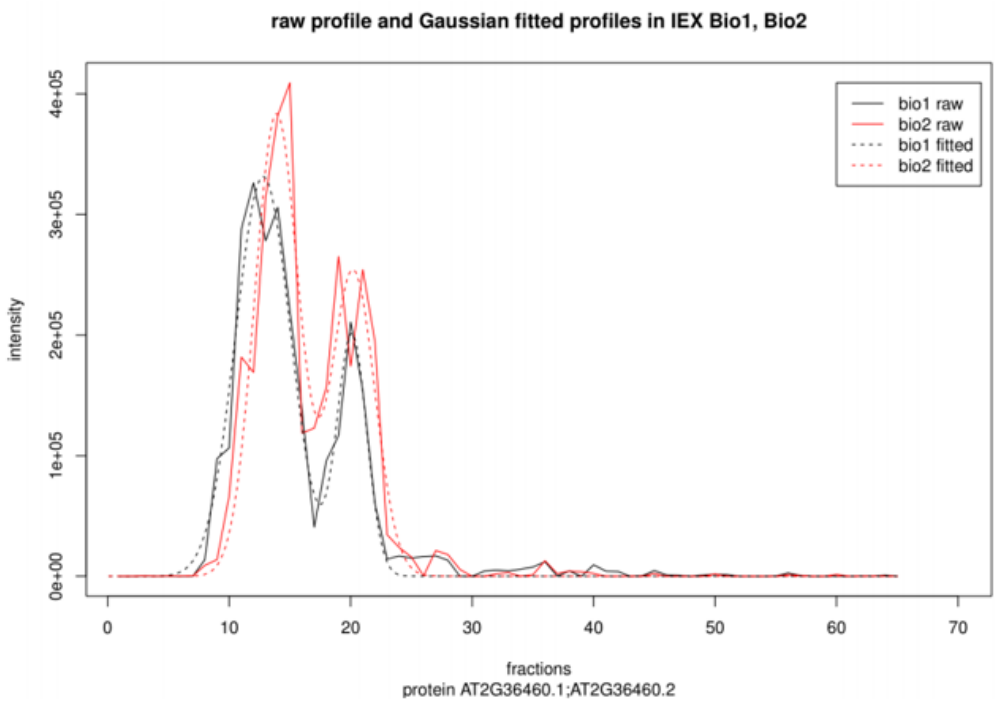


Figure 1.8. Gaussian peak fitting and reproducible peaks

cations in the two biological replicates are within four fractions for the IEX data and two fractions for the SEC data. If there is no Gaussian fitting for a protein, we use its maximum intensity location as the peak location. Proteins with at least one reproducible peak are considered reproducible proteins and will be included in the prediction of protein complexes. With this quality filtering criterion, we obtain 663 proteins as reproducible in both SEC and IEX data. We will conduct a clustering analysis of these proteins for complex prediction.

1.3.4 Standardization of protein profiles

A protein profile shows an elution pattern of the protein across separation fractions, where the protein's abundance amount is observed at each fraction. Profiles have peak intensities ranging from 10^5 to 10^9. (The value is the number of ions detected by the MS machine, which does not have a specific unit. See the discussion of

MS in the previous section.) Intensities represent ions detected by the MS machine, and it is loosely related to the number of peptides or proteins in the sample. We cannot directly cluster raw profiles when the intensities vary with such a large range. Peak locations provide more important information than peak intensities in the protein profile representation. The peak location of a protein implies that the protein is likely to have a specific molecular size or charge, which is a major characteristic of the protein. Intensities across all fractions are proportional to the protein amount observed in the sample, which does not correspond to a protein complex feature and may vary widely. Standardization of the raw profiles is needed to remove the effect of the large variation from the peak intensities. We standardize protein profiles by dividing each profile by its maximum intensity so that profiles of different proteins are comparable regardless of the absolute values of intensities.

1.4 Protein complex prediction through cluster analysis

Peak locations have been used to predict whether a protein belongs to a complex in our published papers McBride, Chen, Reick, Xie, and Szymanski (2017) and Aryal et al. (2014). A peak location in the SEC data corresponds to a molecular mass value, which we termed apparent mass and denoted M_{app}. It represents the overall protein complex mass. Another mass value termed monomer mass, and denoted as M_{mono}, is obtained from the protein amino acid composition. M_{mono} corresponds to the mass of the single protein alone. The ratio of M_{app} and M_{mono} is calculated for each protein, denotes as R_{app}. We also used two replicates of SEC in our previous work. If R_{app} of a protein is larger than 2 in both replicates, we predict this protein forms a complex. However, we do not know the composition of the complex. Our development here is to find the composition of the complex.

Cluster analysis, or clustering, groups observations based on their similarity (Friedman, Hastie, & Tibshirani, 2001). As a result, an observation is more closely related to another member in the same cluster than those in different clusters. It can be used in

an exploratory analysis to discover groups, where each group has a distinct property. We can then select some interesting groups for further analysis. The application of cluster analysis includes identifying groups of consumers (Rajagopal, 2011), identifying communities of a social network (Fortunato, 2010), extracting topics by clustering documents (Steinbach, Karypis, & Kumar, 2000), studying population structure by clustering human genes (Thalamuthu, Mukhopadhyay, Zheng, & Tseng, 2006) and many more.

Clustering algorithms can be categorized into connectivity-based clustering such as hierarchical clustering (Johnson, 1967), centroid-based clustering such as k-means (Ball & Hall, 1965), self-organizing maps (Kohonen, 1990), distribution-based clustering such as Gaussian mixture model (Banfield & Raftery, 1993), density-based clustering such as DBSCAN (Density-based spatial clustering of applications with noise; Ester, Kriegel, Sander, & Xu, 1996). Each type of these clustering methods makes their specific assumptions. Hierarchical clustering assumes objects in the same cluster are connected. K-means assumes spherical clusters and keeps variance small within a cluster. Gaussian mixture model assumes data follows multivariate normal distributions. DBSCAN assumes that if the number of neighbors of a data point is less than a certain value, that point falls at the boundary of a cluster. Each algorithm can detect certain types of clusters and may fail for other cluster types. The choice of clustering methods depends on the dataset and the definition of clusters.

In this project, cluster analysis is used to predict the composition of a protein complex. The rationale is that proteins with the same peak locations may belong to the same complex. In the following subsections, we describe our clustering algorithm. Our developments include a two-round clustering method, clustering with split protein profiles, and tree mining for sub-clusters.

1.4.1 Review of hierarchical clustering

Hierarchical clustering is done by growing a binary tree, whose leaves are the individual data, i.e., protein profiles in our problem. It repeatedly combines two nearest clusters into a larger cluster. A hierarchical tree, called dendrogram, is built. Figure 1.10(b) shows an example of a dendrogram. The height of a node represents the distance of two clusters or objects when they are merged. We can cut the dendrogram at a given height and correspondingly obtain a clustering result with a respective number of clusters.

We briefly describe the hierarchical clustering algorithm used in our analysis. The algorithm starts with the individual data observations, which is the individual protein profiles in our problem. At the first step, each observation belongs to a distinct cluster. Then two clusters with the smallest distance are merged and form a new cluster. Repeat this process until all observations are in one cluster. As an example, in Figure 1.10(b), AT4G33010 and AT3G24503 merge at height 1.21. Subsequently, these two proteins and AT2G364602_2 merge at height 1.71.

The hierarchical clustering algorithm relies on a measure of similarity, or in other words, a distance matrix between two data observations. The choice of similarity measure is typically based on data type and research question and it will affect clustering results. We use the squared Euclidean distance for any pair of protein profiles:

$$\|a - b\|_2^2 = \sum_i (a_i - b_i)^2 \tag{1.4}$$

where a and b are two vectors representing protein profiles and a_i, b_i are the ith elements of a and b. In our datasets, protein profiles have been standardized. If two proteins have the same peak location, their peak heights should be close to each other. Therefore, with the Euclidean distance, the two proteins with similar peak locations have distance close to 0. For clustering of similar datasets, simple peak location distance can also be used as the similarity measure (Kristensen et al., 2012).

The Euclidean distance defines a distance matrix of proteins at the first step of the hierarchical clustering. To merge two clusters, we need an updated similarity

measure for the distance between two clusters. We use Ward's method that defines the distance of cluster A and B as:

$$d(A,B) = \sum_{i \in A \cup B} ||x_i - m_{A \cup B}||^2 - \sum_{i \in A} ||x_i - m_A||^2 - \sum_{i \in B} ||x_i - m_B||^2 = \frac{n_A n_B}{n_A + n_B} ||m_A - m_B||^2 \tag{1.5}$$

where x_i is the i th observation, $|| \cdot ||$ is Euclidean distance, m_j is the center of the cluster j, and n_j is the number of points in it. The distance of two clusters can be viewed as the weighted squared Euclidean distance of the two cluster centers. We choose hierarchical clustering among available clustering methods. First, its dendrogram visualizes the similarity of one object to other clusters or other objects. Second, it has a stable cluster assignment, unlike k-means, whose results are affected by the choice of the initial centroids of clusters, which are randomly chosen. Third, it only needs to run once and the cluster IDs at different numbers of clusters can be done by cutting dendrogram at different heights. We choose Ward's distance because it has tree heights increase as the number of clusters decreases, which is not the case for average or centroid linkage function. Besides, Ward's method penalizes large clusters so that the distribution of cluster sizes is roughly uniform.

1.4.2 Two round clustering for integrative analysis

A challenge of conducting cluster analysis for our protein data is the presence of multiple data types, i.e., SEC and IEX, and two replicates of each. In other words, we need to consider integrative analysis combining four datasets. Integrative analysis is popular in biological studies, as it can remove noises and improve analysis results. There are multiple ways of integrating the two different types of data, i.e., SEC and IEX. One straightforward method is to concatenate multiple datasets. The concatenated profile of one protein combines four profiles from four datasets and uses the combined long profile for clustering. It will be discussed in the next subsection. Another method is to combine cluster results from four datasets. We propose a two-round clustering algorithm to combine clustering results from SEC and IEX data. The

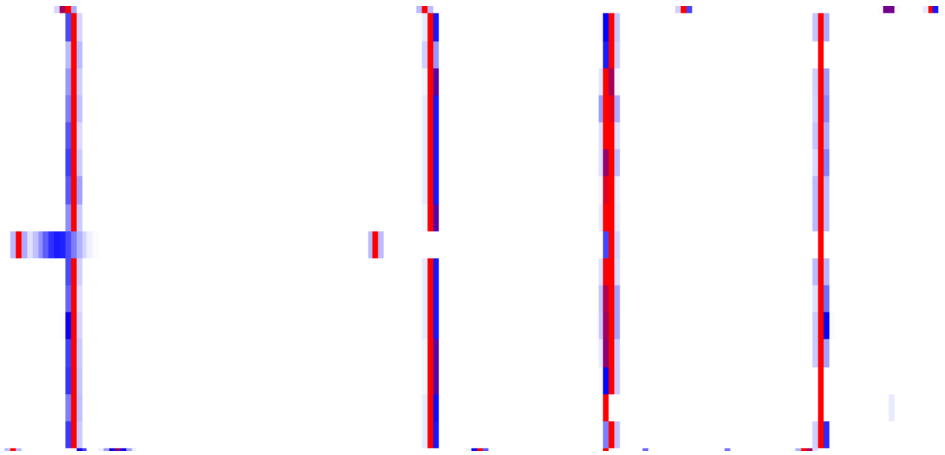


Figure 1.9. Heat map of two round cluster 42. Red tiles represent 1 and white ones represent 0.

basic idea is that proteins that fall into the same clusters in at least three datasets are likely to belong to the same group and form a true complex. Our method does not require proteins to be exactly in the same clusters in different datasets. Proteins in different clusters but with nearby peak locations will be clustered together using our method. More specifically, in the first round, we conduct standard cluster analysis and choose 300 clusters in each dataset. The cluster IDs are ordered according to the peak locations, that is, a low value cluster ID corresponds to peaks on the left of the separation fraction. The reordering assigns cluster IDs according to protein peak locations. The cluster number of 300 in the first round can be changed to an even larger number because it does not correspond to the cluster number of the final clustering result.

In the second round, we cluster proteins using their cluster IDs from the first round as the input data points. The number of clusters can vary and we select 300 as the final number of clusters based on cluster validation measures. In Section 1.6, we propose a tree mining method to refine clustering results and relax the need to decide the number of clusters. Figure 1.9 demonstrates an example of two-round clustering result through the heat map of protein profiles. It shows that all proteins in the

obtained cluster, except one, have the similar peak locations in the four datasets. Table 1.1 shows the cluster IDs of the group of proteins assigned by the first round of clustering in the four datasets. They do not have the same cluster IDs in all four datasets, but cluster IDs are very similar, which means that they have similar peak locations.

Table 1.1.
Cluster IDs in four datasets for proteins in two round cluster 42. Last four columns are cluster IDs for protein profiles in four datasets.

Two round cluster ID	Protein IDs	IEX1	IEX2	SEC1	SEC2
42	AT1G13060	277	226	140	127
42	AT1G16470	276	226	144	129
42	AT1G53850	277	225	141	129
42	AT1G56450	277	225	147	131
42	AT1G79210	277	226	147	127
42	AT2G05840	277	226	147	131
42	AT2G27020	277	226	149	131
42	AT2G37690	286	230	141	129
42	AT3G14290	276	225	149	131
42	AT3G22110	277	226	149	131
42	AT3G51260	277	226	147	131
42	AT3G60820	277	226	149	131
42	AT4G31300	276	226	148	131
42	AT5G35590	276	225	149	131
42	AT5G66140	276	224	141	129
42	AT5G42790	277	226	141	131

1.4.3 Fitted split profiles clustering for multiple peak profiles

From Section 1.2 we know that a protein with multiple peaks may indicate that the protein belongs to multiple complexes. Evidence shows that a protein can belong to multiple complexes and perform different functions (Regev-Rudzki, Karniely, Ben-Haim, & Pines, 2005). However, existing clustering methods do not allow one observation to be assigned to multiple groups. We propose to split a multiple-peak protein profile then run clustering. As a consequence, a multiple-peak protein will be predicted to participate in multiple complexes. We only consider one peak in the SEC data and combine it with each of the multiple peaks in the IEX data. Besides, we use fitted Gaussian curves instead of the original profile data for clustering, as the smooth Gaussian curves can reduce noises in the original data. For proteins that Gaussian curve fitting cannot be done, their standardized profiles are used. More specifically, we first fit Gaussian curves on protein profiles and only keep reproducible peaks. All peaks are standardized to have the same height of 1. Then for each protein, we combine the peak in the SEC data with the largest molecular mass (that is the one on the left) with each of the peaks in the IEX data. We only consider the peak corresponding to large mass in the SEC data, because protein profiles showing large molecular mass are more likely to form a complex. We only use one peak in the SEC data to simplify the problem, when a protein has multiple peaks in both SEC and IEX data. Moreover, since SEC has lower resolution than IEX, SEC produces fewer multiple peak proteins. In other words, only using one peak in the SEC data would not lose much information. We then create multiple profiles for a multiple-peak protein and name each with the original protein ATG number plus a suffix, e.g., "_1", "_2". By splitting profiles, multiple peak proteins can have multiple entries and may have multiple complexes prediction.

Figure 1.10 shows an example of split profile clustering. Protein AT2G36460 has two peaks and becomes two entries AT2G36460_1 and AT2G36460_2, which is also shown in Figure 1.8. Figure 1.10(a) and 1.10(c) show the concatenated profiles of two

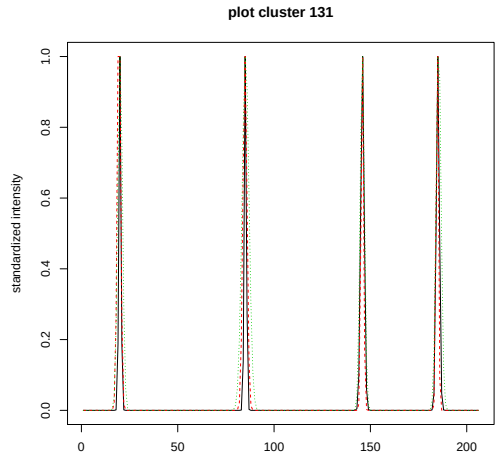
protein groups with two replicates of IEX and two replicates of SEC. The first group corresponds to Cluster 131 in our clustering result and contains the first split entry of AT2G36460_1. The second group corresponds to Cluster 142 and contains the second split entry of AT2G36460_2. AT2G36460_1 has the IEX peak location at fraction number 13, which corresponds to peaks at x-axis value 13 and 78 in Figure 1.10(c). AT2G36460_2 has the IEX peak location at fraction number 20, which corresponds to peaks at x-axis value 20 and 85 in Figure 1.10(a). We can see that proteins that are clustered together have the same peak locations in all four datasets respectively. Figure 1.10(b) and 1.10(d) shows the relatedness of AT2G36460_1 and AT2G36460_2 with other members in the cluster, respectively.

1.5 Cluster validation

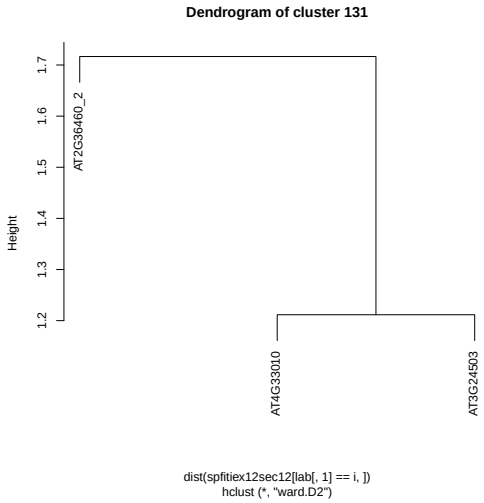
We want to decide the number of clusters and evaluate the cluster results. These are well-known difficult tasks and open problems in cluster analysis. Cluster analysis is unsupervised learning with no clear reference of goodness. Unlike supervised learning, there is no ground truth solution for clustering. In this subsection, we describe several clustering validation methods and use them for our specific situation. More details of cluster validation measures can be found in Handl, Knowles, and Kell (2005).

There are internal measures and external measures for cluster validation. Internal measures only use the information from the clustering results. There are multiple internal measures based on different features of clusters. One type of measures assesses cluster compactness, which is calculated from the average or maximum within-cluster variance. Another type of measures quantifies separation between clusters, for example, calculating the average or minimum distance between cluster centroids. The third type of measures is a combination of the first two measures. One well-known measure is the Dunn index (Dunn, 1974), as specified below

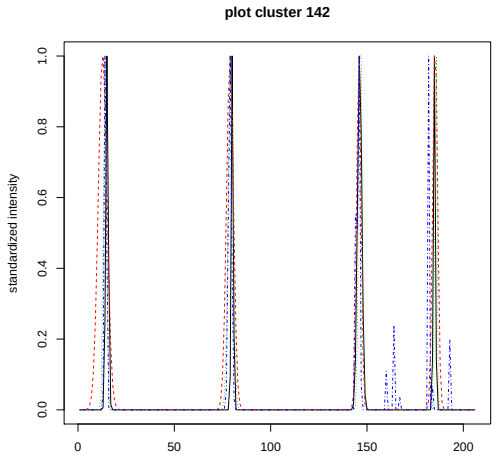
$$D = \frac{\min_{C_k \in C, C_l \in C} dist(C_k, C_l)}{\max_{C_m \in C} diam(C_m)} \tag{1.6}$$



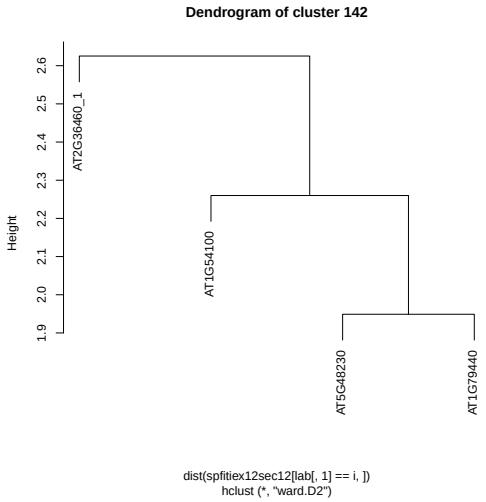
(a) Plot of concatenated protein profiles in cluster 131. The first two peaks represent peaks in two IEX datasets and the last two peaks represent peaks in two SEC datasets. Different lines represent different protein profiles.



(b) Dendrogram of cluster 131



(c) Plot of protein profiles in cluster 142. See Figure 1.10(a) for descriptions.



(d) Dendrogram of cluster 142

Figure 1.10. Example of split profile clustering

where C is the clustering result, C_k is the kth cluster, $dist(C_k, C_l)$ is the distance of cluster C_k and C_l, and $diam(C_m)$ is the within-cluster distance of cluster C_m. There are many definitions of within-cluster distance, for example, the maximum distance of all pairs within a cluster, the average distance of pairs within a cluster, or the maximum distance of objects to cluster centroid. Similarly, there are many definitions of between-cluster distance, for example, the minimum distance of all pairs of objects from two clusters, the average distance of pairs from two clusters, or the distance of two cluster centroids. Larger the Dunn index, better a clustering result. However, using only one number to represent a clustering result is imperfect and may have information loss. Besides, the limitation of internal validation measures is that they are biased toward certain types of clusters, for instance, spherical clusters. Measures based on the minimum or maximum distance may be biased when there are outliers. The first type of measures based on compactness is biased toward small clusters, while the second type based on separation is biased toward large clusters.

External measures require reference group information from additional sources. In other words, an external measure compares the clustering results with true labels or the "golden standard". Intactness and purity are two external validity measures. Intactness and purity are a pair of measures and they are comparable to precision and recall in classification problems. If we consider belonging to a certain cluster as a positive prediction result, intactness is comparable to precision and purity is comparable to recall. The F measure (Rijsbergen, 1979) is a weighted average of intactness and purity. Other measures like the Rand Index (Rand, 1971) and the Jaccard coefficient (Jaccard, 1908) are based on the contingency table of clustering results and the golden standard. They can also be interpreted as the percentage of correctly clustered objects in all objects. The limitation of external validation measures is that the number of known examples is usually very small and may not represent the whole population. External measures can be biased with respect to the number of clusters. Intactness favors larger clusters and purity favors small clusters.

Since each internal or external validation measure only reflects part of the properties of good clusters, it is recommended to compare multiple validation measures.

Finding the number of clusters is another challenge in clustering. For algorithms like k-means, the number of clusters needs to be decided before clustering is conducted. Hierarchical clustering alleviates the need of deciding the number of clusters by using dendrograms, but the dendrogram still needs to be cut at a certain number of clusters to report the final clustering result. Existing literature selects the number of clusters based on a sudden change of measures which is called a kink, a knee or an elbow. One well-known decision criterion is gap statistics (Tibshirani, Walther, & Hastie, 2001). Internal or external validation measures mentioned above can be used to find the number of clusters. Again, those criteria may work well with certain types of clusters like spherical clusters but fail at other types. It is recommended to compare multiple validation measures to get a range of reasonable numbers of clusters.

We use the following measures to decide the number of clusters and to validate the results in this project. We consider two internal measures here, namely compactness and tree height of the dendrogram. We also consider two external measures, intactness and purity. We choose those criteria because they are straightforward and reflect the principle of clustering, that is to maximize between-cluster distance and minimize within-cluster distance. These measures are described below.

Intactness

Intactness is defined as the maximum number of known proteins that belong to one cluster dividing the total number of proteins in the complex. When the number of clusters is 1, intactness is 1. Intactness decreases as the number of clusters increases. It is biased toward large clusters. We will use intactness as one guideline in our choice of the cluster number.

Specifically, assume there are two partitions of the dataset, denoted as $\{C_1, \ldots, C_m\}$ and $\{C'_1, \ldots, C'_p\}$. The first one is from a "golden standard", i.e., either known protein complexes or known information of true groups, and the second one is a clustering result. Intactness is used to measure the consensus of the two. The intactness value is defined for each of the golden standard clusters C_i,

$$\frac{\max_j |C_i \cap C'_j|}{|C_i|} \tag{1.7}$$

Purity

Purity is defined as the maximum number of known proteins that belong to one cluster dividing the total number of proteins in that cluster. It is also used to measure the consensus of two partitions. Purity can be represented as

$$\frac{\max_j |C_i \cap C'_j|}{|C'_j| \text{ where } |C_i \cap C'_j| \text{ gets its maximum}} \tag{1.8}$$

Purity typically increases as the number of clusters increases but can fluctuate because the denominator may change. In contrast to intactness, purity is biased toward small clusters. To get a reasonable number of clusters, we examine purity and intactness together. We select a range of reasonable numbers of clusters such that both measures take reasonably high values.

Compactness

Compactness is defined as mean squared Euclidean distance from its center. Deciding the number of clusters is based on the "elbow" principle, that is, the number of clusters is decided at the value when there is a great change in this measure. Specifically, compactness of cluster A is:

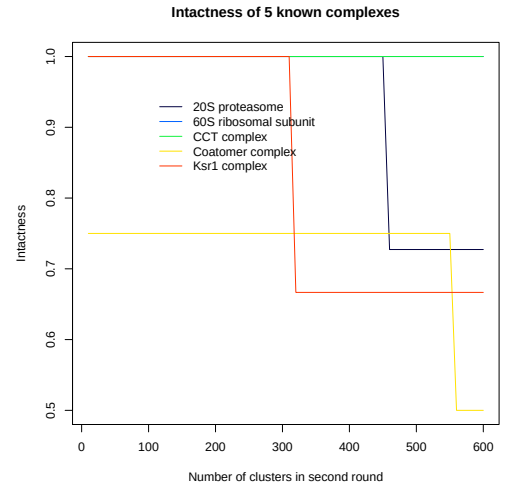
$$\frac{1}{n_A} \sum_{i \in A} ||x_i - m_A||^2 \tag{1.9}$$

where x_i is the i th observation, $||\cdot||$ is a distance measure, typically Euclidean distance, m_j is the center of the cluster j, n_j is the number of points in it. Similarly, we can also use it for a compactness of a subset of a cluster.

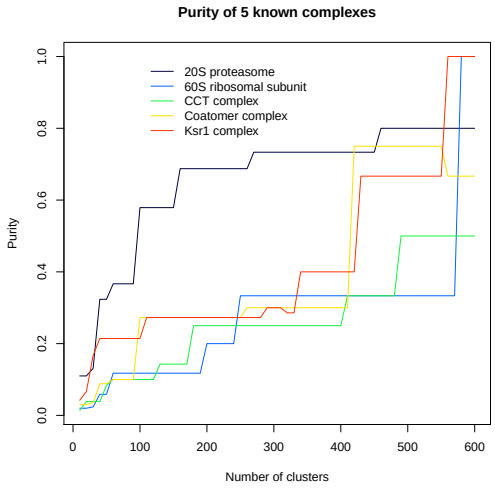
Tree Height

Tree height is the minimum pairwise Ward's distance between clusters at a given number of clusters. Ward's distance is defined in Formula (1.5). Deciding the number of clusters is based on the "elbow" principle. We choose the number of clusters when there is a great change in this measure.

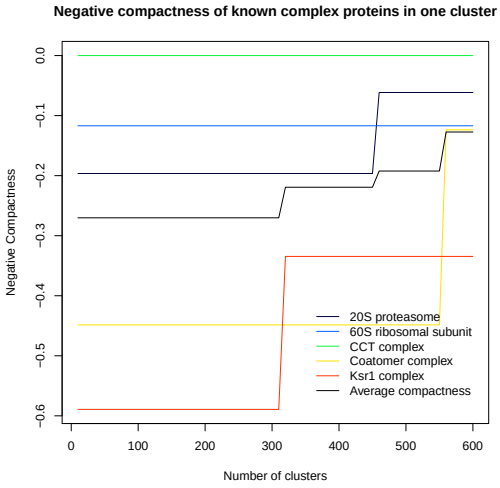
Figure 1.11 is the plot of purity, intactness, negative compactness from the two-round clustering for five known complexes with 300 clusters in the first round and varying numbers of clusters in the second round. In Figure 1.11(a), intactness is decreasing, as the number of clusters increases in the second round, but most known complexes have intactness 1 for up to 300 clusters in the second round, which means most known complexes form tight clusters. In Figure 1.11(b), purity is increasing as the number of clusters increases in the second round. We want to choose the number of clusters such that purity is not too low. No purity reaches 1 except when the number of clusters reaches 600, which means there exist proteins that are falsely clustered together. The reasonable number of clusters would be between 200 and 300. Figure 1.11(c) shows that negative compactness of known proteins in a cluster does not increase much, which shows that clustering results for known complex proteins do not change much. The compactness is plotted on the negative scale, so that larger values correspond to better clustering results Figure 1.11(d) shows the negative compactness of clusters that contain known complex proteins. There is a rapid increase in this measure from cluster number 100 to 250. We choose the cluster number of 300 based on these validation measure plots.



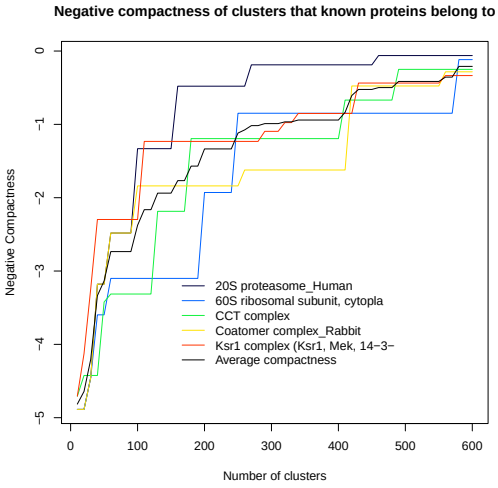
(a) Intactness



(b) Purity



(c) Compactness of known subunit



(d) Compactness of clusters that contain subunit

Figure 1.11. Purity, intactness, compactness of 2 round clusters at different cluster numbers. Each line represent one known complex.

1.6 Tree mining and prediction of protein complexes

Given a specific cluster number, we predict protein complexes based on the clustering results, as described in the previous sections. We improve the prediction results

by a data mining approach that does not require a choice of the number of clusters. The basic idea is to refine a cluster based on a sudden drop in tree heights. In each second-round cluster, we trace the dendrogram and cut the tree to obtain subset clusters. The tree will be cut at the biggest drop in tree height from a top level node to a low level node. This method finds the biggest drops from all drops whose lower level nodes have height less than 2.2. Then it cuts the tree at the lower node of the biggest drop. The value of 2.2 is chosen based on the observation that two protein profiles with a distance larger than 2.2 have different peak locations. This tree mining procedure does not require an exact choice of the number of clusters. Instead, clusters are obtained by tracing the dendrogram trees. Note that cutting the dendrogram tree into refined clusters generates protein complex prediction with strong confidence. Figure 1.12 shows an example of tree mining. The upper left panel shows the dendrogram of one cluster. The blue line represents the height of 2.2 and the red line shows the biggest drop in heights below 2.2. The lower right panel of Figure 1.12 shows the heights of each node of the dendrogram. This cluster is further partitioned into two subgroups by the red line. The lower right panel of Figure 1.12 shows that this subgroup for refined prediction of protein complex is more reliable than the prediction based on the original cluster, as the protein profiles in the predicted complex are very similar to each other.

We provide a table of the predicted complexes, including conservative predictions based on the refined clusters from tree mining Table 1.2 lists proteins in Cluster 19 and the cluster compactness value, which is the same cluster shown in Figure 1.12. We can see that Cluster 19 is further divided into two subgroups. Compactness in the refined clusters is smaller, which means proteins form tighter clusters and the corresponding complex predictions are more reliable.

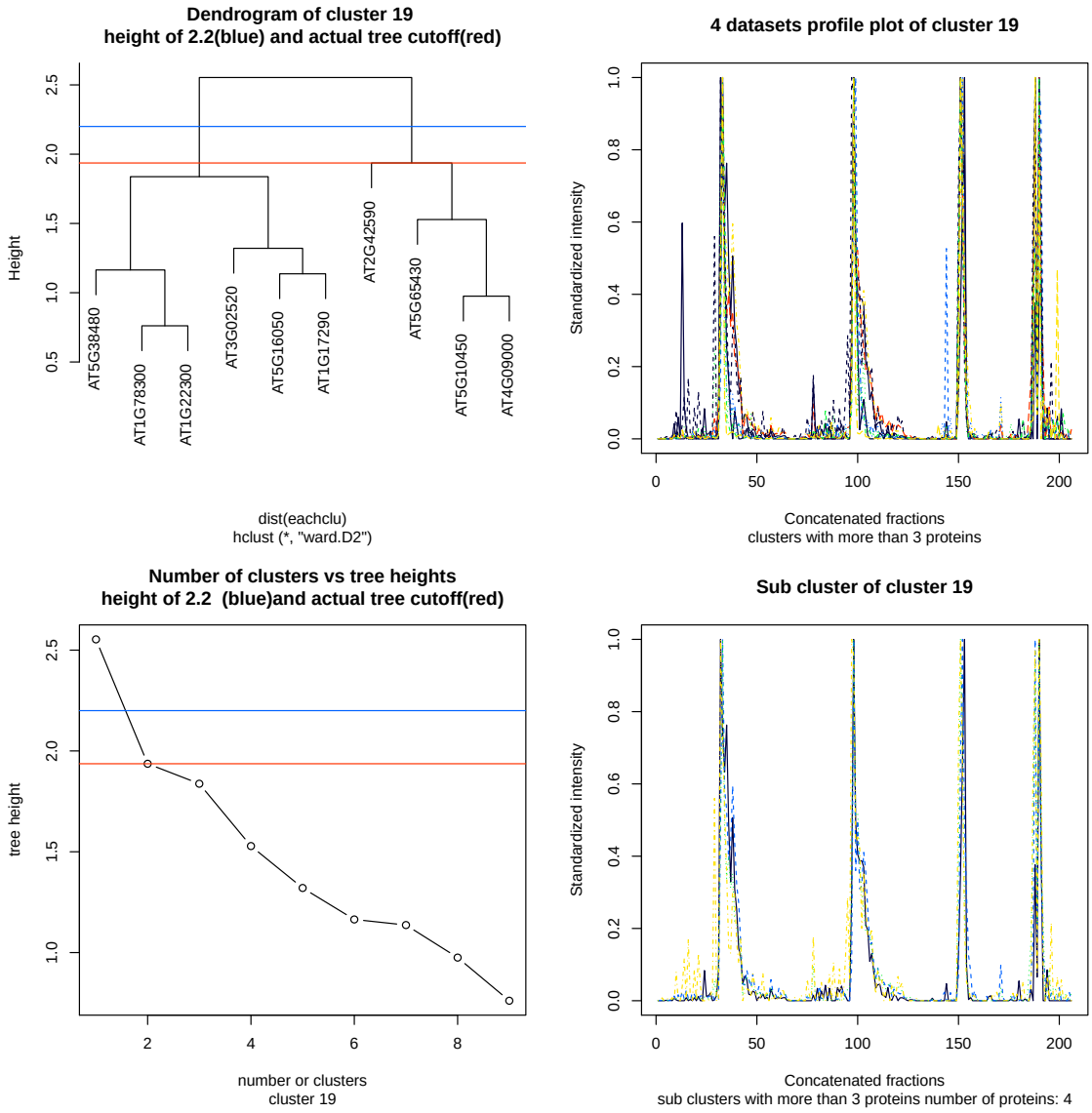


Figure 1.12. Example of tree mining. The blue line means the height of 2.2. The red line means the height the dendrogram is cut. The upper left panel is a dendrogram of Cluster 19. The lower left panel is the plot of sorted tree heights. The upper right panel is the profile plots of all proteins in Cluster 19. The lower right panel is the profile plot of one subcluster of Cluster 19.

Table 1.2.
Table of cluster 19 and refined clusters

Protein ID	Cluster ID	Compactness of Cluster	Cluster ID Refined	Compactness Refined Cluster
AT5G65430	19	1.095187352	191	0.8796217
AT5G10450	19	1.095187352	191	0.8796217
AT4G09000	19	1.095187352	191	0.8796217
AT2G42590	19	1.095187352	191	0.8796217
AT5G38480	19	1.095187352	192	0.695576283
AT5G16050	19	1.095187352	192	0.695576283
AT3G02520	19	1.095187352	192	0.695576283
AT1G78300	19	1.095187352	192	0.695576283
AT1G22300	19	1.095187352	192	0.695576283
AT1G17290	19	1.095187352	192	0.695576283

1.7 List of files

Here is the list of input files, output files, and R codes that can be used to reproduce the analysis results. The Matlab code is at Github `https://github.com/dlchenstat/Gaussian-fitting` as a supplementary file of a published paper McBride et al. (2017). The R code is at Github `https://github.com/dlchenstat/ProteinComplexPredict` as a supplementary file of a submitted manuscript McBride et al. (2018). Users can repeat the analysis results in this chapter using the codes.

1. The IEX input file
   (a) IEX_bio1_common_cytosol.csv
   (b) IEX_bio2_common_cytosol.csv
   (c) CytoContaminents.csv (remove some IEX proteins)

2. The IEX peak detection file (generated from Gaussian fitting Matlab code)
    (a) peakloc-iex-bio1-uma-2015aug.csv
    (b) peakloc-iex-bio2-uma-2015aug.csv
3. The IEX reproducible peak file (obtained from the peak detection file)
    (a) IEX reproducible peaks.csv
4. The SEC input file
    (a) SEC_Bio1_nov.csv
    (b) SEC_Bio2_nov.csv
    (c) SEC_Bio1_Bio2_cytosol_list.csv (select cytosolic proteins in SEC)
5. The SEC peak detection file (generated from Gaussian fitting Matlab code)
    (a) peakloc-sec1-uma-2015nov.csv
    (b) peakloc-sec2-uma-2015nov.csv
6. The SEC reproducible peak file (from the peak detection file)
    (a) SEC reproducible peaks.csv
7. The clustering analysis result with 300 clusters
    (a) Cluster ID SEC+IEX split fitted 300 clusters180131.csv
8. R code for data processing and clustering
    (a) ProteinComplexPredictFunctions.r
    (b) ProteinComplexPredictMain.r
9. The list of known proteins
    (a) Knowns2.0.csv

1.8 Conclusion

We have conducted a data science project that requires meaningful data representation and integration of multiple data types. We applied Gaussian fitting as data quality control for reproducible proteins, deconvoluted multiple peak proteins, and conducted two-round clustering to integrate different data types. We also developed tree mining to refine cluster results and relax the need to decide the number of clusters. We have created computer code for the analysis, which is accessible to broad users. Our collaborator, Dr. Szymanski's group, has been applying this method to other experiment data, such as rice, cotton, and soybean MS data.

2. OPTIMAL TESTS UNDER SPARSE ALTERNATIVE WITH COVARIANCE DEPENDENCE

2.1 Introduction

2.1.1 Global testing

Global hypothesis testing is a fundamental research problem in statistics. It tests whether the global null hypothesis, the intersect of multiple individual null hypotheses, is true. For example, a traditional problem of statistical analysis is to test whether a group of covariates (also called explanatory variables, features, independent variables or predictor variables in different contexts) has any linear relationship with a response variable (also called predicted variable, explained variable, outcome variable or dependent variable), the variable of interest. The application of global testing is widely utilized with the fast development of technology in genomics, neuroscience, finance, engineering, etc., where massive data are collected. For instance, in genome-wide association studies, genetic markers are grouped and are tested if they are associated with the traits, such as drug response or disease status. If the global null hypothesis is rejected, we conclude that at least one individual null hypothesis is false.

Multiple testing (also called multiple comparisons or multiplicity problem) occurs when a large number of tests are conducted simultaneously. It is related to global testing but has different testing criteria. Instead of the type I error, multiple testing controls the false discovery rate (FDR; Benjamini & Hochberg, 1995) or the family-wise error rate (FWER) among the multiple tests. FDR is the expected value of the number of false rejections divided by the number of total rejections. FWER is the probability of having at least one false rejection. Usually, global testing serves as the first step before multiple testing. For example, in comparison of means, the global

ANOVA test is done to determine whether there are any mean differences. If so, multiple testing of pairwise comparisons of means is conducted to determine which pairs have different means. On the other hand, FWER in multiple testing corresponds to the type I error probability of the respective global test. Therefore, the two types of analysis are related. We focus on global testing in this chapter.

Specifically, in this chapter, we study global testing to identify the association of a group of variables with another primary variable. Global tests aggregate information and reduce the burden of multiple testing, as one test instead of multiple tests is conducted. There are two ways to construct a global test. One is based on a joint model, e.g., a linear regression model, with all variables. The well-known F-test is a typical example of a global test. The other approaches combine individual test statistics of the corresponding individual hypotheses. For example, the minimum p-value test (MinP; Tippett, 1931) and higher criticism test (HC; Donoho & Jin, 2008). MinP and HC are particularly optimal and more powerful than the F-test under sparse alternatives, a common situation in modern data analysis. However, arbitrarily strong dependence among variables poses a great challenge towards the p-value calculation of these optimal tests. We will develop a latent variable adjusted method to correct the MinP test under dependence.

2.1.2 The Motivating example of GWAS

In this subsection, we discuss the application of global tests in genome-wide association studies (GWAS) and specifically the need for optimal global tests for sparse effects and correlated variables. GWAS aims to identify single nucleotide polymorphisms (SNP) or genes in an entire genome that are associated with a trait of interest such as disease status or response to treatment (Manolio et al., 2009). SNP is a single base-pair change in a certain location of a genome. SNPs are coded as 0, 1, 2, which represents the copy number of minor alleles. A typical data set contains several hundred thousand to one million SNP variables and one to two thousand subjects.

The most commonly used statistical method in genome-wide data analysis is the single-SNP method that analyzes individual SNPs' effects on the response, for example, using the t-test or the Armitage trend test (Armitage, 1955) for the association of an individual SNP with the response variable. The result is typically demonstrated in a Manhattan plot, as shown in Figure 2.1 below. It displays p-values on $-\log_{10}$ scale of individual SNPs along the whole genome. This simple approach has successfully identified thousands of SNPs that are associated with diseases in the past 20 years. However, the power of individual tests is limited. Since there are hundreds of thousands of SNPs, Bonferroni correction is usually used to adjust multiple p-values to reduce false positive results. SNPs need to have sufficiently strong effects to pass a stringent Bonferroni correction threshold. For example, to test one million SNPs in a dataset, the p-value threshold would be 5×10^{-8}. Figure 2.1 is the Manhattan plot of Crohn's disease (Duerr et al., 2006). There are about 300,000 SNPs, and the p-value threshold is set at $1.67 \times 10^{-7} = 0.05/300000$. Only a few SNPs from two genes are significant at this level, namely IL23R, NOD2. In other words, only a few SNPs have sufficient effects to be identified. Due to the nature of complex diseases, most SNPs may have moderate or small effects and thus cannot be detected. This is related to the problem of "missing heritability" (Manolio et al., 2009) that identified SNPs only contribute a little fraction of the total heritability. Another limitation is that single SNP analysis cannot detect interactions between SNPs.

An improved approach is a SNP-set analysis which analyzes SNPs by sets. SNPs can be grouped by physical locations, such as gene locations, intergenic regions, linkage disequilibrium (LD) blocks or pathways. The SNP-set analysis reduces multiple testing burden from millions of tests to tens of thousands of tests. Another advantage is that global testing of a whole group aggregates sparse and weak effects to improve power. Besides, global testing makes more meaningful interpretations as functions of genes or pathways.

To develop a powerful test in GWAS, we need to consider two issues that can affect type I error and power of global tests. One issue is sparsity. That is, only a few

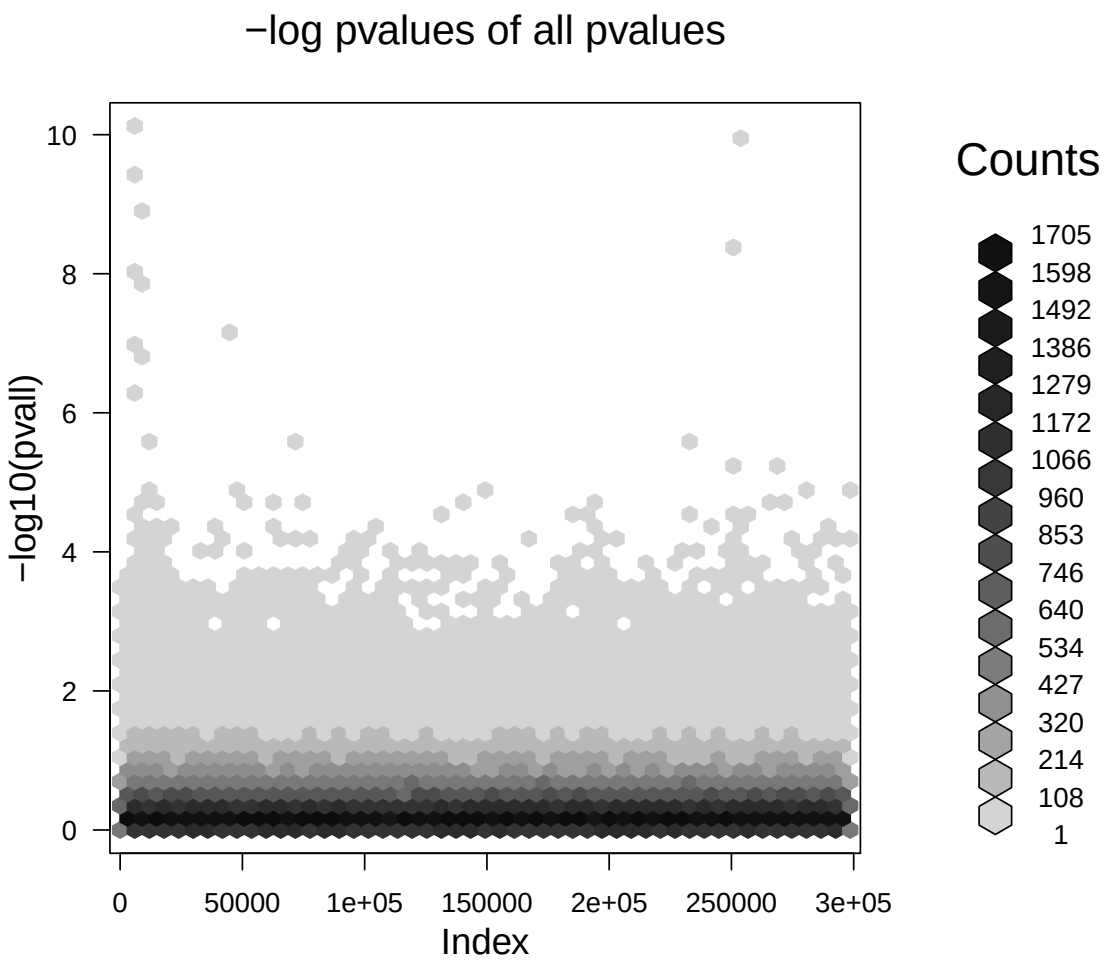


Figure 2.1. Example GWAS Manhattan plot of single SNP P-value in Crohn's disease dataset (Duerr et al., 2006). The y-axis is on $-\log_{10}$(P-values) scale, and the x-axis reflects the physical position of the SNPs by chromosome. Darker tiles represent more SNPs at a given p-value and SNP location.

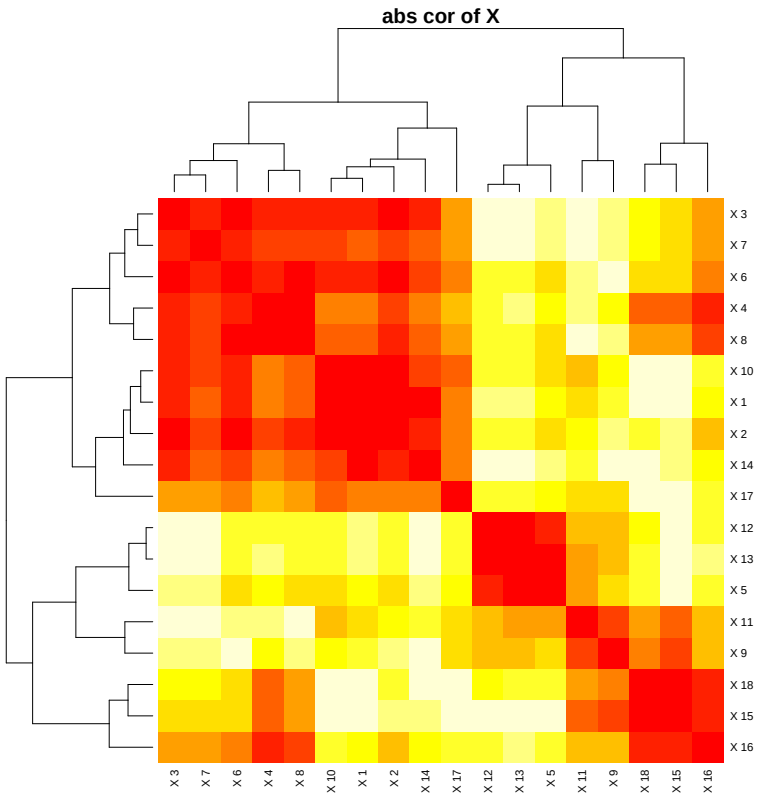


Figure 2.2. Correlation matrix of the gene NIPA1 in the Rheumatoid Arthritis Responder Challenge dataset (Cui et al., 2013). Red tiles mean the absolute correlations between two SNPs are close to 1. White tiles mean the absolute correlation is close to 0. SNPs are reordered so that highly correlated SNPs are together. Tree shaped diagram, also called dendrogram, shows the relatedness of SNPs.

SNPs are effective among a large number of SNPs in the data. As shown in Figure 2.1, the Manhattan plot displays that only SNPs in two genes pass the stringent Bonferroni p-value cutoff, implying that most SNPs have no effects or weak effects. Sparse effects can make F-test less powerful. Tests that would work for sparse effects have been studied, e.g., MinP test and HC (Arias-Castro, Candès, & Plan, 2011). We will review those tests in detail in Section 2.2.2.

Another issue is widespread linkage disequilibrium (LD). LD causes high correlations among genetic markers. It occurs because when a mutation happens, nearby genetic markers are affected together. Figure 2.2 shows an LD pattern in a gene named NIPA1 from the Rheumatoid Arthritis Responder Challenge dataset (Cui et al., 2013). The figure demonstrates that 10 SNPs are highly correlated with each other. This common situation in GWAS data indicates that marginal test statistics from the individual SNPs are also highly correlated. Statistical analysis with correlated variables is a challenging problem. For instance, when we consider a combination test such as MinP and HC, under dependence, there is no guarantee that the null distribution of the test statistics derived under the independence assumptions is still valid.

2.1.3 Existing tests

We review some of the existing global testing methods. Let us first consider a linear regression model

$$\mathbf{Y} = \beta_0 + \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \cdots + \mathbf{X}_d\beta_d + \epsilon, \tag{2.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is the response of n samples, $\mathbf{X}_i = (X_{i1}, \ldots, X_{in})^T$, is the ith variable, $i = 1, \ldots, d$, $\beta_0, \beta_1, \ldots, \beta_d$ are unknown regression coefficients, and ϵ denotes the error term and is assumed to follow $N(0, \sigma^2\mathbf{I})$. If $\mathbf{Y}$ is categorical, e.g., a dichotomous variable representing disease or control in a GWAS, a transformation is often used before applying the regression model.

Denote $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$. We test whether variables are associated with the response:

$$H_0 : \boldsymbol{\beta} = 0 \text{ against } H_1 : \boldsymbol{\beta} \neq 0.$$

The F-test statistic is defined as:

$$F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$$

where $SSE(R) = \mathbf{Y}'(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{Y}$ is the error sum of squares of the null model, e.g., the model only including β_0, $\mathbf{1} = (1, \ldots, 1)'$ is a vector of n 1's, $SSE(F) = \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$ is the error sum of squares of the model including all $\mathbf{X}_i$, with $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d)$ denoting the design matrix, $df_R = n - 1$ and $df_F = n - d - 1$ are degrees of freedom of the reduced and full model error sum of squares. The F-test statistic follows an F distribution under the null hypothesis H_0, and we reject the global null hypothesis when the F statistic is larger than the critical value. The analysis based on the regression model can accommodate other covariates and confounding variables. For example, in a GWAS, we can add clinical covariates, such as age, gender, and baseline variables, in the regression model and then test the effect of $\mathbf{X}$ conditional on the other covariates.

Another global testing approach is to consider the joint hypotheses of individual correlations between $\mathbf{Y}$ and $\mathbf{X}_i$. That is,

$$H_0 : E(\mathbf{X}_i'\mathbf{Y}) = 0 \text{ against } H_1 : E(\mathbf{X}_i'\mathbf{Y}) \neq 0,\ i = 1, \ldots, d$$

The global hypothesis is the intersect of all individual hypotheses. Suppose for each individual hypothesis we have calculated a test statistic with a p-value p_i, $i = 1, 2, \ldots, d$. Under the global null hypothesis that none of the variables $\mathbf{X}_i$ is associated with the response $\mathbf{Y}$ and the individual test statistics are independent, p_i's are iid and follow the uniform distribution $U[0, 1]$. Assume the alternative hypotheses have d^γ non-zeros, where $0 \leq \gamma \leq 1$ is a sparsity parameter. When the alternative hypotheses are sparse, with a small number of non-zero, the F-test will lose power. The F-test statistic can be interpreted as a sum of squares of the marginal test statistics. When the true effects are sparse, the sum of squares may contain too many noises and becomes less powerful.

The Minimum p-value test (MinP; Tippett, 1931) uses the maximum of the individual test statistics or the minimum of p_i's. More specifically, assume each $\mathbf{X}_i$ is standardized and divided by $\sqrt{n}$ so that $\mathbf{X}_i'\mathbf{X}_i = 1$. The marginal test statistic is

$$r_i = \mathbf{X}_i'\mathbf{Y}/s_y, \tag{2.2}$$

where s_y is the standard deviation of $\mathbf{Y}$. This is the sample correlation coefficient of $\mathbf{X}_i$ and $\mathbf{Y}$ multiplied by $\sqrt{n}$. The null distribution of this test statistic is available, e.g., as a t-distribution or a normal distribution. In fact, based on the central limit theorem, the test statistic can be well approximated by a normal distribution as long as the distribution of $\mathbf{X}_i$ and $\mathbf{Y}$ has bounded second moments and the sample size n is moderately large. MinP uses $\max_{1\leq i\leq d}|r_i|$ as the test statistic. Equivalently, we consider the minimum p-value of the marginal test statistics $p_{(1)} = \min_{1\leq i\leq d} p_i$, which follows the beta distribution $Beta(1, d)$ under the null hypothesis and when all marginal test statistics are pairwise independent. The p-value for the global hypothesis is $1 - (1 - p_{(1)})^d$. We reject the global null hypothesis when $p_{(1)}$ is small. It has been proved that under some regularity conditions, MinP is asymptotically powerful with the sparsity level $0 \leq \gamma \leq 1/4$ (Arias-Castro et al., 2011).

The Fisher's combination test statistic (Fisher, 1934) is defined as $-2\sum_{i=1}^{d}\log(p_i)$. Under H_0 and when marginal test statistics are independent, the test statistic is distributed as χ^2_{2d}. This test is powerful when there is a large proportion of non-zero values in the alternative hypotheses (Koziol & Perlman, 1978). By intuition, its test statistic may include too many noises when the alternatives are sparse. Therefore, the Fisher's combination test loses power under sparse alternatives. An adaptive Fisher's combination test (Liang, Wang, Sha, & Zhang, 2016) that combines a few top p-values was developed to adapt different sparsity scenarios. Other combination of marginal p-values are available as follows: $-\sum_{i=1}^{d}\log(1-p_i)$ (Pearson, 1933), $\sum_{i=1}^{d}\log p_i/(1-p_i)$ (Mudholkar & George, 1977), $\sum_{i=1}^{d} p_i$ (Edgington, 1972), $\sum_{i=1}^{d}\Phi^{-1}(p_i)$ (Stouffer, Suchman, DeVinney, Star, & Williams Jr, 1949). Heard and Rubin-Delanchy (2018) compares the power of each combination methods and provides guidance about the choice of combination in practice. The Cauchy combination test (Y. Liu & Xie, 2018b) uses $\sum_{i=1}^{d}\omega_i\tan((0.5-p_i)\pi)$ as the test statistic, where ω_i are non-negative weights and $\sum_{i=1}^{d}\omega_i = 1$. This test statistic's null distribution can be approximated by a Cauchy distribution regardless of the dependence structure of marginal test statistics. It is powerful under strong sparsity.

The Higher Criticism (HC; Donoho & Jin, 2015) test statistic summarizes d marginal p-values p_i for $i = 1, \ldots, d$ and investigates any significance in the whole set. It examines whether the sorted p-values $p_{(i)}$ is close to its expectation i/d. We first sort p-values p_i's in the increasing order $p_{(i)}$ then calculate the HC test statistic as:

$$HC = \max_{1 \leq i \leq d} \sqrt{d} \frac{(i/d) - p_{(i)}}{\sqrt{p_{(i)}(1 - p_{(i)})}} \tag{2.3}$$

Under H_0 and when the individual marginal statistics, or p_i's, are independent, HC has a defined asymptotic distribution as $d \to \infty$. However, for small d, the asymptotic distribution is not accurate. It has been proved that HC is asymptotically powerful under the sparse level $0 \leq \gamma \leq 1/2$ (Arias-Castro et al., 2011). Though it seems to be powerful under a larger range of sparsity than MinP, the result is asymptotic, which means the results only apply for large d. Tests that combine marginal p-values can be used when the raw data is not available or difficult to integrate, especially in meta-analysis.

Sequence kernel association test (SKAT) was proposed by Wu et al. (2011) for testing rare genetic variants in GWAS. SKAT uses a multiple regression model to regress the phenotype on genetic variants. The regression model can be semi-parametric with different kernel functions. By using the linear kernel, SKAT considers the same linear model as in (2.1). On the other hand, SKAT assumes $\boldsymbol{\beta}$ is random and follows an arbitrary distribution. It tests whether the variance of $\boldsymbol{\beta}$ is 0. Its test statistic is

$$\mathbf{Q} = (\mathbf{Y} - \hat{\boldsymbol{\mu}})' \mathbf{K} (\mathbf{Y} - \hat{\boldsymbol{\mu}}) \tag{2.4}$$

where $\mathbf{K} = \mathbf{X}\mathbf{W}\mathbf{X}'$, $\hat{\boldsymbol{\mu}}$ is the fitted response under the null model with only the intercept and $\boldsymbol{W}$ is a diagonal matrix of weights of each variable in $\mathbf{X}$. Different choices of weights can have an impact on the power of the test. Wu et al. (2011) suggested using the function of minor-allele frequency as weights. The test statistic Q follows a mixture of chi-square distributions. Moment matching is used to calculate its null distribution. It can be viewed as a weighted sum of squares of marginal test statistics. The test can be further interpreted as a weighted sum of independent chi-

squares with eigenvalues as their weights. We can decompose $\mathbf{K}$ as $\mathbf{VDV}'$, where $\mathbf{V}$ is a $n \times d$ orthonormal matrix of eigenvectors, $\boldsymbol{D}$ is the diagonal matrix with d non-zero eigenvalues $\lambda_1, \ldots, \lambda_d$. Then we can write Q as

$$Q = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'\mathbf{VDV}'(\mathbf{Y} - \hat{\boldsymbol{\mu}}). \tag{2.5}$$

Under the null hypothesis, Q has the same distribution as $\sum_i^d \lambda_i \chi^2_{1,i}$, the weighted sum of d chi-square of one degree of freedom. An exact method is used to calculate the null distribution (Davies, 1980). From (2.5), if the top eigenvectors, i.e., the first few columns of $\mathbf{V}$ are related to $\mathbf{Y}$, SKAT is powerful. As it is a sum of squares type of tests, it is powerful when effects are dense. Another benefit is that, if $\mathbf{X}$ is not full rank, it can adjust its null distribution with the correct degrees of freedom, so that the test is still valid.

The Neyman Pearson lemma can give a hint on powerful tests for global hypothesis testing. Assume $\mathbf{Z} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ is a vector of marginal test statistics and $\boldsymbol{\Sigma}$ is known. A uniformly most powerful (UMP) test is available for simple alternative hypothesis $H_0 : \boldsymbol{\theta} = 0$ vs $H_1 : \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, where the alternative hypothesis is at a specific vector value. From the Neyman Pearson lemma, the UMP test is $\tilde{\boldsymbol{\theta}}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}$, which is a linear combination of $\mathbf{Z}$ (Bittman, Romano, Vallarino, & Wolf, 2009). In general, the alternative hypothesis $H_1 : \boldsymbol{\theta} \neq 0$ is more complicated and we do not have information about the true value of $\boldsymbol{\theta}$. There is no UMP test for this alternative. The tests mentioned in this section make assumptions on the alternative values $\tilde{\boldsymbol{\theta}}$, for instance, the sparsity assumption for the MinP and HC, so that they can be powerful for certain types of alternative. The UMP test typically cannot be obtained for real data analysis, but the idea of optimal tests is very useful to compare different tests. We can view different kinds of tests as a certain choice of $\tilde{\boldsymbol{\theta}}$. If the choice of $\tilde{\boldsymbol{\theta}}$ is close to the actual one, the test is powerful. For example, the F-test uses $\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}$ as the test statistic and makes $\mathbf{Z}$ as the estimate of $\tilde{\boldsymbol{\theta}}$. On the other hand, if we assume effects have the same strength and the same direction, then the UMP test is $\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}$, which is analogous to Fisher's combination test when $\boldsymbol{\Sigma} = \mathbf{I}$.

2.1.4 Main contributions

Our first contribution is to examine optimal global tests under arbitrary variable dependence structures. We propose a latent factor adjusted MinP test to accommodate arbitrary dependence among variables and obtain accurate p-values for the adjusted MinP. The estimate of the latent variables is from the iterative reweighted surrogate variable analysis, which is an algorithm from a series of surrogate variable analysis (SVA) papers (Leek & Storey, 2008; Leek et al., 2010; Leek, 2014). Our proposed test is to adjust MinP by latent factors, which removes or reduces correlations among the original test statistics. In Section 2.3 we show the correct type I error probability. In Section 2.4 we show that under certain conditions, which assume that the latent variables do not affect the response, our test is more powerful than the original MinP.

Our second contribution is to examine data for appropriate uses of different tests. We characterize the conditions when a testing method is optimal then choose the most powerful method based on the data structure, for example, the sparsity level and correlation structure. As a special demonstration, we consider three tests, the F-test, the original MinP test, and the adjusted MinP test. We explore the data structure using fast and simple summaries then let the data guide the choice of the optimal testing method.

The rest of this chapter is organized as follows: In Section 2.2, we review existing global tests and methods to correct for correlation. We propose a latent variable adjusted MinP in Section 2.3. We provide two versions of proofs of type I error in Section 2.3.2 and 2.3.4. In Section 2.4, we show that under the factor model setting, the power can be improved. Simulation results of type I error and power are in Section 2.3 and Section 2.4, respectively. We propose a test combination strategy in Section 2.5 and report results of a real data analysis.

2.2 Challenges of optimal tests under dependency

2.2.1 MinP is not accurate under dependency

The dependence of marginal test statistics causes problems when calculating p-values for MinP. The issue here is analogous to that of the Bonferroni correction, or more precisely, the Sidak correction (Šidák, 1967). Specifically, Bonferroni uses $dp_{(1)}$ for the smallest adjusted marginal p-value, while Sidak uses $1-(1-p_{(1)})^d$. Those two are similar when $p_{(1)}$ is small and d is large. However, the correction is too conservative when there are high correlations among the marginal test statistics.

The example below shows that when the marginal test statistics are correlated the distribution of the MinP test statistic is much different from the case when the marginal test statistics are pairwise independent. Assume the marginal test statistics under the null hypothesis have a pairwise correlation ρ. Specifically, we express the test statistic as

$$Z_i = \sqrt{\rho}W + \sqrt{1-\rho}K_i, i = 1, \ldots, d \tag{2.6}$$

where W and $K_i, i = 1, \ldots, d$, are iid standard normal, and W can be viewed as the common factor shared by Z_i's. We consider a variety of levels of pairwise correlations among Z_i's, i.e., $\rho = 0, 0.1, \ldots, 0.9$. Let $z_{t/2}$ be the critical value such that $P(max|Z_i| < z_{t/2}) = 0.05$, when Z_i's are iid standard normal variables. It is easy to show that $t = 1 - 0.95^{1/d}$ and $z_{t/2} = \Phi^{-1}((1-0.95^{1/d})/2)$, which gives a type I error probability of 0.05. Figure 2.3 shows the actual type I error $P(max|Z_i| < z_{t/2})$ when Z_i's are correlated at different levels of ρ. We also vary the number of variables d. Note that only when $\rho = 0$, the type I error is controlled at the correct level of 0.05, i.e., $P(max|Z_i| < z_{t/2}) = (1-t)^d = 0.05$.

More specifically, we derive the distribution of the MinP test statistics when all test statistics are equally correlated. The tail probability of each marginal test statistic given the common factor W is:

$$\begin{aligned}
&P(|Z_i| > -z_{t/2}|W) \\
=&P(Z_i > -z_{t/2}|W) + P(Z_i < z_{t/2}|W) \\
=&P(\sqrt{\rho}W + \sqrt{1-\rho}K_i > -z_{t/2}|W) + P(\sqrt{\rho}W + \sqrt{1-\rho}K_i < z_{t/2}|W) \\
=&P(K_i > (-z_{t/2} - \sqrt{\rho}W)/\sqrt{1-\rho}|W) + P(K_i < (z_{t/2} - \sqrt{\rho}W)/\sqrt{1-\rho}|W) \\
=&\Phi((z_{t/2} + \sqrt{\rho}W)/\sqrt{1-\rho}|W) + \Phi((z_{t/2} - \sqrt{\rho}W)/\sqrt{1-\rho}|W)
\end{aligned}$$

The tail probability of the MinP test statistics given the common factor W is:

$$\begin{aligned}
&P(max|Z_i| < -z_{t/2}|W) \\
=&P(|Z_i| < -z_{t/2}, i = 1, \ldots, p|W) \\
=&(1 - \Phi((z_{t/2} + \sqrt{\rho}W)/\sqrt{1-\rho}) - \Phi((z_{t/2} - \sqrt{\rho}W)/\sqrt{1-\rho}))^d
\end{aligned}$$

The tail probability of MinP is:

$$P(max|Z_i| < -z_{t/2}) = \int \left(1 - \Phi\left(\frac{z_{t/2} + \sqrt{\rho}w}{\sqrt{1-\rho}}\right) - \Phi\left(\frac{z_{t/2} - \sqrt{\rho}w}{\sqrt{1-\rho}}\right)\right)^d / \sqrt{2\pi}\exp(-w^2/2)dw$$

In this example, we have an analytical distribution of the MinP test statistic although it is only a special case. Figure 2.3 shows that the actual type I error decreases and is much lower than 0.05 as ρ increases and d increases. When the actual type I error probability is less than 0.04, it is considered as a serious bias of the null distribution. When $\rho \geq 0.4$ we observe that the null distribution of the original MinP test is not correct anymore and it cannot be used for the appropriate type I error control.

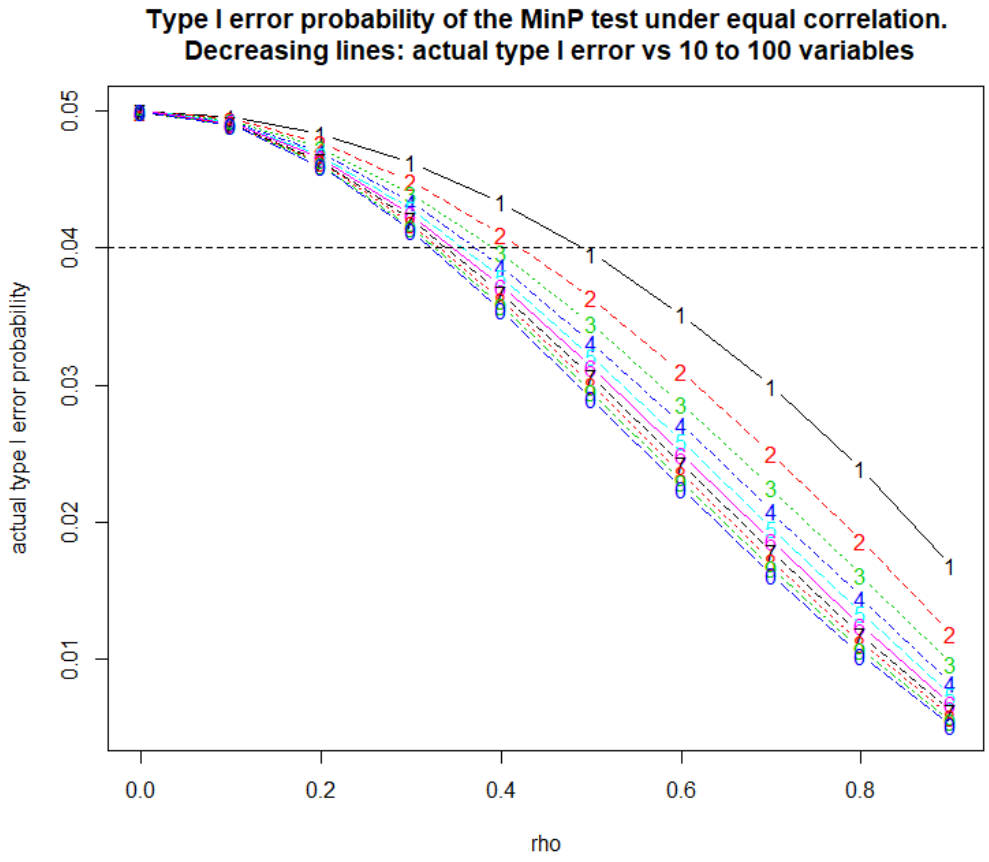


Figure 2.3. Type I error probability of the MinP test under the independence assumption and with equal correlation among any pair of variables. Each decreasing line corresponds to a different number of variables $d = 10, 20, \ldots, 100$ and ρ is the pairwise correlation coefficient.

2.2.2 Existing methods to correct for dependence: methods that calculate p-values under correlation

To correct the null distribution of MinP under dependence, two types of methods were proposed. One calculates p-values under correlation by Monte Carlo sampling, for example, the permutation method used in GWAS. In permutation sampling, the response variable in the sample is shuffled to obtain an empirical distribution of the test statistics under the null hypothesis. Versatile gene-based association study (VEGAS; J. Z. Liu et al., 2010) uses permutation to compute MinP p-values for genes under correlation. However, permutation is very time-consuming. To get a p-value that is accurate at 10^{-6}, we need at least 10^8 permutation samples. To alleviate the computation burden, Barnett, Mukherjee, and Lin (2017) proposed an adaptive permutation. The permutation stops early if the empirical p-value is not significant. One thousand permutation samples are simulated in the first step. If the empirical p-value is greater than 0.1, the simulation stops. Otherwise, we continue to conduct 10,000 simulations. If the empirical p-value is greater than 0.01, the simulation stops. Otherwise, we continue the simulations, etc.

Lamparter, Marbach, Rueedi, Kutalik, and Bergmann (2016) proposed **Pa**thway **sc**oring **al**gorithm (Pascal) to significantly reduce permutation burden using different p-value calculation approaches for different occasions. Pascal uses permutation if none of the marginal p-values is significant. It uses integral over multivariate normal probability density function (Genz, 1992) if any marginal p-value is lower than 10^{-5} and the dimension is less than 1000. If the dimension is larger than 1000 and any marginal p-value is less than 10^{-15}, it uses Bonferroni adjustment with the effective number of tests (Gao, Starmer, & Martin, 2008) to calculate the MinP test under dependence. Using integral of multivariate normal density to calculate MinP p-values is based on the fact that the vector of all test statistics asymptotically follows a multivariate normal distribution (Conneely & Boehnke, 2007). The R package mvtnorm (Genz et al., 2012) uses an efficient multivariate normal integral algorithm.

It calculates the probability of correlated multivariate normal that falls in a high dimension cube with a dimension less than 1000 in less than 1 minute. Its limitation is that the multivariate normal assumption is too strong, and it may not be correct for a small sample size though this assumption holds asymptotically (Conneely & Boehnke, 2007). The effective number of tests is defined as the number of principal components that explain 99.5% of the total variance. Gao et al. (2008) shows it can approximate the results of permutation accurately and efficiently. Pascal can compute about 18,000 gene test statistics in 34 minutes (Lamparter et al., 2016). Conneely and Boehnke (2007) calculates MinP p-values using a random sample of correlated multivariate normal, which is much faster than permutation. Y. Liu and Xie (2018a) proposed Gaussian approximation, which is an accurate and efficient method to calculate MinP p-values under dependence. In each round, it generates a random vector of length d rather than n in the permutation, so it is efficient when $d < n$.

Another type of method for calculation of p-value of MinP under dependence was proposed by R. Sun and Lin (2017). They used an extended beta-binomial distribution to approximate distribution of the number of test statistics that are larger than certain values, to analytically calculate p-values. The method defines $S(t)$ as

$$S(t) = \sum_{i=1}^{d} I\{|Z_i| > t\}, \tag{2.7}$$

where Z_i's are marginal test statistics and $S(t)$ is the number of marginal tests that are larger than the threshold t. The method approximates the distribution of $S(t)$ using an extended beta-binomial distribution, whose variance can be larger than the ordinary binomial distribution. The MinP p-value is

$$P(|Z_i| < \max(z_i), i = 1, \ldots, d|\mathbf{Z} \sim N(0, \mathbf{\Sigma})) = P(S(\max|z_i|) = 0|\mathbf{Z} \sim N(0, \mathbf{\Sigma})), \tag{2.8}$$

where z_i is the observed marginal test statistic. However, in the highly correlated case, like equal correlation, the type I error is not reliable since the extended beta-binomial distribution approximation may not be accurate. The p-value is also not

accurate for very small p-values, which are the significant testing results of interest. Another disadvantage is that computation time is $O(d^3)$, which takes much more time for larger SNP groups.

2.2.3 Existing methods to correct for dependence: methods that transform the data

Transforming the marginal test statistics or variables is another approach to correct for dependence. Innovated higher criticism (Hall & Jin, 2010) transforms test statistics to independent test statistics using the inverse of Cholesky decomposition of the correlation matrix and then apply HC for the transformed test statistics. Specifically, assume the test statistics $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Cholesky method decomposes $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$, where $\boldsymbol{\Sigma}$ is a symmetric and positive definite matrix and $\mathbf{A}$ is a lower triangle matrix. We denote $\boldsymbol{\Sigma}^{-1/2} = \mathbf{A}^{-1}$. The transformed test statistics are $\boldsymbol{\Sigma}^{-1/2}\mathbf{Z} \sim N(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, \mathbf{I})$, which are then mutually independent. The method calculates the HC test statistics from the transformed marginal test statistics. However, the transformation can be unstable. There is no unique solution of $\boldsymbol{\Sigma}^{-1/2}$, and Cholesky decomposition is only one of those. In addition, after the transformation, signals from the transformed test statistics may be diluted (Barnett et al., 2017). It is possible that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$ is not sparse and MinP is not powerful for the transformed marginal test statistics. Besides, the method does not work when the correlation matrix is not full rank.

Transformation through principal component analysis has also been proposed. Those methods use the correlations between principal components (PC) and the response as marginal test statistics. Because PCs are orthogonal to each other, marginal test statistics are independent. MinP can be applied to the marginal test statistics, and its type I error is correct. Aschard et al. (2014) proposed to use principal components to test the relation of multiple traits vs SNPs. They found that the marginal test of the top PC is not the most significant, which is contrary to what most people

think. They proposed combining tests of PCs to increase power. Z. Liu and Lin (2018) studied the power of using MinP of marginal tests with PCs. They found that when the principal angle is 0, the PC is the most powerful and if the principal angle is 90, the PC is powerless. The limitation of the methods that transform the variables is that they do not report the significance of individual variables. The significance of combined variables may be hard to interpret.

2.2.4 Existing methods to correct for dependence: factor models

The following three methods estimate the latent structure of the data. In different papers, it is referred to as "surrogate variables" (Leek & Storey, 2008), "latent factors" (Friguet, Kloareg, & Causeur, 2009), "unwanted variation" (Gagnon-Bartsch & Speed, 2012), or "latent effects" (Y. Sun, Zhang, & Owen, 2012). Different names refer to the same idea that some unobserved variables can explain most structure of the test statistics. Another similar concept is "confounding variables", which are referred to variables that affect both covariates and the response variable (Price et al., 2006). Ignoring confounding variables may cause a spurious association. We use the term of latent variables in our development. In the literature, latent variable methods are used to correct for multiple testing dependence, to provide an accurate estimate of false discovery proportion (Fan, Han, & Gu, 2012), to control FDR (Leek & Storey, 2008), to correct population structure in GWAS (Price et al., 2006), and to correct batch effect (Leek et al., 2010) in gene expression data. We will use latent factor modeling to remove dependence among test statistics and develop an adjusted MinP test. With the latent factor modeling, the adjusted test statistics are weakly dependent. It becomes valid to apply the null distribution of MinP for accurate calculation of p-values.

Singular value decomposition (SVD) and spectral decomposition are important techniques used in the estimation of latent factors. We briefly review them first. Singular value decomposition decomposes a matrix into a product of three matrices,

i.e., $\mathbf{A} = \mathbf{UDV}'$, where columns of $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and have norms of 1. Matrix $\mathbf{D}$ is a diagonal matrix with positive real numbers called singular values, and its diagonals are in decreasing order. Spectral decomposition is like SVD but considers a square matrix. More specifically, the decomposition is $\mathbf{A} = \mathbf{Q\Lambda Q}'$, where columns of $\mathbf{Q}$ are eigenvectors, and they are orthogonal and normalized. A diagonal matrix $\mathbf{\Lambda}$ is the matrix of the corresponding eigenvalues with its diagonals sorted in decreasing order.

Fan et al. (2012) proposed a method called principal factor approximation (PFA) to subtract the dependence among test statistics and weaken the correlation structure. However, it was designed for estimation of false discovery proportion in the problem of multiple testing. It was later proved to have good properties under unknown dependence, i.e., when $\mathbf{X}$ is a random sample, and its covariance matrix $\mathbf{\Sigma}$ is unknown (Fan & Han, 2017). Assume the test statistics have a joint Gaussian distribution

$$(Z_1, \ldots, Z_d)^T \sim N((\mu_1, \ldots, \mu_d)^T, \mathbf{\Sigma}) \tag{2.9}$$

Using principal component analysis, we rewrite $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \sum_{i=1}^{d} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i' \tag{2.10}$$

where $\boldsymbol{\gamma}_i$'s are eigenvectors and λ_i's are their eigenvalues arranged in decreasing order. We consider the top k eigenvalues and the corresponding eigenvectors and write in a matrix

$$\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{A} \tag{2.11}$$

where $\mathbf{L}$, a $d \times k$ matrix with $\mathbf{L} = (\sqrt{\lambda_1}\boldsymbol{\gamma}_1, \ldots, \sqrt{\lambda_k}\boldsymbol{\gamma}_k)$, represents the matrix of the top k eigenvectors, and $\mathbf{A} = \sum_{i=k+1}^{d} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i'$ is the remaining matrix. Diagonals of $\mathbf{A}$ are $(a_1^{-2}, \ldots, a_d^{-2})$, where $a_i^{-2} < 1$, since diagonals of $\mathbf{\Sigma}$ are 1 and diagonals of $\mathbf{LL}'$ are positive. We can write $(Z_1, \ldots, Z_d)^T$ as

$$Z_i = \mu_i + \mathbf{b}_i'\mathbf{W} + K_i \tag{2.12}$$

where $\mathbf{b}_i$ is the ith row of $\mathbf{L}$, $\mathbf{W} = (W_1, \ldots, W_k)$ is a k dimensional latent factors, and $\mathbf{K} = (K_1, \ldots, K_d)$ follows $N(0, \mathbf{A})$. The common factor $\mathbf{W}$ is unknown but can

be estimated. Fan et al. (2012) used L_1 regression to estimate $\mathbf{W}$ because it is robust against outliers. PFA used 90% of the smallest $|z_i|$'s for the regression model (2.12) so that μ_i's were approximately zero.

The adjusted test statistics are:

$$a_i(Z_i - \mathbf{b}_i'\hat{\mathbf{W}}) \tag{2.13}$$

where $\hat{\mathbf{W}}$ is a good estimator of the common factor $\mathbf{W}$ and a_i is a constant greater than 1 (Fan et al., 2012). The adjusted marginal test statistic has a larger mean than the original test statistic Z_i by multiplying a_i and $a_i > 1$. The simulation in Fan et al. (2012) showed the dependence-adjusted test statistics are improved in terms of reducing false non-discovery rate which is referred to as the number of not rejected false null divided by the number of negatives.

EIGENSTRAT (Price et al., 2006) was proposed to control population structure in GWAS. It uses the first few left singular vectors of $\mathbf{X}$ as the latent variables. The method tests the effect of a SNP conditional on the latent variables. However, EIGENSTRAT would lose power when SNP variables are strongly correlated to the response. In this case, removing latent variables leads to removing effects as well. It happens when the effects are strong. On the other hand, in a typical GWAS where effects are weak, this method would work fine. EIGENSTRAT will be used to reduce the correlation structure in one of our factor models in Section 2.3.3. Assume the design matrix $\mathbf{X}$ can be decomposed as $\mathbf{X} = \mathbf{UDV}'$. Let $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ and $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2)$, where $\mathbf{U}_1$ is the first k columns of $\mathbf{U}$ and $\mathbf{V}_1$ is the first k columns of $\mathbf{V}$. EIGENSTRAT uses the correlation of $(\mathbf{I} - \mathbf{U}_1\mathbf{U}_1')\mathbf{X}$ and $(\mathbf{I} - \mathbf{U}_1\mathbf{U}_1')\mathbf{Y}$, that is, the residual regressing on $\mathbf{U}_1$, as the adjusted marginal test statistics.

Surrogate variable analysis (SVA; Leek & Storey, 2008) conducts iterative weighted SVD on the design matrix to estimate latent variables and remove dependence among marginal test statistics. The method is analogous to the usual SVD on the design matrix, but it assigns less weights on variables correlated with the response variable. Conditional on the latent variables, it can be proved that the test statistics become independent and the standard multiple testing criterion can be appropriately applied.

An R package of implementing SVA (Leek, Johnson, Parker, Jaffe, & Storey, 2012) is available at Bioconductor. A limitation is that it is computationally intensive, and it does not always converge. Chen et al. (2017) developed an R package named SmartSVA, which implements efficient estimate algorithm of the number of latent variables, SVD, and F-test in C++ to speed up the computation and make SVA ten times faster. Several variants of SVA were proposed after Leek and Storey (2008). Direct SVA (S. Lee, Sun, Wright, & Zou, 2017) is claimed to improve SVA latent variable estimates when hidden factors are strongly correlated with the primary variables and it does not need iterations. Frozen SVA is used for prediction (Parker, Bravo, & Leek, 2014). Independent SVA uses independent component analysis instead of SVD to better model non-linear independence among variables (Teschendorff, Zhuang, & Widschwendter, 2011). Svaseq (SVA for sequencing data; Leek, 2014) was proposed to use the log transformation to model counts data. We will use SVA in one of our factor models. The detail development is described in Section 2.3.1.

Other latent variable methods include LEAPP (latent effect adjustment after primary projection; Y. Sun et al., 2012), FAMT (factor-analysis-based multiple testing, Friguet et al., 2009), robust FARM test (Fan, Ke, Sun, & Zhou, 2017), which is robust under non-Gaussian data, RUV-2, which uses control probes to estimate latent variables (Gagnon-Bartsch & Speed, 2012). We will show the performance of two adjustment methods in Section 2.3 and 2.4, namely SVA adjusted MinP and EIGENSTRAT adjusted MinP. Overall, the adjusted method through SVA outperforms the original MinP and many other testing methods.

2.3 Adjusted MinP test for arbitrary dependence structures

We have two versions of the proposed tests based on two model settings, namely the inverse regression model (IRM) and regression model (RM). The two models have different assumptions, but both can be used to answer the same question, that is, whether there is any relationship between a group of covariates and the response

variable. IRM regresses covariates on the response variable. Although the idea of inverse regression is less common than the classical regression model, it has certain advantages and has been well studied for multiple testing dependence. A favorable feature of inverse regression is that it is convenient to decompose the design matrix of the covariates as the sum of effects, common factors, and random noises (Leek & Storey, 2008; Y. Sun et al., 2012; Wang, Zhao, Hastie, & Owen, 2015; Friguet et al., 2009). RM is the classical model that regresses the response variable on the covariates. Latent factor modeling under RM has been studied by Fan et al. (2012) for estimating false discovery proportion.

2.3.1 Factor modeling of dependence with an inverse regression model (IRM)

Consider the inverse regression model:

$$\mathbf{X} = \mathbf{YB} + \mathbf{E}, \tag{2.14}$$

where $\mathbf{Y}$ is $n \times 1$ as the vector of the primary variable observed in n subject, n is the sample size, $\mathbf{X}$ is $n \times d$ as the matrix of the covariates, d is the number of covariates, $\mathbf{B}$ is $1 \times d$, a vector of coefficients representing the relationship between $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{E}$ is the $n \times d$ residual matrix with means of 0. For example, $\mathbf{Y}$ represents the disease status or drug response and $\mathbf{X}$ is the matrix of SNPs. Let $\mathbf{X}_i$ be the vector of the ith covariate. There is arbitrary dependence among the covariates, or equivalently, the columns of $\mathbf{E}$ can be highly correlated. Our objective is to examine the relationship between $\mathbf{Y}$ and $\mathbf{X}$. This can be done through global hypothesis testing of

$$H_0 : \mathbf{B} = 0 \text{ against } H_1 : \mathbf{B} \neq 0.$$

To accommodate the dependence among the columns of $\mathbf{E}$, we decompose it and express model (2.14) as

$$\mathbf{X} = \mathbf{YB} + \mathbf{G\Gamma} + \mathbf{U}. \tag{2.15}$$

where $\mathbf{G}$ is $n \times k$, as the matrix of k latent variables, $\mathbf{\Gamma}$ is $k \times d$ as the coefficient of $\mathbf{G}$, $\mathbf{U}$ is $n \times d$ as the random noise. We use the latent variables to characterize the

dependence structure of the covariates. The latent variables can be interpreted as the common factors that the covariates share. We assume the latent variables, or columns of $\mathbf{G}$, are iid. Leek and Storey (2008) used this same model for addressing arbitrarily strong multiple testing dependence. It has been proven that under a reasonable condition, there exist matrices $\mathbf{G}$, $\mathbf{\Gamma}$, and $\mathbf{U}$ such that the columns of $\mathbf{U}$ are jointly independent random vectors.

We can write the inverse regression model for the ith column of $\mathbf{X}$ as:

$$\mathbf{X}_i = \mathbf{Y}B_i + \mathbf{G}\mathbf{\Gamma}_i + \mathbf{U}_i, \tag{2.16}$$

where B_i is ith element of $\mathbf{B}$, $\mathbf{\Gamma}_i$ is ith column of $\mathbf{\Gamma}$, $\mathbf{U}_i$ is ith column of $\mathbf{U}$. To assess the association between $\mathbf{X}$ and $\mathbf{Y}$, we aim to test d hypotheses of the form:

$$H_0 : B_i = 0, \text{ against } H_1 : B_i \neq 0, \, i = 1, \ldots, d.$$

The hypothesis is now on the coefficient B_i given the common latent factors $\mathbf{G}$. The global hypothesis testing is to answer the question that whether there is any relationship between the covariates $\mathbf{X}$ and the primary variable $\mathbf{Y}$. Suppose we have obtained d p-values for the individual hypotheses. We can then conduct the global test by combining p-values using MinP, Fisher's combination, or HC.

We present details of estimating the latent variables using SVA here. SVA first calculates the residual matrix of $\mathbf{X}$ regressing on $\mathbf{Y}$, denoted as $\mathbf{R}$. There are multiple methods to determine the number of latent variables k. The algorithm by Buja and Eyuboglu (1992) is used to determine the number of latent variables, k. It permutes each column independently and gets a null distribution of the matrix's singular values. The first k singular values that are larger than a certain cutoff of the null distribution of the singular values are selected. On the other hand, in SmartSVA, the random matrix theory (Marčenko & Pastur, 1967) is used. It compares the distribution of eigenvalues of a covariance matrix of iid standard normal. Again, it selects k singular values if they are larger than the cutoff of the null distribution of eigenvalues. The latter method does not require permutation and is faster than the first one, although, the latter method has more assumptions on the distribution of $\mathbf{X}$. We adopt the latter method to select the number of latent variables k.

In the first step of SVA, it uses the first k columns of the right decomposition matrix from the SVD of $\mathbf{R}$ as the estimate of $\mathbf{G}$. Let us denote it as $\hat{\mathbf{G}}_{(0)}$. An iterative procedure is conducted to improve the estimation of $\mathbf{G}$. At each iteration b, weights for SVD of $\mathbf{X}$ are calculated using empirical Bayes. SVA examines the hypothesis of $B_i \neq 0$ in the model $\mathbf{X}_i = B_i\mathbf{Y} + \mathbf{\Gamma}_i\hat{\mathbf{G}}_{(b)} + \mathbf{E}_i$ and the hypothesis of $\mathbf{\Gamma}_i \neq 0$ in $\mathbf{X}_i = \mathbf{\Gamma}_i\hat{\mathbf{G}}_{(b)} + \mathbf{E}_i$. A Bayesian probability is calculated as

$$P(B_i = 0|\mathbf{\Gamma}_i \neq 0, \mathbf{X}, \mathbf{Y}, \hat{\mathbf{G}}_{(b)}) = \pi_0 g_0(F_i)/g_1(F_i) \tag{2.17}$$

and similarly for

$$P(\mathbf{\Gamma}_i = 0|\mathbf{X}, \mathbf{Y}, \hat{\mathbf{G}}_{(b)}) \tag{2.18}$$

where π_0 is the prior proportion of the null hypothesis, F_i is the F-test statistic of the hypothesis that $B_i = 0$ in (2.17), g_0 is the uniform distribution density function as the distribution of p-values under the null hypothesis, and g_1 is a mixture of the density of the p-value under the null and alternative hypotheses. The mixture density g_1 can be estimated by the empirical density of the p-values of the data regardless of the null or alternative hypothesis. SVA uses

$$P(B_i = 0, \mathbf{\Gamma}_i \neq 0|\mathbf{X}, \mathbf{Y}, \hat{\mathbf{G}}_{(b)}) = P(B_i = 0|\mathbf{\Gamma}_i \neq 0, \mathbf{X}, \mathbf{Y}, \hat{\mathbf{G}}_{(b)})P(\mathbf{\Gamma}_i \neq 0|\mathbf{X}, \mathbf{Y}, \hat{\mathbf{G}}_{(b)})$$

as weights in SVD of $\mathbf{X}$ (Leek & Storey, 2008). Then we obtain an updated SVD of $\mathbf{X}$ and let us denote the first k columns of the updated SVD right decomposition matrix as $\hat{\mathbf{G}}_{(b+1)}$. The procedure iterates many times until differences in weights in two consecutive iterations are small and the algorithm converges to an estimate of $\mathbf{G}$.

The coefficient estimator of B_i conditioned on $\mathbf{G}$ is

$$\hat{B}_i^{(\mathbf{G})} = (\mathbf{Y}'(\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')\mathbf{Y})^{-1}(\mathbf{Y}'(\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')\mathbf{X}_i)$$

as the ordinary regression estimator of the effect of $\mathbf{Y}$ on $\mathbf{X}_i$ conditioned on $\mathbf{G}$. Our proposed test statistic is the coefficient estimator divided by its standard deviation, that is

$$t_i^{(\mathbf{G})} = \hat{B}_i^{(\mathbf{G})}/sd(\hat{B}_i^{(\mathbf{G})}) \tag{2.19}$$

where

$$sd(\hat{B}_i^{(\mathbf{G})}) = \hat{\sigma}/\sqrt{\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{G}(\mathbf{G}'\mathbf{G})\mathbf{G}'\mathbf{Y}} \tag{2.20}$$

as the estimated standard error of $sd(\hat{B}_i^{(\mathbf{G})})$ and

$$\hat{\sigma}^2 = \frac{\mathbf{X}_i'(\mathbf{I} - (\mathbf{Y},\mathbf{G})((\mathbf{Y},\mathbf{G})'(\mathbf{Y},\mathbf{G}))(\mathbf{Y},\mathbf{G})')\mathbf{X}_i}{n-(k+1)} \tag{2.21}$$

as the error mean square in $\mathbf{X}_i$ regressing on $\mathbf{Y}$ and $\mathbf{G}$. Let p_i denote the p-value of the test statistics, $i = 1, \ldots, d$. Then the MinP test statistic is $\min p_i$ or equivalently $\max |t_i|$. Its p-value is calculated as $1-(1-\min p_i)^d$, and we reject the null hypothesis if $1-(1-\min p_i)^d < \alpha$, where α is the significance level, usually 0.05.

2.3.2 Type I error under IRM

Recall the IRM in Formula (2.14),

$$\mathbf{X} = \mathbf{YB} + \mathbf{E}.$$

We can decompose $\mathbf{E}$ into a low rank dependent component and an independent component as shown in Formula (2.15) and also cited below

$$\mathbf{X} = \mathbf{YB} + \mathbf{G\Gamma} + \mathbf{U}$$

Note that the decomposition can be obtained for an arbitrary distribution for $\mathbf{E}$ and an arbitrary level of dependence across the columns of $\mathbf{E}$. More specifically, we have the following two propositions from Leek and Storey (2008).

Proposition 2.3.1 (Proposition 1 of Leek & Storey, 2008) *Under Model* (2.14)*, suppose for each* $\mathbf{E}_i$*, the* i*th column of* $\mathbf{E}$*, there is no Borel measurable function* g *such that* $\mathbf{E}_i = g(\mathbf{E}_1, \ldots, \mathbf{E}_{i-1}, \mathbf{E}_{i+1}, \ldots, \mathbf{E}_d)$ *almost surely. Then, there exist matrices* $\mathbf{\Gamma}$*,* $\mathbf{G}$ *such that Model* (2.15)

$$\mathbf{X} = \mathbf{YB} + \mathbf{G\Gamma} + \mathbf{U}$$

is valid, and the columns of $\mathbf{U}$ *are jointly independent,*

$$P(\mathbf{U}_1, \ldots, \mathbf{U}_d) = P(\mathbf{U}_1) \times \ldots \times P(\mathbf{U}_d) \tag{2.22}$$

Proposition 2.3.2 (Corollary 1 of Leek & Storey, 2008) *Under the assumption of Proposition 2.3.1, all population-level dependence of the variable set* $\mathbf{X}$ *is removed when conditioning on both* $\mathbf{Y}$ *and* $\mathbf{G}$*. That is, we have the two equations.*

$$P(\mathbf{X}_1, \ldots, \mathbf{X}_d|\mathbf{Y}, \mathbf{G}) = P(\mathbf{X}_1|\mathbf{Y}, \mathbf{G}) \times \ldots \times P(\mathbf{X}_d|\mathbf{Y}, \mathbf{G}) \tag{2.23}$$

$$P(\mathbf{U}_1, \ldots, \mathbf{U}_d|\mathbf{Y}, \mathbf{G}) = P(\mathbf{U}_1|\mathbf{Y}, \mathbf{G}) \times \ldots \times P(\mathbf{U}_d|\mathbf{Y}, \mathbf{G}) \tag{2.24}$$

Based on these two propositions, we obtain the independence of the marginal test statistics below.

Proposition 2.3.3 *Under Model* (2.15)*, the marginal regression coefficient estimators and the marginal test statistics are independent. That is,*

$$P(\hat{B}_1^{(\mathbf{G})}, \ldots, \hat{B}_d^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G}) = P(\hat{B}_1^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G}) \times \ldots \times P(\hat{B}_d^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G})$$

and

$$P(t_1^{(\mathbf{G})}, \ldots, t_d^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G}) = P(t_1^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G}) \times \ldots \times P(t_d^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G})$$

Proof Note that given $\mathbf{Y}$ and $\mathbf{G}$, $\hat{B}_i^{(\mathbf{G})}$ only depends on $\mathbf{U}_i$ and we denote it as $g(\mathbf{U}_i)$:

$$\begin{aligned}
\hat{B}_i^{(\mathbf{G})} =& (\mathbf{Y}'(\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')(\mathbf{Y}B_i + \mathbf{G}\boldsymbol{\Gamma}_i + \mathbf{U}_i) \\
=& (\mathbf{Y}'(\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')(\mathbf{Y}B_i + \mathbf{G}\boldsymbol{\Gamma}_i) \\
&+ (\mathbf{Y}'(\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}')\mathbf{U}_i \\
=& g(\mathbf{U}_i)
\end{aligned}$$

Therefore, $\hat{B}_i^{(\mathbf{G})}$'s are independent conditional on $\mathbf{Y}$ and $\mathbf{G}$.

$$\begin{aligned}
P(\hat{B}_1^{(\mathbf{G})}, \ldots, \hat{B}_d^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G}) =& P(g(\mathbf{U}_1), \ldots, g(\mathbf{U}_d)|\mathbf{Y}, \mathbf{G}) \\
=& P(g(\mathbf{U}_1)|\mathbf{Y}, \mathbf{G}) \times \ldots \times P(g(\mathbf{U}_d)|\mathbf{Y}, \mathbf{G}) \\
=& P(\hat{B}_1^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G}) \times \ldots \times P(\hat{B}_d^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G})
\end{aligned}$$

We know the adjusted test statistic is $t_i^{(\mathbf{G})} = \hat{B}_i^{(\mathbf{G})}/sd(\hat{B}_i^{(\mathbf{G})})$. The denominator $sd(\hat{B}_i^{(\mathbf{G})})$ is estimated from the standard linear regression. Since the only random part of $sd(\hat{B}_i^{(\mathbf{G})})$ is $\mathbf{U}_i$ and $\mathbf{U}_i$'s are independent of each other, we have

$$P(t_1^{(\mathbf{G})}, \ldots, t_d^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G}) = P(t_1^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G}) \times \ldots \times P(t_d^{(\mathbf{G})}|\mathbf{Y}, \mathbf{G})$$

■

The factor model as specified in (2.15) partitions $\mathbf{E}$ into dependent and independent components. Including the latent factors in the modeling removes the dependence of the marginal test statistics. We obtain the following theorem on the correct type I error control.

Theorem 2.3.4 *Under the null hypothesis and the assumption of* (2.15)*, the factor-adjusted marginal test statistics are independent and their p-values, p_i's, are iid and follow a uniform distribution. The factor-adjusted MinP test statistic follows a Beta distribution $Beta(1, d)$ and the factor-adjusted MinP has its type I error correctly controlled at any given α level.*

2.3.3 Factor modeling of dependence with a regression model (RM)

We consider the global testing problem from a different model, where the association of $\mathbf{X}$ and $\mathbf{Y}$ is modeled by regressing $\mathbf{Y}$ on $\mathbf{X}$. For notation convenience, let us assume $\mathbf{Y}$ is the response variable and $\mathbf{X}$ is the design matrix of covariates. Let $\mathbf{Y} \sim N(\mu(\mathbf{X}), \sigma^2\mathbf{I})$, where $\mu(\mathbf{X})$ is the conditional expectation of $\mathbf{Y}$ given $\mathbf{X}$. We assume each $\mathbf{X}_i$ is standardized and divided by $\sqrt{n}$ so that $\mathbf{X}_i'\mathbf{X}_i = 1$. Assume the variance of $\mathbf{Y}$, σ^2, is known and we use σ^2 to replace s_y in (2.2) to calculate the marginal test statistics

$$\mathbf{Z} = \mathbf{X}'\mathbf{Y}/\sigma \tag{2.25}$$

which are the scaled marginal correlations following the distribution

$$\mathbf{Z} \sim N(\mathbf{X}'\mu(\mathbf{X})/\sigma, \mathbf{X}'\mathbf{X}). \tag{2.26}$$

Note that the correlation matrix of the test statistics $\mathbf{Z}$ is $\mathbf{X}'\mathbf{X}$, which we denote as $\mathbf{\Sigma}$. In other words, the correlation structure of the design matrix $\mathbf{X}$ gives the correlation structure of the test statistics $\mathbf{Z}$. Denote $\boldsymbol{\theta} = E\mathbf{Z} = \mathbf{X}'\mu(\mathbf{X})/\sigma$, the mean vector of the marginal test statistics. We test the global hypotheses $H_0 : \boldsymbol{\theta} = 0$ versus $H_1 : \boldsymbol{\theta} \neq 0$. That is, all components of $\boldsymbol{\theta}$ are zero.

Per our discussion before, when there are high correlations among the test statistics $\mathbf{Z}$, existing global testing methods would not work. We propose to decompose the correlation matrix of the test statistics to weaken the correlation structure. Suppose the SVD of $\mathbf{X}$ is $\mathbf{X} = \mathbf{UDV'}$, where the matrix $\mathbf{U}$ is $n \times d$ with orthonormal columns, i.e., $\mathbf{U}'\mathbf{U} = \mathbf{I}$. The diagonal matrix $\mathbf{D}$ is $d \times d$ with the diagonals arranged in a decreasing order. The matrix $\mathbf{V}$ is $d \times d$ with orthonormal columns, i.e., $\mathbf{V}'\mathbf{V} = \mathbf{I}$. We partition $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{D}$ into two parts each, i.e., $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$, $\mathbf{D} = diag(\mathbf{D}_1, \mathbf{D}_2)$, $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2)$, where $\mathbf{U}_1(n \times k)$ is the matrix containing the first k eigenvectors, $\mathbf{D}_1(k \times k)$ contains the eigenvalues that correspond to $\mathbf{U}_1$, and the matrix $\mathbf{V}_1(d \times k)$ contains the right singular vectors that correspond to $\mathbf{U}_1$. Then we have the decomposition of the correlation matrix $\mathbf{\Sigma} = \mathbf{X}'\mathbf{X} = \mathbf{VD}^2\mathbf{V}' = \mathbf{V}_1\mathbf{D}_1^2\mathbf{V}_1' + \mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2'$.

By conducting SVD, we can partition the correlation matrix of the test statistics into two components, one explaining the correlation structure and the other being an independent or weakly dependent matrix. The matrix $\mathbf{U}_1$ can be interpreted as the latent factors that explain the major correlation structure. Following the idea of EIGENSTRAT, we regress both $\mathbf{X}$ and $\mathbf{Y}$ on $\mathbf{U}_1$ and then examine the association of the residuals after adjusting $\mathbf{U}_1$. The covariance of residuals of $\mathbf{X}$ and $\mathbf{Y}$ after regression on $\mathbf{U}_1$ is

$$\mathbf{X}'(\mathbf{I} - \mathbf{U}_1\mathbf{U}_1')\mathbf{Y} \sim N(\mathbf{V}_2\mathbf{D}_2^2\mathbf{U}_2'\mu(\mathbf{X}), \mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2'\sigma^2) \tag{2.27}$$

We standardize each element in the vector $\mathbf{X}'(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1')\mathbf{Y}$ so that each of them has a variance of 1. More specifically, the adjusted test statistic, denoted as $\tilde{\mathbf{t}}$, is:

$$\tilde{\mathbf{t}} = diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2')\mathbf{X}'(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1')\mathbf{Y}/\sigma \sim$$
$$N(diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2')\mathbf{V}_2\mathbf{D}_2^2\mathbf{U}_2'\mu(\mathbf{X})/\sigma, diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2')\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2' diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2')) \tag{2.28}$$

We use $\max \tilde{t}_i^2$ as the adjusted MinP test statistic.

Remark: An advantage of this adjustment method is that it can be applied even if we only have the information of the marginal test statistics and correlation matrix of $\mathbf{X}$, such as some SNP summary statistics and the LD information among SNPs, which are often publicly available, while access of individual-level data is limited. For a review of SNP summary data, see D. J. Liu et al. (2014).

2.3.4 Type I error under RM

Given the definition of the adjusted marginal test statistics in (2.28), their correlation matrix is

$$\tilde{\mathbf{A}} = diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2')\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2' diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2'). \tag{2.29}$$

Our main theorem in this section shows that when the eigenvalues of (2.29) are close to each other, the adjusted MinP test statistic, $\max \tilde{t}_i^2$, has the same null distribution as the standard MinP test statistic. Therefore, the type I error of the adjusted MinP test is correct. We need two lemmas to prove the main theorem. The proofs of the lemmas follow Y. Liu and Xie (2018c) but with relaxed conditions. Lemma 2.3.5 is about the comparison of two multivariate normal distributions with different covariance matrices. Suppose the difference of two correlation matrices $\tilde{\mathbf{A}}$ and $\mathbf{I}$ are controlled by a function of d. Then we obtain that the multivariate normal with the correlation matrix $\tilde{\mathbf{A}}$ converges to the standard multivariate normal. Lemma 2.3.6 provides the asymptotic distribution of d iid squares of standard normal random variables. Based

on these two lemmas, we conclude that $\max \tilde{t}_i^2$ has the same asymptotic distribution as those of the maximum of d iid squares of standard normal random variables. Let us introduce more notations. The Frobenius norm of any matrix $\mathbf{A}$ is $||\mathbf{A}||_F = \sqrt{tr(\mathbf{A}\mathbf{A}')}$, which is the square root of the sum of squares of each element in $\mathbf{A}$.

The following lemma compares the cumulative distribution functions of two multivariate normal distributions with different covariance matrices. Lemma 2.3.5 (i) comes from Theorem 4.2.1 of Leadbetter, Lindgren, and Rootzén (2012). Lemma 2.3.5 (ii) and (iii) are from Lemma 2 (ii) and (iii) of Y. Liu and Xie (2018c) with relaxed conditions.

Lemma 2.3.5 *Suppose $U_1, \ldots, U_d$ are standard normal variables with covariance matrix $\Gamma^1 = (\gamma_{ij}^1)$, and similarly $V_1, \ldots, V_d$ with covariance matrix $\Gamma^0 = (\gamma_{ij}^0)$ and let $b_{ij} = \max\{|\gamma_{ij}^1|, |\gamma_{ij}^0|\}$. Further, let $x_1, \ldots, x_d$ be real numbers. Then we have*

(i)

$$|P(\cap_{j=1}^d\{U_j \le x_j\}) - P(\cap_{j=1}^d\{V_j \le x_j\})| \\ \le (2\pi)^{-1} \sum_{1\le i\le j\le d} |\gamma_{ij}^1 - \gamma_{ij}^0|(1-b_{ij}^2)^{-1/2} \exp\left(-\frac{x_i^2+x_j^2}{2(1+b_{ij})}\right) \quad (2.30)$$

(ii) If $x_1, \ldots, x_d$ are positive then

$$|P(\cap_{j=1}^d\{|U_j| \le x_j\}) - P(\cap_{j=1}^d\{|V_j| \le x_j\})| \\ \le 4\cdot(2\pi)^{-1} \sum_{1\le i\le j\le d} |\gamma_{ij}^1 - \gamma_{ij}^0|(1-b_{ij}^2)^{-1/2} \exp\left(-\frac{x_i^2+x_j^2}{2(1+b_{ij})}\right) \quad (2.31)$$

(iii) Assume that $||\Gamma^1 - \Gamma^2||_F^2 = O(d)$. If $\max_{i\neq j} b_{ij} = c_0 < 1$ and $\min_{1\le i\le d} x_i \ge \sqrt{2a\log d}$ for some constant $a > (1+c_0)/2$, then

$$\lim_{d\to\infty} |P(\cap_{j=1}^d\{U_j \le x_j\}) - P(\cap_{j=1}^d\{V_j \le x_j\})| = 0 \quad (2.32)$$

and

$$\lim_{d\to\infty} |P(\cap_{j=1}^d\{|U_j| \le x_j\}) - P(\cap_{j=1}^d\{|V_j| \le x_j\})| = 0 \quad (2.33)$$

Proof (i) See Theorem 4.2.1 in Leadbetter et al. (2012).

(ii) This is true if Γ^1, Γ^2 are positive semidefinite.

Let $\epsilon = (\epsilon_1, \cdots, \epsilon_d)$, $\epsilon \sim N(0, \sigma^2 I)$. The covariance matrices of $U + \epsilon$ and $V + \epsilon$ are $\Gamma^1 + \sigma^2 I$ and $\Gamma^2 + \sigma^2 I$. Both are positive definite. From Lemma 2 (ii) in Y. Liu and Xie (2018a), we have

$$\begin{aligned}
&|P(\cap_{j=1}^d\{|U_j + \epsilon_j| \le x_j\}) - P(\cap_{j=1}^d\{|V_j + \epsilon_j| \le x_j\})| \\
&\le 4\cdot(2\pi)^{-1} \sum_{1\le i\le j\le d} |\gamma_{ij}^1 - \gamma_{ij}^0|(1-(b_{ij}+I\{i=j\}\sigma^2)^2)^{-1/2} \exp\left(-\frac{x_i^2 + x_j^2}{2(1 + b_{ij} + I\{i=j\}\sigma^2)}\right)
\end{aligned} \tag{2.34}$$

We want to prove $P(\cap_{j=1}^d\{|U_j + \epsilon_j| \le x_j\}) \to P(\cap_{j=1}^d\{|U_j| \le x_j\})$, which means convergence in distribution. Assume the cdf of U, $F_U(x)$ is continuous at $x = (\pm x_1, \cdots, \pm x_d)$. For any δ, there exists r, such that when $||\epsilon|| < r$,

$$|F_U(\alpha x) - F_U(\alpha x + \epsilon)| < \delta \tag{2.35}$$

where $\alpha = (\alpha_1, \cdots, \alpha_d)$, $\alpha_j = \pm 1$, and αx denotes entry wise product.

For $||\epsilon|| < r$, we can find σ^2 such that $\epsilon \sim N(0, \sigma^2 I)$ and

$$\int_{||\epsilon||>r} f_\epsilon(\epsilon) d\epsilon < \delta \tag{2.36}$$

$$\begin{aligned}
&P(\cap_{i=1}^d\{|U_i| < x_i\}) \\
=&P(U_1 < x_1, \cap_{j=2}^d |U_j| < x_j) - P(U_1 < -x_1, \cap_{j=2}^d |U_j| < x_j) \\
=&P(U_1 < x_1, U_2 < x_2, \cap_{j=3}^d |U_j| < x_j) - P(U_1 < x_1, U_2 < -x_2 \cap_{j=2}^d |U_j| < x_j) \\
&- P(U_i < -x_1, U_2 < x_2, \cap_{j=3}^d |U_j| < x_j) + P(U_i < -x_1, U_2 < -x_2, \cap_{j=2}^d |U_j| < x_j) \\
=&\cdots \\
=&\sum_{\alpha_j = \pm 1} (-1)^{\sum \alpha_j} F_U(\alpha x)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&|P(\cap_{i=1}^{d}\{|U_i| < x_i\}) - P(\cap_{i=1}^{d}\{|U_i + \epsilon_i| < x_i\})| \\
=&|P(\cap_{i=1}^{d}\{|U_i| < x_i\}) - \int P(\cap_{i=1}^{d}\{|U_i + \epsilon_i| < x_i\}|\epsilon) f_\epsilon(\epsilon) d\epsilon| \\
=&|\int (\sum_{\alpha_j = \pm 1} (-1)^{\sum \alpha_j} (F_U(\alpha x) - F_U(\alpha x - \epsilon))) f_\epsilon(\epsilon) d\epsilon| \\
<&|\int_{||\epsilon|| < r} 2^d \delta f_\epsilon(\epsilon) d\epsilon| + |\int_{||\epsilon|| > r} 2^d f_\epsilon(\epsilon) d\epsilon| \\
<&\, 2^{d+1}\delta
\end{aligned}$$

Let $\sigma^2 \to 0$, we have

$$\begin{aligned}
&|P(\cap_{j=1}^{d}\{|U_j| \le x_j\}) - P(\cap_{j=1}^{d}\{|V_j| \le x_j\})| \\
&\qquad \le 4 \cdot (2\pi)^{-1} \sum_{1 \le i \le j \le d} |\gamma_{ij}^1 - \gamma_{ij}^0| (1 - b_{ij}^2)^{-1/2} \exp\left(-\frac{x_i^2 + x_j^2}{2(1 + b_{ij})}\right) \quad (2.37)
\end{aligned}$$

(iii) we would like to prove

$$\sum_{1 \le i \le j \le d} |\gamma_{ij}^1 - \gamma_{ij}^0| (1 - b_{ij}^2)^{-1/2} \exp\left(-\frac{x_i^2 + x_j^2}{2(1 + b_{ij})}\right) \to 0 \quad (2.38)$$

$$\begin{aligned}
&\sum_{1 \le i \le j \le d} |\gamma_{ij}^1 - \gamma_{ij}^0| (1 - b_{ij}^2)^{-1/2} \exp\left(-\frac{x_i^2 + x_j^2}{2(1 + b_{ij})}\right) \\
\le&(1 - c_0)^{-1/2} \sum_{1 \le i \le j \le d} |\gamma_{ij}^1 - \gamma_{ij}^0| \exp\left(-\frac{2a \log d}{1 + b_{ij}}\right)
\end{aligned}$$

We can find the number of entries larger than any given constant c_1, say $1/4$, is at most $O(d)$, since $||\Gamma^1 - \Gamma^0||_F^2 = O(d)$. Let the number of entries larger than c_1 is m, $mc_1^2 < O(d)$, $m = O(d)$. Suppose S is a set of (i, j) that $|\gamma_{ij}^1 - \gamma_{ij}^0| < 1/4$. Using the Cauchy-Schwarz inequality, $\sum |\gamma_{ij}^1 - \gamma_{ij}^0| \le d||\Gamma^1 - \Gamma^0||_F$ and $||\Gamma^1 - \Gamma^0||_F^2 = O(d)$, we derive $\sum |\gamma_{ij}^1 - \gamma_{ij}^0| = O(d^{3/2})$.

$$
\begin{aligned}
&\sum_{1\le i\le j\le d} |\gamma_{ij}^1 - \gamma_{ij}^0| \exp\left(-\frac{2a\log d}{1+b_{ij}}\right) \\
=&\sum_{i\le j,(i,j)\in S} |\gamma_{ij}^1 - \gamma_{ij}^0| \exp\left(-\frac{2a\log d}{1+b_{ij}}\right) + \sum_{i\le j,(i,j)\notin S} |\gamma_{ij}^1 - \gamma_{ij}^0| \exp\left(-\frac{2a\log d}{1+b_{ij}}\right) \\
&\le O(d^{3/2}) \exp\left(-\frac{2a\log d}{1+1/4}\right) + O(d) \exp\left(-\frac{2a\log d}{1+c_0}\right) \\
=&O(d^{3/2}) d^{-8/5} + O(d) d^{-2a/(1+c_0)}
\end{aligned}
$$

Since $2a/(1+c_0) > 1$, the above formula converges to 0. ■

Lemma 2.3.6 *Suppose $(U_1, \ldots, U_d)^T$ follows a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}$, where diagonal elements $\sigma_{ii} = 1$ for $1 \le i \le d$. Assume that the eigenvalues of $\mathbf{\Sigma}$ satisfy $\lambda_1 / \sum_{i=1}^d \lambda_i = O(d^{-1})$ and $\max_{1\le i<j\le d} |\sigma_{ij}| \le c_0 < 1$. Then for any $x \in \mathbb{R}$*

$$
P(\max_{1\le i\le d} U_i^2 - 2\log d + \log\log d \le x) \to \exp(-e^{x/2}/\sqrt{\pi}) \tag{2.39}
$$

Proof Let $V_1, \ldots, V_d$ be iid standard normal random variables. From Lemma 2.3.5 (iii) we know that

$$
|P(\max |U_i| \ge \sqrt{x_d}) - P(\max |V_i| \ge \sqrt{x_d})| = o(1). \tag{2.40}
$$

The rest of the proof follows Y. Liu and Xie (2018a). The idea is that since $\max |U_i|$ converges to $\max |V_i|$ in distribution, and $\max |V_i|$ has the extreme value distribution 2.39, we conclude that $\max |U_i|$ also has the same extreme value distribution. ■

Theorem 2.3.7 $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Sigma}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{D}$ *are defined above. Suppose $\mathbf{\Sigma}$ has eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$. The number of latent variables is k. Choose the number k such that $\lambda_{k+1} / \sum_{i=k+1}^d \lambda_i = O(d^{-1})$ and $\min(diag(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2' d / tr(\mathbf{D}_2^2))) > c$ as $d \to \infty$. Then $\max \tilde{\mathbf{t}}_i^2$ defined at* (2.28) *converges to the distribution* (2.39).

Remark: The $\lambda_{k+1}/\sum_{i=k+1}^{d}\lambda_i = O(d^{-1})$ condition also implies $k = O(d)$. It is a much relaxed condition and we even do not require $\lambda_{k+1}/\sum_{i=k+1}^{d}\lambda_i$ converges to 1. In particular, if $\lambda_1 = O(1)$, then $k = 0$. Not all types of correlation matrices have that result. The counterexample is that, if the eigenvalues are aq^i and $q < 1$, then $aq/\sum_{i=1}^{\infty} aq^i = q/(1-q)$ and no k can satisfy the condition. The condition of $\min(diag(A)) > c$ means the variances after adjustment cannot be too close to 0.

Proof Suppose the diagonals of $\mathbf{D}$ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. The number of latent variables is k. We know the adjusted correlation matrix is $\tilde{\mathbf{A}}$ defined in (2.29) We introduce $\mathbf{A}$ to connect $\tilde{\mathbf{A}}$ and $\mathbf{I}$. Let

$$\mathbf{A} = \mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2' d/tr(\mathbf{D}_2^2) \tag{2.41}$$

and the eigenvalues of $\mathbf{A}$ are $\lambda_{k+1}d/\sum_{i=k+1}^{d}\lambda_i \geq \cdots \geq \lambda_d d/\sum_{i=k+1}^{d}\lambda_i$.

We have the following inequalities. $||\tilde{\mathbf{A}} - I||_F \leq ||\tilde{\mathbf{A}}||_F + ||I||_F \leq ||\tilde{\mathbf{A}}||_F + O(d^{1/2})$.

$$||\tilde{\mathbf{A}}||_F = ||diag(\mathbf{A})^{-1/2}\mathbf{A}diag(\mathbf{A})^{-1/2}||_F \tag{2.42}$$

$$\leq c^{-1}||\mathbf{A}||_F = c^{-1}d\sqrt{tr(\mathbf{D}_2^4)}/tr(\mathbf{D}_2^2) \leq c^{-1}\lambda_{k+1}d\sqrt{d-k}/tr(\mathbf{D}_2^2) = O(d^{1/2}) \tag{2.43}$$

Applying Lemma 2.3.5 and Lemma 2.3.6, we obtain the limit distribution of the adjusted test statistics. More specifically, let $\tilde{t}_i$ to be U_i in Lemma 2.3.6. We find $\max \tilde{t}_i^2$ has the limit distribution defined in formula 2.39. ■

2.3.5 Type I error simulation studies

In the simulation studies, we simulate variables $\mathbf{X}$ under different correlation settings. The number of variables is $d = 50$, and the number of observations is $n = 500$. We generate $\mathbf{X}$ from multivariate normal. We vary the correlation matrix of X to assess the type I error of our proposed adjusted MinP including EIGENSTRAT adjusted MinP (Eig MinP for short), and SVA adjusted MinP (SVA MinP for short), where EIGENSTRAT and SVA are two methods of estimating the latent factors, and

they correspond to data modeling of RM and IRM, respectively. We compare the performance with other methods including the original MinP (MinP for short; Tippett, 1931), HC (Donoho & Jin, 2008), correlated MinP (GMinP for short; R. Sun & Lin, 2017), correlated HC (HC-corr for short; R. Sun & Lin, 2017), GHC (generalized HC; Barnett et al., 2017), PFA (principal factor approximation; Fan et al., 2012), SKAT (Wu et al., 2011) and the F-test (F for short). Correlated MinP, correlated HC are methods proposed by R. Sun and Lin (2017) to conduct MinP and HC under arbitrary correlation structures. GHC is like HC, but its denominator uses variance estimate of $S(t)$ under dependence (Barnett et al., 2017). We use the R package named GBJ (R. Sun & Lin, 2017) to implement these methods.

We consider different types of correlation structures, including exponential off-diagonal decay, equal correlation, polynomial off-diagonal decay, and banded correlation matrices. We denote ρ as the correlation parameter. In an exponential off-diagonal decay correlation matrix, the element of the ith row and the jth column is $\rho^{|i-j|}$, where ρ can be $0.1, \ldots, 0.9$. In an equal correlation matrix, the element of the ith row and the jth column is ρ for $i \neq j$, where ρ can be $0.1, \ldots, 0.9$. In a polynomial correlation matrix, the element of the ith row and the jth column is $\rho/(1+|i-j|)^{1.5}$ for $i \neq j$, where ρ can be $0.1, \ldots, 1.4$. In a banded correlation matrix, the element of the ith row and the jth column is $(1-0.4\sqrt{|i-j|})\rho$ if $1 \leq |i-j| \leq 6$ and 0 otherwise, where ρ takes values $\rho = 0.2, 0.3, \ldots, 1.2$.

For simulation of type I error, we generate $\mathbf{Y}$ as n iid normal random variables with $N(0,1)$. We compare type I error of different methods at the significant level $\alpha = 0.05$, where the correct type I error control should be at the same value of 0.05. We simulate $\mathbf{X}$ and $\mathbf{Y}$ 1000 times and show the empirical type I error, which is the proportion of the test statistic larger than the cutoff from the critical value of the null distribution. For latent variable adjusted MinP and original MinP, we use the cutoff under the assumption that marginal p-values are independent.

Table 2.1 and 2.2 show the type I error of tests under exponential decay and equal correlation. In exponential decay, we find SVA MinP has type I error larger than 0.06

Table 2.1.

Type I error of tests under exponential decay. Rows represent correlation parameters. Eig MinP: EIGENSTRAT adjusted MinP, SVA MinP: SVA adjusted MinP: original MinP, GMinP: MinP under correlation by GBJ package, HC-corr: HC under correlation by GBJ package, F: F-test, PFA: method from Fan et al. (2012). Cells with larger than 0.06 are highlighted.

type I, exponential									
	Eig MinP	SVA MinP	GMinP	MinP	HC corr	GHC	F	PFA	SKAT
0.1	0.042	0.047	0.044	0.047	0.039	0.039	0.048	0.199	0.038
0.2	0.056	0.047	0.047	0.049	0.043	0.043	0.052	0.483	0.051
0.3	0.066	0.057	0.039	0.04	0.037	0.037	0.05	0.479	0.044
0.4	0.054	0.052	0.028	0.029	0.026	0.028	0.038	0.437	0.033
0.5	0.057	0.04	0.04	0.037	0.04	0.042	0.058	0.412	0.046
0.6	0.043	0.064	0.052	0.05	0.053	0.055	0.05	0.323	0.039
0.7	0.044	0.066	0.052	0.044	0.05	0.049	0.055	0.297	0.051
0.8	0.044	0.061	0.052	0.036	0.054	0.055	0.049	0.225	0.041
0.9	0.053	0.064	0.051	0.028	0.046	0.048	0.056	0.182	0.046

Table 2.2.

Type I error of tests under equal correlation. Descriptions see Table 2.1

type I equal correlation									
	Eig MinP	SVA MinP	GMinP	MinP	HC corr	GHC	F	PFA	SKAT
0.1	0.051	0.049	0.037	0.039	0.048	0.046	0.049	0.049	0.036
0.2	0.051	0.045	0.045	0.044	0.034	0.04	0.055	0.071	0.04
0.3	0.051	0.053	0.055	0.048	0.031	0.037	0.043	0.059	0.04
0.4	0.045	0.054	0.049	0.038	0.019	0.024	0.045	0.065	0.045
0.5	0.042	0.065	0.068	0.043	0.026	0.031	0.054	0.057	0.061
0.6	0.056	0.082	0.075	0.038	0.018	0.017	0.057	0.062	0.039
0.7	0.043	0.059	0.054	0.021	0.015	0.014	0.059	0.067	0.039
0.8	0.058	0.055	0.061	0.008	0.02	0.02	0.047	0.06	0.052
0.9	0.043	0.036	0.046	0.009	0.026	0.026	0.054	0.06	0.044

Table 2.3.
Type I error of tests under polynomial correlation. See Table 2.1.

Type I polynomial									
	Eig MinP	SVA MinP	GMinP	MinP	HC corr	GHC	F	PFA	SKAT
0.1	0.057	0.051	0.048	0.051	0.048	0.048	0.046	0.068	0.049
0.2	0.056	0.058	0.057	0.06	0.053	0.053	0.039	0.141	0.048
0.3	0.048	0.043	0.041	0.045	0.039	0.038	0.047	0.229	0.044
0.4	0.05	0.05	0.039	0.04	0.038	0.038	0.054	0.249	0.045
0.5	0.056	0.06	0.049	0.052	0.051	0.052	0.051	0.294	0.042
0.6	0.056	0.04	0.041	0.042	0.044	0.044	0.047	0.3	0.048
0.7	0.059	0.055	0.043	0.045	0.04	0.04	0.056	0.286	0.04
0.8	0.051	0.05	0.039	0.042	0.043	0.043	0.045	0.309	0.05
0.9	0.048	0.05	0.045	0.048	0.051	0.05	0.06	0.325	0.054
1	0.06	0.061	0.04	0.04	0.042	0.042	0.046	0.32	0.045
1.1	0.051	0.039	0.034	0.034	0.031	0.031	0.054	0.315	0.034
1.2	0.056	0.063	0.047	0.047	0.045	0.046	0.052	0.303	0.045
1.3	0.048	0.055	0.041	0.04	0.042	0.045	0.05	0.353	0.037
1.4	0.065	0.06	0.046	0.046	0.049	0.047	0.036	0.349	0.047

Table 2.4.
Type I error of tests under banded correlation. See Table 2.1.

Type I Banded									
	Eig MinP	SVA MinP	GMinP	MinP	HC corr	GHC	F	PFA	SKAT
0.1	0.053	0.048	0.045	0.05	0.041	0.04	0.057	0.127	0.04
0.2	0.045	0.052	0.052	0.054	0.045	0.045	0.055	0.233	0.049
0.3	0.049	0.049	0.049	0.05	0.04	0.041	0.053	0.244	0.043
0.4	0.047	0.037	0.025	0.025	0.021	0.021	0.03	0.268	0.045
0.5	0.054	0.069	0.068	0.068	0.059	0.061	0.057	0.265	0.039
0.6	0.046	0.057	0.034	0.035	0.037	0.036	0.047	0.295	0.053
0.7	0.05	0.052	0.053	0.052	0.046	0.047	0.049	0.274	0.051
0.8	0.037	0.051	0.047	0.046	0.042	0.042	0.052	0.259	0.058
0.9	0.039	0.051	0.047	0.04	0.048	0.047	0.043	0.298	0.045
1	0.044	0.052	0.042	0.039	0.046	0.041	0.044	0.283	0.046
1.1	0.041	0.052	0.035	0.03	0.044	0.041	0.046	0.295	0.05
1.2	0.057	0.087	0.046	0.039	0.056	0.054	0.051	0.313	0.054

when $\rho \geq 0.6$ and PFA has type I error much larger than 0.06. PFA estimates too many latent variables when there is no significant latent variable, which is typically the case for exponential decay correlation structures. All other tests have reasonable type I errors except that MinP has a conservative type I error of 0.03 when $\rho \geq 0.8$. In the equal correlation setting, GMinP and SVA MinP have type I error above 0.07 when $\rho = 0.5, 0.6$. The type I error of MinP is less than 0.04 when $\rho \geq 0.6$, which confirms the result derived in Section 2.2.1. The type I error of MinP is much lower under high correlations. HC and GHC are also too conservative under equal correlation.

Table 2.3 and 2.4 show the type I error of tests under polynomial and banded correlation. These two types of correlation matrices are weak, and MinP would work. We would like to see if the factor-adjusted models have too many latent variables. PFA does, and its type I error is much larger than 0.06. SVA MinP and GMinP can be a little sensitive for some correlation parameters. MinP is conservative for banded correlation when $\rho \geq 0.9$.

Overall, Eig MinP, F-test and SKAT would work under arbitrary correlation. SKAT's type I error is controlled because the null distribution of the test statistic under covariance dependence is approximated well. PFA tends to estimate too many latent variables, and its type I error is much larger than 0.06 if there is no significant latent variable, which means there is no large drop in the sorted eigenvalues. SVA MinP can be slightly sensitive.

2.4 Power theory of the factor-adjusted global test statistic

The proposed test not only has the correct type I error, but it can be more powerful than the original MinP. The following simple example shows that removing latent variables can improve power, especially for highly correlated marginal tests. Suppose we have a set of variables $X_i = \sqrt{\rho}W + \sqrt{1-\rho}W_i$, $i = 1, \ldots, d$. Random variable $Y = b_1 W_1 + E$, where W_i, W, E follow iid $N(0, 1)$. Only the correlation between X_1

and Y is non-zero, and other correlations are zero. This setting corresponds to the setting of sparse marginal correlation. The correlation between X_1 and Y is:

$$cor(X_1, Y) = \frac{b_1\sqrt{1-\rho}}{\sqrt{b_1^2+1}} \tag{2.44}$$

We consider W as the common latent factor among X_i, $i = 1, \ldots, d$. If we have a good idea about W, we can remove the latent factor effect and consider the factor-adjusted variables, namely, $\tilde{X}_i = \sqrt{1-\rho}W_i$, $\tilde{Y} = Y$. The correlation between $\tilde{X}_1$ and $\tilde{Y}$ is:

$$cor(\tilde{X}_1, \tilde{Y}) = \frac{b_1}{\sqrt{b_1^2+1}} \tag{2.45}$$

The association between X_i and Y is enhanced after adjusting the latent factor. When we calculate a global test statistic, e.g., using MinP, the factor-adjusted test is more powerful than the original test. When ρ is larger, the improvement of power is larger.

On the other hand, we also have examples that removing latent factors reduces power. Let random variables $X_i = \sqrt{0.9}W + \sqrt{0.1}K_i, i = 1, \ldots, d$, and the true relationship of $\mathbf{X} = (X_1, \ldots, X_d)$ and Y is

$$Y = X_1 + \epsilon = \sqrt{0.9}W + \sqrt{0.1}K_1 + \epsilon$$

where W, $K_i, i = 1, \ldots, d$, and ϵ follows iid standard normal. The correlation between Y and X_1 is

$$cor(Y, X_1) = 1/\sqrt{2} = 0.707$$

Again we consider W as the common latent factor among X_i, $i = 1, \ldots, d$. If we have a good idea about W, we can remove the latent factor effect and consider the factor-adjusted variables, namely, the residuals of X and Y after regressing on W. The correlation of the factor-adjusted variables is

$$cor(\tilde{Y}, \tilde{X}_1) = 0.1/\sqrt{1.1 * 0.1} = 0.3015$$

As the calculation shows, after removing the latent factor, the correlation becomes smaller. Therefore, the power of latent variable adjustment method depends on the underlying model.

2.4.1 Power theory under IRM

We now study the power of the proposed test statistic

$$t_i^{(\mathbf{G})} = \hat{B}_i^{(\mathbf{G})}/sd(\hat{B}_i^{(\mathbf{G})})$$

under factor modeling through the inverse regression model (2.14). Without the latent factor adjustment, the standard regression coefficient estimator from the inverse regression model is $\hat{B}_i = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}_i$. The original test statistic is $\hat{B}_i$ divided by its standard deviation. Conditional on the latent factor $\mathbf{G}$, the adjusted test statistic is based on $\hat{B}_i^{(\mathbf{G})} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{X}_i - \mathbf{G}_i\hat{\mathbf{\Gamma}}_i)$, where $\hat{\mathbf{\Gamma}}_i$ is the coefficient estimator of $\mathbf{G}$. The following proposition shows that the latent factor adjusted test has potential to improve power. Denote the variance of each element of $\mathbf{G}$ as σ_g^2 and that of $\mathbf{U}$ as σ_u^2

Lemma 2.4.1 *For the unadjusted coefficient $\hat{B}_i$, its mean is $E(\hat{B}_i) = B_i$ and its variance is $(\mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{\Gamma}_i'\mathbf{\Gamma}_i\sigma_g^2 + \sigma_u^2)$. For latent-factor-adjusted coefficient $\hat{B}_i^{(\mathbf{G})}$, its mean $E(\hat{B}_i^{(\mathbf{G})}) = B_i$ and its variance is $var(\hat{B}_i^{(\mathbf{G})}) = (\mathbf{Y}'\mathbf{Y})^{-1}\sigma_u^2(1 + k/(n-k-2))$.*

Proof The least square estimate of B_i from the original model is $\hat{B}_i = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}_i$. Its variance is

$$\begin{aligned} var(\hat{B}_i) &= var((\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{G}\mathbf{\Gamma}_i) + var((\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{U}_i) \\ &= (\mathbf{Y}'\mathbf{Y})^{-2}(var(\mathbf{Y}'\mathbf{G}\mathbf{\Gamma}_i) + var(\mathbf{Y}'\mathbf{U}_i)) \\ &= (\mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{\Gamma}_i'\mathbf{\Gamma}_i\sigma_g^2 + \sigma_u^2) \end{aligned}$$

The estimate conditioned on G can be written as: $\hat{B}_i^{(\mathbf{G})} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{X}_i - \mathbf{G}\hat{\mathbf{\Gamma}}_i)$. $\hat{B}_i^{(\mathbf{G})}$ is from the solution of this normal equation

$$\begin{bmatrix} \mathbf{Y}'\mathbf{Y} & \mathbf{Y}'\mathbf{G} \\ \mathbf{G}'\mathbf{Y} & \mathbf{G}'\mathbf{G} \end{bmatrix} \begin{bmatrix} \hat{B}_i^{(\mathbf{G})} \\ \hat{\mathbf{\Gamma}}_i \end{bmatrix} = \begin{bmatrix} \mathbf{Y}'\mathbf{X}_i \\ \mathbf{G}'\mathbf{X}_i \end{bmatrix} \tag{2.46}$$

It is obtained by multiplying $(\mathbf{Y}'\mathbf{Y})^{-1}$ at the first equation of Equation (2.46).

We define the test statistic as the coefficient divided by its standard deviation. The result of an inverse of a block matrix is

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix} \tag{2.47}$$

Here we let $\mathbf{A} = \mathbf{Y}'\mathbf{Y}$, $\mathbf{B} = \mathbf{Y}'\mathbf{G}$, $\mathbf{C} = \mathbf{G}'\mathbf{Y}$, $\mathbf{D} = \mathbf{G}'\mathbf{G}$. The coefficient estimator of $\mathbf{G}$ is $\hat{\mathbf{\Gamma}}_i$.

$$\hat{\mathbf{\Gamma}}_i = (\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G})^{-1}\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{X}_i \tag{2.48}$$

Then $\hat{\mathbf{\Gamma}}_i$ and $\hat{B}_i^{(\mathbf{G})}$ are unbiased estimates.

$$E(\hat{\mathbf{\Gamma}}_i|\mathbf{G}) = \mathbf{\Gamma}_i \tag{2.49}$$

$$E(\hat{B}_i^{(\mathbf{G})}|\mathbf{G}) = E((\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{X}_i - \mathbf{G}\mathbf{\Gamma}_i')|\mathbf{G}) = E((\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{Y}B_i + \mathbf{U}_i')|\mathbf{G}) = B_i \tag{2.50}$$

And

$$E(\hat{B}_i^{(\mathbf{G})}) = B_i \tag{2.51}$$

Its variance is

$$\begin{aligned} & var(\hat{B}_i^{(\mathbf{G})}) \\ =& E(var(\hat{B}_i^{(\mathbf{G})}|\mathbf{G})) + var(E(\hat{B}_i^{(\mathbf{G})}|\mathbf{G})) \\ =& E(var(\hat{B}_i^{(\mathbf{G})}|\mathbf{G})) \\ =& E(var((\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{Y}B_i + G\mathbf{\Gamma}_i' + \mathbf{U}_i - G\hat{\mathbf{\Gamma}}_i')|\mathbf{G})) \\ =& (\mathbf{Y}'\mathbf{Y})^{-2}E(var(\mathbf{Y}'(G\mathbf{\Gamma}_i' - G\hat{\mathbf{\Gamma}}_i')|\mathbf{G})) + (\mathbf{Y}'\mathbf{Y})^{-1}\sigma_u^2 \end{aligned}$$

$$\begin{aligned} & var(\hat{\mathbf{\Gamma}}_i|\mathbf{G}) \\ =& var((\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G})^{-1}\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{U}_i|\mathbf{G}) \\ =& (\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G})^{-1}\sigma_u^2 \end{aligned}$$

$$\begin{aligned} & (\mathbf{Y}'\mathbf{Y})^{-2}E(var(\mathbf{Y}'(G\mathbf{\Gamma}_i - G\hat{\mathbf{\Gamma}}_i)|\mathbf{G})) \\ =& (\mathbf{Y}'\mathbf{Y})^{-2}\mathbf{Y}'E(\mathbf{G}var(\hat{\mathbf{\Gamma}}_i|\mathbf{G})\mathbf{G}')\mathbf{Y} \\ =& (\mathbf{Y}'\mathbf{Y})^{-2}\sigma_u^2\mathbf{Y}'E(\mathbf{G}(\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G})^{-1}\mathbf{G}')\mathbf{Y} \end{aligned}$$

Now we study the distribution of $\mathbf{Y}'\mathbf{G}(\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G})^{-1}\mathbf{G}'\mathbf{Y}$. Note that $\mathbf{G}'\mathbf{Y} \sim N(0, \sigma_g^2(\mathbf{Y}'\mathbf{Y})I)$. Therefore, $\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G}$ follows a Wishart distribution. Its mean is

$$E(\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G}) = n\sigma_g^2 I - (\mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{Y}'\mathbf{Y})\sigma_g^2 I = (n-1)\sigma_g^2 I \tag{2.52}$$

Because $\mathbf{G}'\mathbf{Y}(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G} = 0$, $\mathbf{G}'\mathbf{Y}$ and $(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G}$ are independent. We have $\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G} = \mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G}$, which is a sum of squares of $(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G}$. Then $\mathbf{G}'\mathbf{Y}$ and $\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G}$ are independent.

So by the definition of Hotelling's T^2 distribution,

$$(n-1)(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{G}\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G})^{-1}\mathbf{G}'\mathbf{Y} \sim T^2(k, n-1) \tag{2.53}$$

follows Hotelling's T^2 distribution. We can use the mean of F distribution to calculate the mean of it.

$$\begin{aligned} &E((n-1)(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{G}\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G})^{-1}\mathbf{G}'\mathbf{Y}) \\ =&k(n-1)/(n-k-2) \end{aligned}$$

$$\begin{aligned} &E(\mathbf{Y}'\mathbf{G}\mathbf{G}'(\mathbf{I}-\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}')\mathbf{G})^{-1}\mathbf{G}'\mathbf{Y}) \\ =&(\mathbf{Y}'\mathbf{Y})^{-1}k/(n-k-2) \end{aligned}$$

So

$$var(\hat{B}_i^{(\mathbf{G})}) = (\mathbf{Y}'\mathbf{Y})^{-1}\sigma_u^2(1+k/(n-k-2)) \tag{2.54}$$

■

The means of $\hat{B}_i$ and $\hat{B}_i^{(\mathbf{G})}$ are the same, but the variance of $\hat{B}_i^{(\mathbf{G})}$ is smaller than that of $\hat{B}_i$. The original test statistic is the coefficient divided by its standard deviation. That is

$$t_i = \hat{B}_i / \sqrt{(\mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{\Gamma}_i'\mathbf{\Gamma}_i\sigma_g^2 + \sigma_u^2)} \tag{2.55}$$

On the other hand, the factor-adjusted test statistic is defined as:

$$t_i^{(\mathbf{G})} = \hat{B}_i^{(\mathbf{G})} / \sqrt{(\mathbf{Y}'\mathbf{Y})^{-1}\sigma_u^2(n-2)/(n-k-2)} \tag{2.56}$$

Proposition 2.4.2 *Assume* $\mathbf{\Gamma}_i'\mathbf{\Gamma}_i\sigma_g^2 > k/(n-k-2)\sigma_u^2$. *The mean of the adjusted test statistic* $t_i^{(\mathbf{G})}$ *is larger than the mean of the original test statistic* t_i.

Proof $E\hat{B}_i = E\hat{B}_i^{(\mathbf{G})} = B_i$. The variance of $\hat{B}_i^{(\mathbf{G})}$ is smaller than $var(\hat{B}_i)$ if the following condition is satisfied:

$$(\mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{\Gamma}_i'\mathbf{\Gamma}_i\sigma_g^2) > (\mathbf{Y}'\mathbf{Y})^{-1}k/(n-k-2)\sigma_u^2$$
$$\mathbf{\Gamma}_i'\mathbf{\Gamma}_i\sigma_g^2 > k/(n-k-2)\sigma_u^2$$

When $\mathbf{\Gamma}_i$, k, σ_g^2, σ_u^2 are fixed, and n is sufficiently large, this condition is satisfied, because the length of $\mathbf{\Gamma}_i$ is k, and $\mathbf{\Gamma}_i'\mathbf{\Gamma}_i$ does not increase as n increases. We conclude that

$$|B_i|/\sqrt{(\mathbf{Y}'\mathbf{Y})^{-1}\sigma_u^2(n-2)/(n-k-2)} < |B_i|/\sqrt{(\mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{\Gamma}_i'\mathbf{\Gamma}_i\sigma_g^2+\sigma_u^2)} \qquad (2.57)$$

That is, $|E(t_i^{(\mathbf{G})})| > |E(t_i)|$ ■

The adjusted test statistic has a larger mean value than the original test statistic, while both have asymptotic normal distributions. The adjusted MinP is based on the maximum of the test statistics with larger mean than the original MinP. Therefore, the power of the factor-adjusted test can be improved.

Remark: This theorem is obtained when the latent factors $\mathbf{G}$ are given. Besides the methods of SVA and Eigenstrat, there are many other approaches to estimate $\mathbf{G}$ in practice. There exist matrix factorization methods other than SVD or spectral decomposition, for example, independent component analysis (ICA; Comon, 1994) and non-negative matrix factorization (NMF; D. D. Lee & Seung, 1999). ICA has been used to model non-linear dependence, while NMF has been applied to non-negative matrices. For estimating latent variables in a genotype matrix taking values of 0, 1, 2, Song, Hao, and Storey (2015) proposed to use logistic factor analysis.

Under the regression model of $\mathbf{Y}$ on $\mathbf{X}$, we can obtain a similar result when comparing the adjusted test statistics versus the original unadjusted test statistics.

The original unadjusted test statistics are $\mathbf{t} = \mathbf{X}'\mathbf{Y}/\sigma$. The adjusted test statistics are $\tilde{\mathbf{t}} = diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2')\mathbf{X}'(\mathbf{I} - \mathbf{U}_1\mathbf{U}_1')\mathbf{Y}/\sigma$.

Proposition 2.4.3 *If $E\mathbf{Y} = \mu(\mathbf{X})$ is orthogonal to $\mathbf{U}_1$, that is $\mathbf{U}_1'\mu(\mathbf{X}) = 0$, the adjusted test statistics have larger absolute means.*

Proof The mean of adjusted test statistics $\tilde{\mathbf{t}}$ is

$$diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2')\mathbf{X}'(\mathbf{I} - \mathbf{U}_1\mathbf{U}_1')\mu(\mathbf{X})/\sigma = diag^{-1/2}(\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2')\mathbf{X}'\mu(\mathbf{X})/\sigma \quad (2.58)$$

Because $\mathbf{\Sigma} = \mathbf{V}_1\mathbf{D}_1^2\mathbf{V}_1' + \mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2'$ all diagonals of $\mathbf{\Sigma}$ are 1 and diagonals of $\mathbf{V}_1\mathbf{D}_1^2\mathbf{V}_1'$ are positive, we have the diagonals of $\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2'$ less than 1. Compared to $E\mathbf{Z} = E\mathbf{X}'\mathbf{Y}/\sigma = \mathbf{X}'\mu(\mathbf{X})/\sigma$, the mean of the unadjusted marginal tests, the adjusted one has larger means by multiplying a diagonal matrix with elements larger than 1. ■

2.4.2 Power simulation studies

For the simulation of power, the correlation structure of $\mathbf{X}$ is the same as in the simulation of type I error. We consider the signal strength $d_0 = (1-\sqrt{\gamma})\sqrt{4\log d}$. We consider sparsity levels $\gamma = 1/4$ and $1/2$. Let S denote the set of variables associated with the response, then the size of S is $|S| = d^\gamma$ and the associated variables start from the first variable. To simulate $\mathbf{Y}$, we consider two scenarios. Let $\boldsymbol{\theta}$ be a coefficient vector, $\theta_i \neq 0$ if the ith variable is associated with $\mathbf{Y}$. We draw θ_i uniformly from $[1/2d_0, 3/2d_0]$ for each $i \in S$. In the first scenario named theta-sparsity, we simulate $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\theta} + \epsilon$, where ϵ is a vector of n iid normal random variables with $N(0,1)$. In this scenario, $\boldsymbol{\theta}$ is a vector of the expectation of marginal test statistics. In the second scenario name beta-sparsity, we simulate $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \epsilon$, where $\boldsymbol{\theta}$ represents the vector of regression coefficients. We simulate $\mathbf{X}$ and $\mathbf{Y}$ 1000 times. We calculate the empirical power as the proportion of rejecting the null hypothesis out of the total number of simulations. Since PFA has inflated type I error, we do not include PFA in the simulation.

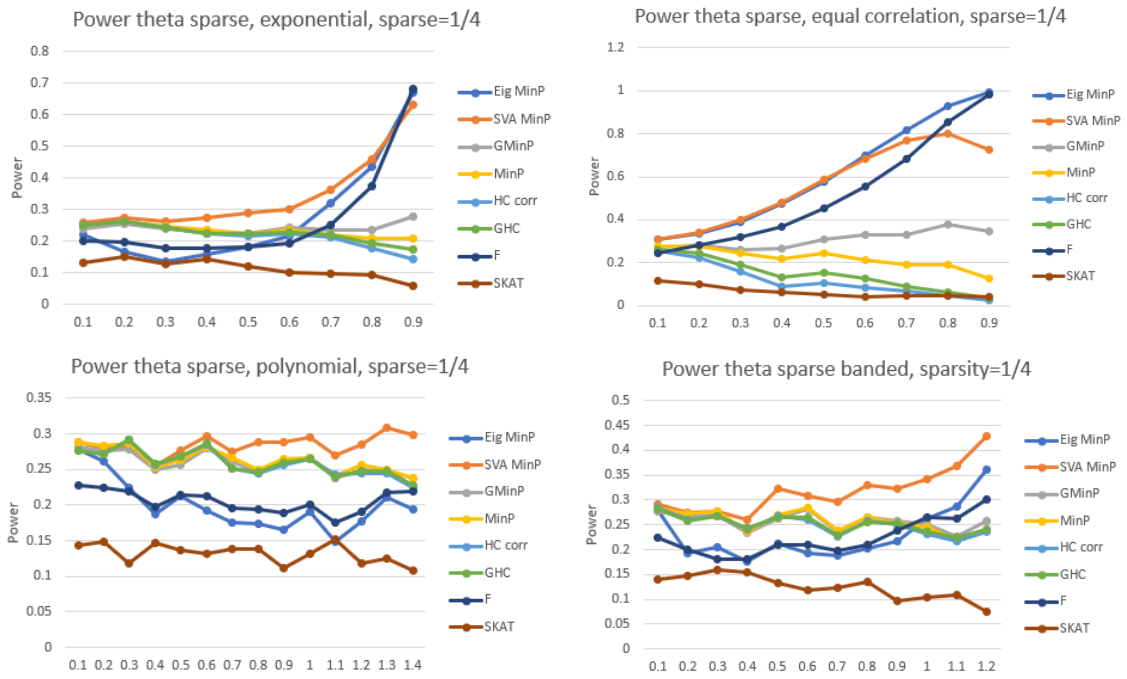


Figure 2.4. Power of tests under four types of correlation and theta sparse with sparsity 1/4. The X-axis represents the correlation parameter, and the y-axis represents power. Eig MinP: EIGENSTRAT adjusted MinP, SVA MinP: SVA adjusted MinP: original MinP, GMinP: MinP under correlation by GBJ package, HC-corr: HC under correlation by GBJ package, F: F-test, PFA: method from Fan et al. (2012).

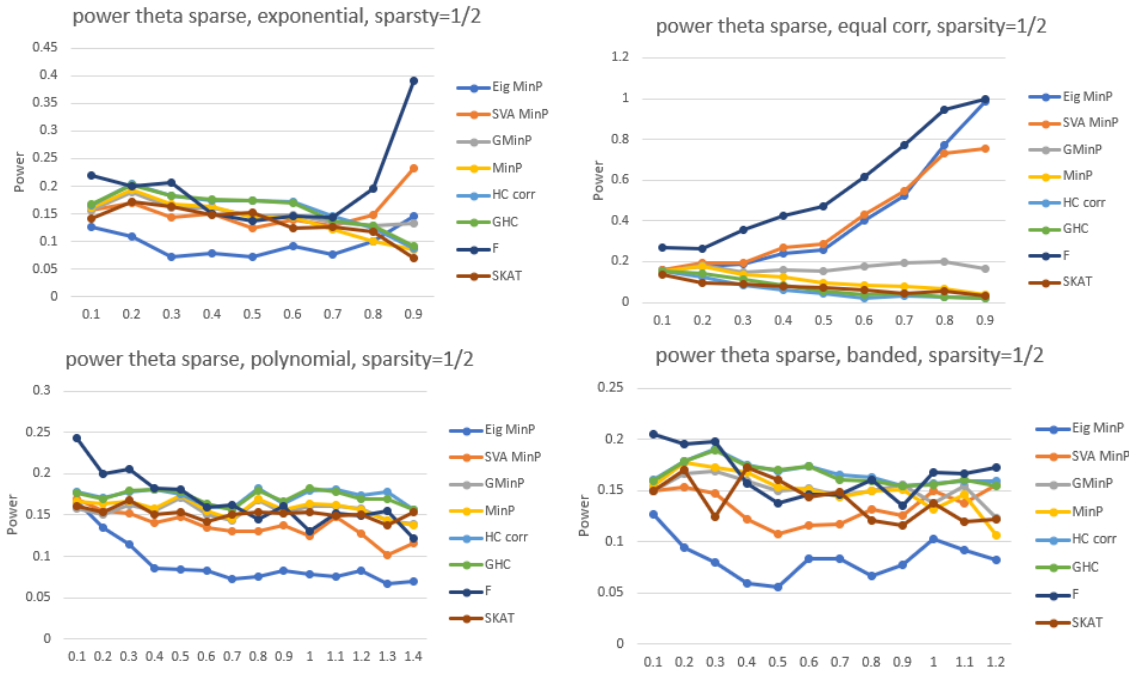


Figure 2.5. Power of tests under four types of correlation and theta sparse with sparsity 1/2. See Figure 2.4 for legend.

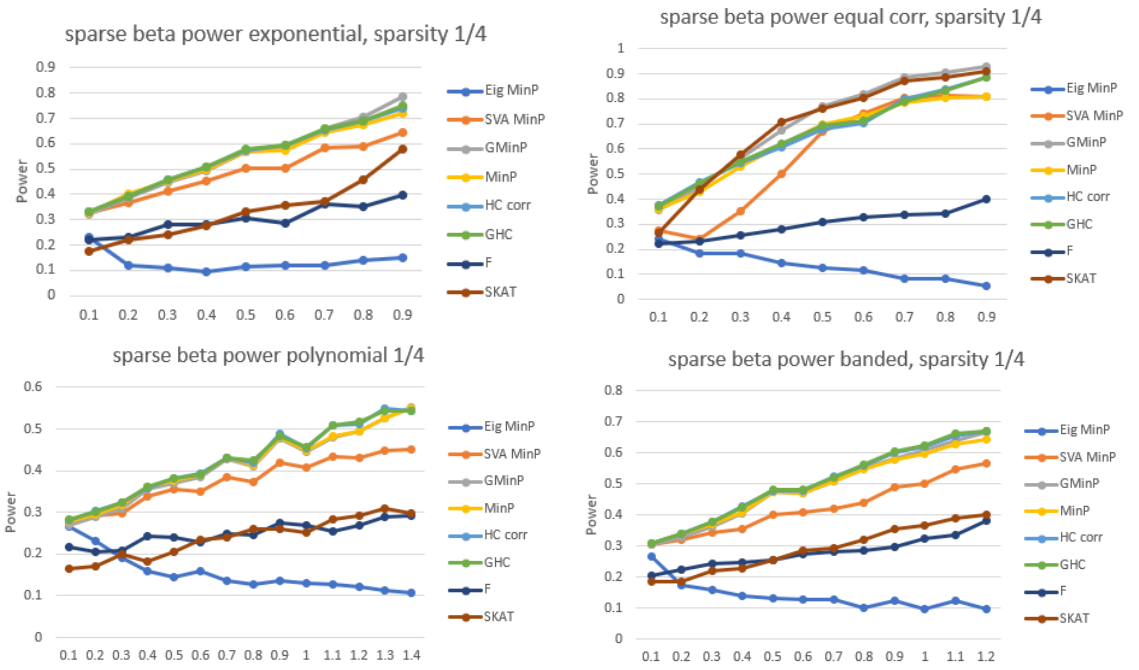


Figure 2.6. Power of tests under four types of correlation and beta-sparsity with sparsity 1/4. See Figure 2.4 for legend.

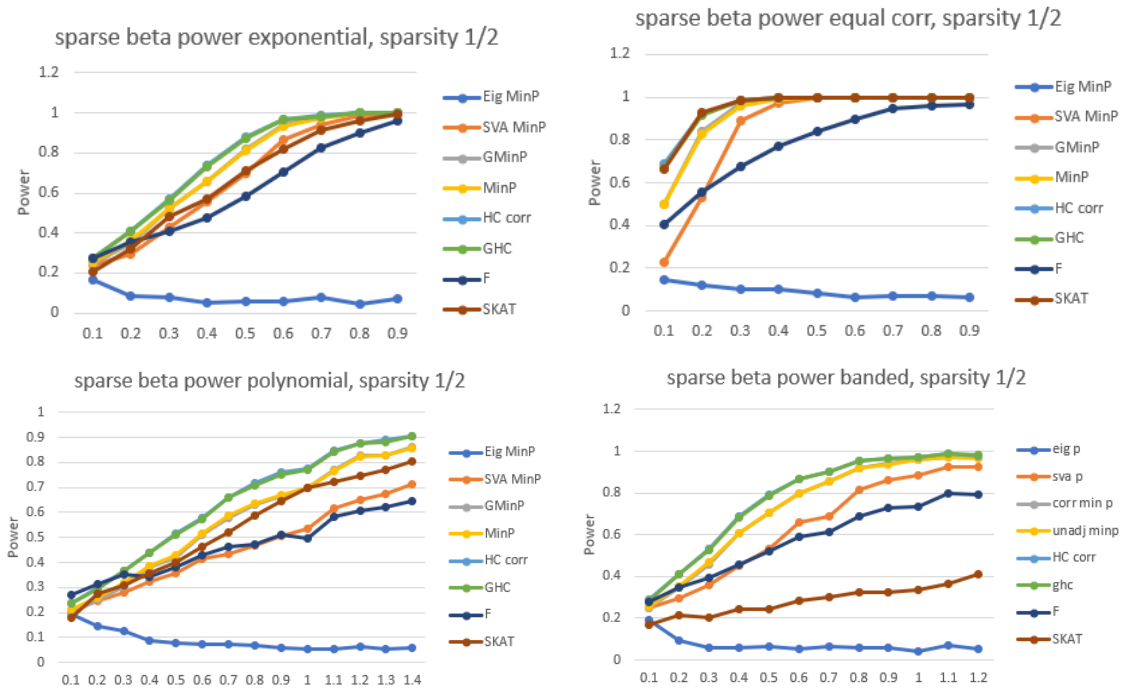


Figure 2.7. Power of tests under four types of correlation and beta-sparsity with sparsity 1/2. See Figure 2.4 for legend.

We first report the results for the scenario of theta-sparsity. Figures 2.4 shows the simulation results under four types of correlation matrices for the theta-sparsity scenario. The sparsity level is at 1/4. In this setting, the results show that the latent variable adjusted tests are more powerful than other existing tests. Specifically, SVA adjusted MinP is the most powerful test in most situations. In the polynomial and banded correlation structures, the rank according to the power is 1. SVA MinP, 2. GHC, HC-corr, MinP, 3. F and Eig MinP, 4. SKAT. In the exponential decay correlation structure, Eig MinP and F are less powerful than GMinP, MinP, HC-corr, GHC when $\rho \leq 0.5$ and but more powerful than those four tests when $\rho > 0.5$.

Figures 2.5 shows power with theta-sparsity under sparsity of 1/2. Overall, F-test performs the best. The HC method and its relevant approaches, HC-corr and GHC, performs in the second place, while MinP, adjusted MinP, and SKAT perform worse with less power. As Arias-Castro et al. (2011) discussed, MinP is powerful when the sparsity is up to 1/4. It is not surprising that MinP is less powerful when the sparsity is at 1/2.

We next report the results for the beta-sparsity scenario. Figures 2.6 shows power with beta-sparsity at the sparsity level 1/4. The rank of power is clearly shown. In exponential, polynomial, and banded correlation structures, GMinP, GHC, HC-corr, MinP take the first four places. SVA MinP is the fifth. SKAT and F take the sixth and seventh, and Eig MinP is powerless. In equal correlation, SKAT and GMinP take the first two places. GHC, HC-corr, MinP take third to fifth place. SVA MinP is the sixth followed by F and Eig MinP.

Figures 2.7 shows power with beta-sparsity under sparsity of 1/2, where MinP is supposed to be less powerful. In exponential, polynomial and banded, GHC and HC-corr take the first two places, followed by MinP, GMinP. SKAT is more powerful than F in exponential and polynomial but not in banded. SVA MinP is close to SKAT in exponential, and it is close to F in polynomial and banded. In equal correlation, SKAT, GHC, and HC-corr take the first three places, followed by MinP, GMinP. F takes seventh place.

In the beta-sparsity scenario and equal correlation, SKAT is the best, and Eig MinP almost has no power. As the second example in Section 2.4 shows, if the top latent vector is associated with the response variable, the latent variable adjustment would lose power. SVA MinP is robust under this situation by avoiding the direction of the top eigenvectors, but it still loses some power. On the other hand, SKAT would be powerful because it is a weighted sum of marginal correlation of eigenvectors when the top eigenvectors are associated.

2.5 Combination of multiple tests

2.5.1 The combination strategy

We propose a combination of multiple tests based on the data structure, i.e., sparsity and correlation, to improve the power of tests, since no test is powerful under all data conditions. The idea of combining different testing approaches to improve performance has been explored by Barnett et al. (2017) and R. Sun and Lin (2017). Different tests are conducted, and a combination test strategy is used to combine the multiple testing results. Simulation results show that the combined strategy may have a slight loss of power compared to the optimal one, but it is robust under all data settings. On the other hand, its limitation is that it adds multiple testing burden as all different testing methods are conducted for the same data set. As for our combination strategy, our method does not apply multiple global tests but let data decide which test to use.

Arias-Castro et al. (2011) compared the power of MinP, HC and the F-test under different sparsity levels. MinP is powerful when the sparsity level is $0 < \gamma < 1/4$. HC is powerful when the sparsity level is $0 < \gamma < 1/2$. The F-test is powerful when $1/2 < \gamma < 1$. We use the sparsity level to choose MinP or the F-test. To estimate the sparsity, we use marginal test statistics $r_i = \mathbf{X}_i'\mathbf{Y}/s_y$ and count the number of those statistics that are larger than a cutoff, e.g., $\sqrt{2\log d}$. Another issue is correlations among the covariates, which decide whether we need an adjusted MinP.

According to our simulations, if the mean off-diagonal correlation is larger than 0.5, the p-value of the original MinP is biased. Under this situation, adjusted MinP is needed. Barnett et al. (2017) used a summary statistic of a correlation matrix to quantify the bias of the distribution of MinP or HC when variables are correlated. Let the ith row and jth column of the correlation matrix, denoted as $\mathbf{\Sigma}$, be Σ_{ij}. Define $\bar{\rho}^r = 2/(d(d-1)) \sum_{1 \le i < l \le d} (\Sigma_{ij})^r$ as the mean of rth moments of off-diagonal correlations. The summary statistic is defined as $\sum_{r=1}^{\infty} \bar{\rho}^r / r!$. When variables have equal pairwise correlation ρ, this summary statistic is $\exp(\rho) - 1$. Though it is a summation of infinite terms, the sum of the first few terms, e.g., the first ten terms is accurate enough in practice.

We propose the following rules to choose tests: We define the proportion of the covariates whose marginal test statistics are larger than the cutoff $\sqrt{2 \log d}$ as the signal proportion. If the signal proportion is less than 0.1, or the number of covariates is less than 10, we choose to use F-test. Otherwise, we use MinP. The cutoff of $\sqrt{2 \log d}$ is also suggested by Arias-Castro et al. (2011) though it is an asymptotic conclusion. When we decide to use MinP, if the mean off-diagonal correlation is less than 0.5, we use the original MinP. Otherwise, we choose to use the SVA adjusted MinP. The specific cutoff values can be changed for different data and tasks.

2.5.2 Real data analysis

We apply our proposed test strategy to analyze the Rheumatoid Arthritis responder dataset (Cui et al., 2013). The data is from Rheumatoid Arthritis Responder Challenge organized by DREAM and Sage Bionetworks (`https://www.synapse.org/RA_Challenge`). The data contains about 2 million SNPs and several clinical covariates for about two thousand RA patients who took three anti-TNF drugs. The response variable is the change of disease activity score 3-12 months after taking an anti-TNF therapy, denoted as ΔDAS28. The clinical covariates include Batch (genotyping batch), Cohort (Name of the cohort from which the individual was ascertained), Drug

(One of three drugs received), baselineDAS28 (baseline Disease Activity Score), Gender, Mtx (whether the patient has cotherapy). Genotypes were imputed as dosages, which mean the expected number of minor allele and they range from 0 to 2.

We do not include a clinical covariate, i.e., age since there are a large number of missing values. We remove samples if they have at least one missing value of the above six clinical covariates. There are 1869 samples in our analysis. Regression of the six clinical covariates shows that all of them are significant with p-values less than 0.05. The baseline DAS28 is strongly significant with a p-value less than 10^{-16}.

We use global testing conditioning on clinical covariates to select genes that may contribute to drug response. We group SNPs into genes using human genome reference hg38 (`https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg38`) and apply global tests to gene groups. We remove SNPs that are highly correlated with clinical covariates and other SNPs, i.e., we remove one SNP in each pair of SNPs with larger than 0.99 absolute correlation. We choose SVA as the estimation method for latent factors since it performs the best among the adjusted methods in most cases as shown in the simulation study. We only apply latent factor adjusted methods when there are more than 10 SNPs in a gene. We have 37543 tests. The significance cut off is about 1.3×10^{-6}.

Figure 2.8 to Figure 2.12 show the top 30 p-values of SNP groups (or genes) conditioning on clinical covariates. Genes with ten or fewer SNPs are all tested by F-test. The combined test selects 17572 genes for MinP and 19970 genes for F-test, but only one gene for SVA adjusted MinP. We find that most of the combined tests are from F-test because the signal proportion is not sparse, or the gene has less than 10 SNPs. Overall, the combined test has comparable performance to MinP and F-test. The combined test chooses 23rd significant unadjusted MinP gene, CNST at p-value 7.31×10^{-4} because there are only 1 SNP that pass the $\sqrt{2 \log d}$ threshold among 43 SNPs in the gene. The combined test chooses F-test for the fourth significant F-test gene, ACTR3C at p-value 4.31×10^{-5}, with 7 SNPs passing the $\sqrt{2 \log d}$ threshold among 22 SNPs in the gene. Only one gene is tested by SVA MinP because very few

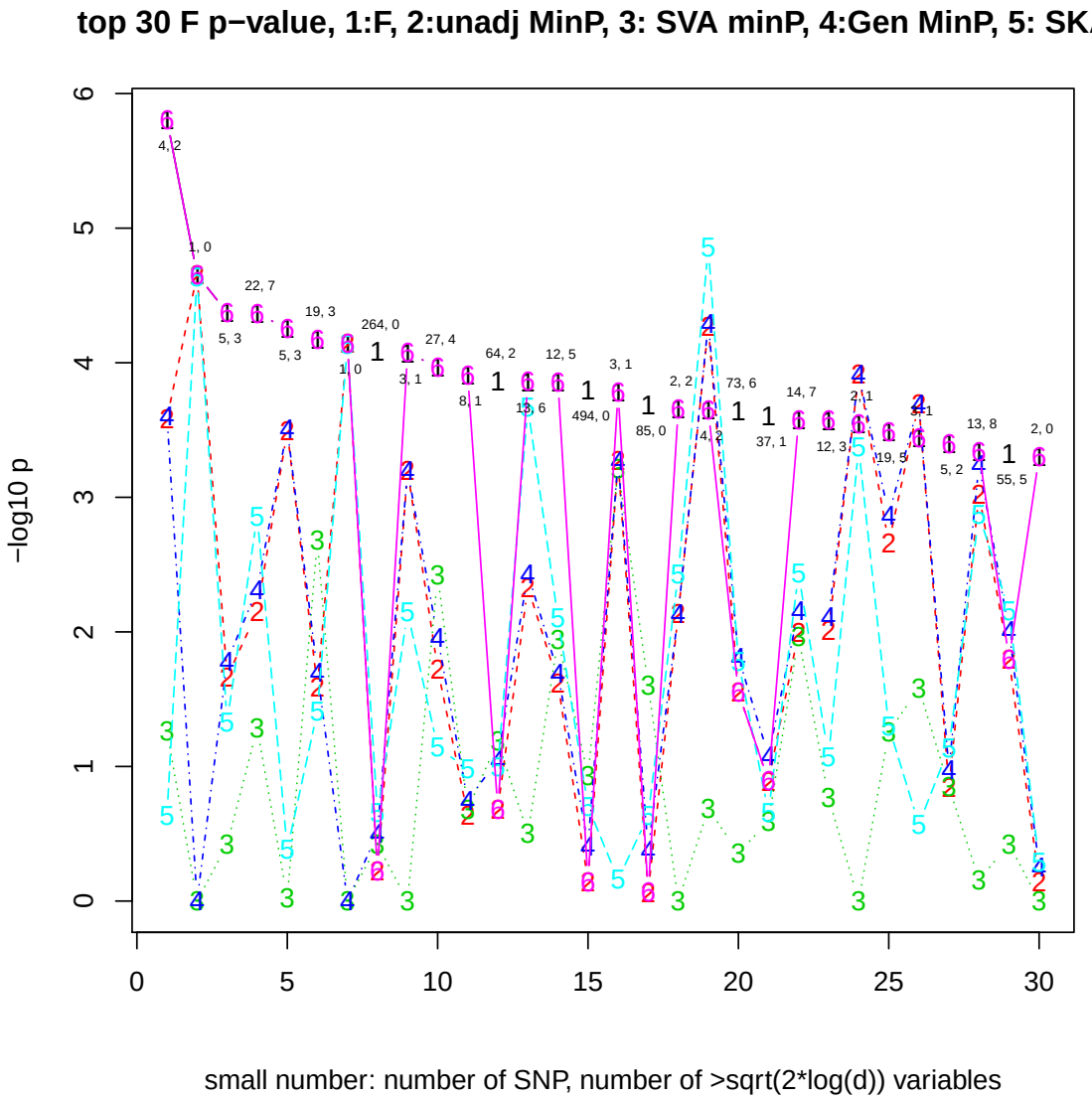


Figure 2.8. Top 30 F p-values of RA challenge data on $-\log_{10}$ scale conditioning on clinical covariates. P-values are sorted by the significance of F-test. For each line, 1 represents F-test, 2 represents unadjusted MinP, 3 represents SVA adjusted MinP, 4 represents SKAT, 5 represents the combined test. Small numbers represent the number of SNPs in a gene and the number of SNPs that have absolute marginal test statistics larger than $\sqrt{2\log(d)}$.

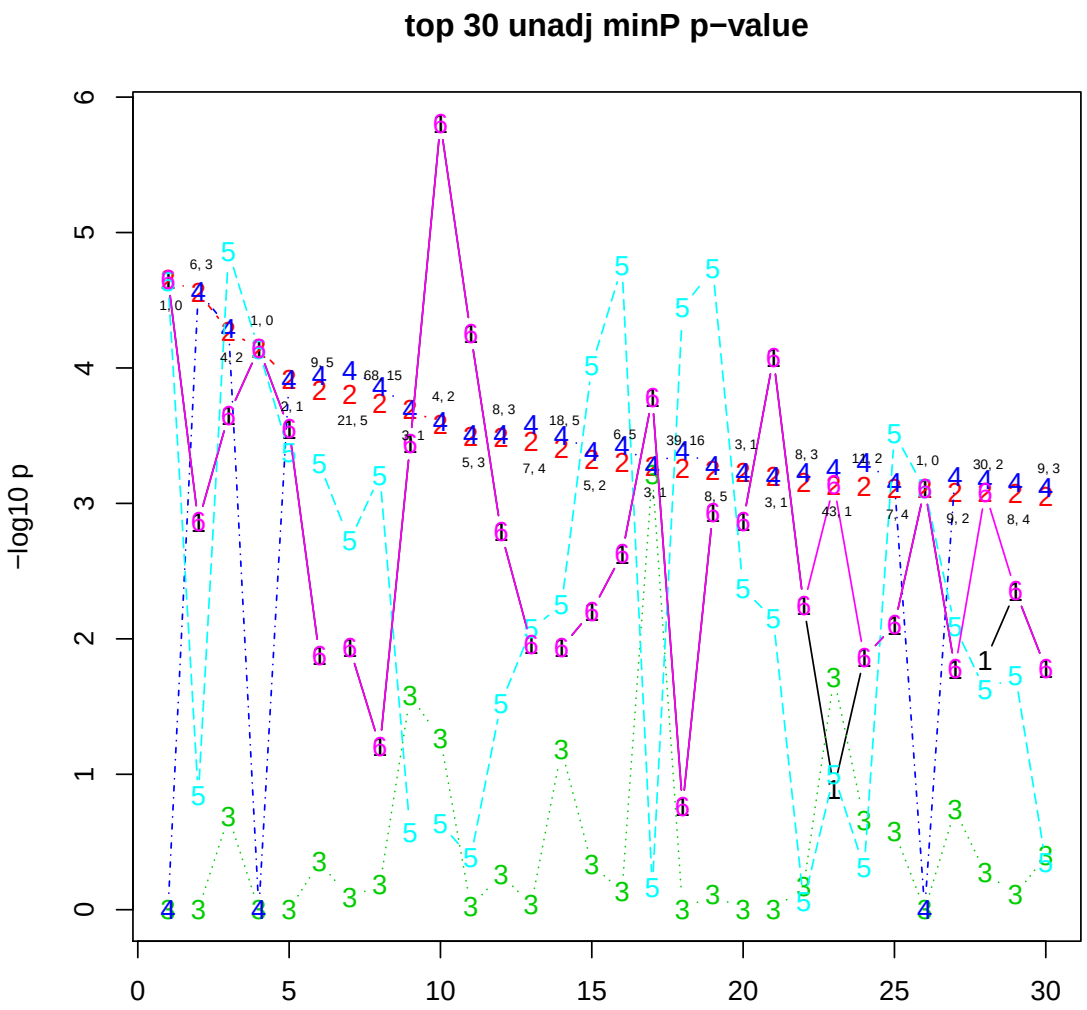


Figure 2.9. Top 30 unadjusted MinP p-values of RA challenge data on $-\log_{10}$ scale conditioning on clinical covariates. P-values are sorted by the significance of unadjusted MinP. See Figure 2.8 for legend.

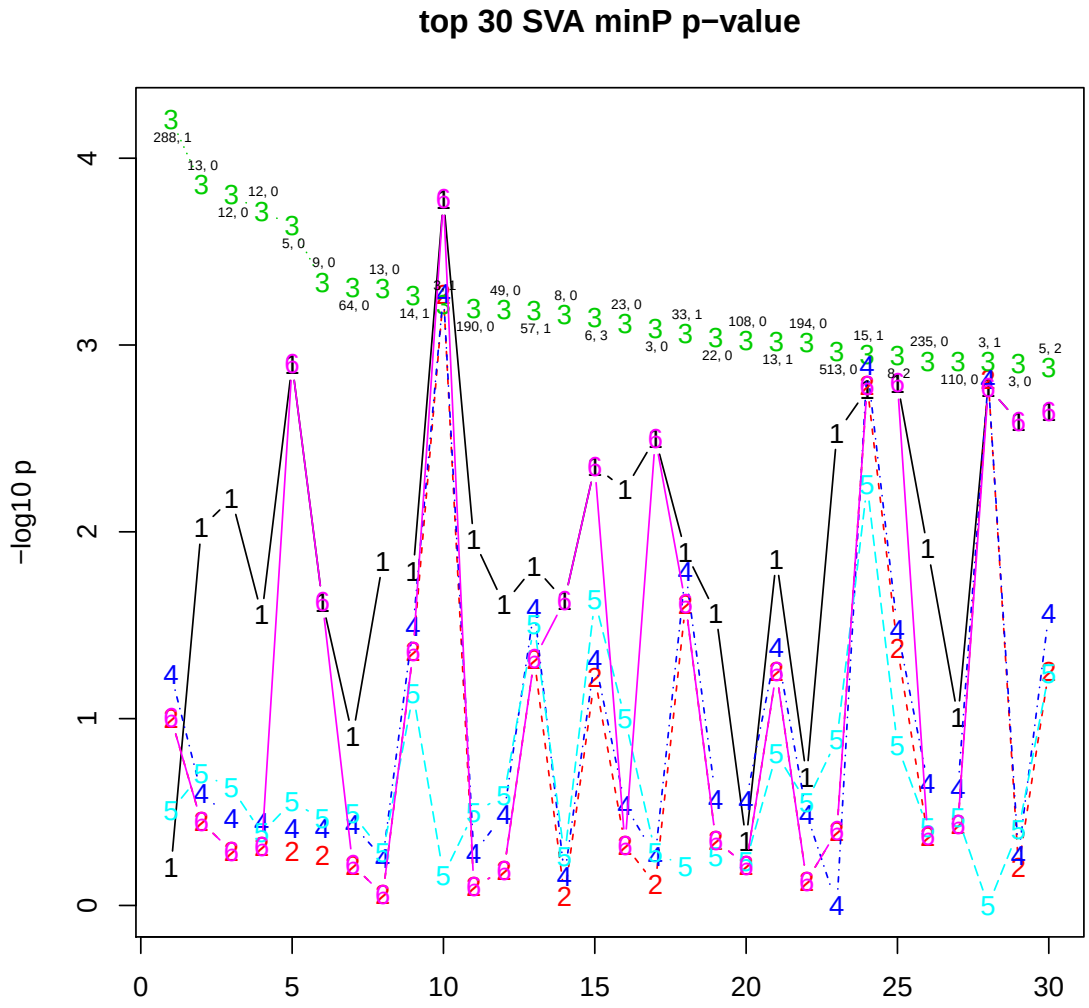


Figure 2.10. Top 30 SVA adjusted MinP p-values of RA challenge data on $-\log_{10}$ scale conditioning on clinical covariates. P-values are sorted by the significance of SVA adjusted MinP. See Figure 2.8 for legend.

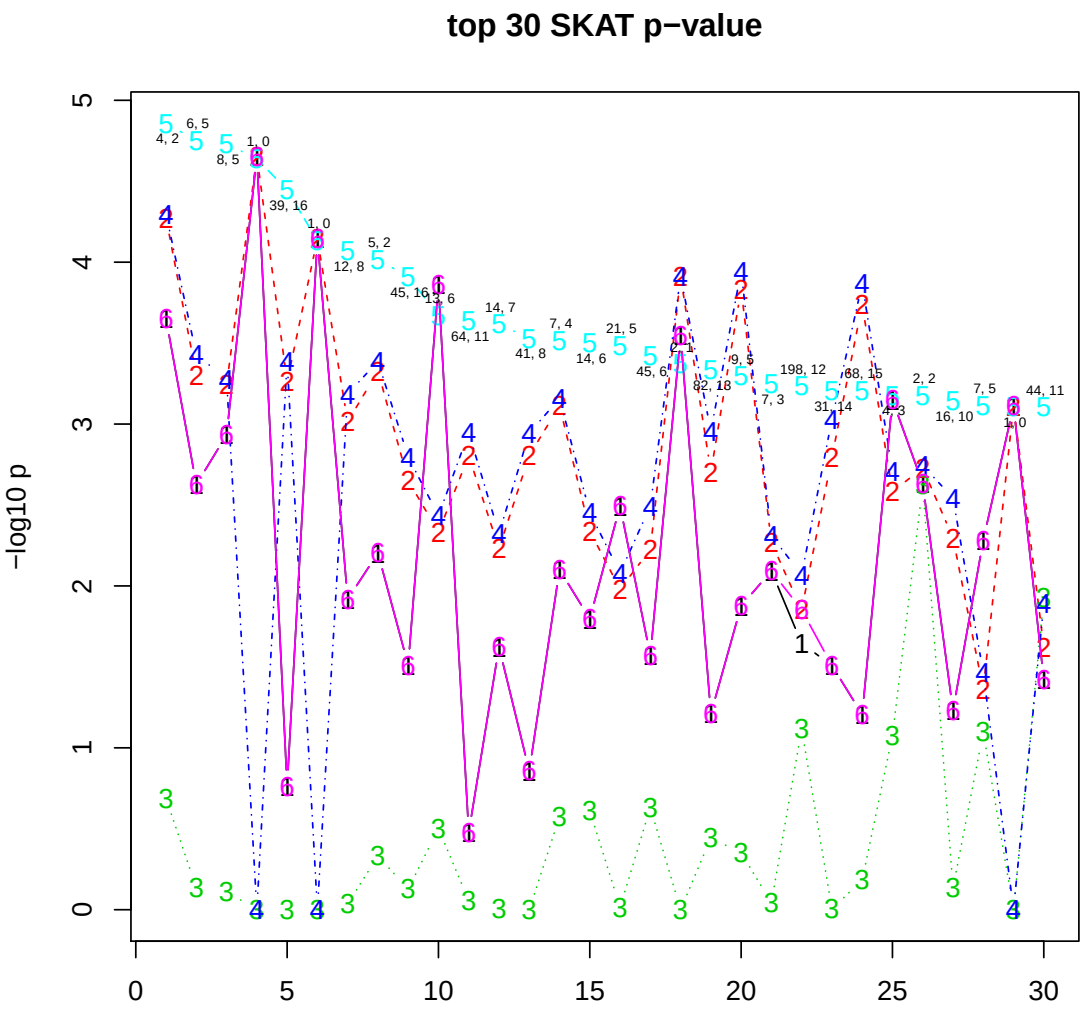


Figure 2.11. Top 30 SKAT p-values of RA challenge data on $-\log_{10}$ scale conditioning on clinical covariates. P-values are sorted by the significance of SKAT. See Figure 2.8 for legend.

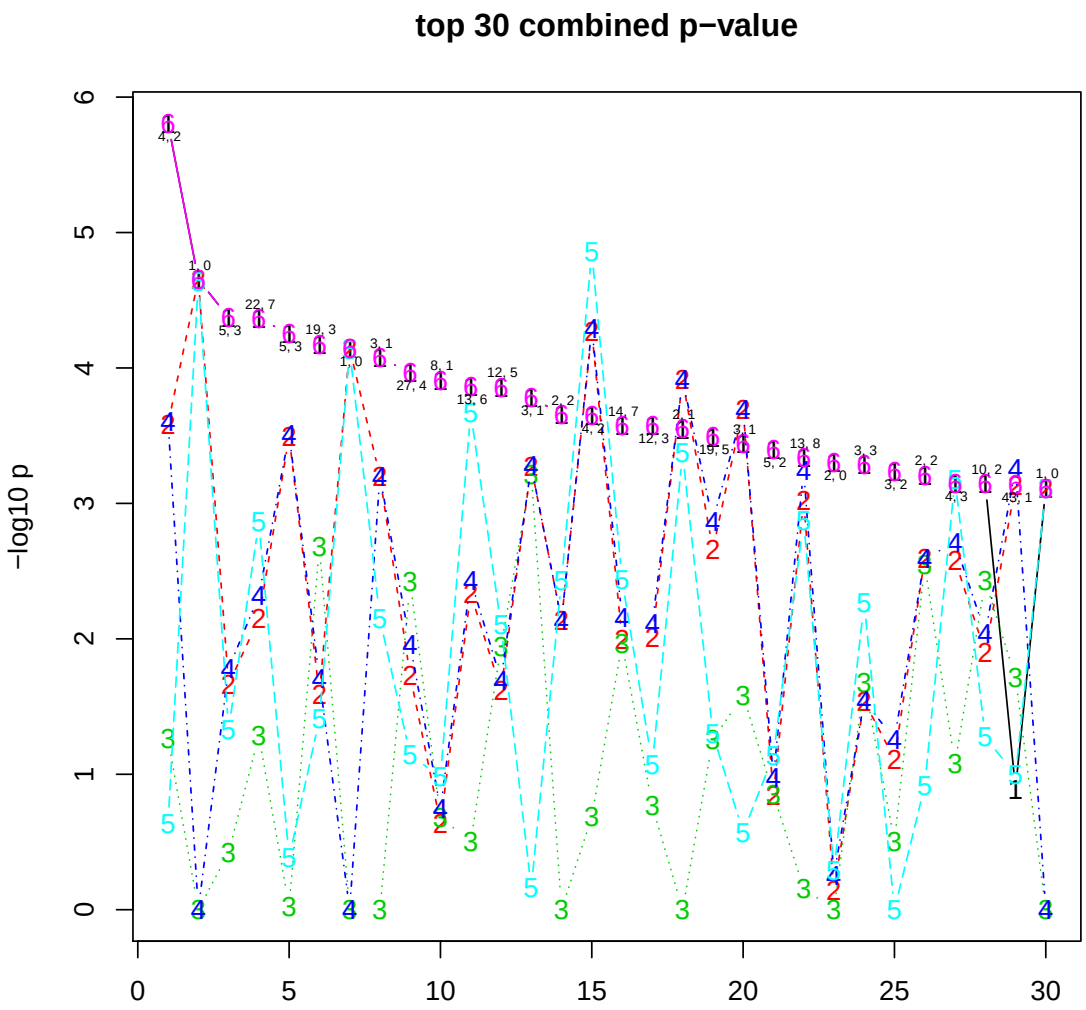


Figure 2.12. Top 30 combined test p-values of RA challenge data on $-\log_{10}$ scale conditioning on clinical covariates. P-values are sorted by the combined test. See Figure 2.8 for legend.

genes are both highly correlated and have more than 10 SNPs. An explanation is that including clinical variates in the analysis reduces dependence among SNPs. We find many genes that have more significant latent factor adjusted MinP p-values than other tests. Since they are below the Bonferroni cutoff, we do not know whether they truly associate with the response variable.

We also report the global testing results without clinical covariates. In the Dream challenge, one of the tasks is to build a prediction model of drug response using only SNP data, so we apply global tests using SNP data alone as this can be the first step to screen significant SNPs. In the list of top MinP genes, latent factor adjusted MinP can increase the significance of the third significant MinP gene, NIPA1, from 1.29×10^{-6} to 1.23×10^{-7}.

We further examine the correlation structure of the third significant gene NIPA1. It has 18 SNPs. SVA finds five latent factors in this gene. Figure 2.13 shows that the first latent factor has high correlations with ten SNPs. Figure 2.2 shows there is a correlation block of ten SNPs. Figure 2.14 shows the correlation after adjustment. We can see the high correlations among SNPs are removed.

In this real data example, most significant genes have either a small number of SNPs or the correlation of SNPs within a gene is not very high. As a result, factor-adjusted MinP is not employed for the top significant genes. We consider that the factor-adjusted test is powerful under certain conditions, but it is the best practice to use different tests under different conditions.

2.6 Discussion

We develop a latent factor adjusted MinP test for global sparse alternative hypotheses, which works in the presence of arbitrarily strong dependence. Simulation studies show that our method is powerful when effects are sparse, and covariates are highly correlated. We prove that under certain conditions, its type I error is correct and the power can be improved. In addition, we propose a method to select optimal

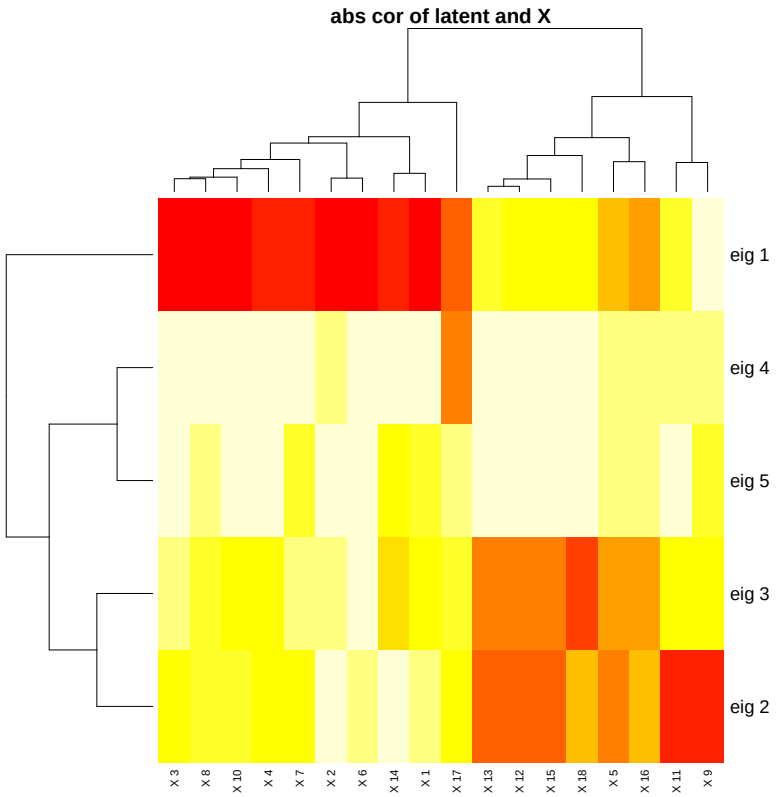


Figure 2.13. Absolute correlation of latent variables and genes in NIPA1. Red tiles represent high correlations. White ones represent no correlation. eig 1 means the first latent factor etc. X 1 means the first SNP etc.

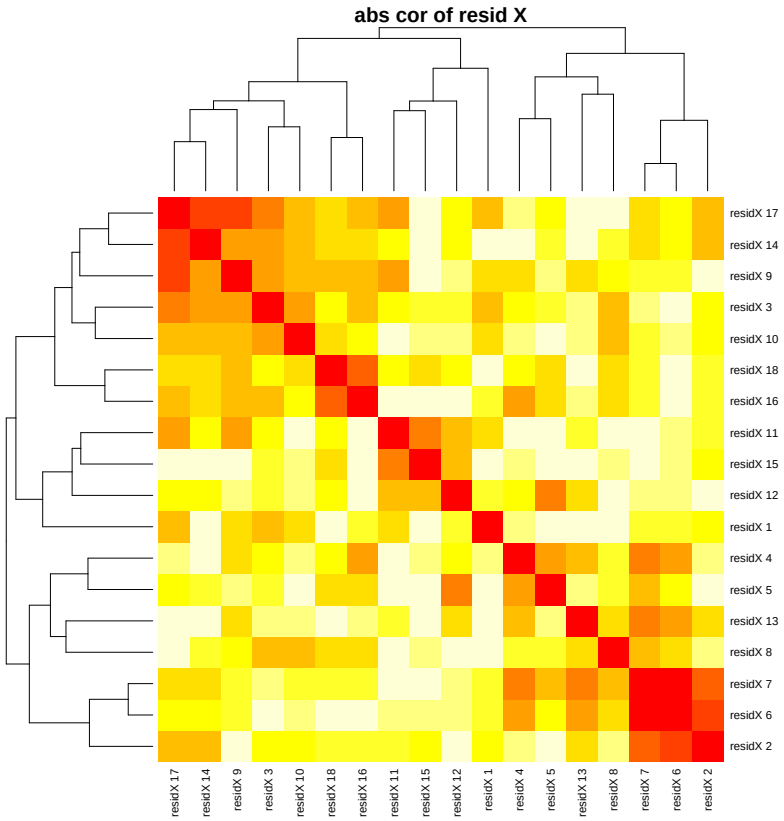


Figure 2.14. Correlation of genes after adjustment. resid X 1 means the residual of the first SNP after regressing latent factors. See Figure 2.13 for legend

tests using summary statistics such as marginal correlations and the average pairwise correlations among variables. This method does not increase the total number of tests performed but uses simple data information to guide the choice of an appropriate test method.

Our theorems assume that the latent factors are correctly estimated, but in practice, many aspects affect the effectiveness of the latent factor adjusted tests, for instance, the specific data distributions of discrete, non-negative, or heavy tail will make factor estimation difficult. Another practical issue is that the data matrix is from a sample and the correlation matrix is estimated. Fan and Han (2017) discussed the accuracy of factor-adjusted methods under unknown dependence. The distributions of top eigenvalues affect the estimate of the number of latent variables. The difficult case would be that there is no clear drop in the ordered eigenvalues. In a global test, the goal is to have correct type I error and improve power. The dataset has large n and small d, while in multiple testing, the data has large d and small n, and there is at least one significant marginal test statistic. The current literature of factor adjusted testing methods is for the multiple testing problem. Existing factor-adjusted methods may work well under their original settings but not for global testing, especially when we consider sparse and weak alternatives. Some of the future work is to study different factor estimation methods and how they can be combined with optimal global testing approaches.

REFERENCES

REFERENCES

Arias-Castro, E., Candès, E. J., & Plan, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 2533–2556.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*(3), 375–386.

Aryal, U. K., Xiong, Y., McBride, Z., Kihara, D., Xie, J., Hall, M. C., & Szymanski, D. B. (2014). A proteomic strategy for global analysis of plant protein complexes. *The Plant Cell*, *26*(10), 3867–3882.

Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P.-E., Trégouët, D.-A., & Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, *94*(5), 662–676.

Ball, G. H., & Hall, D. J. (1965). *Isodata, a novel method of data analysis and pattern classification* (Tech. Rep.). Menlo Park CA: Stanford research inst.

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 803–821.

Barnett, I., Mukherjee, R., & Lin, X. (2017). The generalized higher criticism for testing snp-set effects in genetic association studies. *Journal of the American Statistical Association*, *112*(517), 64–76.

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

Bittman, R. M., Romano, J. P., Vallarino, C., & Wolf, M. (2009). Optimal testing of multiple hypotheses with common effect direction. *Biometrika*, *96*(2), 399–410.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate behavioral research*, *27*(4), 509–540.

Chen, J., Behnam, E., Huang, J., Moffatt, M. F., Schaid, D. J., Liang, L., & Lin, X. (2017). Fast and robust adjustment of cell mixtures in epigenome-wide association studies with smartsva. *BMC genomics*, *18*(1), 413.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, *36*(3), 287–314.

Conneely, K. N., & Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, *81*(6), 1158–1168.

Coskun, O. (2016). Separation techniques: chromatography. *Northern clinics of Istanbul*, *3*(2), 156.

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., & Mann, M. (2014). Maxlfq allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction. *Molecular & Cellular Proteomics*, mcp–M113.

Cox, J., & Mann, M. (2008). Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, *26*(12), 1367.

Cox, J., & Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. *Annual Review of Biochemistry*, *80*, 273–299.

Cui, J., Stahl, E. A., Saevarsdottir, S., Miceli, C., Diogo, D., Trynka, G., ... others (2013). Genome-wide association study and gene expression analysis identifies cd84 as a predictor of response to etanercept therapy in rheumatoid arthritis. *PLoS Genetics*, *9*(3), e1003394.

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(3), 323–333.

Donoho, D., & Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, *105*(39), 14790–14795.

Donoho, D., & Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, *30*(1), 1–25.

Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., ... others (2006). A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *Science*, *314*(5804), 1461–1463.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95–104.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *The Journal of Psychology*, *80*(2), 351–363.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, pp. 226–231).

Fan, J., & Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(4), 1143–1164.

Fan, J., Han, X., & Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, *107*(499), 1019–1035.

Fan, J., Ke, Y., Sun, Q., & Zhou, W.-X. (2017). Farm-test: Factor-adjusted robust multiple testing with false discovery control. *arXiv preprint arXiv:1711.05386*.

Ferrer, I., & Thurman, E. M. (2003). *Liquid chromatography/mass spectrometry, ms/ms and time of flight ms: analysis of emerging contaminants*. ACS Publications.

Fields, S., & Sternglanz, R. (1994). The two-hybrid system: an assay for protein-protein interactions. *Trends in Genetics*, *10*(8), 286–292.

Fisher, R. A. (1934). *Statistical methods for research workers*. Edinburgh.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York, NY, USA:.

Friguet, C., Kloareg, M., & Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, *104*(488), 1406–1415.

Gagnon-Bartsch, J. A., & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, *13*(3), 539–552.

Gao, X., Starmer, J., & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *32*(4), 361–369.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*(2), 141–149.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2012). *mvtnorm: Multivariate normal and t distributions. r package version 0.9-9996.*

Gross, J. H. (2017). Tandem mass spectrometry. In *Mass spectrometry* (pp. 539–612). Springer.

Hall, P., & Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, *38*(3), 1686–1732.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21*(15), 3201–3212.

Heard, N. A., & Rubin-Delanchy, P. (2018). Choosing between methods of combining-values. *Biometrika*, *105*(1), 239–246.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, *44*, 223–270.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., ... Gerstein, M. (2003). A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, *302*(5644), 449–453.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Jungbauer, A., & Hahn, R. (2009). Ion-exchange chromatography. In *Methods in enzymology* (Vol. 463, pp. 349–371). Elsevier.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Koulman, A., Woffendin, G., Narayana, V. K., Welchman, H., Crone, C., & Volmer, D. A. (2009). High-resolution extracted ion chromatography, a new tool for metabolomics and lipidomics using a second-generation orbitrap mass spectrometer. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, *23*(10), 1411–1418.

Koziol, J. A., & Perlman, M. D. (1978). Combining independent chi-squared tests. *Journal of the American Statistical Association*, *73*(364), 753–763.

Kristensen, A. R., Gsponer, J., & Foster, L. J. (2012). A high-throughput approach for measuring temporal changes in the interactome. *Nature Methods*, *9*(9), 907–909.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., ... Garcia-Hernandez, M. (2011). The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic Acids Research*, *40*(D1), D1202–D1210.

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., & Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS Computational Biology*, *12*(1), e1004714.

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (2012). *Extremes and related properties of random sequences and processes*. Springer Science & Business Media.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788.

Lee, S., Sun, W., Wright, F. A., & Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika*, *104*(2), 303–316.

Lee, Y. (2016, 11). *Rice project report.*

Leek, J. T. (2014). Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, *42*(21), e161–e161.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882–883.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733.

Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, *105*(48), 18718–18723.

Liang, X., Wang, Z., Sha, Q., & Zhang, S. (2016). An adaptive fisher's combination method for joint analysis of multiple phenotypes in association studies. *Scientific Reports*, *6*, 34323.

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., ... Zhang, H. (2014). Meta-analysis of gene-level tests for rare variant association. *Nature Genetics*, *46*(2), 200.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., ... Martin, N. G. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, *87*(1), 139–145.

Liu, Y., & Xie, J. (2018a). Accurate and efficient p-value calculation via gaussian approximation: a novel monte-carlo method. *Journal of the American Statistical Association*, 1–9.

Liu, Y., & Xie, J. (2018b). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*(just-accepted), 1–29.

Liu, Y., & Xie, J. (2018c). Powerful test based on conditional effects for genome-wide screening. *The Annals of Applied Statistics*, *12*(1), 567.

Liu, Z., & Lin, X. (2018). A geometric perspective on the power of principal component association tests in multiple phenotype studies. *Journal of the American Statistical Association*(just-accepted), 1–36.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... others (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747.

Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*(4), 457.

McBride, Z., Chen, D., Lee, Y., Aryal, U., J., Xie, & Szymanski, D. (2018). *A label free mass spectrometry method to predict endogenous protein complex composition.* (Manuscript submitted for publication)

McBride, Z., Chen, D., Reick, C., Xie, J., & Szymanski, D. B. (2017). Global analysis of membrane-associated protein oligomerization using protein correlation profiling. *Molecular & Cellular Proteomics*, mcp–000276.

Mori, S., & Barth, H. G. (2013). *Size exclusion chromatography.* Springer Science & Business Media.

Mudholkar, G. S., & George, E. O. (1977). *The logit statistic for combining probabilities-an overview* (Tech. Rep.). Rochester, NY: ROCHESTER UNIV DEPT OF STATISTICS.

Murray, K. (2006). *Schematic of tandem mass spectrometry.* `https://commons.wikimedia.org/wiki/File:MS_MS.png`. (Online; accessed 5-December-2018)

Parker, H. S., Bravo, H. C., & Leek, J. T. (2014). Removing batch effects for prediction problems with frozen surrogate variable analysis. *PeerJ*, *2*, e561.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, *25*(3/4), 379–410.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904.

Rajagopal, S. (2011). Customer data clustering using data mining technique. *arXiv preprint arXiv:1112.2663*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Regev-Rudzki, N., Karniely, S., Ben-Haim, N. N., & Pines, O. (2005). Yeast aconitase in two locations and two metabolic pathways: seeing small amounts is believing. *Molecular biology of the cell*, *16*(9), 4163–4171.

Revez, K., Landwehr, J., & Keybl, J. (2001). *Measurement of 13c and 18o isotopic ratios of caco3 using a thermoquest finnigan gas bench ii delta plus xl continuous flow isotope ratio mass spectrometer with application to devils hole core dh-11 calcite.* United States Geological Survey.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, *17*(10), 1030.

Rijsbergen, V. (1979). *Cj information retrieval.* London, Butterworths.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*(318), 626–633.

Song, M., Hao, W., & Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature genetics*, *47*(5), 550.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *Kdd workshop on text mining* (Vol. 400, pp. 525–526).

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). *The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.* Princeton, NJ: Princeton Univ. Press.

Sun, R., & Lin, X. (2017). Set-based tests for genetic association using the generalized berk-jones statistic. *arXiv preprint arXiv:1710.02469*.

Sun, Y., Zhang, N. R., & Owen, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics*, *6*(4), 1664–1688.

Teschendorff, A. E., Zhuang, J., & Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, *27*(11), 1496–1505.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., & Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, *22*(19), 2405–2412.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tippett, L. H. C. (1931). *Methods of statistics*. Williams Norgate: London.

Wang, J., Zhao, Q., Hastie, T., & Owen, A. B. (2015). Confounder adjustment in multiple hypothesis testing. *arXiv preprint arXiv:1508.04178*.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82–93.

Yoder, N. (2011). Peakfinder. *Internet: http://www. mathworks. com-/matlabcentral/fileexchange/25500*.

VITA

VITA

Donglai Chen was brought up in Shenzhen, China. He received a Bachelor's Degree in statistics in Peking university in 2013. He began PhD study at Purdue in 2013. He got Master's degree in 2016 at Purdue University. His major professor is Jun Xie. After graduation, he will work as a data scientist at Sam's Club.

Publications:

Aryal, U. K., McBride, Z., **Chen, D.**, Xie, J., & Szymanski, D. B. (2017). Analysis of protein complexes in Arabidopsis leaves using size exclusion chromatography and label-free protein correlation profiling. *Journal of Proteomics*. 166, 8-18.

McBride, Z., **Chen, D.**, Reick, C., Xie, J., & Szymanski, D. B. (2017). Global Analysis of Membrane-associated Protein Oligomerization Using Protein Correlation Profiling. *Molecular & Cellular Proteomics*, 16(11), 1972-1989.

Liu, M., Qu, H., Bu, Z., **Chen, D.**, Jiang, B., Cui, M., ..., Di, J. (2015). Validation of the Memorial Sloan-Kettering Cancer Center Nomogram to Predict Overall Survival After Curative Colectomy in a Chinese Colon Cancer Population. *Annals of Surgical Oncology*, 22(12), 3881–3887.

Chen, D., Jiang, B., Xing, J., Liu, M., Cui, M., Liu, Y., ..., Zhang, C. (2013). Validation of the Memorial Sloan Kettering Cancer Center nomogram to predict disease-specific survival after R0 resection in a Chinese gastric cancer population. *PloS One*, 8(10), e76041.

Chen, D., Liu, C., & Xie, J. (2016). Multi-locus Test and Correction for Confounding Effects in Genome-Wide Association Studies. *The International Journal of Biostatistics*, 12(2).