

ON LATENT SYSTEMIC EFFECTS IN MULTIPLE HYPOTHESES

A DISSERTATION  
SUBMITTED TO THE DEPARTMENT OF STATISTICS  
AND THE COMMITTEE ON GRADUATE STUDIES  
OF STANFORD UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

Yunting Sun  
August 2011

© 2011 by Yunting Sun. All Rights Reserved.

Re-distributed by Stanford University under license with the author.



This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.

<http://creativecommons.org/licenses/by-nc/3.0/us/>

This dissertation is online at: <http://purl.stanford.edu/fg181ks0498>

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

**Art Owen, Primary Adviser**

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

**Nancy Zhang, Co-Adviser**

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

**Bradley Efron**

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

**Wing Wong**

Approved for the Stanford University Committee on Graduate Studies.

**Patricia J. Gumport, Vice Provost Graduate Education**

*This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.*

# Abstract

This dissertation deals with two closely related topics of latent systemic effect in multiple hypothesis testing in addition to supplying an overview of the growing literature in the field.

The first part aims at searching for associations with a primary variable among a great many candidate variables in high throughput settings. High throughput hypothesis testing can be made difficult by the presence of systemic effects and other latent variables. It is well known that those variables alter the level of tests and induce correlations between tests. It is less well known that dependencies can change the relative ordering of significance levels among hypotheses. Poor rankings lead to wasteful and ineffective followup studies. The problem becomes acute for latent variables that are correlated with the primary variable. We propose a two stage analysis to counter the effects of latent variables on the ranking of hypotheses. Our method, called LEAPP, statistically isolates the latent variables from the primary one. In simulations it gives better ordering of hypotheses than competing methods such as SVA and EIGENSTRAT. For an illustration, we turn to data from the AGEMAP study relating gene expression to age for 16 tissues in the mouse. LEAPP generates rankings with greater consistency across tissues than the rankings attained by the other methods.

The second part studies the detection of DNA copy number variation (CNV) across samples. Experimental artifacts, such as local trends, if not carefully removed, may be misconstrued as significant recurrent regions. We develop an alternating algorithm to adjust the effects of latent variables on the detection of recurring CNVs. Our method, called CNVlatent, improves accuracy in detecting CNVs for simulated data

compared to methods without adjustments for latent effects. We resort to two data sets for illustration. One is from the chromosome 9p region in 44 pediatric leukemia samples and the other is from a region on cytoband 11 on the q-arm of chromosome 22. There are a few visible CNVs in the data sets and CNVlatent successfully detects those visible changes and adjusts for the latent effects.

There are many studies regarding segmentation of CNVs, but incorporating copy number information into association tests remains an open problem for lack of accuracy of copy number genotyping. We proposed a statistical framework for genotyping CNVs on a detected genomic region encompassing the putative CNVs in the analysis of both inherited and somatic copy number variants. To pool information across SNPs, we take into account the different response rates and noise properties of each SNP. We carry out the model calibration with an Expectation-Maximization (EM) based algorithm. Our method achieves higher estimation precision in synthetic data and generate estimators with greater consistency across a data set with replicate samples than existing methods such as CNVtools.

# Acknowledgement

First and foremost, I would like to thank my advisors, Professor Art Owen and Professor Nancy Zhang. They not only have provided me with guidance and encouragement but have also made me understand what makes a good researcher and teacher. I have been privileged and fortunate to work with them on this thesis. I am indebted to Art for dedicating so much effort to my technical reports, correcting my English with infinite patience and sharing his knowledge and intuition about Statistics. I am very grateful to Nancy for her extremely helpful suggestions and tremendous encouragement in my research through the years.

I would also like to thank Professor Bradley Efron and Professor Wing Wong for providing insightful comments and invaluable feedback to my thesis work as members of both my oral committee and reading committee. I am also grateful to Professor Hua Tang for serving on my oral committee.

I would like to thank all the faculty members and the staff in the Department of Statistics for providing such a wonderful environment for statistical education and research. In addition, I would like to thank my friends and peer students at Stanford for valuable discussions and happy hours together.

Last but not least, I would like to thank my parents who have been incredibly supportive and patient throughout my twenty-one years of education.

# Contents

|  |           |
|--|-----------|
| Abstract   | iv        |
| Acknowledgement  | vi        |
| <b>I Multiple hypothesis testing, adjusting for latent variables</b> | <b>1</b>  |
| <b>1 Introduction</b>  | <b>2</b>  |
| <b>2 Background</b>  | <b>5</b>  |
| 2.1 Large Scale Multiple Comparison . . . . .                        | 5         |
| 2.2 Latent Factor Model of Large Dimension . . . . .                 | 10        |
| <b>3 Models and Notations</b>  | <b>15</b> |
| 3.1 Statement of The Problem . . . . .                               | 15        |
| 3.2 Review of Existing Approaches . . . . .                          | 17        |
| 3.2.1 SVA . . . . .  | 17        |
| 3.2.2 EIGENSTRAT . . . . .   | 18        |
| 3.2.3 Other Methods . . . . .  | 19        |
| <b>4 Latent Effects Adjustment</b>                                   | <b>21</b> |
| 4.1 LEAPP Method . . . . .   | 22        |
| 4.1.1 Model Reduction . . . . .                                      | 22        |
| 4.1.2 Alternatives for non-sparse $\gamma$ . . . . .                 | 25        |

|       |  |    |
|-------|--|----|
| 4.1.3 | Rank Estimation . . . . .                              | 26 |
| 4.2   | Identifiability Conditions . . . . .                   | 26 |
| 4.3   | Theory . . . . .                                       | 29 |
| 4.4   | Performance On Synthetic Data . . . . .                | 44 |
| 4.4.1 | Multiple testing with known primary variable . . . . . | 44 |
| 4.4.2 | Simulated SNP association study . . . . .              | 51 |
| 4.5   | Real Data . . . . .                                    | 55 |
| 4.5.1 | Agemap mice data . . . . .                             | 55 |
| 4.5.2 | Breast Cancer Microarray Study . . . . .               | 57 |
| 4.6   | Conclusions . . . . .                                  | 62 |

## **II Copy number variation detection, adjusting for latent variables** **63**

|          |  |           |
|----------|--|-----------|
| <b>5</b> | <b>Modeling Copy Number Variation</b>                  | <b>64</b> |
| 5.1      | Introduction . . . . .                                 | 64        |
| 5.2      | Total Copy Number Estimation for One Sample . . . . .  | 65        |
| 5.2.1    | CBS Based Change Point Detection . . . . .             | 66        |
| 5.2.2    | Fused Lasso . . . . .                                  | 67        |
| 5.3      | Cross Sample Copy Number Variation Detection . . . . . | 68        |
| 5.4      | Copy Number Variation Genotyping . . . . .             | 70        |
| <b>6</b> | <b>Cross Sample CNV Detection</b>                      | <b>72</b> |
| 6.1      | Model and Notations . . . . .                          | 72        |
| 6.2      | Performance On Synthetic Data . . . . .                | 78        |
| 6.3      | Real Data . . . . .                                    | 81        |
| 6.4      | Discussion . . . . .                                   | 83        |
| <b>7</b> | <b>EM Based CNV Genotyping</b>                         | <b>87</b> |
| 7.1      | Model . . . . .  | 87        |
| 7.2      | Performance on Synthetic Data . . . . .                | 92        |
| 7.3      | Real Data . . . . .                                    | 93        |

|                          |    |
|--------------------------|----|
| 7.4 Discussion . . . . . | 96 |
|--------------------------|----|

# List of Tables

|     |   |    |
|-----|---|----|
| 4.1 | This table shows the number of samples required for SVA to attain the same AUC that LEAPP attains with $n = 60$ samples. For example with $\text{SNR} = 2$ and $\text{LNR} = 0.5$ , and $\rho = 0.25$ , SVA requires 66 samples or 10% more. The entries of 100% denote settings where the increase needed was $\geq 100\%$ . . . . . | 49 |
| 4.2 | This table shows the improvement in AUC and MSE for the LEAPP method relative to SVA. Here $\rho$ is the correlation between the primary and latent variables and SLR,SNR and LNR are defined in the text. . . . .  | 53 |
| 4.3 | This table shows the improvement in estimation of $U, V$ for the LEAPP method relative to SVA. Here $\rho$ is the correlation between the primary and latent variables and LSR, SNR and LNR are defined in the text. . . . .  | 54 |
| 4.4 | This table shows the AUC comparison of methods SVA, LEAPP and EIGENSTRAT for 2 simulated SNP association studies, where the relative risk $R$ is set to be 1.5 and 3 respectively. . . . .  | 54 |
| 4.5 | Histogram of 3170 $z$ -values from filtered breast cancer microarray study Hedenfalk (2001) without any adjustment, with adjustment using LEAPP, SVA and EIGENSTRAT, with left to right, top to bottom respectively (by R package <i>locfdr</i> ). . . . .  | 61 |
| 4.6 | This table shows the MLE mean and variance for each of the four methods: empirical null without any adjustment, empirical null adjusted by LEAPP, SVA and EIGENSTRAT by <i>locfdr</i> R package. . . . .  | 62 |

|     |  |    |
|-----|--|----|
| 4.7 | This table shows the number of genes identified as true discoveries, 0.2 control level for 5 methods, which are fdr statistics with theoretical null without adjustment, empirical null without any adjustment, empirical null adjusted by LEAPP, SVA and EIGENSTRAT. Local fdr statistics are calculated via <i>locfdr</i> R package. . . . . | 62 |
| 6.1 | This table shows the counts of number of latent factors estimated by cross validation in category $\{0, 1, 2, 3\}$ . . . . .   | 80 |
| 6.2 | This table shows the loss incurred $\text{Loss}(\hat{S}, S)$ . . . . .   | 80 |
| 7.1 | Comparison of our method(EM) and CNVtools in percentage of correct identification. . . . .   | 92 |

# List of Figures

|     |   |    |
|-----|---|----|
| 4.1 | This figure shows the knee of the ROC curves for a simulation with $\rho = 1/2$ , SLR= 1/2 and SNR=1. The best (highest) results are for an oracle that was given the latent variables. The second best are for the proposed LEAPP method. A raw method making no adjustment gives ROCs just barely larger than SVA. EIGENSTRAT did quite poorly in this setting. The relative performance for SVA, EIGENSTRAT and the raw method were different in other settings. . . . . | 47 |
| 4.2 | This figure shows the improvement in AUC for the LEAPP method relative to SVA. Here $\rho$ is the correlation between the primary and latent variables. The signal to noise ratio and latent to noise ratio are described in the text. The color scheme encodes $(AUC_{\text{lea}} - AUC_{\text{sva}})/AUC_{\text{sva}}$ . . . . .  | 48 |
| 4.3 | This figure shows the improvement in AUC for the LEAPP method relative to EIGENSTRAT. The simulation conditions are as described in Figure 4.2. The color scheme encodes $(AUC_{\text{rot}} - AUC_{\text{eig}})/AUC_{\text{eig}}$ . . . . .   | 50 |
| 4.4 | This figure shows the improvement in precision for the LEAPP method relative to SVA. Precision is the fraction of truly affected genes among the top $H = 50$ ranked genes. The simulation conditions are as described in Figure 4.2. The color scheme encodes $(PRE_{\text{lea}} - PRE_{\text{sva}})/PRE_{\text{sva}}$ . . . . .   | 51 |

|     |   |    |
|-----|---|----|
| 4.5 | This figure shows the improvement in precision for the LEAPP method relative to EIGENSTRAT. Precision is the fraction of truly affected genes among the top $H = 50$ ranked genes. The simulation conditions are as described in Figure 4.2. The color scheme encodes $(\text{PRE}_{\text{rot}} - \text{PRE}_{\text{eig}})/\text{PRE}_{\text{eig}}$ . . . . .   | 52 |
| 4.6 | This figure shows the ROC curves of methods SVA, (ROTATE)LEAPP and EIGENSTRAT in two simulated SNP association studies where the relative risk $R$ is set to be 1.5 and 3 respectively. . . . .   | 55 |
| 4.7 | This figure shows the resemblance among significant gene sets from 16 tissues in the AGEMAP study. We plot $I_\alpha$ versus $U_\alpha$ (from equation (4.31)) increasing $\alpha$ from 0 until $U_\alpha = 700$ . The greatest self-consistency among lists is from LEAPP. EIGENSTRAT is second best. The baseline curve is computed assuming that the rankings for all 16 tissues are mutually independent. . . . .   | 58 |
| 4.8 | This figure shows the histogram of $z$ -values from breast cancer microarray study Hedenfalk (2001), comparing 7 breast cancer patients having BRCA1 mutation to 8 with BRCA2 mutation $N = 3226$ genes. Green solid curve is the fitted mixture density $f$ and blue dashed curve is the fitted null subdensity $p_0 f_0$ and both are output from R package <i>locfdr</i> ). . . . .  | 60 |
| 6.1 | Recovered signal matrix $\hat{S}$ from 4 methods and the true signal matrix simulated in the poisson setting. . . . .   | 82 |
| 6.2 | Chromosome 9p in 44 Leukemia samples . . . . .  | 85 |
| 6.3 | Chromosome 22 . . . . .   | 86 |
| 7.1 | This figure shows the clustering performance between CNVtools and our EM based method. In the left panel, blue line is the principal component direction and red line is orthogonal to the blue line. In the right panel, the blue line is the optimal projection direction and the red line is the direction that splits two clusters found by EM based method. Detected clusters are colored in red and black respectively for CNVtools and our method. . . . . | 93 |

|     |   |    |
|-----|---|----|
| 7.2 | Classification Error: x axis is the error rate for CNVtools and y axis is the error rate for our EM based method . . . . .  | 95 |
| 7.3 | x coordinate: probe intensity of SNP 1; y coordinate: probe intensity of SNP 2. Clusters detected are colored. CNVtools detected 2 clusters (red, black) and our method detected 3 clusters (green, red,black). . . . . | 96 |

# Part I

Multiple hypothesis testing,  
adjusting for latent variables

# Chapter 1

## Introduction

There has been considerable progress in multiple testing methods for high throughput applications. A common example, coming from biology, is testing which of  $N$  genes' expression levels correlate significantly with a scalar variable, which we'll call the primary variable. The primary variable may be an experimentally applied treatment or it may be a covariate such as a phenotype. It is observed and thus known. We will use the gene expression example for concreteness, although it is just one of many instances of this problem.

High throughput experiments may involve thousands or even millions of hypotheses. Because  $N$  is so large, serious problems of multiplicity arise. For independent tests, methods based on the false discovery rate (Dudoit and van der Laan, 2008) have been very successful. Attention has turned more recently to dependent tests (Efron, 2010).

One prominent cause of dependency among test statistics is the presence of latent variables. For example, in microarray-based experiments, it is well known that samples processed in the same batch are correlated. Batch, technician, and other sources of variation in sample preparation can be modeled by latent variables. Another example comes from genetic association studies, where differences in ancestral history among subjects can lead to false or inaccurate associations. Price et al. (2006) used principal components to extract and correct for ancestral history, in effect modeling the genetic background of the subjects as latent variables. A third example comes

from copy number data, where local trends along the genome cause false positive copy number calls (Olshen et al., 2004). Diskin et al. (2008) conducted experiments showing that these local trends correlate with the variation of GC content along the genome, and are caused by differences in the quantity and handling of DNA. These laboratory effects are hard to measure, but can be quantified using a latent variable model. In this part of the dissertation, we assume that primary variable is known and consider latent variables that might even be correlated with the primary variable. Latent variables in copy number data will be discussed in the second part of the dissertation where we consider situations without a known primary variable but with smoothness assumptions on signals.

When the primary variable is an experimentally applied treatment, then problematic latent variables are those that are partially confounded with the treatment. Randomization reduces the effects of such confounding but randomization is not always perfectly applied and batch or other effects may be imbalanced with respect to the treatment (Leek et al., 2010).

These latent variables have some severe consequences. They alter the level of the hypothesis tests and they induce correlations among multiple tests. Another consequence, that we find especially concerning, is that the latent variables may affect the rank ordering among the  $N$   $p$ -values. When high throughput methods are used to identify candidates for further followup it is important that the highly ranked items contain as many non-null cases as possible.

Our approach to this problem uses a rotated model in which we separate the latent variables from the primary variable. We do this by creating two data sets, one in which both primary and latent variables are present and one in which the primary variables are absent. We use the latter data set to estimate the latent variables and then substitute their estimates into the former. Since each gene has its own effect size in relation to the primary variable, the former model is supersaturated. We conduct inference under the setting where the parameter vector relating the genes to the primary variable is sparse, as is commonly assumed in multiple testing situations. Each non-null hypotheses behaves as an additive outlier, and we then apply an outlier detection method from She and Owen (2011) to find them. We call the method

LEAPP, for *latent effect adjustment after primary projection*.

Chapter 2 reviews large scale multiple hypotheses and latent factor analysis, which are useful for understanding later chapters. Chapter 3 presents our notation and introduces several existing models, including SVA (Leek and Storey, 2008) and EIGENSTRAT (Price et al., 2006), to which we make comparisons. Chapter 4 presents the LEAPP method and shows via simulation and real data examples that LEAPP generates better rankings of the non-null hypotheses than one would get by either ignoring the latent variables, by SVA, or by EIGENSTRAT.

# Chapter 2

## Background

In this chapter, we review large-scale multiple comparison methods in section 2.1 and latent factor model inference in section 2.2 to prepare for further studies in later chapters.

### 2.1 Large Scale Multiple Comparison

Multiple comparisons problem occurs when a set of statistical inferences are simultaneously considered. An unguarded use of single-inference procedures leads to a greatly increased false positive rate. It is particularly destructive in the current era of scientific mass production, as often thousands of cases are considered simultaneously to draw a conclusion. To control the multiplicity effect, classical multiple comparison procedures aim to control the probability of committing any type I error in families of comparisons under simultaneous consideration. Let us now consider the general problem of simultaneously testing a finite numbers of hypotheses  $H_i$  ( $i = 1, \dots, M$ ) and the corresponding p-values are available and denoted as  $p_i$  ( $i = 1, \dots, M$ ). The probability of one or more false rejection is called *family-wise error rate* (FWER) and we shall require that

$$\text{FWER} \leq \alpha$$

for all possible constellations of true and false hypotheses. To control FWER, recent research advances Bonferroni-type procedures such as Simes (1986), Hochberg (1988) and Rom (1990). Those procedures assume independence among test statistics, control the FWER in the strong sense and often tend to have substantially less power than the per comparison procedure of the same levels. Westfall and Young (1993) provide resampling-based multiple testing procedures for controlling the FWER. Resampling methods are suggested to capture the impact of dependence on the joint null distribution of the test statistics but they are shown to have shortcomings in controlling the Type-I error rate (Yifan et al. (2006), Xu and Hsu (2007)).

Benjamini and Hochberg (1995) suggests a new point of view that the number of erroneous rejections should be taken into account and not only the question whether any error was made. They proposed the false discovery rate (FDR), which is shown to admit more powerful procedure. Let  $\mathcal{R}$  be the total number of rejections and let  $\mathcal{F}$  be the number of false rejections, i.e., the number of rejections among the  $\mathcal{R}$  rejections corresponding to true null hypotheses. Define  $Q$  to be  $\mathcal{F}/\mathcal{R}$  (and defined to be 0 if  $\mathcal{R} = 0$ ). Thus  $Q$  is the proportion of rejected hypotheses that are rejected erroneously. The *false discovery rate* (FDR) is

$$\text{FDR} = \mathbb{E}(Q).$$

The paper mentioned above has sparked many discussions about possible improvement of the initial proposition, hereafter referred to as the BH procedure. Let us consider a two-groups model. Suppose that the  $N$  cases (“genes” if in the microarray studies) are each either *null* or *non-null* with prior probability  $p_0$  or  $p_1 = 1 - p_0$  and each is represented by its own  $z$  value “ $z_i$ ”, for  $i = 1, \dots, N$ , with density either  $f_0(z)$  or  $f_1(z)$ :

$$\begin{aligned} p_0 &= \mathbb{P}(\text{null}) & f_0(z) & \text{density if null} \\ p_1 &= \mathbb{P}(\text{non-null}) & f_1(z) & \text{density if non-null.} \end{aligned} \tag{2.1}$$

The  $z_i$ 's, theoretically, should yield standard  $\mathcal{N}(0, 1)$  normal distributions under a classical null hypothesis. For example, the  $z$ -values can be obtained by transforming

the usual two-sample  $t$  statistic  $t_i$  comparing the 2 groups' expression levels for gene  $i$ , to a standard normal scale via

$$z_i = \Phi^{-1}(F(t_i)).$$

Here  $\Phi$  and  $F$  are the cumulative distribution functions of standard normal and  $t$  distributions with the same degrees of freedom as  $t_i$ . Let  $F_0(z)$  and  $F_1(z)$  denote the cumulative distribution functions (cdf) of  $f_0(z)$  and  $f_1(z)$  in (2.1) and define the mixture cdf  $F(z) = p_0F_0(z) + p_1F_1(z)$ . Then Bayes rule yields the *a posteriori* probability of a gene being in the null group of (2.1) given that its  $z$ -value  $Z$  is less than some threshold  $z$ , say “Fdr( $z$ )”, as

$$Fdr(z) \equiv \mathbb{P}(\text{null} | Z > |z|) = p_0F_0(z)/F(z).$$

Benjamini and Hochberg (1995)'s false discovery rate control rule begins by estimating  $F(z)$  with the empirical cdf

$$\bar{F}(z) = \#\{z_i \leq z\}/N$$

yielding  $\bar{F}dr(z) = p_0F_0(z)/\bar{F}(z)$ . The rule selects a control level “ $q$ ”, say  $q = 0.1$ , and then declares as non-null those genes having  $z$ -values  $z_i$  satisfying  $z_i \leq z_0$ , where  $z_0$  is the maximum value of  $z$  satisfying

$$\bar{F}dr(z_0) \leq q. \tag{2.2}$$

Assume that  $p_0 \geq 0.9$ , which is true in many circumstances. Usually  $p_0$  is taken to be 1 and  $F_0(z)$  the theoretical null, i.e. the standard normal cdf  $\Phi(z)$ . Benjamini and Hochberg (1995) shown that the expected proportion of null genes reported following rule (2.2) will be no greater than  $q$ . Their method assumes independence among  $z_i$ 's and it is essentially equivalent to declaring non-null those genes whose estimated tail-area posterior probability of being null is no greater than  $q$ .

As densities are more natural than tail areas for Bayesian fdr interpretation, Efron

et al. (2001) introduced *local false discovery rate* as an empirical Bayes version of Benjamini and Hochberg (1995) focusing on densities rather than tail areas. To estimate the local false discovery rate, Efron (2008) proposed an empirical Bayes based algorithm, which will be detailed below. Define the mixture density from (2.1).

$$f(z) = p_0 f_0(z) + p_1 f_1(z).$$

Bayes rule gives

$$\text{fdr}(z) \equiv \mathbb{P}(\text{null} | Z = z) = p_0 f_0(z) / f(z)$$

for the probability of a gene being in the null group given  $z$ -values  $z$ . Here  $\text{fdr}(z)$  is the *local false discovery rate*. To calculate the local false discovery rate, Efron (2005) set  $f_0(z)$  to the theoretical null distribution

$$f_0(z) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2},$$

took  $p_0 = 1$  and estimated  $f(z)$  by the algorithm *locfdr*, (R function available from the CRAN library) through standard Poisson GLM software. When theory provides the only information available for null behavior, we have no choice but the use of the theoretical null. However, in large-scale simultaneous testing situations, serious defects in the theoretical null may become obvious (for example, unobserved covariates, correlations across arrays and genes), while empirical Bayes methods can provide more realistic null distribution. Efron (2005) proposed using an estimated empirical null distribution  $\mathcal{N}(\delta, \sigma^2)$  instead of the theoretical null  $\mathcal{N}(0, 1)$  in the calculation and got totally different true discoveries from using theoretical null for BRCA data of Hedenfalk (2001).

Dependence among test statistics remains an issue (see Efron (2007), Gordon et al. (2007) and Dudoit and van der Laan (2008) for a comprehensive review). Recent studies suggest that high correlations among test statistics affect a strong control of the actual proportion of false discoveries. Although current methods of multiple comparison are generally shown to control expected Type-I error rates, they suffer from high instability in the presence of correlation. For example, for a high amount of

dependence, the BH thresholding method tends to over-control FDR leading to more conservative rules than expected under the assumption of independence. Many papers have focused on the control of FDR under various patterns of dependence between test statistics. Benjamini and Yekutieli (2001) showed the BH procedure still controls the FDR under positive regression dependency on each of the test statistics corresponding to the true null hypotheses. For all other forms of dependency, a simple conservative modification of the procedure controls the false discovery rate. Efron (2008) proposed the aforementioned empirical null method to account for dependence. The above methods more or less tackle the dependence through the test statistics alone instead of utilizing the entire data matrix across samples and variables.

Apart from FWER and FDR controlling method, recent literatures have also advanced gFWER-controlling procedures, including the marginal single-step and step-down procedures of Lehmann and Romano (2005), the joint single-step procedures of Dudoit et al. (2004) and joint resampling-based empirical Bayes approach of van der Laan et al. (2006). Dudoit and van der Laan (2008) introduce a general statistical framework for multiple hypothesis testing for controlling a range of Type I error rates that are broadly defined as parameter  $\Theta(F_{\mathcal{F}_n, \mathcal{R}_n})$  of the joint distribution  $F_{\mathcal{F}_n, \mathcal{R}_n}$  of the number of Type I errors  $\mathcal{F}_n$  and the rejected hypotheses  $\mathcal{R}_n$ . Their procedure featured *test statistics null distribution* (rather than a data generating null distribution) used to obtain rejection regions for the test statistics, confidence regions for the parameter and adjusted p-values (Dudoit et al. (2004); van der Laan et al. (2004b); van der Laan et al. (2006)). Procedures that take into account the joint distribution of the test statistics may have better power than those based solely on the marginal distribution of the test statistics (see Pollard and van der Laan (2004)). Their procedures provide asymptotic control of the Type I error rate for arbitrary dependence under asymptotic null domination assumption, i.e., the number of Type I errors  $\mathcal{F}_n$  under the true distribution for the test statistics is asymptotically stochastically smaller than the corresponding number of Type I errors  $\mathcal{F}_0$  under the assumed null distribution.

In contrast to FDR-controlling approaches, that focus on the *expected value* of the proportion of false positives among the rejected hypotheses, Lehmann and Romano (2005) and Romano and Wolf (2005) propose procedures that control *tail probability*

for this proportion. The key idea is that FDR-controlling approaches control the proportion of false positives on average but they do not preclude large variations in the proportion of false positives. Their procedures control the tail probability for the proportion of false positives (TPFP) among the rejected hypotheses. Those procedures rely on a number of assumptions concerning the joint distribution of the test statistics, such as, independence, positive regression dependence, ergodic dependence, or normality. Augmentation procedure of van der Laan et al. (2004a) provides exact asymptotic control of the TPFP and allows general dependence among test statistics but it may lack power in finite samples.

## 2.2 Latent Factor Model of Large Dimension

A factor model has the following representation:

$$\Xi = \Lambda F^T + E. \tag{2.3}$$

|                                       |                            |
|---------------------------------------|----------------------------|
| $\Xi \in \mathbb{R}^{N \times n}$     | response values            |
| $\Lambda \in \mathbb{R}^{N \times k}$ | latent, nonrandom rows     |
| $F \in \mathbb{R}^{n \times k}$       | latent, independent rows   |
| $E \in \mathbb{R}^{N \times n}$       | idiosyncratic noise matrix |

with dimensions

|            |                     |
|------------|---------------------|
| $n$        | number of samples   |
| $N$        | number of variables |
| $k \geq 1$ | latent dimension.   |

Let  $\Xi_j = (\Xi_{1j}, \dots, \Xi_{Nj}), j = 1, \dots, n$  be  $N \times 1$  vectors,  $F_j = (F_{j1}, \dots, F_{jk}), j = 1, \dots, n$  be the  $k \times 1$  vectors and  $\Lambda_i = (\Lambda_{i1}, \dots, \Lambda_{ik}), i = 1, \dots, N$  be the  $k \times 1$  vectors.

Factor models have a wide range of applications. In asset pricing,  $\Xi_{ij}$  represents asset  $i$ 's return in period  $j$ ;  $F_j$  is a vector of factor returns which are common drivers for price variations in period  $j$ ; and  $E_{ij}$  is the idiosyncratic return. In biology,  $\Xi_{ij}$  represents the gene expression value of the  $i$ th gene and  $j$ th sample;  $F_j$  can be batch effect or ancestral factor for the  $j$ th sample.

Classical factor analysis assume that the number of variables  $N$  is fixed and much smaller than the sample size  $n$ , factors are independent and identically distributed (iid) and are independent of noise  $E$ . Components of the idiosyncratic noise matrix  $E_{ij}$  are independently and identically distributed over variables and across samples. Although not essential, normality of  $E_{ij}$  is often assumed and maximum likelihood estimation is used in estimation. In addition, inferential theory is based on the basic assumption that the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (\Xi_i - \bar{\Xi}) (\Xi_i - \bar{\Xi})^\top$$

is  $\sqrt{n}$  consistent for the covariance matrix of  $\Xi_j$ :  $\Sigma = \mathbf{cov}(\Xi_j)$  and the factor loadings  $\Lambda$  can be consistently estimated (Anderson, 1984).

The assumptions are restrictive for many problems, such as in genetic association study and economic analysis. The number of variables ( $N$ ) is often comparable or even larger than the number of samples ( $n$ ). As a result, the distribution of  $\sqrt{n} (\hat{\Sigma} - \Sigma)$  may not be asymptotic normal. The most popular technique for estimating factor models is the principal components (PC) analysis. Some estimation theory of large factor models has been obtained recently. Stock and Watson (1999) studied the uniform consistency of estimated factors and derived some rates of convergence for large  $n$  and  $N$ . The rate of convergence was also studied by Bai and Ng (2002). Bai (2003) derived the rate of convergence and the limiting distributions for the estimated factors, factor loadings and common components, estimated by the principal components method. The results were derived under more general assumptions than

classical factor analysis and it allowed time and cross section dependence and weak dependence between factors and idiosyncratic errors, but it required the factors to strongly dominate the idiosyncratic terms. Onatski (2009) considered the asymptotic distribution of the principle component estimator when the factor is “weak” compared to the idiosyncratic part.

We simplify the assumption of Bai (2003) and summarize the result as follows. The method of principal components minimizes

$$V(k) = \min_{\Lambda, F^T F/n = I_k} \frac{1}{Nn} \sum_{i=1}^N \sum_{j=1}^n (\Xi_{ij} - \Lambda_i^T F_j)^2.$$

This problem is identical to maximizing  $\text{tr}(F^T(\Xi^T \Xi)F)$ . The estimated factor matrix, denoted by  $\hat{F}$  is  $\sqrt{n}$  times eigenvectors corresponding to the  $k$  largest eigenvalues of the  $n \times n$  matrix  $\Xi^T \Xi$ , and  $\hat{\Lambda}^T = (\hat{F}^T \hat{F})^{-1} \hat{F}^T \Xi^T = \hat{F}^T \Xi^T / n$  are the corresponding factor loadings. Let  $\|\mathcal{A}\| = (\text{tr}(\mathcal{A}^T \mathcal{A}))^{1/2}$ .

**Assumption 1.**  $\mathbb{E}\|F_j\|^4 \leq M < \infty$ ,  $F_j, j = 1, \dots, n$  are independent and identically distributed and  $\frac{1}{n} F^T F \xrightarrow{p} I_k$  for a  $k \times k$  identity matrix  $I_k$ .

**Assumption 2.**  $\|\Lambda_i\| \leq \bar{\lambda} < \infty$  and  $\|\Lambda^T \Lambda / N - \Sigma_\Lambda\| \rightarrow 0$  for some  $k \times k$  positive definite matrix  $\Sigma_\Lambda$ .

**Assumption 3.** Idiosyncratic noise  $E \sim \mathcal{N}(0, \sigma^2 I_N \otimes I_n)$  and independent of factors  $F$ .

**Assumption 4.** For each  $i$ , as  $n \rightarrow \infty$ ,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n F_j E_{ij} \xrightarrow{d} \mathcal{N}(0, \sigma^2 I_k),$$

**Assumption 5.** The eigenvalues of the  $k \times k$  matrix  $\Sigma_\Lambda$  are distinct.

Assumption 5 guarantees a unique limit for  $\hat{F}^T F/n$ , which appears in the limiting distribution. Otherwise, its limit can only be determined up to orthogonal transformations. This assumption is not needed for determining the number of factors as it

depends on the projection matrix  $\mathcal{P}_{\hat{F}}$  which is invariant to orthogonal transformations.

Let  $\mathcal{V} = \text{diag}(\underline{\Xi}_1, \underline{\Xi}_2, \dots, \underline{\Xi}_k)$ , where  $\underline{\Xi}_1 > \dots > \underline{\Xi}_k > 0$  are the eigenvalues of  $\Sigma_\Lambda$ , and  $\zeta$  is the corresponding eigenvector matrix such that  $\zeta^\top \zeta = I_k$ . We have the following proposition for the consistency of estimated factors.

**Proposition 1.** Under Assumptions 1-5,

$$\text{plim}_{n, N \rightarrow \infty} \frac{\hat{F}^\top F}{n} = \zeta^\top.$$

*Proof.* Under the assumptions 1-5, the assumptions A-D and G in Bai (2003) are satisfied and in particular  $\Sigma_F = I_k$ . By proposition 1 of Bai (2003),

$$\text{plim}_{n, N \rightarrow \infty} \frac{\hat{F}^\top F}{n} = \mathcal{Q}.$$

where  $\mathcal{Q}$  is given by  $\mathcal{Q} = \mathcal{V}^{1/2} \zeta^\top \Sigma_\Lambda^{-1/2}$ . As  $\Sigma$  is positive definite, we have the following:

$$\begin{aligned} \Sigma_\Lambda &= \zeta \mathcal{V} \zeta^\top \\ \Sigma_\Lambda^{-1/2} &= \zeta \mathcal{V}^{-1/2} \zeta^\top \\ \mathcal{Q} &= \mathcal{V}^{1/2} \zeta^\top \Sigma_\Lambda^{-1/2} \\ \mathcal{Q} &= \mathcal{V}^{1/2} \zeta^\top \zeta \mathcal{V}^{-1/2} \zeta^\top \\ &= \zeta^\top. \end{aligned}$$

□

When  $\Sigma_\Lambda$  is a diagonal matrix,  $\zeta$  is simply a  $k \times k$  identity matrix  $I_k$  up to a sign change on the diagonal.

**Lemma 1.** Under Assumptions 1-5, as  $n, N \rightarrow \infty$ :

$$\mathcal{V}_{Nn} = n^{-1} \hat{F}^\top \left( \frac{1}{nN} \Xi^\top \Xi \right) \hat{F} \xrightarrow{p} \mathcal{V}.$$

*Proof.* It is a direct result from Lemma A.3 in Bai (2003). □

**Theorem 1.** Under Assumptions 1-5, as  $n, N \rightarrow \infty$ , if  $\sqrt{n}/N \rightarrow 0$ , then for each  $i$ ,

$$\begin{aligned} \sqrt{n} \left( \hat{\Lambda}_i - \zeta^\top \Lambda_i \right) &= \mathcal{V}_{Nn}^{-1} \left( \frac{\hat{F}^\top F}{n} \right) \left( \frac{\Lambda^\top \Lambda}{N} \right) \frac{1}{\sqrt{n}} \sum_{j=1}^n F_j E_{ij} + o_p(1) \\ &\xrightarrow{d} \mathcal{N}(0, \sigma^2 I_k). \end{aligned} \quad (2.4)$$

*Proof.* Let  $H = (\Lambda^\top \Lambda / N) (F^\top \hat{F} / n) V_{Nn}^{-1}$  be a  $k \times k$  matrix. Assumptions 1 and 2 together with  $\hat{F}^\top \hat{F} / n = I_k$ , Proposition 1 and Lemma 1 imply that  $\|H\| = O_p(1)$  and  $\|H^{-1} - \zeta^\top\|_2 \xrightarrow{p} 0$ . We decompose the left hand side of (2.4) as

$$\hat{\Lambda}_i - \zeta^\top \Lambda_i = \left( \hat{\Lambda}_i - H^{-1} \Lambda_i \right) + \left( H^{-1} \Lambda_i - \zeta^\top \Lambda_i \right).$$

According to the proof of Theorem 1 and 2 in Bai (2003),

$$\begin{aligned} \hat{\Lambda}_i - H^{-1} \Lambda_i &= H^\top \frac{1}{n} \sum_{j=1}^n F_j E_{ij} + o_p(1) \\ &\xrightarrow{d} \mathcal{N}(0, \sigma^2 I_k). \end{aligned} \quad (2.5)$$

As  $\|\Lambda_i\| \leq \bar{\lambda} < \infty$ ,  $\|H^{-1} \Lambda_i - \zeta^\top \Lambda_i\| \leq \bar{\lambda} \|H^{-1} - \zeta^\top\|_2 \xrightarrow{p} 0$ . Combine it with the result from (2.5) and the desired limiting distribution is obtained.

□

# Chapter 3

## Models and Notations

In this chapter, we introduce a specific statistical model of the latent-variable induced dependence structure in the context of multiple hypotheses in section 3.1. From there we move to review some recently proposed approaches such as SVA and EIGENSTRAT in section 3.2. Since latent factor modeling in multiple hypotheses has been an exciting and popular research area, many papers have been written over the past few years. We do not intend to give a full literature review, and only selectively include a few algorithms that are representative. This is to prepare for the relevant comparisons in Chapter 4.

### 3.1 Statement of The Problem

Now we return to the latent systemic effects in multiple hypotheses problem. The data we observe are a response matrix  $Y \in \mathbb{R}^{N \times n}$  and a variable of interest  $g \in \mathbb{R}^n$ , which we call the primary variable. In an expression problem  $Y_{ij}$  is the expression level of gene  $i$  for subject  $j$ . Very often the primary variable  $g$  is a group variable taking just two values, such as  $\pm 1$  for a binary phenotype, then linearly transformed to have mean 0 and norm 1. The quantity  $g_j$  can also be a more general scalar, such as the age of subject  $j$ .

We are interested to know which genes, if any, are linearly associated with the variable  $g$ . We capture this linear association through the  $N \times n$  matrix  $\gamma g^T$  where  $\gamma$

is a vector of  $N$  coefficients. When most genes are not related to  $g$ , then  $\gamma$  is sparse.

Often there are covariates  $X$  other than  $g$  that we should adjust for. The covariate term is  $\beta X^T$  where  $\beta$  contains coefficients. The latent variables that cause tests to be mutually correlated are assumed to take an outer product form  $UV^T$ . Neither  $U$  nor  $V$  is observed. Finally, there is observational noise with a variance that is allowed to be different for each gene, but assumed to be constant over subjects.

The full data model is

$$Y = \gamma g^T + \beta X^T + UV^T + \Sigma E \quad (3.1)$$

for variables

|   |   |
|---|---|
| $Y \in \mathbb{R}^{N \times n}$                   | response values                                 |
| $g \in \mathbb{R}^{n \times r}$                   | primary predictor, e.g. treatment               |
| $\gamma \in \mathbb{R}^{N \times r}$              | primary parameter, possibly sparse              |
| $X \in \mathbb{R}^{N \times s}$                   | $s$ covariates (e.g. sex) per subject           |
| $\beta \in \mathbb{R}^{N \times s}$               | $s$ coefficients, including per gene intercepts |
| $U \in \mathbb{R}^{N \times k}$                   | latent, nonrandom rows (e.g. genes)             |
| $V \in \mathbb{R}^{n \times k}$                   | latent, random rows (e.g. subjects)             |
| $E \sim \mathcal{N}(0, I_N \otimes I_n)$          | noise, and,                                     |
| $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_N)$ | standard deviations,                            |

with dimensions

|            |   |
|------------|---|
| $n$        | number of arrays/subjects               |
| $N \gg n$  | number of genes                         |
| $s \ll n$  | number of covariates                    |
| $k \geq 1$ | latent dimension, often one, and        |
| $r \geq 1$ | primary parameter dimension, often one. |

After adjusting for  $X$ , the genes are correlated through the action of the latent portion  $UV^T$  of the model. They may have unequal variances, through both  $\Sigma$  and  $U$ . It is possible to generalize the model to have a primary variable  $g$  of dimension  $r \geq 1$  but we focus on the case with  $r = 1$ .

We pay special attention to the case of  $k = 1$  latent variable. When  $k = 1$ , the dependence between the variable  $g$  of interest and the latent variable  $V$  can be summarized by a single correlation coefficient  $\rho = \frac{g^T V}{\|g\| \|V\|}$ .

Writing (3.1) in terms of indices yields

$$Y_{ij} = \gamma_i g_j + \beta_i^T X_j + U_i^T V_j + \sigma_i \varepsilon_{ij}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n. \quad (3.2)$$

Here  $\beta_i$  and  $U_i$  are the  $i$ 'th rows of  $\beta$  and  $U$  respectively, while  $X_j$  and  $V_j$  are the  $j$ 'th rows of  $X$  and  $V$ ,  $\sigma_i$  is the  $i$ 'th diagonal element of  $\Sigma$  and  $\varepsilon_{ij}$  is the  $ij$  element of  $E$ .

## 3.2 Review of Existing Approaches

### 3.2.1 SVA

Leek and Storey (2008) proposed surrogate variable analysis (SVA) method which is an iteratively reweighted surrogate variable analysis algorithm adjusting for latent variables before doing a regression. But it does not isolate them.

A full and precise description of SVA appears in the supplementary information and online software for Leek and Storey (2008). Here we present a brief outline. In our notation, their model is

$$Y = \gamma g^T + UV^T + \Sigma E.$$

The SVA algorithm uses iteratively reweighted singular value decompositions (SVDs) to estimate  $U$ ,  $V$  and  $\gamma$ . The weights are empirical Bayes estimates of  $\Pr(\gamma_i = 0, U_i \neq 0 \mid Y, g, V)$  from Storey et al. (2005). Their method seeks to remove the primary term  $\gamma g^T$  by downweighting rows with  $\gamma_i \neq 0$ . Our method creates columns that are free of the primary variable by rotation.

The SVA iteration is as follows. First, they fit a linear model without any latent variables, getting estimates  $\hat{\gamma}$  and the residual  $R = Y - \hat{\gamma}g^\top$ . Second, they apply the simulation method of Buja and Eyuboglu (1992) to  $R$  to estimate the number  $k$  of factors, and then take the top  $k$  right eigenvectors of  $R$  as the initial estimator  $\hat{V}$ . Third, they form the empirical Bayes estimates  $w_i = \Pr(\gamma_i = 0, U_i \neq 0 \mid Y, g, \hat{V})$  from Storey et al. (2005). Fourth, based on those weights, they perform a weighted singular value decomposition of the original data matrix  $Y$ , where row  $i$  is weighted by  $w_i$ . The weighted SVD gives them an updated estimator  $\hat{V}$ . They repeat steps 3 and 4, revising the weights  $w_i$  and then the matrix  $\hat{V}$ , until  $\hat{V}$  converges. They perform significance analysis on  $\gamma$  through the multivariate linear regression model

$$Y = \gamma g^\top + U\hat{V} + \Sigma E,$$

where  $\hat{V}$  is treated as known covariates to adjust for the primary effect  $g$ .

To estimate the number  $k$  of factors in the SVD, they use a simulation method of Buja and Eyuboglu (1992). That algorithm uses Monte Carlo sampling to adjust for the well known problem that the largest singular value in a sample covariance matrix is positively biased. That method has two parameters: the number of simulations employed and a significance threshold. The default significance threshold was 0.1 and the default uses 20 permutations.

### 3.2.2 EIGENSTRAT

EIGENSTRAT (Price et al., 2006) was developed to control for differences in ancestry in genetic association studies, where the matrix  $Y$  represent the alleles carried by the subjects at the genetic markers (e.g.  $Y_{ij} \in \{0, 1, 2\}$  counts the number of one of the alleles). The primary variable can be case versus control, disease status, or other clinical traits.

In our notation, they begin with a principal components analysis approximating  $Y$  by  $\hat{U}\hat{V}^\top$  for  $\hat{U} \in \mathbb{R}^{N \times k}$  and  $\hat{V} \in \mathbb{R}^{n \times k}$ . Then for  $i = 1, \dots, N$  they test whether  $Y_{i,1:n}$  is significantly related to  $g$  in a regression including the  $k$  columns of  $\hat{V}$ , or equivalently whether the partial correlation of  $Y_{i,1:n}$  on  $g$  adjusted for  $\hat{V}$ , is significant. Although

the data are discrete and the method resembles one for Gaussian data, the results still clearly obtain latent variables showing a natural connection to the geographical region of the subjects' ancestors.

EIGENSTRAT has an apparent weakness. If the signal  $\gamma g^\top$  is large then its presence will corrupt the estimates of  $\hat{U}$  and  $\hat{V}$ . The estimate  $\hat{V}$  will be correlated with the effect  $g$  that we are trying to estimate a coefficient for. Indeed, we find in our simulations of Section 4.4, that EIGENSTRAT performs poorly when the signal is large compared to the latent variable.

EIGENSTRAT also requires the choice of a rank  $k$  for the latent term. Price et al. (2006) describe a default choice of  $k = 10$ . Patterson et al. (2006) apply a spiked covariance model test of Johnstone (2001) using the Tracy-Widom distribution (Tracy and Widom, 1994).

### 3.2.3 Other Methods

A number of other methods have been proposed for this problem, which we have not included in our numerical comparisons. Here we mention several of them, relating their approaches to the notation of Section 3.1.

Friguet et al. (2009) model their data as  $Y = \gamma g^\top + UV^\top + \Sigma E$ . They assume the latent  $V$  is normally distributed (independent of  $E$ ) and that  $U$  is nonrandom. They do not assume sparsity for  $\gamma$ . They estimate  $U$ ,  $V$ ,  $\gamma$  and  $\Sigma$  by an EM algorithm. They find that using  $\hat{V}$  in an FDR procedure is an improvement compared to a model that does not employ latent variables.

Lucas et al. (2010) take  $Y = \beta X^\top + UV^\top + \Sigma E$  and make extensive use of sparsity priors. They include the primary variable  $g$  as one of the columns of  $X$ , instead of singling it out as we do. Under their sparsity priors, a coefficient is either 0 or it is  $\mathcal{N}(0, \tau^2)$ . The probability of a nonzero coefficient is  $\pi$  which in turn has a Beta distribution with a small mean. They apply sparsity priors to the elements of both the coefficient matrix  $\beta$  and the latent variables  $U$ . The parameters  $\pi$  and  $\tau$  are different for each column of  $\beta$ . They use Markov chain Monte Carlo for their inferences.

Allen and Tibshirani (2010) model the data as  $Y = \gamma g^T + E$  where  $E \sim \mathcal{N}(0, \Sigma \otimes \Gamma)$ . That is, the noise covariance is of Kronecker form which models dependence between rows and between columns. Our model has a different variance equal to the sum of two Kronecker matrices, one from  $UV^T$  and one from  $\Sigma E$ . They estimate  $\Sigma$  and  $\Gamma$  by maximum likelihood with a penalty on the norm of the inverses of  $\Sigma$  and  $\Gamma$ . Their  $L_1$  penalties encourage sparsity in  $\hat{\Sigma}^{-1}$  and  $\hat{\Gamma}^{-1}$ . They then whiten  $Y$  using  $\hat{\Gamma}$  and  $\hat{\Sigma}$  and apply false discovery rate methods. They also show that correlations among different columns leads to incorrect estimates of FDR while correlated rows do not much affect the estimates of FDR.

Carvalho et al. (2008) consider similar problems but apply a very different formulation. They treat the primary variable (our  $g$ ) as the response and use the data matrix (our  $Y$ ) as predictors.

# Chapter 4

## Latent Effects Adjustment

In this chapter, we propose the LEAPP method to adjust latent effects for multiple hypothesis testing and evaluate its performance. Section 4.1 presents the two-stage LEAPP method in detail. It consists of a special rotation and outlier detection. Section 4.2 discusses the conditions under which the general model is identifiable such that the decomposition into a sum of a sparse signal matrix and a low rank latent effect matrix is unique. Section 4.3 contains some theoretical results about the method: The specific rotation matrix used does not affect our answer. For the case of one latent variable and no covariates, LEAPP consistently estimates the latent structure. We also get a bound for the sum of squared coefficient errors when the effects are sparse as well as conditions for estimated coefficients to be sign consistent. Section 4.4 shows via simulation that LEAPP generates better rankings of the non-null hypotheses than one would get by either ignoring the latent variables, by SVA, or by EIGENSTRAT. EIGENSTRAT estimates the latent variables (by principal components) without first adjusting for the primary variable. LEAPP outperforms it when the latent variable is weaker than the primary. EIGENSTRAT does well in simulations with weak primary variables, which matches the setting that motivated it. Still it is interesting to learn that it does not extend well to problems with strong primary variables. SVA estimates the primary variable's coefficients without first adjusting for correlation between the primary and latent variables. LEAPP outperforms it when the latent and primary variables are correlated. Section 4.5 compares the methods on the AGEMAP data

of Zahn et al. (2007) and the breast cancer data of Hedenfalk (2001). In the AGEMAP data, the primary variable is age. While we don't know the truly non-null genes for this problem, we have a proxy. The data set has 16 subsets, each from one tissue type. We find that LEAPP gives gene lists with much greater overlap among tissues than the gene lists achieved by the other methods. In the breast cancer data, the primary variable is cancer type. Though we don't know the true non-null genes for this problem as well, we estimate the empirical null distribution according to Efron (2008) and find that the estimated empirical null distribution of z-scores adjusted by LEAPP is closer to the theoretical null than other methods. Our conclusions are in Section 4.6. Some background material are given in the appendix.

## 4.1 LEAPP Method

In this section, we present the LEAPP method. Subsection 4.1.1 reduces the model by a special rotation to an outlier detection and robust regression problem. Subsection 4.1.3 discusses in detail how we can choose the number of latent factors.

### 4.1.1 Model Reduction

Our LEAPP (latent effects adjustment after primary projection) proposal is based on a series of reductions described here. First we choose an orthogonal matrix  $O \in \mathbb{R}^{n \times n}$  such that  $g^T O^T = (\eta, 0, 0, \dots, 0) \in \mathbb{R}^{1 \times n}$  where  $\eta = \|g\| > 0$ . Without loss of generality, we assume that the primary predictor has been scaled so that  $\eta = 1$ .

Using  $O$  we construct the **rotated model**

$$\tilde{Y} \equiv Y O^T = \gamma g^T O^T + \beta X^T O^T + UV^T O^T + \Sigma E O^T \quad (4.1)$$

$$\equiv \gamma \tilde{g}^T + \beta \tilde{X}^T + U \tilde{V}^T + \Sigma \tilde{E}, \quad (4.2)$$

where  $\tilde{g}$ ,  $\tilde{X}$ ,  $\tilde{V}$  and  $\tilde{E}$  are rotated versions of their counterparts without the tilde. Notice that  $\tilde{E} = E O^T \stackrel{d}{=} E$ , because  $E \sim \mathcal{N}(0, I_N \otimes I_n)$ . By construction,  $\tilde{g}^T = (1, 0, \dots, 0)$ . Therefore the model for  $\tilde{Y}_{ij}$  is different depending on whether  $j = 1$  or

$j \neq 1$ :

$$\tilde{Y}_{i1} = \beta_i^\top \tilde{X}_1 + U_i^\top \tilde{V}_1 + \gamma_i + \sigma_i \varepsilon_{i1}, \quad \text{and} \quad (4.3)$$

$$\tilde{Y}_{ij} = \beta_i^\top \tilde{X}_j + U_i^\top \tilde{V}_j + \sigma_i \varepsilon_{ij}, \quad j = 2, \dots, n. \quad (4.4)$$

The rotated model concentrates the primary coefficients  $\gamma_i$  in the first column of  $\tilde{Y}$ . Our approach is to base tests and estimates of  $\gamma_i$  on equation (4.3). We need to substitute estimates for unknown quantities  $\sigma_i$ ,  $\beta_i$  and  $U_i$  in (4.3). The estimates come from the model in equation (4.4).

This rotated approach has some practical advantages: First, we do not need to iterate between applying equations (4.3) and (4.4). Instead we use (4.4) once to estimate unknowns  $U$ ,  $\sigma$  and  $\beta$  and then use (4.3) once to judge  $\gamma_i$ . Second, the last  $n - 1$  columns of  $\tilde{Y}$ , and hence estimates  $\hat{\sigma}$ ,  $\hat{\beta}$ , and  $\hat{U}$ , are statistically independent of the first column. Third, problems (4.3) and (4.4) closely match settings for which there are usable methods as described next.

Assume that estimates  $\hat{\sigma}_i$ ,  $\hat{U}_i$ , and  $\hat{\beta}_i$  from (4.4) are given. We may then write (4.3) as

$$\tilde{Y}_{i1} - \hat{\beta}_i^\top \tilde{X}_1 = \hat{U}_i^\top \tilde{V}_1 + \gamma_i + \hat{\sigma}_i \varepsilon_{i1}. \quad (4.5)$$

The right hand side of equation (4.5) is a regression with measurement errors in the predictors  $\hat{U}_i$ , mean-shift outliers  $\gamma_i$  and unequal error variances. We use the  $\Theta$ -IPOD algorithm of She and Owen (2011), adjusting it to handle unequal  $\sigma_i$ , to estimate  $\gamma_i$ . See Appendix A.1.

We don't know  $\beta_i$ ,  $U_i$  and  $\sigma_i$ , but we may estimate them from the data for  $j \geq 2$ . In the process we will also estimate  $\tilde{V}_j$  for  $j \geq 2$ . Let  $\bar{X}$ ,  $\bar{Y}$ ,  $\bar{V}$  and  $\bar{E}$  be the last  $n - 1$  columns of  $\tilde{X}$ ,  $\tilde{Y}$ ,  $\tilde{V}$  and  $\tilde{E}$ , respectively. Then the model for the last  $n - 1$  columns of the data is

$$\bar{Y} = \beta \bar{X}^\top + U \bar{V}^\top + \Sigma \bar{E}.$$

We adopt an iterative approach, where to update  $\hat{\sigma}_i$  from the other variables, we

take

$$\hat{\sigma}_i^2 = \frac{1}{n-1} \sum_{j=2}^n (\tilde{Y}_{ij} - \hat{\beta}_i^\top \tilde{X}_j - \hat{U}_i^\top \hat{V}_j)^2. \quad (4.6)$$

Given  $\hat{\sigma}_i$  we write the model as

$$Y^* = \beta^* \bar{X}^\top + U^* \bar{V}^\top + \bar{E} \quad (4.7)$$

where  $Y^* = \hat{\Sigma}^{-1} \bar{Y}$ ,  $\beta^* = \hat{\Sigma}^{-1} \beta$ , and  $U^* = \hat{\Sigma}^{-1} U$ . We use the criss-cross regression of Gabriel and Zamir (1979) (see Appendix A.2) to estimate  $\beta^*$ ,  $U^*$  and  $\bar{V}$  and then multiply those estimates by  $\hat{\Sigma}$  to get  $\hat{\beta}$  and  $\hat{U}$ . To start the iteration, we set  $\hat{\sigma}_i = 1$ , for  $i = 1, \dots, N$ .

To fit our model, we need to choose the rank  $k$  for the latent term  $UV^\top$ . We follow Leek and Storey (2008) in using the method of Buja and Eyuboglu (1992), as described in Section 3.2.1.

Given estimates  $\hat{\sigma}_i$  we apply  $\Theta$ -IPOD to the regression

$$\frac{\tilde{Y}_{i1} - \hat{\beta}_i^\top \tilde{X}_1}{\hat{\sigma}_i} = \frac{\hat{U}_i^\top}{\hat{\sigma}_i} \tilde{V}_1 + \frac{\gamma_i}{\hat{\sigma}_i} + \frac{\sigma_i}{\hat{\sigma}_i} \varepsilon_{i1}. \quad (4.8)$$

We write the final form of this model as

$$\underline{Y}_i = \underline{U}_i^\top \tilde{V}_1 + \underline{\gamma}_i + \underline{\varepsilon}_i, \quad (4.9)$$

for response  $\underline{Y}_i = (\tilde{Y}_{i1} - \hat{\beta}_i^\top \tilde{X}_1)/\hat{\sigma}_i$ , predictors  $\underline{U}_i = \hat{U}_i/\hat{\sigma}_i$ , unknown coefficients  $\tilde{V}_1$ , additive outliers  $\underline{\gamma}_i = \gamma_i/\hat{\sigma}_i$  and errors  $\underline{\varepsilon}_i = \varepsilon_{i1}\sigma_i/\hat{\sigma}_i$ .

Our statistic for testing  $H_{i0} : \gamma_i = 0$  is

$$T_i = \frac{\underline{Y}_i - \underline{U}_i^\top \hat{V}_1}{\hat{\tau}}, \quad (4.10)$$

where  $\hat{V}_1$  is the  $\Theta$ -IPOD estimate of  $\tilde{V}_1$  and  $\hat{\tau}$  is an estimate of the error standard deviation  $\tau$  from (4.9). We use the median absolute deviation from the median (MAD) to estimate  $\hat{\tau}$ . For  $p$ -values we use  $\Pr(|Z| \geq |T_i|)$  where  $Z \sim \mathcal{N}(0, 1)$ . Candidate

hypotheses are ranked from most interesting to least interesting by taking the  $p$  values from smallest to largest.

### 4.1.2 Alternatives for non-sparse $\gamma$

We have emphasized the setting in which  $\gamma$  is a sparse vector. When  $\gamma$  is not a sparse vector, then its large components may not be flagged as outliers because the MAD estimate of  $\tau$  would be inflated due to contamination by  $\gamma$ . In this case however we can fall back on a simpler approach to estimating  $\tau$ . The error  $\underline{\varepsilon}_i$  has variance  $\mathbb{E}(\sigma_i^2/\hat{\sigma}_i^2)$ . This variance differs from unity only because of estimation errors in  $\hat{\sigma}_i$ . We can then use  $\tau^2 = 1$ . We can account for fitting  $s$  regression parameters to the  $n - 1$  samples in each row of  $\bar{Y}$  by taking  $\tau^2 = \mathbb{E}((n - 1 - s)/\chi_{n-1-s}^2) = (n - s - 1)/(n - s - 3)$ . A further approximate adjustment for estimating  $k$  latent vectors is to take  $\tau^2 = (n - s - k - 1)/(n - s - k - 3)$ . This estimate of  $\tau$  can be used in (4.10) for ranking of hypotheses if  $\gamma$  is not suspected to be sparse.

Now suppose  $\gamma$  is not sparse but  $\tau$  is known. We can instead use ridge regression to fit the additive outlier model (4.9) as follows.

$$L(\tilde{V}_1, \underline{\gamma}) = \sum_{i=1}^N \left( \underline{Y}_i - \underline{U}_i^T \tilde{V}_1 - \underline{\gamma}_i \right)^2 + \lambda \sum_{i=1}^N \underline{\gamma}_i^2 \quad (4.11)$$

for regression with additive outliers  $\underline{\gamma}$ , and tuning parameter  $\lambda > 0$ . As we know the true value of error variance  $\tau^2$ , we can use this information to select the tuning parameter  $\lambda$  by the method of moments.

We can first fix  $\underline{\gamma}$  and minimize the criterion over  $\tilde{V}_1$  and then minimize the criterion over  $\underline{\gamma}$ . Let the regression hat matrix be  $H = \underline{U}(\underline{U}^T \underline{U})^{-1} \underline{U}^T$  and the solution to (4.11) is

$$\hat{\underline{\gamma}} = (1 + \lambda)^{-1} (I_N - H) \underline{Y},$$

and the residual is then

$$\underline{Y} - \underline{U} \hat{V}_1 - \hat{\underline{\gamma}} = \frac{\lambda}{1 + \lambda} (I_N - H) \underline{Y} = \frac{\lambda}{1 + \lambda} \hat{\underline{\varepsilon}} \quad (4.12)$$

where  $\hat{\mathcal{E}}$  is the vector of residuals from an ordinary regression of  $\underline{Y}$  on  $\underline{U}$ . Let the residual mean square of  $\hat{\mathcal{E}}$  be  $\hat{\tau}_{\mathcal{E}}^2$ . Suppose the variance of error is known to be  $\tau = 1$ . We can solve for  $\lambda$ , based on the method of moments,

$$\left(\frac{\hat{\tau}_{\mathcal{E}}\lambda}{1+\lambda}\right)^2 = 1,$$

that is  $\lambda = 1/(\hat{\tau}_{\mathcal{E}} - 1)$  for  $\hat{\tau}_{\mathcal{E}} > 1$ . When  $\hat{\tau}_{\mathcal{E}} \leq 1$ , the solution is degenerate and we may conclude that none of the observations are outliers.

Though we have a method for non-sparse  $\gamma$ , we will emphasize the sparse case in the following discussions.

### 4.1.3 Rank Estimation

The problem of choosing the number  $k$  of latent variables is a difficult one that arises for all the methods we used. The Tracy-Widom strategy is derived for the case with  $\Sigma = \sigma I_N$  while our motivating applications have heteroscedasticity.

Even for  $\Sigma = \sigma I_N$  it is known that the best rank for estimating  $UV^T$  is not necessarily the true rank. There is a well known threshold strength below which a factor is not detectable and Perry (2009) shows that there is a still higher threshold below which estimating that factor worsens the estimate of  $UV^T$ . Owen and Perry (2009) present a cross-validatory estimate for the rank  $k$  and Perry (2009) shows how to tune it to choose a rank  $k$  that gives the best reconstruction as measured by Frobenius norm.

In our numerical comparisons, LEAPP, SVA and EIGENSTRAT were all given the same rank  $k$  to use. Sometimes  $k$  was fixed at a default value. Other times we used the method of Buja and Eyuboglu (1992).

## 4.2 Identifiability Conditions

In this section we show the necessary and sufficient conditions for identifiability of the model. We focus on a simpler target decomposition of  $Y$  into  $Y = \gamma g^T + UV^T$

where  $\|g\| = 1$ ,  $\gamma$  is  $s$ -sparse and  $UV^\top$  is of rank  $k$ . As we are most interested in  $\gamma$ , possible ambiguity in  $\beta X^\top + UV^\top$  is not of concern. Let

$$\Omega = \Omega(\gamma g^\top) = \{\bar{\gamma} g^\top \in \mathbb{R}^{N \times n} : \text{supp}(\bar{\gamma}) \subseteq \text{supp}(\gamma)\}$$

be the space of matrices of rank 1, with row space spanned by  $g$  and row supports a subset of the row support of  $\gamma g^\top$ . The row support is the set of rows that have at least 1 nonzero entry. Let  $\mathcal{P}_\Omega$  be the orthogonal projector to  $\Omega$  under the inner product  $\langle \mathcal{A}, \mathcal{B} \rangle = \text{tr}(\mathcal{A}^\top \mathcal{B})$ ; this projection is given by

$$\mathcal{P}_\Omega(M) = \bar{\gamma} g^\top, \bar{\gamma}_i = \begin{cases} [Mg]_i, & \text{if } i \in \text{supp}(\gamma) \\ 0, & \text{otherwise} \end{cases}$$

for all  $i \in \{1, 2, \dots, N\}$ . Furthermore, let

$$T = T(UV^\top) := \tag{4.13}$$

$$\{Q_1 + Q_2 \in \mathbb{R}^{N \times n} : \text{range}(Q_1) \subseteq \text{range}(UV^\top) \text{ and } \text{range}(Q_2^\top) \subseteq \text{range}(VU^\top)\} \tag{4.14}$$

be the span of matrices either with row-space contained in that of  $UV^\top$ , or with column-space contained in that of  $UV^\top$ . Let  $\mathcal{P}_T$  be the orthogonal projector to  $T$ . Under the inner product  $\langle \mathcal{A}, \mathcal{B} \rangle = \text{tr}(\mathcal{A}^\top \mathcal{B})$ , Hsu et al. (2011) shows that this projection is given by

$$\mathcal{P}_T(M) = LL^\top M + MRR^\top - LL^\top MRR^\top$$

where  $L \in \mathbb{R}^{N \times k}$  and  $R \in \mathbb{R}^{n \times k}$  are, respectively, matrices of left and right orthonormal singular vectors corresponding to the non-zero singular values of  $UV^\top$  and  $k$  is the rank of  $UV^\top$ .

Before analyzing whether  $(\gamma g^\top, UV^\top)$  can be recovered in general, we ask a simpler question. Suppose we had prior information about the matrices space  $\Omega$  and  $T$ , in

addition to being given  $Y = \gamma g^\top + UV^\top$ , can we uniquely recover  $\gamma g^\top \in \Omega$  and  $UV^\top \in T$  from  $Y$ ? If there exists  $M \in \mathbb{R}^{N \times n} \in T \cap \Omega$ ,  $Y$  can be decomposed as  $(\gamma g^\top + M) + (UV^\top - M)$  where  $\gamma g^\top + M \in \Omega$  and  $UV^\top - M \in T$ . Thus the decomposition is not unique. If there is no such  $M$ , the decomposition is unique. When  $\gamma = \mathbf{0}$ , the decomposition is trivial, so we only consider the case when  $\gamma \neq \mathbf{0}$ .

**Theorem 2.** *Suppose  $\|g^\top R\| < 1$  and  $\min_{\text{supp}(\tilde{\gamma}) \subseteq \text{supp}(\gamma), \|\tilde{\gamma}\|=1} \|(I - LL^\top)\tilde{\gamma}\| > 0$ . Then  $\Omega \cap T = \{\mathbf{0}\}$ .*

*Proof.* Suppose  $\exists M \in \Omega \cap T$ , we must have

$$\mathcal{P}_T(\mathcal{P}_\Omega(M)) = M \quad (4.15)$$

Let

$$\mathcal{P}_\Omega(M) = \bar{\gamma} g^\top, \bar{\gamma}_i = \begin{cases} [Mg]_i, & \text{if } i \in \text{supp}(\gamma) \\ 0, & \text{otherwise} \end{cases}$$

for all  $i \in \{1, 2, \dots, N\}$ . By the definition of  $\mathcal{P}_\Omega, \mathcal{P}_T$  and equation (4.15),

$$\begin{aligned} M &= \mathcal{P}_T(\mathcal{P}_\Omega(M)) = LL^\top \bar{\gamma} g^\top + \bar{\gamma} g^\top RR^\top - LL^\top \bar{\gamma} g^\top RR^\top \\ M &= \bar{\gamma} g^\top - (I - LL^\top) \bar{\gamma} g^\top (I - RR^\top) \end{aligned} \quad (4.16)$$

$$Mg - \bar{\gamma} = (I - LL^\top) \bar{\gamma} g^\top (I - RR^\top) g. \quad (4.17)$$

Left multiply  $\bar{\gamma}^\top$  on both sides of equation (4.17), and we get

$$\bar{\gamma}^\top (I - LL^\top) \bar{\gamma} g^\top (I - RR^\top) g = 0.$$

By assumption,  $g^\top (I - RR^\top) g \neq 0$  and  $\bar{\gamma}^\top (I - LL^\top) \tilde{\gamma} > 0$  for any  $\|\tilde{\gamma}\| = 1$ ,  $\text{supp}(\tilde{\gamma}) \subseteq \text{supp}(\gamma)$ . As a result, we must have  $\bar{\gamma} = \mathbf{0}$ . Substitute  $\bar{\gamma}$  with 0 in the equation (4.16) and we obtain  $M = \mathbf{0}$ .

□

The conditions  $\|g^\top R\| < 1$  and  $\min_{\text{supp}(\tilde{\gamma}) \subseteq \text{supp}(\gamma), \|\tilde{\gamma}\|=1} \|(I - LL^\top)\tilde{\gamma}\| > 0$  are also

necessary. If  $\|g^\top R\| = 1$ , for any  $\bar{\gamma}$  such that  $\text{supp}(\bar{\gamma}) = \text{supp}(\gamma)$ ,  $M = \bar{\gamma}g^\top$  enjoys a row space which is a subset of the row space of  $UV^\top$  and thus  $M \in \Omega \cap T$ . If  $\exists \tilde{\gamma} \neq 0, \text{supp}(\tilde{\gamma}) \subseteq \text{supp}(\gamma)$  such that  $\|(I - LL^\top)\tilde{\gamma}\| = 0$ ,  $M = \tilde{\gamma}g^\top$  enjoys a column space which is a subset of column space of  $UV^\top$  and thus  $M \in \Omega \cap T$ .

### 4.3 Theory

In this section we prove some properties of our approach to testing many hypotheses in the presence of latent variables. We focus on a simpler version of the model that is more tractable:

$$Y = \gamma g^\top + UV^\top + \sigma E \quad (4.18)$$

where  $g \in \mathbb{R}^{n \times 1}, U \in \mathbb{R}^{N \times k}, V \in \mathbb{R}^{n \times k}$  and  $E \sim \mathcal{N}(0, I_N \otimes I_n)$ . Compared to the full model (3.1), equation (4.18) has no covariate term  $X^T$ , and has constant variance  $\Sigma = \sigma I_N$ . This simplification allows us to apply results from the literature to our model. It removes the Monte Carlo based rank estimation step and the alternation between estimating  $\Sigma$  and using the estimate  $\hat{\Sigma}$ . Our algorithm requires the choice of a rotation matrix  $O$  such that  $Og/\|g\| = e_1$ . There are multiple possibilities for this matrix. We show that our algorithm is invariant to the choice of  $O$ .

**Theorem 3.** *Given the number  $k$  of latent factors, our estimates of  $U$  and  $\gamma$  do not depend on the rotation  $O$  used as long as  $Og = e_1$ .*

*Proof.* We just prove the result when there are no known covariates  $X$  and a common noise variance across genes. The proof can be extended to the full model easily. Without loss of generality, we assume  $\|g\| = 1$ . Let  $O_0 \in \mathbb{R}^{n \times n}$  be a fixed orthogonal matrix such with  $O_0g = e_1$ . Suppose  $O \neq O_0$  is any other orthogonal matrix in  $\mathbb{R}^{n \times n}$  such that  $Og = e_1$ . There exists an orthogonal matrix  $P$  such that  $O = PO_0$ . Now

$$e_1 = Og = PO_0g = Pe_1$$

and so the first column of  $P$  is  $e_1$ . As the rest of columns of  $P$  are all orthogonal to

the first column, we have  $P_{1j} = 0, j = 2, \dots, n$ , so we can write

$$P = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & P^* \end{pmatrix},$$

for an orthogonal matrix  $P^* \in \mathbb{R}^{(n-1) \times (n-1)}$ . Let  $\tilde{Y}^0 = YO_0^\top$  and  $\tilde{Y} = YO^\top$ . We can write  $\tilde{Y}$  in terms of  $\tilde{Y}^0$  via

$$\tilde{Y} = YO^\top = YO_0^\top P^\top = \tilde{Y}^0 P^\top = \tilde{Y}^0 \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & P^* \end{pmatrix}. \quad (4.19)$$

From (4.19), we know that the first column of  $\tilde{Y}$  equals that of  $\tilde{Y}^0$ . Let the last  $n - 1$  columns of  $\tilde{Y}$  and  $\tilde{Y}^0$  be  $\bar{Y}$  and  $\bar{Y}^0$  respectively. Then from (4.19), we notice that  $\bar{Y} = \bar{Y}^0 P^{*\top}$ .

As  $P^*$  is an orthogonal matrix, the left singular vectors of  $\bar{Y}$  are the same as those of  $\bar{Y}^0$  up to a sign change. These left singular vectors are the columns of  $\hat{U}$  and  $\hat{U}^0$  respectively. Therefore  $\hat{U} = \hat{U}^0 S$  where  $S = \text{diag}(\pm 1, \dots, \pm 1) \in \mathbb{R}^{k \times k}$ .

The outlier detection algorithm uses  $\tilde{Y}_{i1}$  and  $\hat{U}$  (for rotation  $O$ ) or  $\tilde{Y}_{i1}^0$  and  $\hat{U}^0$  (for rotation  $O^0$ ). Because  $\tilde{Y}_{i1} = \tilde{Y}_{i1}^0$  the only change can be through the choice of  $\hat{U}$  or  $\hat{U}^0 = \hat{U}S$ . The matrix  $\hat{U}$  enters (4.5) through  $\hat{U}^\top \tilde{V}_1 = (\hat{U}^0)^\top S \tilde{V}_1$ . Changing the sign of the  $j$ 'th column of  $\hat{U}$  results in a sign change for the  $j$ 'th estimated coefficient in  $\tilde{V}_1$ , so that  $\hat{\gamma} = \hat{\gamma}^0$ .  $\square$

The following theorem provides a sufficient condition for our estimate  $\hat{U}$  to consistently estimate  $U$  (the angle between  $\hat{U}$  and  $U$  tends to zeros as sample size grows). We study the case with no  $\beta X^\top$  term and we also assume that  $k = 1$  is known. Two scenarios of assumptions (A and B) on  $V$  and  $g$  are discussed.

Scenario A:

$g$  is assumed to be fixed and without loss of generality we set  $\|g\| = 1$ .  $V$  is assumed to be random and there exists a random variable  $\bar{W} \in \mathbb{R}^{(n-1) \times 1}$  such

that  $\bar{W}$  is independent of noise  $E$ ,  $\mathbb{E}(\bar{W}^\top \bar{W}) = 1$  and

$$V = \rho_n g + \sqrt{1 - \rho_n^2} O_0^\top \begin{pmatrix} 0 \\ \bar{W} \end{pmatrix}$$

for an arbitrary orthogonal matrix  $O_0 : O_0 g = e_1$ . The entries of  $V$  are weakly correlated and  $\mathbb{E}(V^\top V) = 1$ .  $\rho_n$  measures the angle between  $V$  and  $g$ .

Scenario B:

Both  $g$  and  $V$  are assumed to be random and

$$\begin{pmatrix} g_i \\ V_i \end{pmatrix} \text{ i.i.d. } \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_n \\ \rho_n & 1 \end{pmatrix} \right)$$

for  $i = 1, \dots, n$ . Also  $g, V$  are independent of noise  $E$ .

We will show that in both scenarios, as long as the latent factor  $U$  is large enough compared to the noise level, we will be able to detect and estimate  $U$  fairly well. Our size measure  $\|U\|^2(1 - \rho^2)$  takes account of the correlation. With a higher  $\rho$ , more of the latent factor is removed from  $\bar{Y}$ .

We measure error by the cosine  $\Phi(\hat{U}, U) = \hat{U}^\top U / (\|\hat{U}\| \|U\|)$  of the angle between  $\hat{U}$  and  $U$ . The estimate  $\hat{U}$  is determined only up to sign. Replacing  $\hat{U}$  by  $-\hat{U}$  causes a change from  $\hat{V}$  to  $-\hat{V}$  and leaves the model unchanged. We only need  $\max(\Phi(\hat{U}, U), \Phi(-\hat{U}, U)) = |\Phi(\hat{U}, U)| \rightarrow 1$  for consistency.

Before we show the condition for the consistency of  $\hat{U}$ , we introduce a consistency result under “strong factor” assumption from proposition 9 of Harding (2009). Consider model (2.3) in Chapter 2 and number of latent factor  $k = 1$ . Suppose Assumptions 1,3,5 are satisfied. Let  $\hat{\Lambda}$  be the principal component estimate for loadings  $\Lambda$ . Let  $\text{Sp}(\mathcal{A})$  be the spectrum of an arbitrary matrix  $\mathcal{A}$ . We present a measure of the spectral gap  $\tilde{\mu}$  (Harding (2009) page 27), which corresponds to the ratio of the minimum eigenvalue of the spectrum due to the factors over the maximum eigenvalue due to the noise term.

**Definition 1.**

$$\tilde{\mu} = \frac{\min (Sp(\lim_{n \rightarrow \infty} \Lambda \Lambda^T))}{\max (Sp(\lim_{n \rightarrow \infty} \frac{1}{n} E E^T))}$$

**Lemma 2** (Harding, 2009). *The degree of inconsistency in the estimates of the factor loadings  $\Lambda$  as  $N/n \rightarrow c \in (0, \infty), n \rightarrow \infty$  is given by*

$$\sqrt{\frac{1 - c/\tilde{\mu}^2}{1 + c/\tilde{\mu}}} \leq |\Phi(\hat{\Lambda}, \Lambda)| \leq 1.$$

If  $\tilde{\mu} \rightarrow \infty$  then the principal component estimate of  $\hat{\Lambda}$  is consistent,  $\Phi(\hat{\Lambda}, \Lambda) \rightarrow 1$ .

**Theorem A 1.** *Suppose that  $Y$  follows the simple model (4.18) with  $k = 1$  latent variable,  $\Sigma = \sigma^2 I_N$ , no  $\beta X^T$  term and  $g, V$  are in scenario A. Assume that  $\|U\|^2(1 - \rho_n^2)/n \rightarrow \infty$  and  $N(n)/n \rightarrow c \in (0, \infty)$  as  $n \rightarrow \infty$ . Let  $\hat{U}$  be our estimator for  $U$  using rank  $k = 1$ . Then  $|\Phi(\hat{U}, U)| \rightarrow 1$  as  $n \rightarrow \infty$  with probability 1.*

*Proof.* By assumption,  $V = \rho_n g + \sqrt{1 - \rho_n^2} W$  and  $W = O_0^T \begin{pmatrix} 0 \\ \bar{W} \end{pmatrix}$ . By construction  $\tilde{g} = O g = (1, 0, \dots, 0)$ . Similar to the proof of theorem 1, we can show that there exists  $P$  such that  $O = P O_0$  and

$$P = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & P^* \end{pmatrix}$$

for an orthogonal matrix  $P^* \in \mathbb{R}^{(n-1) \times (n-1)}$ . Hence

$$\tilde{W} = O W = P O_0 O_0^T (0, \bar{W}^T)^T = \begin{pmatrix} 0 \\ P^* \bar{W} \end{pmatrix}.$$

Let  $\bar{W} = P^* \tilde{W}$ . As  $\tilde{W}$  satisfies  $\mathbb{E}(\tilde{W}^T \tilde{W}) = 1, \tilde{W}^T \tilde{W} \xrightarrow{p} 1$  and  $P^*$  is an orthogonal matrix,  $\bar{W}$  also has  $\mathbb{E}(\bar{W}^T \bar{W}) = 1, \bar{W}^T \bar{W} \xrightarrow{p} 1$ . In this setting our estimator  $\hat{U}$ , from criss-cross regression, is the top left singular vector of  $\bar{Y}$ , multiplied by the top singular value, where

$$\bar{Y} = \sqrt{1 - \rho_n^2} U \bar{W}^\top + \Sigma \bar{E}. \quad (4.20)$$

For this factor model (4.20), spectral gap measure  $\tilde{\mu} = O(\|U\|^2(1 - \rho_n^2)/n)$  by definition. The assumption  $\|U\|^2(1 - \rho_n^2)/n \rightarrow \infty$  implies  $\tilde{\mu} \rightarrow \infty$ . As an immediate result of Lemma 2,  $|\Phi(\hat{U}, U)| \rightarrow 1$  as  $n \rightarrow 1$ .  $\square$

Under scenario B, the result is similar except for a scaling of  $n$ .

**Theorem B 1.** *Suppose that  $Y$  follows the simple model (4.18) with  $k = 1$  latent variable and assumptions in scenario B are satisfied.  $\Sigma = \sigma^2 I_N$ , and no  $\beta X^\top$  term. Assume that  $\|U\|^2(1 - \rho_n^2) \rightarrow \infty$  and  $N(n)/n \rightarrow c \in (0, \infty)$  as  $n \rightarrow \infty$ . Let  $\hat{U}$  be our estimator for  $U$  using rank  $k = 1$ . Then  $|\Phi(\hat{U}, U)| \rightarrow 1$  as  $n \rightarrow \infty$  with probability 1.*

*Proof.* Let  $g^\circ = g/\|g\|$ . By construction,  $Og^\circ = e_1$  and orthogonal matrix  $O$  can be written as  $O = \begin{pmatrix} g^{\circ\top} \\ \bar{O} \end{pmatrix}$  where  $\bar{O}g = 0$ . Hence we have

$$OV = \begin{pmatrix} g^{\circ\top} V \\ \bar{O}V \end{pmatrix}.$$

By assumption,  $\begin{pmatrix} g_i \\ V_i \end{pmatrix}$  i.i.d.  $\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_n \\ \rho_n & 1 \end{pmatrix}\right)$ . Hence

$$V|g \sim \mathcal{N}(\rho_n g, (1 - \rho_n^2)I_n)$$

and since  $\bar{O}g = 0$ ,

$$\bar{O}V|g \sim \mathcal{N}\left(0, (1 - \rho_n^2)I_{n-1}\right).$$

As a result,  $\bar{O}V \sim \mathcal{N}(0, (1 - \rho_n^2)I_{n-1})$  and we can simply write  $\bar{O}V = \sqrt{1 - \rho_n^2} \bar{W}$  for some  $\bar{W} \sim \mathcal{N}(0, I_{n-1})$ . It is clear that  $\mathbb{E}\left(\frac{1}{n-1} \bar{W}^\top \bar{W}\right) = 1$ . In this setting our estimator  $\hat{U}$ , from criss-cross regression, is the top left singular vector of  $\bar{Y}$ , multiplied

by the top singular value, where

$$\bar{Y} = \sqrt{1 - \rho_n^2} U \bar{W}^\top + \Sigma \bar{E}. \quad (4.21)$$

For this factor model (4.21), spectral gap measure  $\tilde{\mu} = O(\|U\|^2(1 - \rho_n^2))$  by definition. By assumption,  $\|U\|^2(1 - \rho_n^2) \rightarrow \infty$  implies  $\tilde{\mu} \rightarrow \infty$ . As an immediate result of Lemma 2,  $|\Phi(\hat{U}, U)| \rightarrow 1$  as  $n \rightarrow 1$ .

□

The consistency for  $k > 1$  can also be proved similarly except that the estimated loadings can be off by a rotation though the space spanned by  $\hat{U}$  will be a consistent estimator for the space spanned by  $U$  such that  $\|U(U^\top U)^{-1}U^\top - \hat{U}(\hat{U}^\top \hat{U})\hat{U}^\top\|_2 \rightarrow 0$  as  $n \rightarrow \infty$ .

Next we give conditions for the final step of LEAPP to accurately estimate  $\gamma$ , that is, for  $\|\hat{\gamma} - \gamma\|$  to be small. To do this we combine results on random matrix theory from Bai (2003) with methods used to prove results on compressed sensing in Candes and Randall (2006). We consider the case of  $k = 1$  latent variable, constant variance noise  $E \sim \mathcal{N}(0, \sigma^2 I_N \otimes I_n)$  and discuss the result in scenario A and B respectively.

In our simulations we found little difference between robust and non-robust versions of the  $\Theta$ -IPOD algorithm. This is not surprising, since our simulations did not place nonzero  $\gamma_i$  preferentially at high leverage points (extreme  $u_i$ ). For our analysis we replace the robust  $\Theta$ -IPOD algorithm by the Dantzig selector for which strong results are available.

Our algorithm was designed assuming that the primary variable  $g$  is not too strongly correlated with the latent variable  $V$ . In our analysis we also impose a separation between the effects  $\gamma$  and the latent quantity  $U$ . Specifically, we assume that  $\gamma$  is  $s$ -sparse and that  $U$  is not.

After the rotation, the first column can be written as

$$y = \tilde{Y}_1 = UV^\top g / \|g\| + \|g\| \gamma + \sigma \epsilon$$

where  $\epsilon \sim \mathcal{N}(0, I_n)$ .

A vector  $x$  is  $s$ -sparse if it has at most  $s$  nonzero components. Following Candès and Randall (2006), we define the sequences  $a_s(A)$  and  $b_s(A)$  as the largest and smallest numbers (respectively) such that

$$a_s(A)\|x\| \leq \|Ax\| \leq b_s(A)\|x\|$$

holds for all  $s$ -sparse  $x$ .

**Theorem A 2.** *Suppose that  $Y$  follows the model (4.18) with  $k = 1$  latent factor and a fixed correlation  $\rho \in (-1, 1)$  between fixed  $g$  and random  $V$  in scenario  $A$ , no  $\beta X^\top$  term and a parameter vector  $\gamma$  that is  $s$ -sparse. Let our estimated  $U$  be  $\hat{U}$  and set  $U^* = \hat{U}/\|\hat{U}\|$ . Writing  $|U_{(1)}^*| \geq |U_{(2)}^*| \geq \dots \geq |U_{(N)}^*|$  for the ordered components of  $U^*$ , assume that there is a constant  $0 < B < 1$  such that*

$$\sum_{i=1}^{2s} (U_{(i)}^*)^2 + \frac{1}{2} \sum_{i=1}^{3s} (U_{(i)}^*)^2 \leq B.$$

and  $\|(I - U^*U^{*\top})(\tilde{Y}_1 - \gamma)\|_\infty \leq \lambda$ . Then the Dantzig estimator  $\hat{\gamma}$ , which minimizes

$$\|\hat{\gamma}\|_1 \quad \text{subject to} \quad \|(I - U^*U^{*\top})(\tilde{Y}_1 - \hat{\gamma})\|_\infty \leq \lambda$$

satisfies

$$\|\hat{\gamma} - \gamma\| \leq \frac{4\sqrt{s}\lambda}{1-B}.$$

*Proof.* We write the first column of the rotated data

$$\begin{aligned} y &= \tilde{Y}_1 = U\rho + \gamma + \sigma\epsilon \\ y &= \hat{U}\rho + \gamma + \sigma\epsilon + (U - \hat{U})\rho \\ y &= \hat{U}\rho + \gamma + \tilde{\epsilon} \end{aligned}$$

By assumption,

$$\|(I - U^*U^{*\top})(\tilde{\epsilon})\|_\infty = \|(I - U^*U^{*\top})(\tilde{Y}_1 - \gamma)\|_\infty \leq \lambda. \quad (4.22)$$

Then the proof of Theorem 3.1 in (Candes 2006) showed that the Dantzig estimator  $\hat{\gamma}$  satisfies

$$\|\hat{\gamma} - \gamma\| \leq 4\sqrt{s} \frac{\lambda}{D}$$

where  $D = 1 - b_{2s}^2(U^{\star\top}) - b_{3s}^2(U^{\star\top})/2 + a_{3s}^2(U^{\star\top})/2$ . As  $U^{\star}$  is a vector, we have for any  $k \geq 1$ ,  $a_k^2(U^{\star\top}) \geq 0$  and  $b_k^2(U^{\star\top}) \leq \sum_{i=1}^k (U_{(i)}^{\star})^2$ . Hence as long as there exists a constant  $0 < B < 1$  such that  $\sum_{i=1}^{2s} (U_{(i)}^{\star})^2 + \frac{1}{2} \sum_{i=1}^{3s} (U_{(i)}^{\star})^2 \leq B$ , then  $D \geq 1 - B$  and

$$\|\hat{\gamma} - \gamma\| \leq 4\sqrt{s} \frac{\lambda}{1 - B}$$

and thus our result is proved.  $\square$

Similarly, we can obtain the condition for Scenario B.

**Theorem B 2.** *Suppose that  $Y$  follows the model (4.18) with  $k = 1$  latent factor and a fixed correlation  $\rho \in (-1, 1)$  between random  $g$  and random  $V$  in scenario B, no  $\beta X^\top$  term and a parameter vector  $\gamma$  that is  $s$ -sparse. Let our estimated  $U$  be  $\hat{U}$  and set  $U^{\star} = \hat{U}/\|\hat{U}\|$ . Writing  $|U_{(1)}^{\star}| \geq |U_{(2)}^{\star}| \geq \dots \geq |U_{(N)}^{\star}|$  for the ordered components of  $U^{\star}$ , assume that there is a constant  $0 < B < 1$  such that*

$$\sum_{i=1}^{2s} (U_{(i)}^{\star})^2 + \frac{1}{2} \sum_{i=1}^{3s} (U_{(i)}^{\star})^2 \leq B.$$

and  $\|(I - U^{\star}U^{\star\top})(\tilde{Y}_1 - \|g\|\gamma)\|_{\infty} \leq \lambda$ . Then the Dantzig estimator  $\hat{\gamma}$ , which minimizes

$$\|\hat{\gamma}\|_1 \quad \text{subject to} \quad \|(I - U^{\star}U^{\star\top})(\tilde{Y}_1 - \|g\|\hat{\gamma})\|_{\infty} \leq \lambda$$

satisfies

$$\|g\| \|\hat{\gamma} - \gamma\| \leq \frac{4\lambda\sqrt{s}}{1 - B}.$$

*Proof.* We write the first column of the rotated data

$$\begin{aligned} y &= \tilde{Y}_1 = Ug^T V / \|g\| + \|g\| \gamma + \sigma \epsilon \\ y &= \hat{U} g^T V / \|g\| + \|g\| \gamma + \sigma \epsilon + \sqrt{n}(U - \hat{U}) \frac{g^T V}{\sqrt{n} \|g\|} \\ y &= \hat{U} g^T V / \|g\| + \|g\| \gamma + \tilde{\epsilon}. \end{aligned}$$

The first column in the Scenario B is similar to that of Scenario A, except that instead of  $\rho$  in the proof of Theorem A 2, we have  $g^T V / \|g\|$ . And also because of the change of scaling from Theorem A 2, we have an  $\|g\|$  in front of  $\gamma$ . The rest of the proof is essentially the same as that of Theorem A 2. □

**Corollary 1.** *Suppose  $U/\|U\|$  is uniformly distributed on the set of vectors of unit length in  $\mathbb{R}^N$  independent of noise  $\epsilon$ ,  $V$  and  $g$  are in Scenario A and  $U^T U / (Nn) \rightarrow \sigma_u^2 > 0$ ,  $n/N \rightarrow c \in (0, \infty)$  as  $n \rightarrow \infty$ . There exists a positive constant  $C > 2$ , such that  $\|(I - U^* U^{*\top})(\tilde{Y}_1 - \gamma)\|_\infty \leq C \sqrt{\log(N)}$  occurs with very large probability.*

*Proof.* For the following proof, let  $u = \|U\| = O(\sqrt{Nn})$ ,  $U^\circ = U/\|U\|$  and  $\sigma = 1$  as the general cases is treated by a simple rescaling. We just show the proof assuming Scenario A occurs. The proof for Scenario B is similar. Let  $\tilde{\epsilon} = \tilde{Y}_1 - \gamma$ .

$$\begin{aligned} \|(I - U^* U^{*\top})(\tilde{Y}_1 - \gamma)\|_\infty &= \|(I - U^* U^{*\top})\tilde{\epsilon}\|_\infty \\ &= \|(I - U^\circ U^{\circ\top})\tilde{\epsilon}\|_\infty + \|(U^* U^{*\top} - U^\circ U^{\circ\top})\tilde{\epsilon}\|_\infty \\ &\leq \|(I - U^\circ U^{\circ\top})\tilde{\epsilon}\|_\infty + \|(U^* U^{*\top} - U^\circ U^{\circ\top})\epsilon\|_\infty \\ &\quad + \|(U^* U^{*\top} - U^\circ U^{\circ\top})U^\circ \rho u\|_\infty. \end{aligned}$$

By a modification of the proof of corollary 3.2 in Candes and Randall (2006), we have

$$\|(I - U^\circ U^{\circ\top})\tilde{\epsilon}\|_\infty = \|(I - U^\circ U^{\circ\top})\epsilon\|_\infty \leq \sqrt{\frac{3c'(N-1)\log(N)}{N}} \quad (4.23)$$

with probability at least  $1 - N^{-(c'-2)/2} \frac{\sqrt{2}}{\sqrt{c'\pi \log(N)}}$  for some constant  $c' > 2$ .

As  $\epsilon$  is a vector of independent normal,  $((U^*U^{*\top} - U^\circ U^{\circ\top})\epsilon)_i \sim \mathcal{N}(0, \sigma_i'^2)$ , where  $\sigma_i'^2 = U_i^{\circ 2} + U_i^{*2} - 2U_i^\circ U_i^* U^{\circ\top} U^*$ . This implies  $\epsilon'_i = ((U^*U^{*\top} - U^\circ U^{\circ\top})\epsilon)_i / \sigma_i'$  is standard normal with density  $\phi(t) = (2\pi)^{-1/2} e^{-t^2/2}$ . For each  $i$ ,  $\mathbb{P}(|\epsilon'_i| > t) < 2\phi(t)/t$  and thus

$$\mathbb{P}(\sup_{1 \leq i \leq N} |\epsilon'_i| \geq t) \leq 2N\phi(t)/t.$$

With  $t = \sqrt{c' \log(N)}$ , this gives  $\mathbb{P}(\sup_{1 \leq i \leq N} |\epsilon'_i| \geq \sqrt{c' \log(N)}) \leq N^{-(c'-2)/2} \frac{\sqrt{2}}{\sqrt{c'\pi \log(N)}}$ . Better bounds are possible but we will not pursue these refinements here. By the assumption  $U^\top U / (Nn) \rightarrow \sigma_u^2 > 0$ , the  $\tilde{\mu}$  defined in Theorem A 1 is of  $O(N)$  and thus  $1 - U^{\circ\top} U^* = O(\frac{1}{N})$ . Hence

$$\begin{aligned} \|(U^*U^{*\top} - U^\circ U^{\circ\top})\epsilon\|_\infty &\leq \sqrt{c' \log(N)} \max_{1 \leq i \leq N} \sqrt{U_i^{\circ 2} + U_i^{*2} - 2U_i^\circ U_i^* U^{\circ\top} U^*} \\ &\leq \sqrt{c' \log(N)} \sqrt{2(1 - U^{\circ\top} U^*)} \\ &\leq O(\sqrt{2c' \log(N)/N}) \end{aligned}$$

with probability at least  $1 - N^{-(c'-2)/2} \frac{\sqrt{2}}{\sqrt{c'\pi \log(N)}}$

$$\begin{aligned} \|(U^*U^{*\top} - U^\circ U^{\circ\top})U^\circ \rho u\|_\infty &= \|(U^*U^{*\top} - U^\circ U^{\circ\top})U^\circ\|_\infty \|\rho U\| \\ &\leq (\|U^* - U^\circ\|_\infty + \|U^*\|_\infty |1 - U^{\circ\top} U^*|) \|\rho U\| \end{aligned}$$

As  $U^*$  is a unit vector,  $\|U^*\|_\infty \leq 1$ . Therefore  $\|U^*\|_\infty |1 - U^{\circ\top} U^*| \|\rho U\| \leq c_2$  for some positive constant  $c_2$ . Next we will show that  $\|U^* - U^\circ\|_\infty \|\rho U\| \leq c_3 \sqrt{\log N}$  for some positive constant  $c_3$ .

As we saw in the proof of Theorem A 1, the last  $n - 1$  columns of the rotated data are

$$\bar{Y} = \sqrt{1 - \rho^2} U \bar{W}^\top + \bar{E}$$

where  $\mathbb{E} \bar{W}^\top \bar{W} = 1$ ,  $\bar{W}^\top \bar{W} \xrightarrow{p} 1$  and  $\bar{E}_{ij} \sim \mathcal{N}(0, 1)$ . Our setting is a special case of that in Theorem 2(i) of Bai (2003): we have just one latent variable and no time

series structure. Let  $\widehat{W}$  be the top principal component factor of  $\bar{Y}$ ,  $\|\widehat{W}\| = 1$  and  $\hat{U}$  be the corresponding principal component loadings estimated from  $\bar{Y}$ . By Theorem 1 of Chapter 2,

$$\xi_i \equiv \hat{U}_i - U_i \xrightarrow{d} \mathcal{N}\left(0, \frac{\Psi_i}{(1 - \rho^2)Q^2}\right) \quad i = 1, \dots, N \quad \text{as } n \rightarrow \infty$$

where

$$\Psi_i \equiv \mathbf{var}\left(\sum_{j=1}^{n-1} \bar{W}_j \bar{E}_{ij}\right) = 1, \quad \text{and} \quad Q \equiv \text{plim}_{n \rightarrow \infty} \widehat{W}^\top \bar{W} = 1 \text{ or } -1.$$

The proof of Theorem 2(i) of Bai (2003) further shows that  $\xi_i = \eta_i + O_p(1/\sqrt{n})$ , with  $\eta_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/(1 - \rho^2))$ . Hence

$$\|\hat{U} - U\|_\infty \leq \frac{\sqrt{c' \log(N)}}{\sqrt{1 - \rho^2}}$$

with probability at least  $1 - N^{-(c'-2)/2} \frac{\sqrt{2}}{\sqrt{c' \pi \log(N)}}$ . As a result,

$$\|U^* - U^\circ\|_\infty \|\rho U\| \leq 2|\rho| \|\hat{U} - U\|_\infty \leq 2 \frac{|\rho| \sqrt{c' \log(N)}}{\sqrt{1 - \rho^2}}$$

with high probability.

Hence there exists a positive constant  $C > 2$  such that  $\|(I - U^* U^{*\top})(\tilde{Y}_1 - \gamma)\|_\infty \leq C \sqrt{\log(N)}$  with very high probability.

□

Suppose  $U$  can be estimated accurately, we can simply study the final step of LEAPP as if  $U$  is known and give the conditions for estimate  $\hat{\gamma}$  from a non-robust version of  $\Theta$ -IPOD to be sign consistent, i.e.

$$\mathbb{P}(\text{sgn}(\hat{\gamma}) = \text{sgn}(\gamma)) \rightarrow 1 \quad \text{as } N \rightarrow \infty$$

where

$$\text{sgn}(x) = \begin{cases} 1 & x > 0, \\ 0 & x = 0, \\ -1 & x < 0 \end{cases}.$$

We write the first column of the rotated data

$$y = \tilde{Y}_1 = U\rho + \gamma + \sigma\epsilon.$$

In the following discussion, we normalize the data by  $\rho = \rho\sqrt{U^TU}$ ,  $U = U/\sqrt{U^TU}$  and assume  $\sigma = 1$ . Combine unknown parameters  $\rho$  and  $\gamma$  and we obtain

$$y = (U, I_{N \times N}) \begin{pmatrix} \rho \\ \gamma \end{pmatrix} + \epsilon. \quad (4.24)$$

We can write  $A = (U, I_{N \times N}) \in \mathbb{R}^{N \times (N+1)}$ ,  $\theta = (\rho, \gamma^T)^T \in \mathbb{R}^{N+1}$ . Then the equation (4.24) can be rewritten as

$$y = A\theta + \epsilon \quad (4.25)$$

As  $\gamma$  is  $s$ -sparse,  $\theta$  is at most  $s + 1$ -sparse. Recovering the sparse  $\theta$  is essentially a linear regression model selection problem with sample size  $N$  and number of variables  $p_N = N + 1$ . Lasso is a popular approach to tackle the problem. The lasso estimator for (4.25) is defined as

$$\hat{\theta}(\lambda) = \arg \min_{\theta} \|y - A\theta\|_2^2 + \lambda\|\theta\|_1, \quad (4.26)$$

where  $\|\cdot\|_1$  stands for the  $L_1$  norm of a vector which equals the sum of absolute values of the vectors entries. The non-robust version of  $\Theta$ -IPOD algorithm is the same as lasso without penalizing the  $\rho$  term. For our analysis, we replace the robust  $\Theta$ -IPOD algorithm by lasso to get strong result.

Consider a general linear regression model

$$y = A\theta + \epsilon \quad (4.27)$$

where  $y \in \mathbb{R}^{N \times 1}$  is a response,  $A = (A_1, \dots, A_{p_N}) \in \mathbb{R}^{N \times p_N}$  is a design matrix and  $A_i$  is its  $i$ th column.  $\theta \in \mathbb{R}^{p_N \times 1}$  is the vector of model variables and let  $O = \text{supp}(\theta)$ ,  $q(N) = |O|$  be the number of non-zeros in  $\theta$ . We can rearrange the order of variables and assume the first  $q(N)$  elements of  $\theta$  are non zeros and the rest of  $\theta$  are zeros. Then  $\theta = (\theta_1, \dots, \theta_q, \theta_{q+1}, \dots, \theta_{p_N})^\top$ ,  $\theta_i \neq 0, i = 1, \dots, q$  and  $\theta_i = 0, i = q + 1, \dots, p_N$ . Let  $\theta(1) = (\theta_1, \dots, \theta_q)$  and  $A(1), A(2)$  be the first  $q$  and last  $(N - q)$  columns of  $A$  respectively. Let  $C_{11} = A(1)^\top A(1)$ ,  $A_{21} = A(2)^\top A(1)$ .

Zhao and Yu (2006) shows that the sign consistency of  $\hat{\theta}$  is closely related to the *Strong Irrepresentable Condition* which is defined below.

**Definition 2** (Strong Irrepresentable Condition). *There exists a positive constant vector  $\eta$*

$$|C_{21}C_{11}^{-1}\text{sgn}(\theta(1))| \leq \mathbf{1} - \eta$$

where  $\mathbf{1}$  is an  $N - q$  vector of 1's and the inequality holds element-wise.

**Lemma 3** (Zhao and Yu (2006)). *Consider model (4.27) and assume  $\epsilon$  are i.i.d. random variables with finite  $2k$ 'th moment  $E(z^{2k}) < \infty$  for an integer  $k > 0$ . There exists  $0 \leq c_1 < c_2 \leq 1$  and  $M_1, M_2, M_3 > 0$  such that*

$$\frac{1}{N}A_i^\top A_i \leq M_1, \forall i = 1, \dots, p_N \quad (1)$$

$$\alpha^\top C_{11}\alpha \geq M_2, \forall \|\alpha\|_2 = 1 \quad (2)$$

$$q(N) = O(N^{c_1}) \quad (3)$$

$$N^{\frac{1-c_2}{2}} \min_{i \in O} |\theta_i| \geq M_3 \quad (4)$$

*Strong Irrepresentable condition implies that lasso has sign consistency for  $p_N = o(N^{(c_2-c_1)k})$ . In particular,  $\forall \lambda_N$  such that  $\frac{\lambda_N}{\sqrt{N}} = o(N^{\frac{c_2-c_1}{2}})$  and  $\frac{1}{p_N}(\frac{\lambda_N}{\sqrt{N}})^{2k} \rightarrow \infty$ , we*

have

$$P(\text{sgn}(\hat{\theta}) = \text{sgn}(\theta)) \geq 1 - O\left(\frac{p_N N^k}{\lambda_N^{2k}}\right) \rightarrow 1 \quad \text{as } N \rightarrow \infty.$$

Now we move back to our problem (4.24). Let  $O = \text{supp}(\gamma)$ ,  $s = |O|$ ,  $G = \{i, \gamma_i = 0\}$ . Let  $U_O, U_G$  be the vector of  $U$  restricted to the indices in  $O, G$  respectively. For any matrix  $\mathcal{A}$ ,  $\mathcal{A}_O$  is defined as the matrix  $\mathcal{A}$  with columns restricted to the indices in  $O$ . Same can be defined for subscript  $G$ . The following theorem shows the conditions for sign consistency of the estimate  $\hat{\gamma}$  from a non-robust version of  $\Theta$ -IPOD.

**Theorem 4.** *Consider the model (4.25). Assume the parameter vector  $\gamma$  is  $s$ -sparse,  $U$  is known and there exists a positive constant  $\eta_0 \in \mathbb{R}$ ,*

$$\frac{|-\text{sgn}(\rho) + \sum_{i \in O} U_i \text{sgn}(\gamma_i)|}{\|U_G\|^2} \|U_G\|_\infty \leq 1 - \eta_0.$$

Assume there exists  $0 \leq c_1 < c_2 \leq 1$  and  $M > 0$ ,

$$s = O(N^{c_1}), N^{\frac{1-c_2}{2}} \min_{i \in O} |\gamma_i| \geq M.$$

Let  $\hat{\theta}$  be the estimate from (4.26). Then for  $\forall \lambda_N$  that satisfies  $\frac{\lambda_N}{\sqrt{N}} = O(N^{\frac{c_2 - c_1 - \delta}{2}})$ ,  $c_2 - c_1 > \delta > 0$ , there exists an integer  $k > 0$  such that  $k(\frac{c_2 - c_1 - \delta}{2}) > 1$  and then we have

$$P(\text{sgn}(\hat{\theta}) = \text{sgn}(\theta)) \geq 1 - O\left(\frac{N^{k+1}}{\lambda_N^{2k}}\right) \rightarrow 1 \quad \text{as } N \rightarrow \infty.$$

*Proof.* For our problem,  $p_N = N + 1$ . We can rearrange the order of variables and assume the first  $s$  elements of  $\gamma$  are non zeros and the rest of  $\gamma$  are zeros. Then  $\theta = (\theta_1, \dots, \theta_s, \theta_{s+1}, \dots, \theta_{N+1})^\top$  where  $\theta_1 = \rho, \theta_i \neq 0, i = 2, \dots, s + 1$  and  $\theta_i = 0, i = s + 2, \dots, N + 1$ ,  $\theta(1) = (\rho, \gamma_1, \gamma_2, \dots, \gamma_s)$  and  $A = (U, I_O, I_G) \in R^{N \times (N+1)}$ . Let  $A(1), A(2)$  be the first  $(s + 1)$  and last  $(N - s - 1)$  columns of  $A$  respectively and  $\theta(1) = (\theta_1, \dots, \theta_{s+1})$ . Let  $C_{11} = A(1)^\top A(1), C_{21} = A(2)^\top A(1)$ . As  $A(1) = (U, I_O)$ ,  $A(2) = I_G, C_{11} = \begin{pmatrix} 1 & U_O^\top \\ U_O & I \end{pmatrix}, C_{21} = (U_G, \mathbf{0})$ . By a simple matrix multiplication

and inversion, we obtain

$$C_{21}C_{11}^{-1} = \frac{1}{\|U_G\|^2} U_G (-1, U_O^\top)$$

$$\begin{aligned} |C_{21}C_{11}^{-1}\text{sgn}(\theta(1))| &= \left| \frac{1}{\|U_G\|^2} U_G (-1, U_O^\top) \text{sgn}(\theta(1)) \right| \\ &\leq \frac{|-\text{sgn}(\rho) + \sum_{i \in O} U_i \text{sgn}(\gamma)|}{\|U_G\|^2} |U_G| \\ &\stackrel{(a)}{\leq} 1 - \eta_0. \end{aligned}$$

Inequality (a) holds element-wise and is an immediate result from the assumption that

$$\frac{|-\text{sgn}(\rho) + \sum_{i \in O} U_i \text{sgn}(\gamma)|}{U_G^\top U_G} \|U_G\|_\infty \leq 1 - \eta.$$

Hence the strong irrepresentable condition holds. Conditions for Lemma 3 are all satisfied. As the elements of vector  $\epsilon$  are distributed as independent standard normal and thus they have any finite even moments. For  $\forall \lambda_N$  that satisfies  $\frac{\lambda_N}{\sqrt{N}} = O(N^{\frac{c_2 - c_1 - \delta}{2}})$ ,  $c_2 - c_1 > \delta > 0$ , we can pick a large enough  $k$  such that  $\frac{1}{N+1} \left( \frac{\lambda_N}{\sqrt{N}} \right)^{2k} \rightarrow \infty$ . According to Lemma 3, we have

$$\mathbb{P}(\text{sgn}(\hat{\theta}(\lambda_N)) = \text{sgn}(\theta)) \geq 1 - O\left(\frac{(N+1)N^k}{\lambda_N^{2k}}\right) \rightarrow 1 \quad \text{as } N \rightarrow \infty.$$

□

**Corollary 2.** *In particular, Let  $L_G = \max_{i \in G} |U_i|$ ,  $L_O = \max_{i \in O} |U_i|$ , if  $L_G < 1/2$ ,  $L_O \leq \frac{2}{s + \sqrt{s^2 + 8s}}$ , the strong representable condition holds and the lasso estimator  $\hat{\theta}$  is sign consistent.*

*Proof.*

$$\begin{aligned}
\frac{|-\operatorname{sgn}(\rho) + \sum_{i \in O} U_i \operatorname{sgn}(\gamma)|}{U_G^\top U_G} \|U_G\|_\infty &\leq \frac{1 + \|U_O\|_1}{1 - U_O^\top U_O} \|U_G\|_\infty \\
&\leq \frac{1 + sL_O}{1 - sL_O^2} L_G \\
&\stackrel{(b)}{<} \frac{1 + s \frac{2}{s + \sqrt{s^2 + 8s}}}{1 - s \left(\frac{2}{s + \sqrt{s^2 + 8s}}\right)^2} \frac{1}{2} \\
&\leq 1
\end{aligned}$$

Inequality (b) is by assumption. And therefore the strong irrerepresentable condition holds. The result follows from Theorem 4. □

## 4.4 Performance On Synthetic Data

In this section, we generate data from the model (3.1) with a numeric primary variable and a SNP association study from Price et al. (2006). LEAPP yields consistently better performance on those synthetic data sets than other methods such as SVA and EIGENSTRAT.

### 4.4.1 Multiple testing with known primary variable

In this subsection, we generate data from the model (3.1) and compare the results from the algorithms to each other, to an oracle which is given the latent variable, and to a raw regression method which makes no attempt to adjust for latent variables.

We choose  $s = 0$ , omitting the  $\beta X^\top$  covariate term, so the simulated data satisfy

$$Y = \gamma g^\top + UV^\top + \Sigma E. \quad (4.28)$$

Our simulations have  $n = 60$  (subjects) and  $N = 1000$  (genes). Our primary covariate is a binary treatment vector  $g \propto (1, \dots, 1, -1, \dots, -1)$ , with equal numbers of 1 and  $-1$ , normalized so that  $g^\top g = 1$ .

The vector  $\gamma$  of treatment effects has independent components  $\gamma_i$  taking the values 0 and  $c > 0$  with probability 0.9 and 0.1 respectively. We will choose  $c$  in order to attain specific signal to noise ratios. The matrix  $\Sigma$  is a diagonal with nonzero entries  $\sigma_i$  sampled independently from an inverse gamma distribution:  $1/\sigma_i^2 \sim \text{Gamma}(10)/9$ . Note that  $\mathbb{E}(\sigma_i^2) = 1$ .

We use  $k = 1$  latent variable that has correlation  $\rho$  with  $g$ . The latent vector  $U = (u_1, \dots, u_N)$  is generated as independent  $U(-a, a)$  random variables. We will choose  $a$  to obtain specific latent to noise variance ratios. The latent vector  $V$  is taken to be  $\rho g + \sqrt{1 - \rho^2}W$  where  $W$  is uniformly distributed on the set of unit vectors orthogonal to  $g$ .

The model (4.28) gives  $Y$  three components: the signal  $\mathcal{S} = \gamma g^\top$ , the latent structure  $\mathcal{L} = UV^\top$ , and the noise  $\mathcal{N} = \Sigma E$ . The relative sizes of these components affect the difficulty of the problem. We use Frobenius and spectral norms to describe the size of these matrices.

The noise matrix is constructed so that  $\mathbb{E}(\sigma_i^2 \epsilon_{ij}^2) = \mathbb{E}(\sigma_i^2) = 1$ , so that  $\mathbb{E}(\|\mathcal{N}\|_F^2) = Nn$ . Because the signal and latent matrices have rank 1,

$$\mathbb{E}(\|\mathcal{S}\|_F^2) = \mathbb{E}(\|\mathcal{S}\|_2^2) = \mathbb{E}(\|\gamma\|_2^2) = N\pi c^2 \quad (4.29)$$

$$\mathbb{E}(\|\mathcal{L}\|_F^2) = \mathbb{E}(\|\mathcal{L}\|_2^2) = \mathbb{E}(\|U\|_2^2) = Na^2/3. \quad (4.30)$$

For our simulation, we specified the ratios

$$\text{SNR} \equiv \pi c^2, \quad \text{and} \quad \text{LNR} \equiv a^2/3$$

and varied them over a wide range. We also use  $\text{SLR} = 3\pi c^2/a^2$ .

We also varied the level of  $\rho$ , the correlation between the latent and primary variables. For each setting of SNR, SLR, LNR and  $\rho$  under consideration, we simulated the process 100 times and prepared ROC curves, from the pooled collection of 100,000 predictions.

The methods that we applied are as follows:

- true** an oracle given  $UV^T$  which then does regression of  $Y - UV^T$  on  $g$ ,
- raw** multivariate regression of  $Y$  on  $g$  ignoring latent variables,
- eig** EIGENSTRAT of Price et al. (2006),
- sva** surrogate variable analysis from Leek and Storey (2008), and
- lea** our proposed LEAPP method.

The ROC curves for one set of conditions are shown in Figure 4.1. There, the best performance is from the oracle. The next best method is LEAPP. After that comes the raw method making no adjustment for latents, then SVA and finally EIGENSTRAT has the worst performance in this setting.

Because the ROC curves from the simulations have few if any crossings, we can reasonably summarize each one by a single number. We have used the area under the curve (AUC) for a global comparison as well as a precision measure for the quality of the most highly ranked values. That measure is the fraction of truly different genes among the highest ranking  $H$  genes. We use  $H = 50$ .

When  $\rho = 0$ , EIGENSTRAT, SVA and LEAPP have almost equivalent performance. For  $\rho > 0$ , the oracle always had the highest AUC and LEAPP was always second. The ordering among the other three methods varied. Sometimes EIGENSTRAT was the best of those three, and other times SVA was best of those three.

Figure 4.2 shows a heatmap of the improvement in AUC for LEAPP versus SVA. The improvements are greatest when  $\rho$  is large. This is reasonable because SVA is not designed to account for correlation between the latent and primary variables. At each correlation level, the greatest differences arise when SNR is small and LNR is about 2.

Figure 4.3 shows the improvement in AUC for LEAPP versus EIGENSTRAT. The improvements are largest when the primary effect is large.

The improvements versus SVA are smaller than those versus EIGENSTRAT. To judge the practical significance of the improvement we repeated some of these simulations for SVA, increasing  $n$  until SVA achieved the same AUC that LEAPP did. Sometimes SVA required only 2 more observations (one treatment and one control) to match the AUC of the LEAPP methods. Sometimes it was unable to match the AUC even given double the sample size, that is  $n = 120$  observations instead of  $n = 60$ .

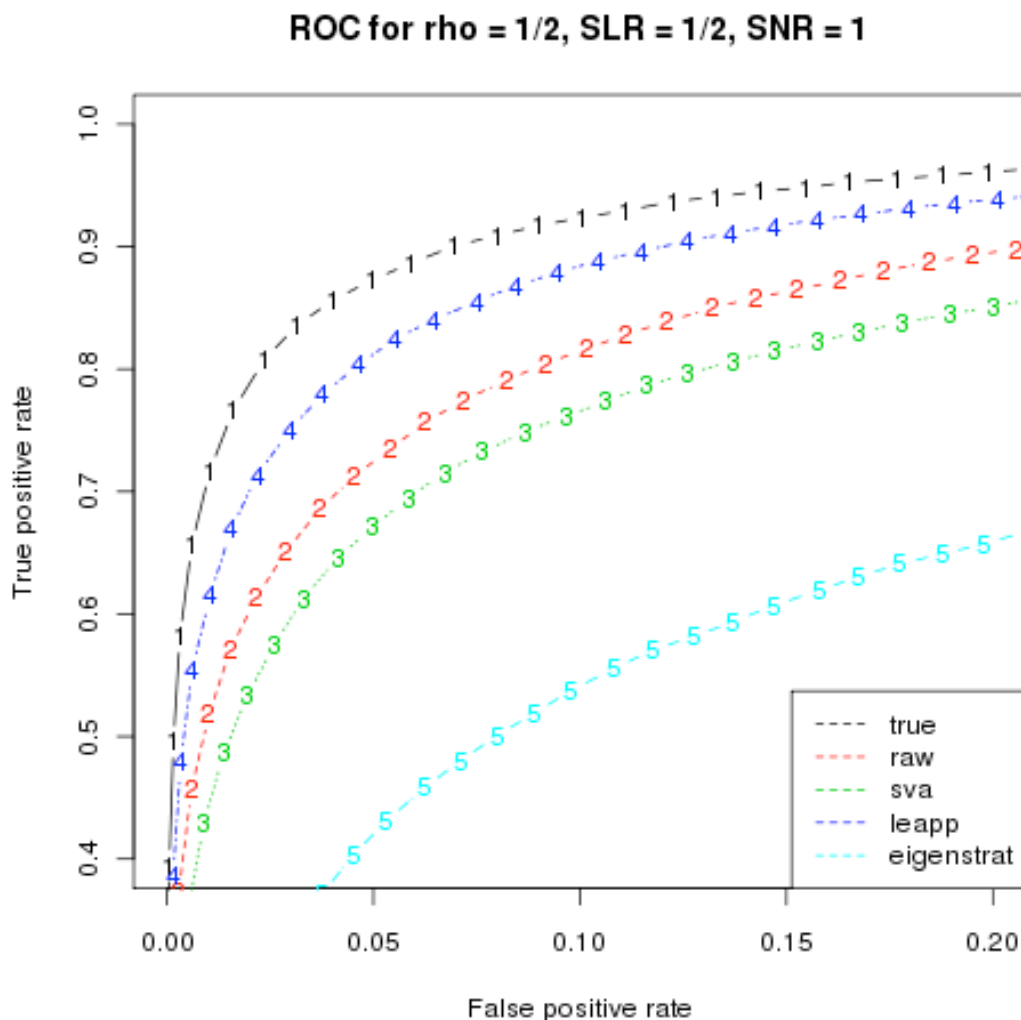


Figure 4.1: This figure shows the knee of the ROC curves for a simulation with  $\rho = 1/2$ , SLR= 1/2 and SNR=1. The best (highest) results are for an oracle that was given the latent variables. The second best are for the proposed LEAPP method. A raw method making no adjustment gives ROCs just barely larger than SVA. EIGEN-STRAT did quite poorly in this setting. The relative performance for SVA, EIGEN-STRAT and the raw method were different in other settings.

Not surprisingly, the advantage of LEAPP is greatest when the latent variable is most strongly correlated with the primary.

Table 4.1 shows a feature of this problem that we also see in the Figures. The

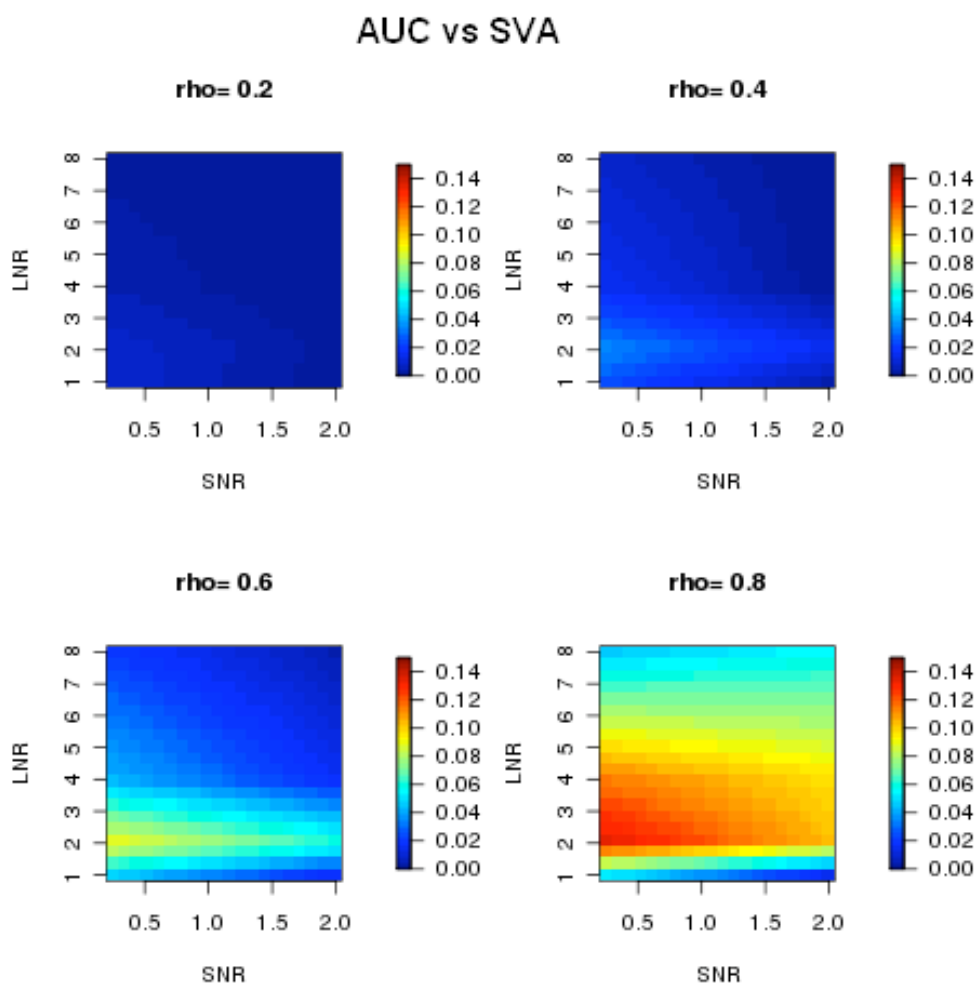


Figure 4.2: This figure shows the improvement in AUC for the LEAPP method relative to SVA. Here  $\rho$  is the correlation between the primary and latent variables. The signal to noise ratio and latent to noise ratio are described in the text. The color scheme encodes  $(AUC_{lea} - AUC_{sva}) / AUC_{sva}$ .

improvement over SVA is quite small when  $LNR = 0.5$ . A small enough latent effect becomes undetectable, both methods suffer and there is little difference. Similarly a very large latent effect ( $LNR = 8$ ) is easy to detect by both methods. The largest differences arise for medium sized latent effects.

High throughput methods are often used to identify candidates for future followup investigation. In that case we value high precision for the most highly ranked

| Conditions |     | $\rho = 0.25$ |    | $\rho = 0.5$ |    | $\rho = 0.75$ |     |
|------------|-----|---------------|----|--------------|----|---------------|-----|
| SNR        | LNR | n             | %  | n            | %  | n             | %   |
| 2          | 0.5 | 66            | 10 | 66           | 10 | 62            | 3   |
| 2          | 1   | 68            | 13 | 92           | 53 | 120           | 100 |
| 2          | 2   | 66            | 10 | 74           | 23 | 114           | 90  |
| 2          | 4   | 62            | 3  | 66           | 10 | 88            | 47  |
| 2          | 8   | 62            | 3  | 66           | 10 | 72            | 20  |
| 1          | 0.5 | 64            | 7  | 64           | 7  | 62            | 3   |
| 1          | 1   | 66            | 10 | 90           | 50 | 120           | 100 |
| 1          | 2   | 64            | 7  | 76           | 27 | 120           | 100 |
| 1          | 4   | 64            | 7  | 66           | 10 | 90            | 50  |
| 1          | 8   | 62            | 3  | 66           | 10 | 76            | 27  |
| 0.5        | 0.5 | 64            | 7  | 64           | 7  | 62            | 3   |
| 0.5        | 1   | 66            | 10 | 84           | 40 | 120           | 100 |
| 0.5        | 2   | 66            | 10 | 78           | 30 | 110           | 83  |
| 0.5        | 4   | 66            | 10 | 68           | 13 | 88            | 47  |
| 0.5        | 8   | 62            | 3  | 68           | 13 | 72            | 20  |

Table 4.1: This table shows the number of samples required for SVA to attain the same AUC that LEAPP attains with  $n = 60$  samples. For example with SNR = 2 and LNR = 0.5, and  $\rho = 0.25$ , SVA requires 66 samples or 10% more. The entries of 100% denote settings where the increase needed was  $\geq 100\%$ .

hypotheses. Figure 4.4 shows the improvement of LEAPP over SVA, as measured by precision. Figure 4.5 shows the improvement of LEAPP over EIGENSTRAT, as measured by precision.

To understand why our method can do better, we compare the mean squared error (MSE) of estimator  $\hat{\gamma}$ , the estimator  $\hat{U}$  at non null genes and the estimator  $\hat{V}$  of SVA with those of LEAPP method. The result is summarized in Table 4.2 and Table 4.3. The numbers in the table give the percentage of improvement of our method over SVA. For AUC, it is relative increase of AUC:  $(\text{AUC}_{\text{lea}} - \text{AUC}_{\text{SVA}})/\text{AUC}_{\text{SVA}}$ ; for MSE, it is relative decrease of MSE:  $(\text{MSE}_{\text{SVA}} - \text{MSE}_{\text{lea}})/\text{MSE}_{\text{SVA}}$ ; for the estimation of  $U$ , it is the relative increase in the correlation of  $\hat{U}, U$  at non null genes ( $C^U = \text{Cor}(\hat{U}_0, U_0)$  where  $\hat{U}_0, U_0$  are  $\hat{U}, U$  restricted to non null genes), that is,  $(C_{\text{lea}}^U - C_{\text{SVA}}^U)/C_{\text{SVA}}^U$  and

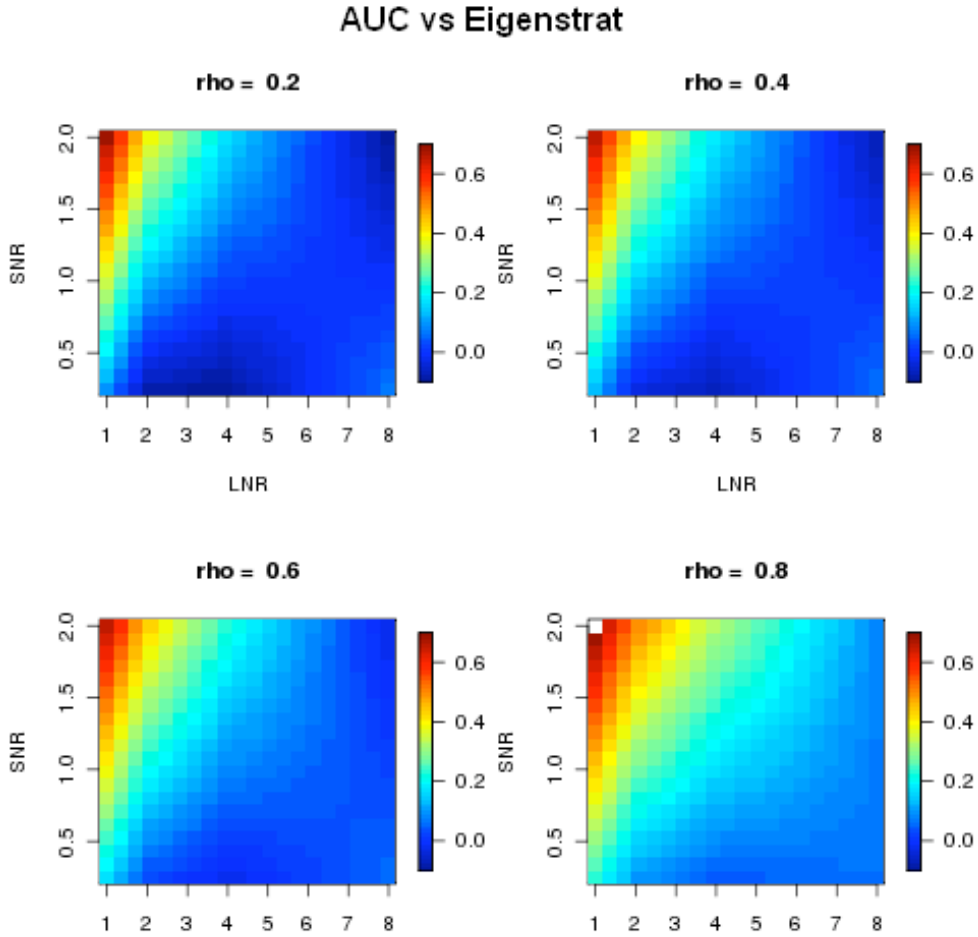


Figure 4.3: This figure shows the improvement in AUC for the LEAPP method relative to EIGENSTRAT. The simulation conditions are as described in Figure 4.2. The color scheme encodes  $(AUC_{\text{rot}} - AUC_{\text{eig}})/AUC_{\text{eig}}$ .

for the estimation of  $V$ , it is the relative increase in the correlation of  $\hat{V}, V$  ( $C^V = \text{Cor}(\hat{V}, V)$ ), that is  $(C_{\text{lea}}^V - C_{\text{sva}}^V)/C_{\text{sva}}^V$ . We observe that in most cases our estimator  $\hat{U}, \hat{V}$  have better precision and result in higher AUC and lower MSE than those of sva.

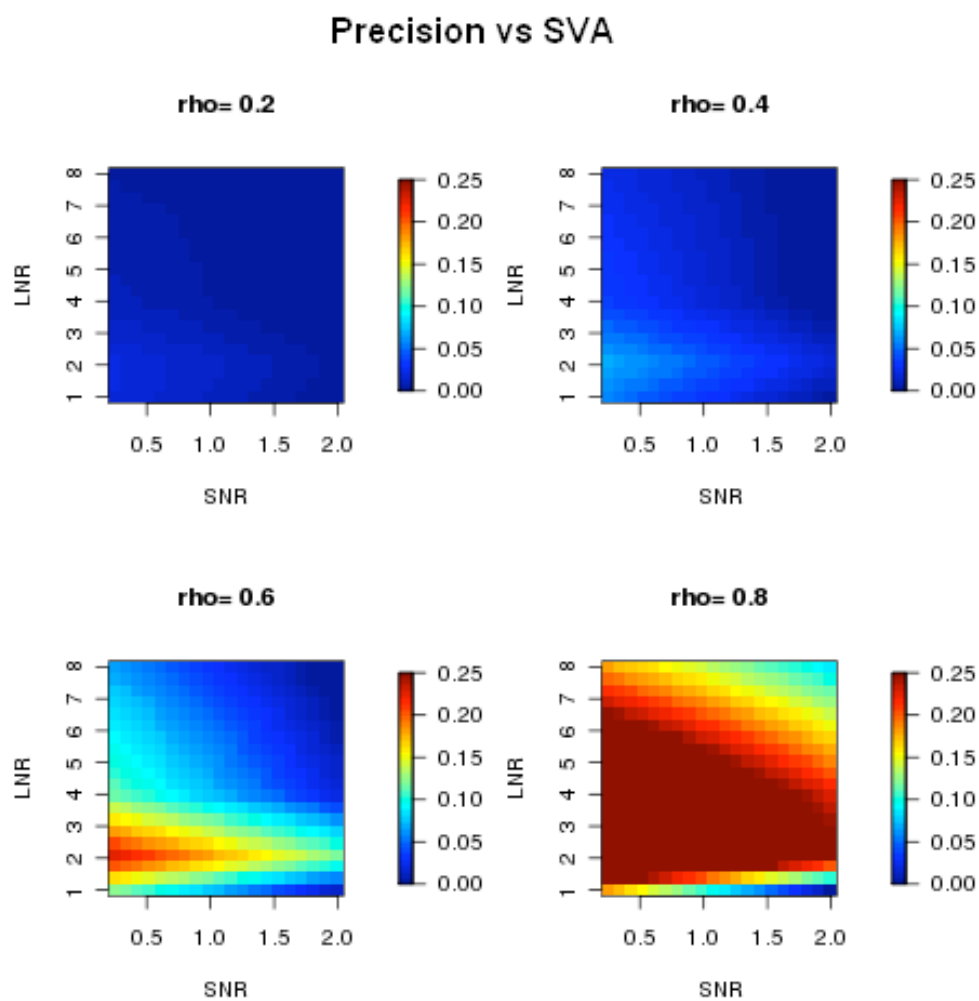


Figure 4.4: This figure shows the improvement in precision for the LEAPP method relative to SVA. Precision is the fraction of truly affected genes among the top  $H = 50$  ranked genes. The simulation conditions are as described in Figure 4.2. The color scheme encodes  $(PRE_{lea} - PRE_{sva}) / PRE_{sva}$ .

#### 4.4.2 Simulated SNP association study

In this subsection, we generate data following the simulation example in Price et al. (2006) and compare the results from methods LEAPP, SVA and EIGENSTRAT.

In specific, we generated data at 1000 random SNPs for 50 cases and 50 controls, with 60% of the cases and 40% of the controls sampled from population 1 and the

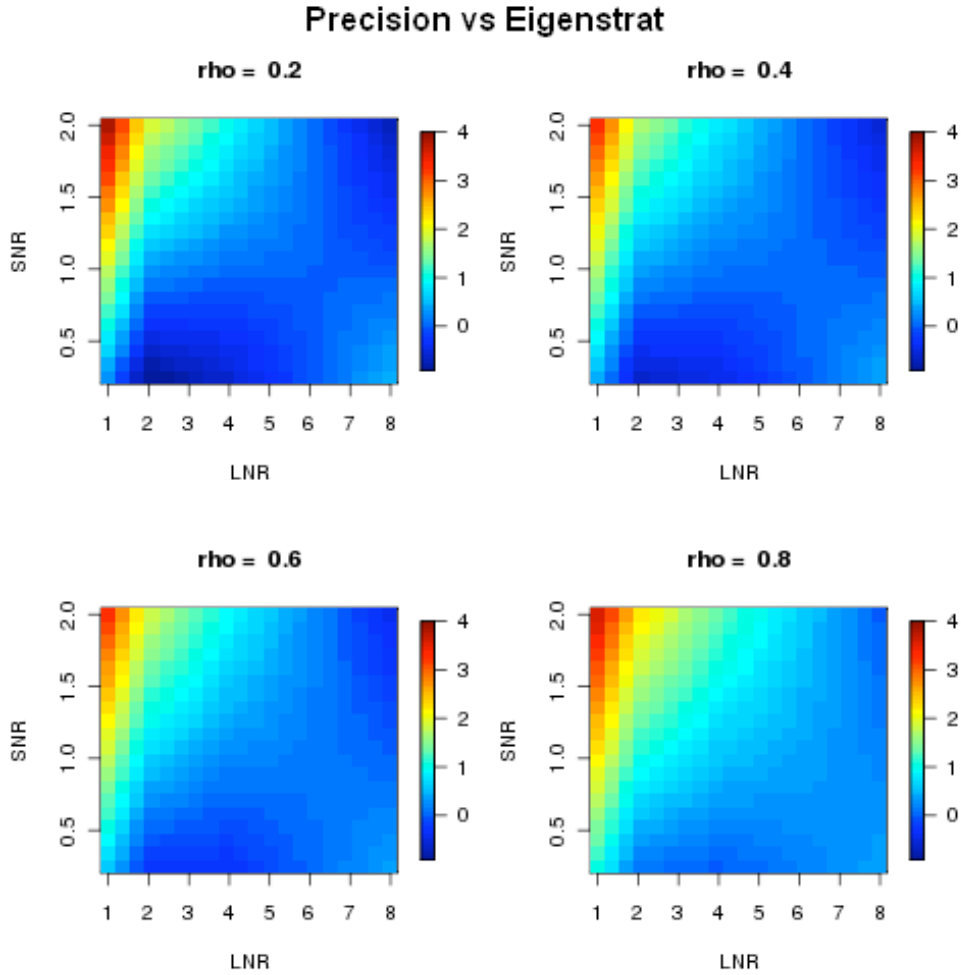


Figure 4.5: This figure shows the improvement in precision for the LEAPP method relative to EIGENSTRAT. Precision is the fraction of truly affected genes among the top  $H = 50$  ranked genes. The simulation conditions are as described in Figure 4.2. The color scheme encodes  $(\text{PRE}_{\text{rot}} - \text{PRE}_{\text{eig}})/\text{PRE}_{\text{eig}}$ .

remaining cases and controls sampled from population 2. Allele frequencies for population 1 and population 2 were generated using the Balding-Nichols model (Balding and Nichols, 1995) with  $F_{ST} = 0.01$ . For each SNP, an ancestral population allele frequency  $p$  was drawn from the uniform distribution on  $[0.1, 0.9]$ . The allele frequencies for populations 1 and 2 were each drawn from a beta distribution with parameters  $p(1 - F_{ST})/F_{ST}$  and  $(1 - p)(1 - F_{ST})/F_{ST}$ . This distribution has mean  $p$  and

| Conditions |      |     | $\rho = 0.25$ |       | $\rho = 0.5$ |       | $\rho = 0.75$ |       |
|------------|------|-----|---------------|-------|--------------|-------|---------------|-------|
| SNR        | SLR  | LNR | AUC %         | MSE % | AUC %        | MSE % | AUC %         | MSE % |
| 2          | 4    | 0.5 | 0.14          | 3.16  | 0.18         | 2.63  | 0             | 0.04  |
| 2          | 2    | 1   | 0.21          | 4.54  | 1.4          | 15.27 | 0.78          | 7.91  |
| 2          | 1    | 2   | 0.15          | 4.56  | 1.27         | 16.85 | 5.11          | 27.08 |
| 2          | 0.5  | 4   | 0.04          | 1.79  | 0.39         | 7.85  | 6.73          | 26.92 |
| 2          | 0.25 | 8   | 0.03          | 0.98  | 0.19         | 4.06  | 2.29          | 12.92 |
| 1          | 2    | 0.5 | 0.47          | 3.69  | 0.36         | 2.39  | 0.04          | 0.39  |
| 1          | 1    | 1   | 1.13          | 5.21  | 4.15         | 12.16 | 2.43          | 5.23  |
| 1          | 0.5  | 2   | 0.65          | 4.47  | 4.27         | 13.47 | 19.86         | 21.08 |
| 1          | 0.25 | 4   | 0.27          | 1.69  | 1.61         | 6.68  | 9.51          | 12.73 |
| 1          | 0.13 | 8   | 0.14          | 0.82  | 0.82         | 3.65  | 4.18          | 6.27  |
| 0.5        | 1    | 0.5 | 1             | 0.98  | 0.72         | 0.97  | 0.05          | 0.11  |
| 0.5        | 0.5  | 1   | 1.55          | 2.32  | 3.99         | 4.63  | 4.39          | 4.54  |
| 0.5        | 0.25 | 2   | 1.26          | 2     | 5.62         | 5.96  | 15.29         | 13.15 |
| 0.5        | 0.13 | 4   | 0.6           | 1.08  | 2.48         | 3.18  | 8.19          | 8.09  |
| 0.5        | 0.06 | 8   | 0.3           | 0.43  | 1.27         | 1.94  | 4.27          | 4.66  |

Table 4.2: This table shows the improvement in AUC and MSE for the LEAPP method relative to SVA. Here  $\rho$  is the correlation between the primary and latent variables and SLR,SNR and LNR are defined in the text.

variance  $FSTp(1-p)$ . It follows that the quantity FST agrees with its usual measure of genetic distance between two populations. The risk model with a relative risk of R for the causal allele was implemented as follows: for individuals from population 1 with population allele frequency  $pl$ , control individuals were assigned genotype 0, 1 or 2 with probabilities  $(1-pl)^2$ ,  $2pl(1-pl)$ , or  $pl^2$ , respectively, and case individuals were assigned genotype 0, 1 or 2 with relative probabilities  $(1-pl)^2$ ,  $2Rpl(1-pl)$ , or  $R^2pl^2$ , respectively, each scaled to sum to 1. 10% of the SNPs are causal allele. We simulate the process 100 times for each parameter setting.

Table (4.6) shows the ROC curves for those 3 methods when the relative risk is 1.5 and 3 respectively. As those curves never cross, it makes sense to compare the area under the curve alone. Table (4.4) shows the area under the curve corresponding to those ROC curves. It shows that when the relative risk is low, i.e., the signal of causal SNPs is weak, the three methods don't differ much in AUC (Area under the

| Conditions |      |     | $\rho = 0.25$ | $\rho = 0.5$ | $\rho = 0.75$ |        |       |       |
|------------|------|-----|---------------|--------------|---------------|--------|-------|-------|
| SNR        | SLR  | LNR | U %           | V %          | U %           | V %    | U %   | V %   |
| 2          | 4    | 0.5 | 88.6          | 116.3        | 104.38        | 136.6  | 30.16 | 74.01 |
| 2          | 2    | 1   | 18.86         | 30.24        | 31.75         | 40.62  | 81.84 | 95.48 |
| 2          | 1    | 2   | 1.35          | 2.72         | 2.79          | 3.62   | 87.92 | 84.55 |
| 2          | 0.5  | 4   | 0.17          | 0.47         | 0.36          | 0.57   | 2.99  | 2.83  |
| 2          | 0.25 | 8   | 0.04          | 0.13         | 0.08          | 0.15   | 0.3   | 0.2   |
| 1          | 2    | 0.5 | 89.63         | 114.86       | 100.43        | 132.38 | 23.27 | 61.88 |
| 1          | 1    | 1   | 22.16         | 33.81        | 28.68         | 36.57  | 76.88 | 92.69 |
| 1          | 0.5  | 2   | 1.62          | 2.69         | 2.48          | 3.28   | 13.41 | 18.23 |
| 1          | 0.25 | 4   | 0.2           | 0.47         | 0.34          | 0.55   | 0.4   | 0.76  |
| 1          | 0.13 | 8   | 0.04          | 0.12         | 0.08          | 0.14   | 0.15  | 0.17  |
| 0.5        | 1    | 0.5 | 98.04         | 121.53       | 88.72         | 125.32 | 26.28 | 60.36 |
| 0.5        | 0.5  | 1   | 22.53         | 30.53        | 30.54         | 38.13  | 50.69 | 68.1  |
| 0.5        | 0.25 | 2   | 1.37          | 2.49         | 1.49          | 3.04   | 1.47  | 5.19  |
| 0.5        | 0.13 | 4   | 0.21          | 0.44         | 0.26          | 0.51   | -0.12 | 0.59  |
| 0.5        | 0.06 | 8   | 0.04          | 0.11         | 0.06          | 0.12   | -0.03 | 0.13  |

Table 4.3: This table shows the improvement in estimation of  $U, V$  for the LEAPP method relative to SVA. Here  $\rho$  is the correlation between the primary and latent variables and LSR, SNR and LNR are defined in the text.

| R   | SVA    | LEAPP  | EIGENSTRAT |
|-----|--------|--------|------------|
| 1.5 | 0.6878 | 0.6924 | 0.6817     |
| 3   | 0.9655 | 0.9670 | 0.7609     |

Table 4.4: This table shows the AUC comparison of methods SVA, LEAPP and EIGENSTRAT for 2 simulated SNP association studies, where the relative risk  $R$  is set to be 1.5 and 3 respectively.

curve) though SVA and LEAPP are slightly better than EIGENSTRAT and when the relative risk is high, both SVA and LEAPP have much higher area under the curve than EIGENSTRAT.

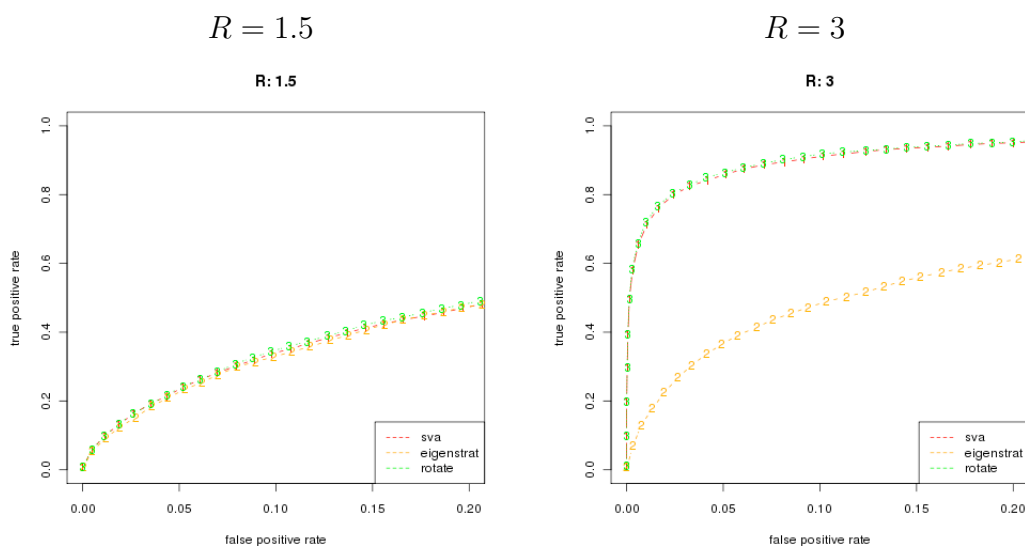


Figure 4.6: This figure shows the ROC curves of methods SVA, (ROTATE)LEAPP and EIGENSTRAT in two simulated SNP association studies where the relative risk  $R$  is set to be 1.5 and 3 respectively.

## 4.5 Real Data

In this section, we present the performance of LEAPP on two data sets. Subsection 4.5.1 focuses on the AGEMAP study Zahn et al. (2007) and it shows that the LEAPP method achieves greater concordance across tissues. Subsection 4.5.2 discusses the breast cancer study Hedenfalk (2001) and the estimated empirical null distribution of z-scores adjusted by LEAPP is closer to the theoretical null distribution than SVA, EIGENSTRAT or simply no adjustment.

### 4.5.1 Agemap mice data

The AGEMAP study (Zahn et al., 2007) investigated age-related gene expression in mice. Ten mice at each of four age groups were investigated. From these 40 mice, samples were taken of 16 different tissues, resulting in 640 microarray data sets. A small number of those 640 microarrays were missing. From each microarray 8932 probes were sampled. Perry and Owen (2010) found that many of the tissues in this dataset exhibited strong latent variables. Their approach assumed that the latent

variables were orthogonal to the treatment.

Our underlying assumption is that aging should have at least mildly consistent results from tissue to tissue. That should in turn show up as overlap in gene lists computed from multiple tissues, whereas a noisier estimation method should tend to have less overlap among tissues.

For any two tissues, we can measure the overlap between their sets of highly ranked genes. For two sets  $A$  and  $B$ , their resemblance (Broder, 1997) is

$$\text{res}(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where  $|\cdot|$  denoted cardinality. Given two tissues and a significance level  $\alpha$  we can compute the resemblance of the genes identified as age-related in the tissues. Resemblance is then a function of  $\alpha$ . Plotting the numerator  $|A \cap B|$  versus the denominator  $|A \cup B|$  as  $\alpha$  increases we obtain curves depicting the strength of the overlap.

In our setting with 16 tissues there are  $\binom{16}{2} = 120$  resemblances to consider. To keep the comparison manageable as well as to pool information from all tissues, we computed the following quantities

$$I_\alpha = \sum_{1 \leq j < j' \leq 16} |A_j^\alpha \cap A_{j'}^\alpha|, \quad \text{and} \quad U_\alpha = |\cup_{j=1}^{16} A_j^\alpha|, \quad (4.31)$$

where  $A_j^\alpha$  is the set of statistically significant genes at level  $\alpha$  for tissue  $j$ . We can think of  $I_\alpha/U_\alpha$  as a pooled resemblance. We would like to see large  $I_\alpha$  at each given level of  $U_\alpha$ .

Figure 4.7 plots  $I_\alpha$  versus  $U_\alpha$  for the methods we are comparing. To make a precise comparison we arranged for each method that estimated latent structure to employ the same estimate for the rank of the latent component. That rank is either 1, 2, 3, or the value chosen by the method of Buja and Eyuboglu (1992). At any rank LEAPP generates the most self-consistent gene lists over almost the entire range. EIGENSTRAT is usually second. SVA beats a raw method that makes no adjustments. LEAPP retains its strong performance when the rank is chosen from the data while the other two methods become poorer in that case.

Resemblance across tissues could also be high if there exist latent variables strongly correlated with age which are repeated across tissues. For example, consider a scenario where all tissues from young mice are in one batch, and all tissues from elder mice are in a different batch. If there are strong batch biases, then “age-related” genes would be reported by the raw method, and the same genes would be ranked high across all tissues. However, note that raw performs the worst of all methods in Figure 4.7, which gives some reassurance that the high resemblance of the other methods is due to successful removal of latent variables.

Given what we have learned from simulations, the relative performance of EIGENSTRAT and SVA gives us some insight into this data. Since EIGENSTRAT has done well, it is more likely that the signal is not very strong. Since SVA has done poorly, it is more likely that the latent variables in this data are correlated with age. There is also the possibility that they are correlated with sex (the covariate). Our simulations did not include a covariate.

### 4.5.2 Breast Cancer Microarray Study

Figure 4.8 concerns the “BRCA data” from Hedenfalk (2001) and Efron (2008), which can be obtained at <http://research.nhgri.nih.gov/microarray/NEJMSupplement>. It is a microarray experiment comparing two classes of breast cancer patients,  $N = 3226$  genes,  $n = 15$  microarrays,  $n_1 = 7$  for class 1,  $n_2 = 8$  for class 2. A two sample  $t$ -statistics  $t_i$  was computed and converted to  $z$ -value  $z_i = \Phi^{-1}(F_{13}(t_i))$ . The  $N$   $z$ -values have the following MLE empirical null using R package *locfdr* (see Efron (2008)).

$$z \sim \mathcal{N}(0.027, 1.583).$$

In this case, the theoretical  $\mathcal{N}(0, 1)$  null, substantially underestimates the  $z$ -values’ variability and there are 107 genes significant at 0.1 control level based on local  $\text{fdr}$  statistics under theoretical null. Their histogram looks like an overdispersed but short-tailed normal. In addition, we find that there are a few genes that have extremely large observations. Storey and Tibshirani (2003) also mentioned the presence of those outliers. We follow the approach in Storey and Tibshirani (2003) and

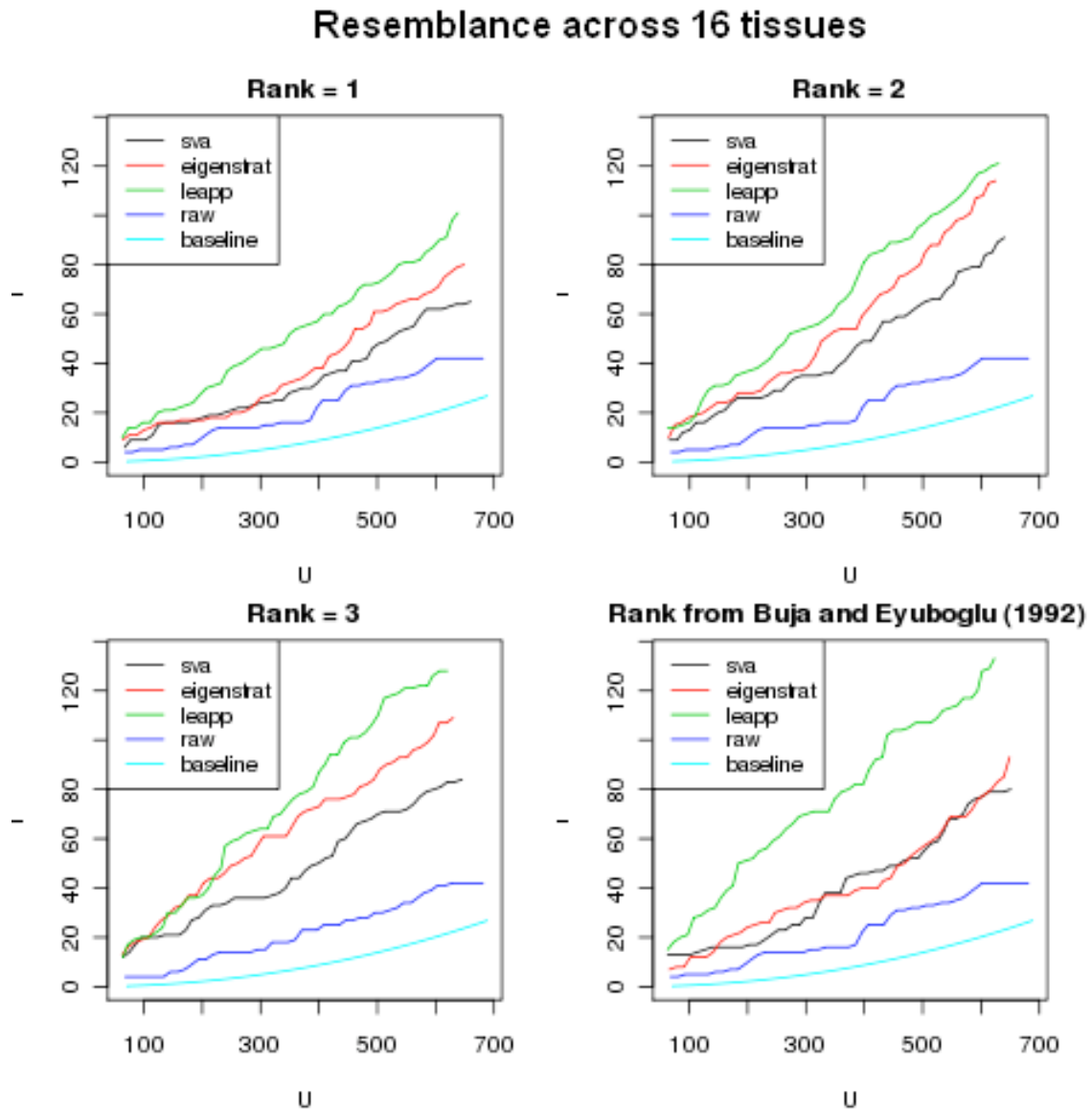


Figure 4.7: This figure shows the resemblance among significant gene sets from 16 tissues in the AGEMAP study. We plot  $I_\alpha$  versus  $U_\alpha$  (from equation (4.31)) increasing  $\alpha$  from 0 until  $U_\alpha = 700$ . The greatest self-consistency among lists is from LEAPP. EIGENSTRAT is second best. The baseline curve is computed assuming that the rankings for all 16 tissues are mutually independent.

eliminate genes that have one or more measurement exceeding 20. A value of 20 is several interquartile ranges away from the interquartile range of all of the data and

did not seem trustworthy for this example. This left us 3170 genes. However, the  $z$ -scores for those 3170 genes seem still far away from the theoretical null as shown in Table 4.5. Suspecting the existence of latent effects, we apply LEAPP, SVA and EIGENSTRAT to the gene expression data respectively and plot the adjusted  $z$ -values in Table 4.5. Table 4.6 shows the MLE mean and variance for each of the four methods. For SVA, EIGENSTRAT and LEAPP methods, algorithm according to Buja and Eyuboglu (1992) all select 5 latent factors. It seems that the empirical null distribution of  $z$ -values adjusted by LEAPP is the closest to standard normal distribution. It has mean 0.012 and standard deviation 1.018. EIGENSTRAT is the second and SVA performs poorly with the estimated empirical null  $\mathcal{N}(-0.009, 1.425)$ .

Table 4.7 shows the number of genes identified as true discoveries at 0.2 control level by 5 methods, which are local  $fdr$  statistics using theoretical null, empirical null, empirical null adjusted by SVA, empirical null adjusted by LEAPP and empirical null adjusted by EIGENSTRAT. The results were computed using R package *locfdr*. At 0.2 control level, the statistics by theoretical null found 127 positive and the statistics by empirical null adjusted by LEAPP identified 3 positives. All other methods based on empirical null found no positives. At 0.1 control level, none of the methods identified any true discovery.

Among the 3 significant genes at 0.2 control level detected by local  $fdr$  with empirical adjusted by LEAPP in Table 4.7, one gene called “HV5D3” with  $fdr$  statistics 0.11 is NAP1L4 nucleosome assembly protein 1-like 4. NCBI website describe this gene as follows:

NAP1L4 nucleosome assembly protein 1-like 4 is found to encode a member of the nucleosome assembly protein (NAP) family which can interact with both core and linker histones. It can shuttle between the cytoplasm and nucleus, suggesting a role as a histone chaperone. This gene is one of several located near the imprinted gene domain of 11p15.5, an important tumor-suppressor gene region. Alterations in this region have been associated with the Beckwith-Wiedemann syndrome, Wilms tumor, rhabdomyosarcoma, adrenocortical carcinoma, and lung, ovarian, and breast cancer.

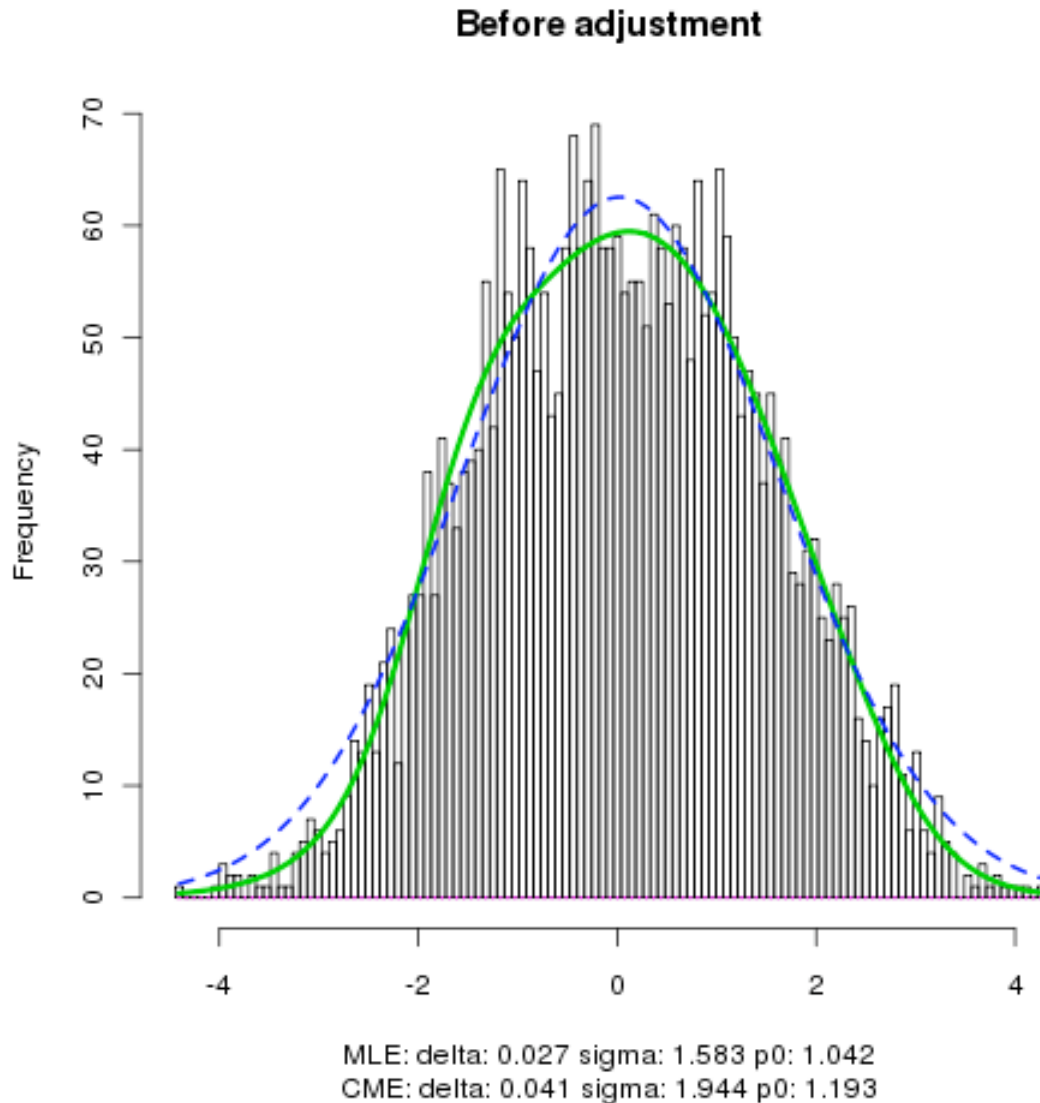


Figure 4.8: This figure shows the histogram of  $z$ -values from breast cancer microarray study Hedenfalk (2001), comparing 7 breast cancer patients having BRCA1 mutation to 8 with BRCA2 mutation  $N = 3226$  genes. Green solid curve is the fitted mixture density  $f$  and blue dashed curve is the fitted null subsdensity  $p_0 f_0$  and both are output from R package *locfdr*).

More details can be found in NCBI website database. The local  $\text{fdr}$  statistics based on theoretical null for “HV5D3” is 0.21, and it is preceded by 137 more significant

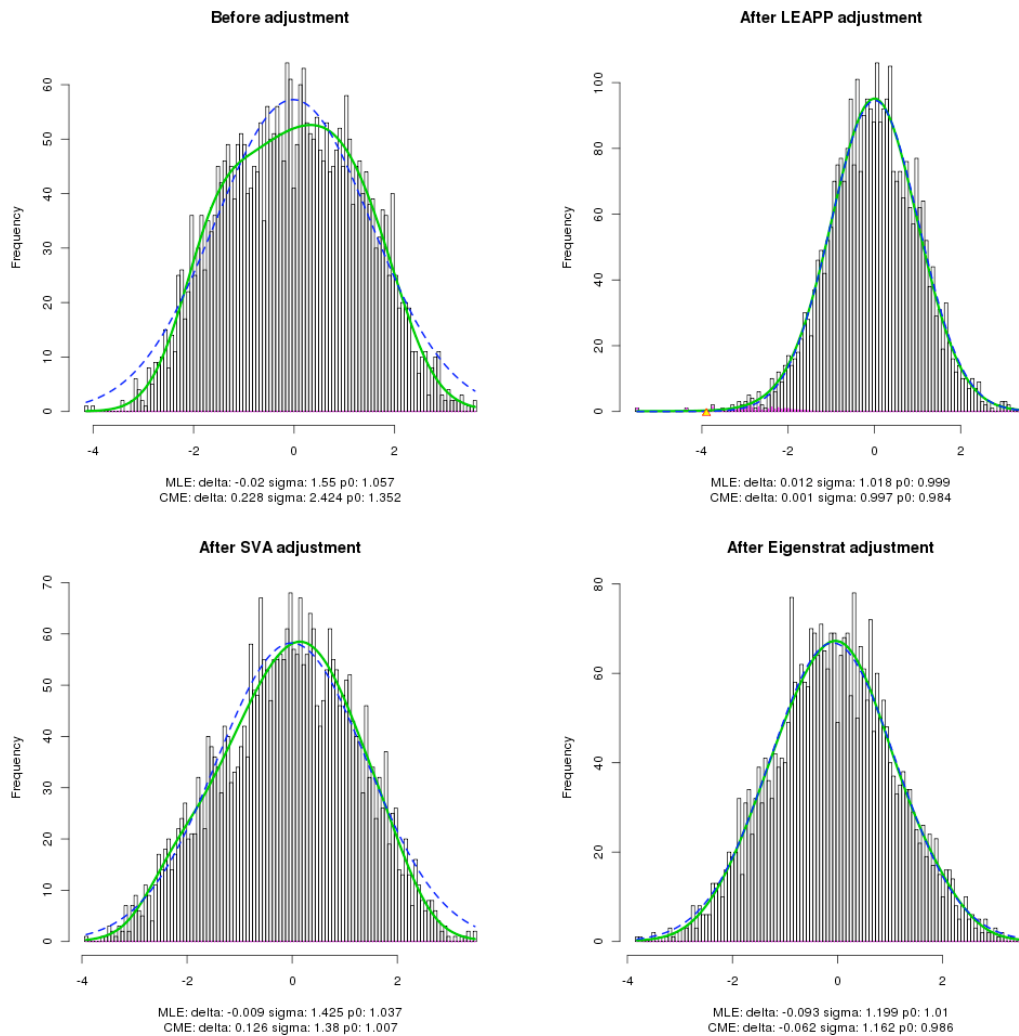


Table 4.5: Histogram of 3170  $z$ -values from filtered breast cancer microarray study Hedenfalk (2001) without any adjustment, with adjustment using LEAPP, SVA and EIGENSTRAT, with left to right, top to bottom respectively (by R package *locfdr*).

genes. We check the top 20 of those genes in terms of significance through the Google search engine, none of them have a reported association with breast cancer. The local  $\text{fdr}$  statistics based on empirical null without adjustment, adjusted by SVA or EIGENSTRAT are all 1.

|      | Empirical Null | LEAPP | SVA     | EIGENSTRAT |
|------|----------------|-------|---------|------------|
| mean | -0.02          | 0.012 | - 0.009 | -0.093     |
| std  | 1.55           | 1.018 | 1.425   | 1.199      |

Table 4.6: This table shows the MLE mean and variance for each of the four methods: empirical null without any adjustment, empirical null adjusted by LEAPP, SVA and EIGENSTRAT by *locfdr* R package.

|            | Theoretical | Empirical | LEAPP | SVA | EIGENSTRAT |
|------------|-------------|-----------|-------|-----|------------|
| BRCA data: | 127         | 0         | 3     | 0   | 0          |

Table 4.7: This table shows the number of genes identified as true discoveries, 0.2 control level for 5 methods, which are *fdr* statistics with theoretical null without adjustment, empirical null without any adjustment, empirical null adjusted by LEAPP, SVA and EIGENSTRAT. Local *fdr* statistics are calculated via *locfdr* R package.

## 4.6 Conclusions

High throughput testing has performance that deteriorates in the presence of latent variables. Latent variables that are correlated with the treatment variable of interest can severely alter the ordering of *p*-values. Our LEAPP method separates the latent variable from the treatment variable, making an adjustment possible. We have found in simulations that the adjustment brings about a better ordering among hypotheses than is available from either SVA or EIGENSTRAT. The improvement over SVA is largest when the latent variable is correlated with the primary one. The improvement over EIGENSTRAT is largest when the primary variable has a large effect. On the AGEMAP data we found better consistency among tissues for significance estimated by LEAPP than for either SVA or EIGENSTRAT. On the breast cancer data, the *z* scores adjusted by LEAPP has an empirical distribution closer to the theoretical null than either SVA or EIGENSTRAT.

## Part II

# Copy number variation detection, adjusting for latent variables

# Chapter 5

## Modeling Copy Number Variation

### 5.1 Introduction

For a biological sample, the DNA copy number of a genomic region is defined as the number of copies of the DNA in that region within the genome of the sample, relative to either a single control sample or a pool of population reference samples. Microarray technology has advanced the genome-wide fine scale measurement of DNA copy number in the past decade. Systematic studies are made possible to achieve a better understanding of the role of DNA copy number changes in human disease and in phenotypic variation in the human population (Zhang, 2010). This chapter reviews the computational and statistical problems that arise in DNA copy number data and surveys recent development in their treatment.

A copy number variant (CNV) is defined as a genomic region where the DNA copy number differs between two or more individuals from a population. Within the last five years, many studies ((McCarroll and Altshuler, 2007);(Cooper and Zerr, 2008)) have shown that CNVs are a common type of genetic variation in the human population and they also contribute to phenotypic variation.

Given the raw DNA copy number data from a single sample, an immediate challenge lies in recovering the true underlying copy number from the noisy measurements. This problem, often referred to as segmentation of total copy number, has drawn considerable attention and is reviewed in Section 5.2. A more complex problem is the

joint analysis of multiple copy number profiles, each coming from a different biological sample. There can be many different goals in such cross-sample analyses, which deserve different statistical approaches. Section 5.3 reviews the modeling issues and recent developments in cross-sample models for DNA copy number. The locations of rare and common copy number variants can be determined using segmentation or scanning strategies, but how to incorporate estimated CNVs to downstream association analysis is still an open problem. Section 5.4 reviews current approaches in CNV genotyping and areas for improvement.

## 5.2 Total Copy Number Estimation for One Sample

The total DNA copy number data for any given sample comes in the form of a sequence  $\{(x_i; y_i), i = 1, \dots, n\}$ , where  $n$  is the number of probes and  $x_i$  and  $y_i$  are the genome location and normalized intensity for probe  $i$  respectively. “Probe” and “normalized intensity” can be referred to Peiffer et al. (2006) for more details. The “total copy number” is the sum of the copy numbers for the chromosomes inherited from the two biological parents. It can vary over a continuous scale (see Zhang (2010)). The data from most platforms is in the form of a log ratio of the DNA quantity in the target sample versus the DNA quantity in an appropriate control. The normal state, where the copy number in the target agrees with that in the control, should have mean 0. A contiguous segment of measurements that are on average higher (or lower) than 0 indicates a gain (or loss) in copy number (Zhang (2010)).

The observed intensities are noisy surrogates of the true copy number at the measured positions. Since chromosomes are gained and lost in segments, adjacent positions in the genome are highly likely to have the same underlying copy number. This is why change-point models (Olshen et al. (2004) Zhang and Siegmund (2007)); smoothing methods (Lai et al. (2007), Tibshirani and Wang (2008)), Haar-based wavelets Hsu et al. (2005), spatially restricted clustering Xing et al. (2007), and various formulations of hidden Markov models (Guha et al. (2006); Colella et al.

(2007)) have been proposed for the estimation of DNA copy number. It is impossible to review in this chapter all of the above approaches and we focus on CBS Olshen et al. (2004) and fused Lasso Tibshirani and Wang (2008). In Subsection 5.2.1, We introduce the change-point formulation for this problem that underlies the Circular Binary Segmentation (CBS) algorithm (Olshen et al. (2004); Venkatraman and Olshen (2007)) which is one of the simplest and most transparent methods. We discuss the fused Lasso method Tibshirani and Wang (2008) in Subsection 5.2.2.

### 5.2.1 CBS Based Change Point Detection

Since the location of the probes, at a coarse global scale, is approximately uniformly distributed in the genome, the location information  $\{x_i, i = 1, \dots, n\}$  is often ignored in the segmentation process. Then, a simple changepoint model for the sequence of intensities is

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \dots, n$$

where  $\mu = \{\mu_i : i = 1, \dots, n\}$  is a piecewise constant function of  $i$ , and  $\epsilon_i, i = 1, \dots, n$  are i.i.d. errors. We assume that there exists a series of change-points  $0 = \tau_0 < \tau_1 < \dots < \tau_m < \tau_{m+1} = n$  such that

$$\mu_i = \theta_t, i \in [\tau_t, \tau_{t+1}), t = 0, \dots, m.$$

The errors are usually assumed to be Gaussian, although this assumption is not crucial if the distances between successive  $\tau_j$ 's are large. Under this model, the segmentation problem becomes estimating the changepoints and the means within each segment. The number of change-points  $m$  is also not known and has been observed to range from below 10 to above 100 in some tumor samples. If the values of the change-points  $\tau$  are known, then  $\theta_j$  can be estimated by the mean of the observations that fall in the  $j$ -th segment. To estimate  $\tau$ , the CBS algorithm uses a greedy top-down approach that recursively applies the generalized likelihood ratio statistic for testing a change. To be specific, for any interval  $1 \leq a < b \leq n$ , let the null hypothesis be that the observations are i.i.d. Gaussian and let the alternative be that there is a sub-interval

with a change in mean and no change in variance. The generalized likelihood ratio statistic is

$$\max_{a < s < t < b} Z_{s,t}, \quad \text{where} \quad Z_{s,t} = \frac{S_t - S_s - \frac{t-s}{b-a}(S_b - S_a)}{\hat{\sigma} \sqrt{(t-s)[1-(t-s)/(b-a)]}}, \quad (5.1)$$

and  $S_j = y_1 + \dots + y_j$ . CBS starts by setting  $a = 1, b = n$ . Let  $z^{\text{obs}}$  be the observed maximum of  $Z_{s,t}$ , and  $(s^*, t^*)$  be the maximizing interval. If the p-value of the scan,  $\mathbb{P}(\max_{a < s < t < b} Z_{s,t} > z^{\text{obs}})$ , is smaller than some preset threshold  $\alpha$ , then the maximizing interval is reported and the intervals  $[a, s^*), [s^*, t^*), [t^*, b]$  are recursively scanned using the same procedure. The recursion stops when none of the subregions contain a change that is significant at the level  $\alpha$ .

The p-value for the scan statistic in (5.1) can be computed using asymptotic approximations given by James et al. (1987), which are quite accurate for tail probabilities. Alternatively, Zhang and Siegmund (2007) proposed a modified BIC criterion for estimating  $m$ , and showed that, when used in conjunction with CBS, has more accurate off-the-shelf performance than p-value based thresholds.

## 5.2.2 Fused Lasso

The “fused lasso” (Tibshirani et al., 2004) is a generalization of the lasso (Tibshirani, 1996) and is defined by

$$\hat{\beta} = \arg \min \sum_i \left( y_i - \sum_j x_{ij} \beta_j \right)^2 \quad (5.2)$$

$$\text{subject to } \sum_{j=1}^p |\beta_j| \leq s_1, \quad (5.3)$$

$$\sum_{j=2}^p |\beta_j - \beta_{j-1}| \leq s_2. \quad (5.4)$$

In the copy number variation detection, Tibshirani et al. (2004) applies the fused lasso in the special case when  $\{x_{ij}, i = 1, \dots, n, j = 1, \dots, p\}$  is the identity matrix.

$$\hat{\beta} = \arg \min \sum_i (y_i - \beta_i)^2 \quad (5.5)$$

$$\text{subject to } \sum_{j=1}^p |\beta_j| \leq s_1, \quad (5.6)$$

$$\sum_{j=2}^p |\beta_j - \beta_{j-1}| \leq s_2. \quad (5.7)$$

The resulting coefficient vector  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_n)$  is piecewise constant and serves as an estimate for the true copy number variation like  $\hat{\mu}$  in Subsection 5.2.1. Here,  $s_1$  controls the overall DNA copy number alteration amount of the target chromosome (or chromosome arm), while  $s_2$  controls the frequency of the alterations in the target region.  $s_1, s_2$  are estimated from pre-smoothed data. They threshold  $|\hat{\beta}_i|$  by a value  $a$  to obtain the final regions of gain or loss.  $a$  is varied over a range to seek the solution with estimated false discovery rate closest to the target, such as 0.01. Their simulation result suggests that fused lasso better captures the true DNA copy number alterations than CBS, especially when the aberration width is small.

### 5.3 Cross Sample Copy Number Variation Detection

In an integrated analysis of copy number data across multiple biological samples, it is often interesting to know the differences across samples as well as the similarities. The underlying signal is not shared and usually, only a fraction of the samples are carriers of any given CNV.

One goal in copy number studies over a cohort of tumor samples is to find regions of recurrent aberration. Such regions, where a large number of samples of the same type of tumor have gained or lost copies, may contain genes that are key drivers in the development of the tumor. For example, Figure 6.2 shows a set of tumor

samples, with many samples carrying overlapping deletions covering chromosome 9 of Schiffman et al. (2009). Such commonly deleted regions may carry genes that play a role in cell proliferation or delay apoptosis. Similarly, commonly amplified regions may contain tumor suppressor genes. The boundary of a CNV region may not be shared across samples. In post-segmentation procedures (Newton et al. (1998), Rouveirol et al. (2006)), each sample is segmented on its own, and the cross sample analysis sees only the segmented data. However, in some cases, such as inherited CNVs, the change-points are shared across samples for instances of the same CNV. In such cases, aggregating information across samples can improve the power of detecting shared weak signals. Zhang et al. (2010) shows that joint segmentation methods for detecting inherited CNVs can boost power.

It is often assumed that it is more likely for biologically significant aberrations to recur across samples than experimental or statistical errors. However, many errors in segmentation are due to experimental artifacts, such as local trends, which also recur across samples at the same locations. Also one commonly used assumption in the aforementioned methods is that the measurement errors are independent and identically distributed. However, this assumption may not be valid as Diskin et al. (2008) shows that the whole-genome microarrays with large-insert clones designed to determine DNA copy number often show variation in hybridization intensity that is related to the genomic position of the clones, called “genomic waves”, and it correlates best with GC content. Unfortunately, the causes of genomic waves are not well understood and those systemic effects can not be completely adjusted for. Consequently, they may prevent accurate inference of copy number variations (CNVs). Diskin et al. (2008) proposed a GC-wave factor(GCWF) measure and developed a computational approach by fitting regression models with GC content included as a predictor variable, which is shown to improve the accuracy of CNV detection. The GC content may not exactly capture the latent systemic effects which appear as genomic waves and regression on GC content may also run the risk of removing part of the signals of CNVs. We propose an alternating algorithm in Chapter 6 to adjust the latent systemic effects for copy number variation detection for tumor samples where change-points may not be shared across samples for instances of the same CNV.

## 5.4 Copy Number Variation Genotyping

In the analysis of both inherited and somatic copy number variants, it is often useful to obtain a sparse cross-sample summary of a complex region for use in downstream analyses. For example, in clinical studies we may have, along with the copy number data, variables such as survival outcome or status of other biomarkers. We may want to find chromosomal regions whose copy number status is correlated with these variables. These types of analyses are often done with gene or protein expression data, but for copy number data it is unclear what to use as the explanatory variables. If each probe were considered as a variable, then the smoothness of the underlying signal is ignored. Since copy number studies now routinely use platforms containing hundreds of thousands to over a million probes, if each probe were considered a variable we would be faced with a very large number of highly correlated variables, which would reduce the sensitivity of downstream analyses. Some studies take the average copy number over each chromosome, chromosome arm, or cytoband as the variables for downstream analysis. This clearly is a coarse method that sacrifices sensitivity.

Barnes et al. (2008) proposed a principal component based method for weighting the probes and obtained summary statistics for downstream association study. Their method is incorporated into R package CNVtools and we give a brief description below. Given the observed log intensity matrix  $X \in \mathbb{R}^{n \times N}$ , where  $N$  is the number of probes and  $n$  is the number of samples, they computed the first principal component score of  $X$  and then fit a mixture normal distribution on this principal component score with a user specified number of clusters  $K$ . Following an initial fit of the Gaussian mixture model, they obtained an  $n \times K$  matrix of posterior copy number probabilities  $P$ .  $P_{ik}, i = 1, \dots, n, k = 1, \dots, K$  is the probability of  $i$ th sample in cluster  $k$ . They then found the first canonical correlation, i.e., the unit vector of  $a$  and  $b$  which maximize the correlation between  $Xa$  and  $Pb$ .

$$\hat{a}, \hat{b} = \underset{a, b}{\operatorname{argmax}} \operatorname{cor}(Xa, Pb) \quad \text{s.t.} \quad \|a\| = \|b\| = 1.$$

The first canonical variate  $X\hat{a}$  then provided their improved composite score and they used it as a surrogate for the true copy number in the downstream association study. For details, see the probe weighting section of the supplementary materials of Barnes et al. (2008).

One possible issue is that their method assumes implicitly that response rates and noise levels are the same across SNPs but this may not be true in practice as the measurement quality of SNPs can vary significantly. In Chapter 7, we propose an Expectation-Maximization based algorithm for CNV genotyping in the analysis of both inherited and somatic copy number variants and shows via simulation and real data example that our method achieve better accuracy than CNVtools Barnes et al. (2008).

# Chapter 6

## Cross Sample CNV Detection

In this chapter, we consider a model that takes into account latent systemic effects and propose an alternating algorithm to calibrate the model. Section 6.1 introduces our model and notations. In our model, the latent systemic effects are captured by a low rank matrix. We calibrate the model by an alternating algorithm, called CNVlatent, estimating the latent effect and signal matrix together. Section 6.2 shows the simulation result of our model on synthetic data compared with a modified fused lasso method without adjusting for latent effects, adjusting for the true latent effect matrix revealed by an oracle and adjusting for latent effects with known number of latent factors.

### 6.1 Model and Notations

The model is

$$Y = S + L + E \tag{6.1}$$

for variables

|   |  |
|---|--|
| $Y \in \mathbb{R}^{n \times N}$                   | response values  |
| $S \in \mathbb{R}^{n \times N}$                   | scaled copy numbers, value in each row is piecewise constant |
| $L \in \mathbb{R}^{n \times N}$                   | latent systemic effects, of rank $k$                         |
| $E \sim \mathcal{N}(0, \sigma^2 I_N \otimes I_n)$ | noise  |

with dimensions

|            |                              |
|------------|------------------------------|
| $n$        | number of samples            |
| $N \gg n$  | number of SNPs, and          |
| $k \geq 1$ | latent dimension, often one. |

As we have mentioned in Chapter 5, the underlying signal is not shared across samples and usually, only a fraction of the samples are carriers of any given CNV. Let  $D \in \mathbb{R}^{N \times (N-1)}$ .

$$D = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots \\ -1 & 1 & 0 & \cdots & \cdots \\ 0 & -1 & 1 & 0 & \cdots \\ \cdots & 0 & -1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & -1 & 1 \\ \cdots & \cdots & \cdots & 0 & -1 \end{pmatrix}$$

$$D_{ij} = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = j + 1 \\ 0 & \text{otherwise} \end{cases}$$

Then  $SD$  is the differenced intensity across positions. We assume that  $SD$  is a sparse matrix such that the changes in copy number across SNPs and samples are few. We

estimate  $L$  and  $S$  by the following optimization:

$$\begin{aligned} & \text{minimize}_{S,L} \|Y - L - S\|_F^2 + \lambda \|SD\|_1 \\ & \text{subject to: } \text{rank}(L) \leq r \end{aligned} \tag{6.2}$$

where  $r, \lambda$  are tuning parameters.  $r$  specifies the rank of the latent effect matrix or the number of latent factors and  $\lambda$  controls the number of change points. This is not a convex problem but it is bi-convex in the sense that if we fix one variable, the problem is convex in the other variable. We can obtain a local optimum by an alternating algorithm called CNVlatent, which is detailed in Algorithm 1.

---

**Algorithm 1** CNVlatent Algorithm

---

**Require:**  $Y, \lambda, r, \text{Max.Iter}$  and  $\text{TOL}$

$k \leftarrow 0, L_0 \leftarrow \mathbf{0}$ .

**while**  $L_k, S_k$  have not converged at  $\text{TOL}$  and  $k < \text{Max.Iter}$  **do**

$S_k = \text{argmin}_S \|Y - S - L_{k-1}\|_F^2 + \lambda \|SD\|_1$  (Step  $S_k$ : Sparse Update)

$L_k = \text{argmin}_{\text{rank}(L) \leq r} \|Y - S_k - L\|_F^2$  (Step  $L_k$ )

$k \leftarrow k + 1$

**end while**

---

We stop when the relative change is small such that  $\|L_k - L_{k+1}\| / \|L_k\| + \|S_k - S_{k+1}\| / \|S_k\| < \text{TOL}$ . Step ( $L_k$ ) can be solved by a singular value decomposition truncated at rank  $r$  so that  $L_k$  is the best rank  $r$  approximation to  $Y - S_k$ . Step ( $S_k$ ) is a generalized lasso problem and we can carry out a change of variable and reduce it to a regular lasso problem, which is detailed in Lemma 4.

**Lemma 4.** *Consider the following generalized lasso problem*

$$\text{minimize}_S \|Y - S\|_F^2 + \lambda \|SD\|_1, \tag{6.3}$$

where  $Y \in \mathbb{R}^{n \times N}$  is the response matrix,  $S \in \mathbb{R}^{n \times N}$  is the variable matrix and matrix  $D \in \mathbb{R}^{N \times (N-1)}$  has full column rank. There exists  $A \in \mathbb{R}^{N \times 1}$  such that  $\tilde{D} = (D, A)$  is invertible. Let  $\tilde{D}^{-1} = \begin{pmatrix} \mathcal{X}_1^{(N-1) \times N} \\ \mathcal{X}_2^{1 \times N} \end{pmatrix}$  and  $P = \mathcal{X}_2^T (\mathcal{X}_2 \mathcal{X}_2^T)^{-1} \mathcal{X}_2$ . Let  $\hat{\Theta}_1$  be the

solution to the following minimization problem,

$$\hat{\Theta}_1 = \underset{\Theta_1}{\operatorname{argmin}} \|Y(I - P) - \Theta_1 \mathcal{X}_1(I - P)\|_F^2 + \lambda \|\Theta_1\|_1 \quad (6.4)$$

and  $\hat{\Theta}_2$  be

$$\hat{\Theta}_2 = (Y - \hat{\Theta}_1 \mathcal{X}_1) \mathcal{X}_2^\top (\mathcal{X}_2 \mathcal{X}_2^\top)^{-1}.$$

Then  $\hat{S} = (\hat{\Theta}_1, \hat{\Theta}_2) \tilde{D}^{-1}$  is the solution to (6.3).

*Proof.* Since  $D$  has full column rank, there exists  $A \in \mathbb{R}^{N \times 1}$  such that  $\tilde{D} = (D, A)$  is invertible. We change variable to

$$\Theta = S\tilde{D} = S(D, A) = (SD, SA).$$

Let  $\Theta_1 = SD, \Theta_2 = SA$ . Then

$$S = \Theta \tilde{D}^{-1} = \Theta_1 \mathcal{X}_1 + \Theta_2 \mathcal{X}_2,$$

so that optimization problem (6.3) becomes

$$\operatorname{minimize}_{\Theta} \|Y - \Theta_1 \mathcal{X}_1 - \Theta_2 \mathcal{X}_2\|_F^2 + \lambda \|\Theta_1\|_1. \quad (6.5)$$

It is clear that at the solution the second block of the variable  $\Theta$  is given by a linear regression:

$$\hat{\Theta}_2 = (Y - \hat{\Theta}_1 \mathcal{X}_1) \mathcal{X}_2^\top (\mathcal{X}_2 \mathcal{X}_2^\top)^{-1}$$

Therefore by substituting  $\Theta_2$  with  $\hat{\Theta}_2$  in (6.5), we can rewrite (6.3) as

$$\operatorname{minimize}_{\Theta_1} \|Y(I - P) - \Theta_1 \mathcal{X}_1(I - P)\|_F^2 + \lambda \|\Theta_1\|_1, \quad (6.6)$$

where  $P = \mathcal{X}_2^\top (\mathcal{X}_2 \mathcal{X}_2^\top)^{-1} \mathcal{X}_2$ . Let its solution be  $\hat{\Theta}_1$  and  $\hat{\Theta} = (\hat{\Theta}_1, \hat{\Theta}_2)$ . We can back-transform to get solution for  $\hat{S} = \hat{\Theta} \tilde{D}^{-1}$ .  $\square$

Since Step  $S_k$  is the same as solving the problem (6.3) and thus by Lemma 4, we have the following algorithm.

**Algorithm 2** Algorithm for  $S_k$ **Require:**  $\tilde{D}, \tilde{D}^{-1}$ 

$$\hat{\Theta}_1 = \operatorname{argmin}_{\Theta_1} \|Y(I - P) - L_{k-1}(I - P) - \Theta_1 \mathcal{X}_1(I - P)\|_F^2 + \lambda \|\Theta_1\|_1$$

$$\hat{\Theta}_2 = (Y - L_{k-1} - \hat{\Theta}_1 \mathcal{X}_1) \mathcal{X}_2^\top (\mathcal{X}_2 \mathcal{X}_2^\top)^{-1}$$

$$S_k = (\hat{\Theta}_1, \hat{\Theta}_2) \tilde{D}^{-1}$$

For Algorithm 1, we just need to compute  $\tilde{D}, \tilde{D}^{-1}$  once. The calculation of  $\tilde{D}$  and  $\tilde{D}^{-1}$  is detailed below:

**Compute  $\tilde{D}, \tilde{D}^{-1}$ :** There exists  $A \in \mathbb{R}^{N \times 1}$  such that  $\tilde{D} = (D, A)$  is invertible. For example,  $A$  can be chosen as  $A = (1, 1, \dots, 1)^\top$ , then each row of  $\tilde{D}^{-1}$  should be orthogonal to all the columns of  $\tilde{D}$ .

Let the  $i$ th row of  $\tilde{D}^{-1}$  be  $\tilde{D}_i^{-1} = (\tilde{D}_{i,1}^{-1}, \dots, \tilde{D}_{i,N}^{-1})$ .  $\tilde{D}_i^{-1} \tilde{D} = e_i$  implies for  $1 \leq i \leq N - 1$ ,

$$\begin{aligned} \tilde{D}_{i,1}^{-1} &= \tilde{D}_{i,2}^{-1} = \dots = \tilde{D}_{i,i}^{-1} \\ \tilde{D}_{i,i}^{-1} - \tilde{D}_{i,i+1}^{-1} &= 1 \\ \tilde{D}_{i,i+1}^{-1} &= \tilde{D}_{i,i+2}^{-1} = \dots = \tilde{D}_{i,N}^{-1} \\ \sum_{j=1}^N \tilde{D}_{i,j}^{-1} &= 0, \end{aligned}$$

and for  $i = N$ ,

$$\begin{aligned} \tilde{D}_{N,1}^{-1} &= \tilde{D}_{N,2}^{-1} = \dots = \tilde{D}_{N,N}^{-1} \\ \sum_{j=1}^N \tilde{D}_{N,j}^{-1} &= 1. \end{aligned}$$

A bit of calculation shows that

$$\tilde{D}_{ij}^{-1} = \begin{cases} \frac{N-i}{N} & \text{if } j \leq i \\ -\frac{i}{N} & \text{if } j > i \end{cases}$$

for  $1 \leq i \leq N-1, 1 \leq j \leq N$  and  $D_{Nj} = \frac{1}{N}, j = 1, \dots, N$ . Let  $\tilde{D}^{-1} = \left( \mathcal{X}_1^{(N-1) \times N}, \mathcal{X}_2^{1 \times N} \right)$ , then  $\mathcal{X}_2 = \left( \frac{1}{N}, \dots, \frac{1}{N} \right)$ .

There are many algorithm for solving a regular lasso problem, including coordinate descent (Friedman et al., 2007), LARS (Efron et al., 2004) and sophisticated convex optimization procedures such as Alternating Direction Method Multipliers (ADMM) (Boyd et al., 2010). We choose to use pathwise coordinate descent to solve the regular lasso problem in Algorithm 2 mainly because it is faster than its competitors and easy to implement. Tseng (2001) shows that the coordinate descent algorithm converges to a minimizer of the objective function  $f$  if  $f(x_1, x_2, \dots, x_p) = g(x_1, \dots, x_p) + \sum_j h_j(x_j)$  where  $g(\cdot)$  is convex and differentiable and  $h_j(\cdot)$  is convex. This condition holds for the Step  $S_k$  but not the whole algorithm.

To determine the tuning parameter  $\lambda$  and  $r$ , we devise a special kind of cross validation. For cross validation, We need to divide samples into a training set and a test set and because of the ‘‘smoothness’’ of the copy numbers across SNPs, we can take ordered odd numbered SNPs as ‘‘predictors’’ and even numbered SNPs as ‘‘response’’ as adjacent signals should be similar. It would be beneficial to think of the blocks after rearranging columns as

$$\begin{pmatrix} Y_{\text{train:predictor}} & Y_{\text{train:response}} \\ Y_{\text{test:predictor}} & Y_{\text{test:response}} \end{pmatrix}$$

Following (Owen and Perry, 2009), we use the hold-in set

$$\begin{pmatrix} Y_{\text{train:predictor}} & Y_{\text{train:response}} \\ Y_{\text{test:predictor}} & \star \end{pmatrix}$$

to predict the hold-out  $Y_{\text{test:response}}$ . We fit CNVlatent algorithm to  $Y_{\text{train:predictor}}$

and get the estimated signal matrix  $\hat{S}_{\text{train:predictor}}$  and the estimated latent effect matrix  $\hat{L}_{\text{train:predictor}}$ . Take SVD of  $\hat{L}_{\text{train:predictor}}$  and use its principal component loadings as predictors to fit a regression function to the “cleaned” response columns  $Y_{\text{train:response}} - \hat{S}_{\text{train:predictor}}$ . The underlying assumption is that the signals  $S_{\text{train:predictor}}$  and  $S_{\text{train:response}}$  should be similar, so  $\hat{S}_{\text{train:predictor}}$  is a good substitute for  $S_{\text{train:response}}$ . Then, to evaluate the function on the hold-out set, we apply the CNVlatent to  $Y_{\text{test:predictor}}$  and get the estimated signal matrix  $\hat{S}_{\text{test:predictor}}$  and the estimated latent effect matrix  $\hat{L}_{\text{test:predictor}}$ . And then we apply estimated regression function to the principal component loadings of  $\hat{L}_{\text{test:predictor}}$  to get a prediction  $\hat{L}_{\text{test:response}}$  for the latent structure. The final prediction will be  $\hat{Y}_{\text{test:response}} = \hat{L}_{\text{test:response}} + \hat{S}_{\text{test:predictor}}$ . Let the estimated prediction error from model at  $(r, \lambda)$  be  $\text{RSS}(r, \lambda)$ . We choose

$$(\hat{r}, \hat{\lambda}) = \min_{r, \lambda} \text{RSS}(r, \lambda)$$

$\tilde{D}, \tilde{D}^{-1}, P$  will be computed only once. Step  $S_k$  requires  $O(n(N-1)^2)$  and Step  $L_k$  requires  $O(Nnr)$  where  $r$  is the number of factors selected.

## 6.2 Performance On Synthetic Data

In this section, we generate data from the model (6.1) and compare the results from the algorithms to an oracle which is given the latent variable, to an oracle which is given the rank of the latent effect matrix and to a raw method which makes no attempt to adjust for latent variables.

We simulate an intensity matrix  $Y$  with sample size  $n = 80$  and number of SNPs  $N = 100$ . The difference signal matrix  $\Theta_1 = SD \in \mathbb{R}^{n \times (N-1)}$  has independent and identically distributed entries such that

$$\Theta_{1ij} = \begin{cases} 1 & \text{w.p. } N^{-\gamma} \\ -1 & \text{w.p. } N^{-\gamma} \\ 0 & \text{w.p. } 1 - 2N^{-\gamma}, \end{cases}$$

where  $\gamma = 1$ . And let the signal matrix  $S \in \mathbb{R}^{n \times N}$  simply be the cumulative sum of  $\Theta_1$ :  $S_{i1} = 0, S_{ij} = \sum_{l=1}^j \Theta_{1il}, i = 1, \dots, n, j = 2, \dots, N$ . We use  $k = 2$  latent factors. The latent effect matrix  $L$  is simulated by  $L = UV^T$  where  $U \in \mathbb{R}^{n \times 2}, V \in \mathbb{R}^{N \times 2}, U_{il}$  i.i.d.  $\sim \mathcal{N}(0, 1), V_{jl}$  i.i.d.  $\sim 0.5\mathcal{N}(0, 1), i = 1, \dots, n, j = 1, \dots, N, l = 1, \dots, k$ . We add noise  $E \sim 0.1\mathcal{N}(0, 1)$ . We carry out the simulation 100 times and apply the cross validation detailed in the last section 6.1 with rank  $r$  ranging from 0 to 3 and  $\lambda$  ranging from 0.01 to 0.3. We use Max.Iter = 100 and TOL =  $10^{-4}$ . The average signal to latent ratio  $\frac{\|S\|_F^2}{\|L\|_F^2}$  is 0.81. We will denote this simulation setting as “uniform”.

It is more difficult to detect the change points when the range of the copy number variation is short. We can control the difficulty of the problem by simulating the interarrival time between change points in each sample instead of the location of change points. We have the same settings as the previous simulation except that the signal matrix is no longer independently and identically distributed taking values in  $\{-1, 0, 1\}$ . For row  $i$  of  $S$ , we simulate the location of the first change point  $f_i$  following Poisson distribution with mean  $N/5$  and the interval between the location of the second change point and that of the first change point  $\text{dur}_i$  following Poisson distribution with mean  $N/2$  and let the location of the second change point be  $s_i = \min(f_i + \text{dur}_i, N)$ . For each sample, it starts at 0 and can gain 1 or lose 1 at the first change point and returns to 0 at the second change point. The signal to latent ratio  $\frac{\|S\|_F^2}{\|L\|_F^2}$  is 0.41. We will denote this simulation setting as “poisson”.

The methods that we applied are as follows:

- oracle** an oracle given  $UV^T$  which then uses our algorithm with rank  $r = 0$ ,
- raw** use our algorithm on  $Y$  ignoring latent variables with rank  $r = 0$ ,
- rankknown** use our algorithm on  $Y$  with rank  $r = 2$ ,
- cnvlatent** use our algorithm on  $Y$  and cross validation to choose  $\lambda, r$ .
- pca** use our algorithm on  $Y - \hat{Y}_k$ ,  
where  $\hat{Y}_k$  is the best rank  $k$  approximation to  $Y$ ,  
and  $k$  is determined by cross validation (Owen and Perry, 2009).

Table 6.1 summarizes the estimated number of latent factors using cross validation for two simulation settings mentioned above. The simulation was conducted 100

times. In the “uniform” setting, we can see that 45 out of 100 times we obtained the correct rank 2, highest frequency among the 4 possible ranks. In the “poisson” setting, we obtained the correct rank 2, 55 out of 100 times. When there are no latent factors, we got the rank right 99 out of 100 times for both simulation settings. There

| Simulation | True rank | Estimated rank |    |    |    |
|------------|-----------|----------------|----|----|----|
|            |           | 0              | 1  | 2  | 3  |
| uniform    | $r = 2$   | 13             | 23 | 45 | 19 |
|            | $r = 0$   | 99             | 1  | 0  | 0  |
| poisson    | $r = 2$   | 8              | 18 | 55 | 19 |
|            | $r = 0$   | 99             | 1  | 0  | 0  |

Table 6.1: This table shows the counts of number of latent factors estimated by cross validation in category  $\{0, 1, 2, 3\}$ .

are many ways to evaluate the accuracy of estimated  $\hat{S}$  and we base our evaluation on the accuracy of estimation of the location of change points. The following is the loss function we use:

$$\text{Loss}(\hat{S}, S) = \| (\hat{S} - S) D \|_1$$

where for  $\mathcal{A} \in \mathbb{R}^{n \times N}$ ,  $\|\mathcal{A}\|_1 = \sum_{i=1}^n \sum_{j=1}^N |\mathcal{A}_{ij}|$ . This loss function tells us how accurate our estimator  $\hat{S}$  is. Table 6.2 summarizes the result of losses of estimate  $S$  by the four methods described above when the true number of latent factor is 2. Our method for known latent effect matrix (true) achieves the lowest loss and method for known number of latent factor (knownrank) the second, our method (CNVlatent) the third, pca method follows and the method without adjusting for latent factors (raw) has the highest loss. It seems knowing the correct number of latent factors can much improve the result. The Figure 6.1 shows the recovered signal matrix  $S$  in the poisson simulation setting. CNVlatent reconstructs the signal quite well. It captures

| simulation | oracle | knownrank | cnvlatent | pca    | raw    |
|------------|--------|-----------|-----------|--------|--------|
| uniform    | 21.32  | 30.61     | 70.56     | 73.35  | 107.10 |
| poisson    | 39.60  | 60.98     | 121.24    | 157.32 | 222.38 |

Table 6.2: This table shows the loss incurred  $\text{Loss}(\hat{S}, S)$ .

the major signals while adjusting for the latent effects. The raw method without any adjustment was severely affected by latent effects and had many false discoveries especially at both ends of the region simulated.

### 6.3 Real Data

We consider first the chromosome 9p region in 44 pediatric leukemia samples, which were analyzed using Molecular Inversion Probe technology in Schiffman et al. (2009). Figure 6.2 shows this region, which was covered by 276 probes in the assay. It is a very noisy data set with a lot of stripes across samples around probe 65, which indicate the presence of latent systemic effects. This region harbors obvious overlapping deletions. We applied CNVlatent method to this data set with cross validation. The number of latent factor selected by CNVlatent is 1. We find that CNVlatent reconstructs this region quite well. The homozygous deletion around probe number 140 encompasses 15,027 bp in length and maps to chr9:21,962,445-21,977,472 (NCBI Build 35.1). This region belongs to the CDKN2A locus, which encodes a tumor suppressor gene. Many of the hemizygous and homozygous deletions of CDKN2A found in these data have been validated by RT-PCR. See Schiffman et al. (2009) for a complete analysis of these data. Also in the reconstructed region, the stripes around probe 65 were removed by CNVlatent successfully. The only problem is that CNVlatent seems to overfit and picked a lot of small segments at both ends of the region.

As a second example, we analyze the region on cytoband 11 on the q-arm of chromosome 22. As documented in the Database of Genomic Variants (Iafate et al., 2004), this region has nested deletions with different break-points between individuals in the population. Our data come from a set of 62 Illumina 550K beadchips described in Zhang et al. (2010). We focus on a 2000 marker segment of the data covering the region of interest. The CNVlatent algorithm was applied to this region to yield results summarized in Figure 6.3. The number of latent factor was found to be zero. The segmentation in Figure 6.3 captures the obvious variant regions around probe number 1000-1300 and 1800-1900. The latent variables seem to be present but quite weak and CNVlatent missed it.

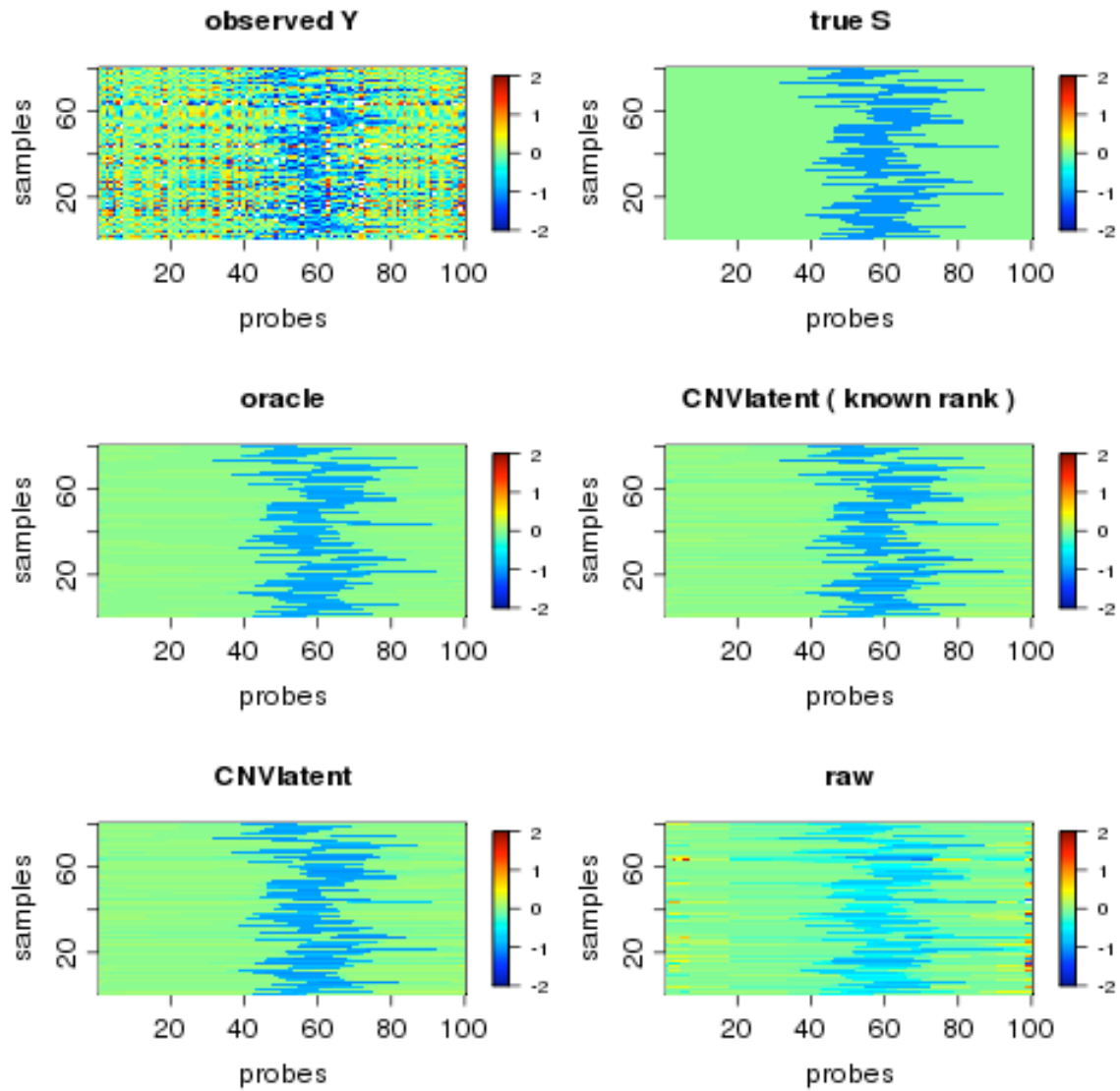


Figure 6.1: Recovered signal matrix  $\hat{S}$  from 4 methods and the true signal matrix simulated in the poisson setting.

## 6.4 Discussion

We introduced a general framework for detecting copy number variants in the presence of latent systemic effects. We find that the optimization problem is biconvex, so we may end up with local optimum instead of the global optimum. To find the global optimum, a multistart version of CNVlatent can be used. It would be also be interesting to know when the decomposition of response matrix  $Y$  to be the sum of matrix  $S$  such that  $SD$  is sparse and a low rank matrix  $L$  is identifiable. We can rewrite the decomposition as

$$\begin{aligned} Y &= S + L \\ YD &= SD + LD \end{aligned}$$

Let  $Y' = YD, S' = SD, L' = LD$ , then

$$Y' = S' + L'$$

where  $S'$  is a sparse matrix and  $L'$  is a low rank matrix. This is a robust matrix decomposition problem with sparse corruptions, see Hsu et al. (2011), Candes et al. (2009) for more details. Hsu et al. (2011) consider the situation that  $S'$  is a fixed matrix and show that when the maximum number of non-zero entries in any row or column of  $S'$  is not too large and the singular vectors of  $L'$  are not too sparse, the decomposition is identifiable. And Hsu et al. (2011) also shows that such a decomposition is possible via a combination of  $l_1$  norm and nuclear norm minimization (6.7) under mild conditions on the number of non-zero entries in  $S'$  and sparsity of singular vectors in  $L'$ . The solution is robust under small perturbation of  $Y'$ .

$$\begin{aligned} &\text{minimize } \lambda \|S'\|_1 + \|L'\|_* \\ &\text{subject to } Y' = S' + L' \end{aligned} \tag{6.7}$$

Candes et al. (2009) considers the situation when the sparsity pattern of  $S'$  is

selected uniformly at random. To ensure identifiability, they assume the right and left singular vectors of the low rank matrix  $L'$  satisfy the incoherence condition such that the singular vectors are reasonably spread out and thus not sparse. Candes et al. (2009) also shows that exact recovery is possible for a given  $\lambda = 1/\sqrt{\max(n, N)}$  under mild conditions on the rank of matrix  $L'$  and the number of non-zeros in  $S'$ .

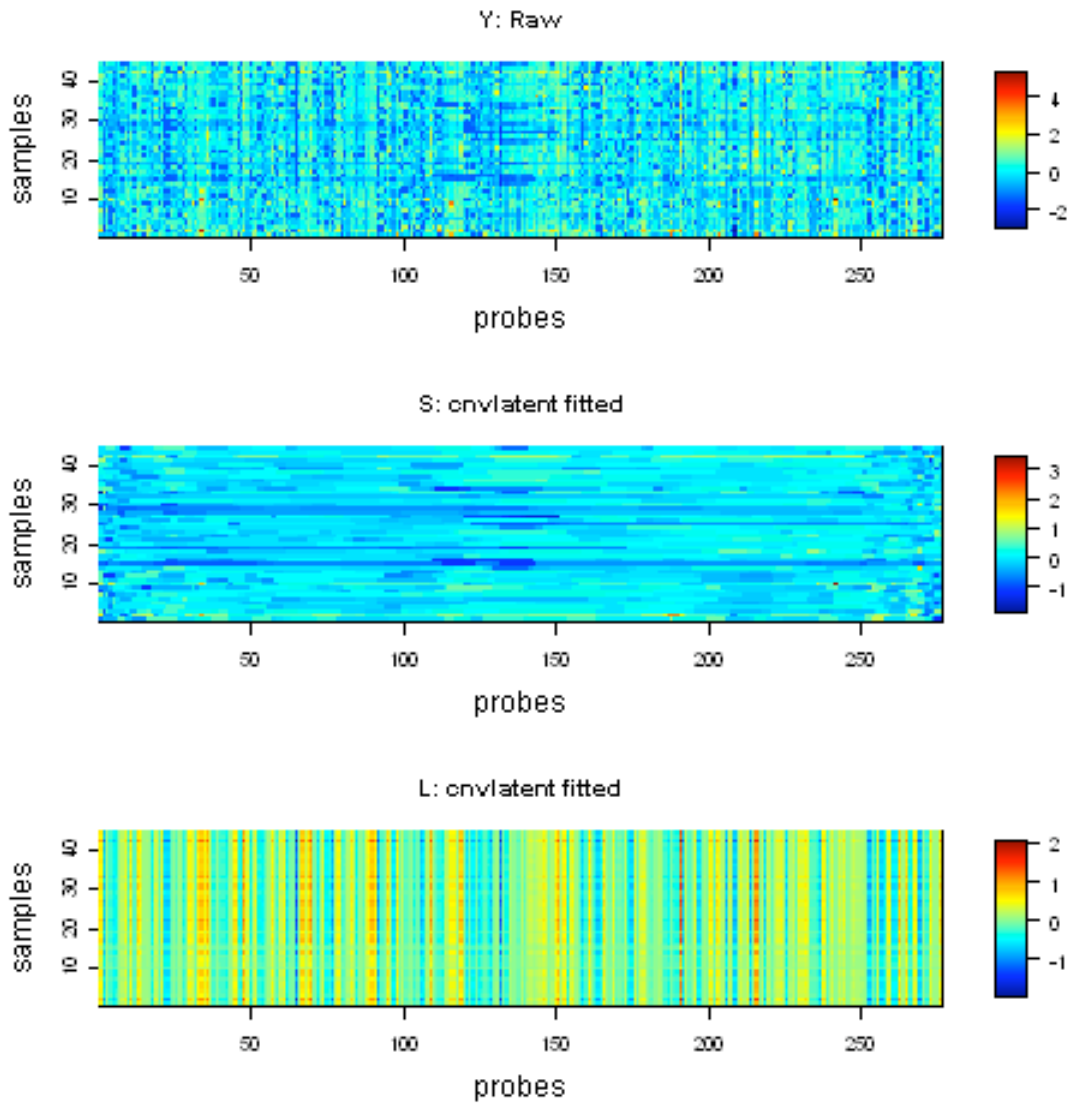


Figure 6.2: Chromosome 9p in 44 Leukemia samples

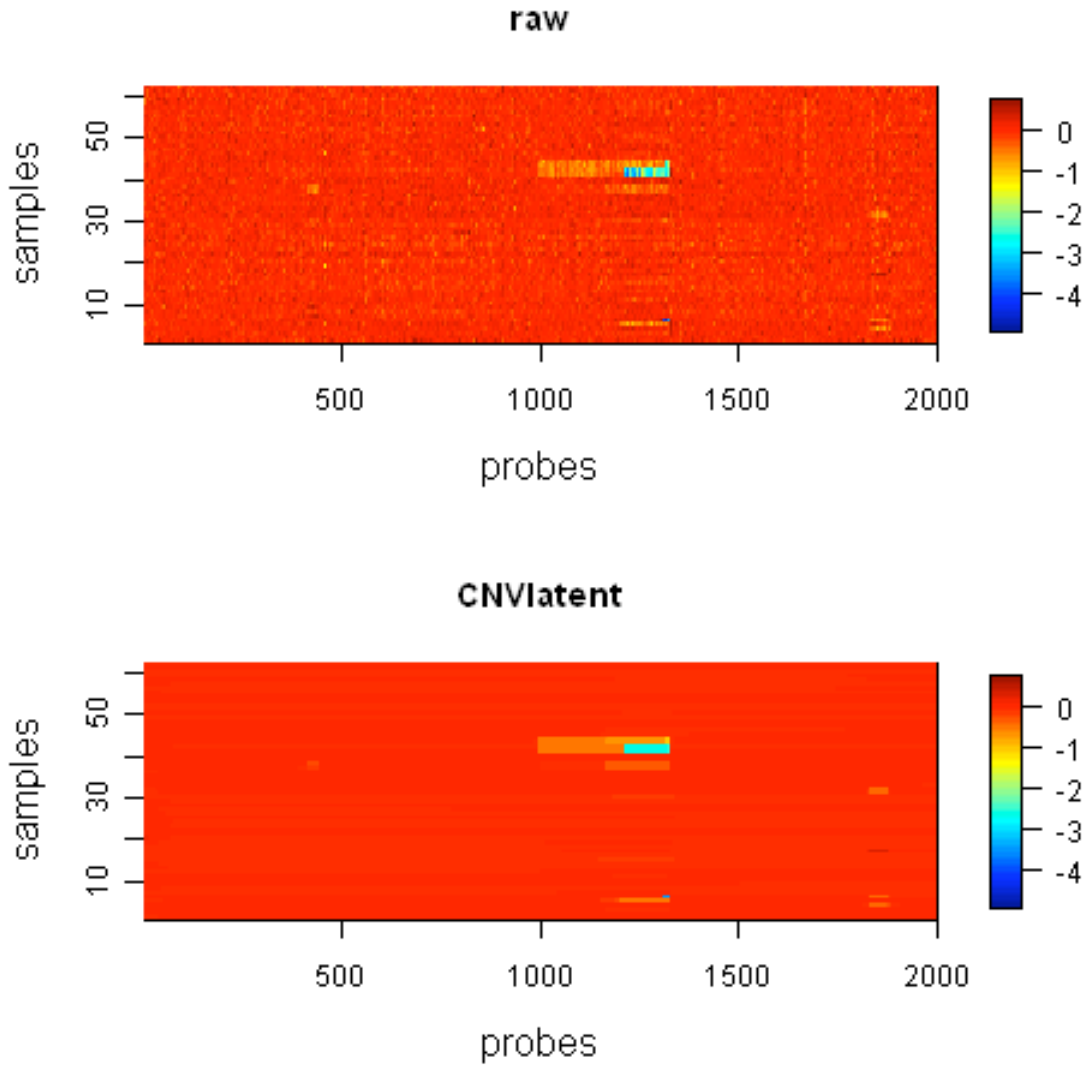


Figure 6.3: Chromosome 22

# Chapter 7

## EM Based CNV Genotyping

In this chapter we present an Expectation-Maximization based algorithm for CNV genotyping in the analysis of both inherited and somatic copy number variants discussed in Section 5.4. Section 7.1 introduces our model and notations. Numerical examples in Section 7.2 and real data in Section 7.3 show that our method achieves better accuracy than Barnes et al. (2008). This is a model for CNV detection with shared boundary between subjects, which is completely different from the model of CNVlatent in Chapter 6.

### 7.1 Model

We introduce a model for total probe intensity alone as follows:

$$X = \alpha\mu^\top + \Sigma E \tag{7.1}$$

where

|   |   |
|---|---|
| $p$   | number of probes  |
| $n$   | number of subjects  |
| $X \in \mathbb{R}^{p \times n}$                   | observed probe log intensity values                           |
| $\mu \in \mathbb{R}^n$                            | true log intensity, i.i.d. entries                            |
|   | $P(\mu_i = u_k) = \pi_k, k = 1, \dots, K,$                    |
|   | $\sum_{k=1}^K \pi_k = 1, \pi = (\pi_1, \dots, \pi_K)$         |
| $\alpha \in \mathbb{R}^p$                         | response rate of SNPs   |
|   | $\sum_{i=1}^p \alpha_i = 1, \alpha_i \geq 0, i = 1, \dots, p$ |
| $E \sim \mathcal{N}(0, I_p \otimes I_n)$          | noise, and,   |
| $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_p)$ | standard deviations of noise.                                 |

The motivation of the model comes from the fact that though the number of SNPs in the detected region may range from tens to hundreds, the center of clusters (subjects with same putative CNV) often reside in a low dimensional space as a result of smoothness of signal and segmentation. For each subject, each SNP in the detected region gives a noisy measurement of the same putative CNV in the form of probe intensity. The quality of the measurement is determined by the response rates of the SNPs. If the response rate is low, we may not be able to discriminate gains and losses of the CNV across samples. We model  $\mu_i$  as a discrete random variable and there are K levels of probe intensity ( $u = (u_1, \dots, u_K)$ ) which correspond to K possible scenarios of copy number gains or losses. The response rate of the jth SNP is modeled as  $\alpha_j$ . To avoid identifiability issues, we assume  $\alpha_j \geq 0, \sum_{j=1}^p \alpha_j = 1$ . The positivity assumption is attributed to the fact that the information across probes should be consistent. If  $\alpha_j$  is close to zero, the jth SNP has a low response rate. The model allows a SNP specific noise level.

Suppose we know that the number of levels of probe intensity is  $K$ , our goal is to estimate  $\mu_i, i = 1, \dots, n$  for putative CNVs.

If we have the information of  $(X, I(\mu_i = u_k), i = 1, \dots, n, k = 1, \dots, K)$ , the full log likelihood function  $l(u, \alpha, \sigma, \pi)$  is:

$$l = \sum_{i=1}^n \sum_{j=1}^p \sum_{k=1}^K \left[ -\frac{(x_{ij} - u_k \alpha_j)^2}{2\sigma_j^2} - \log(\sigma_j) \right] I(\mu_i = u_k) \quad (7.2)$$

$$+ \sum_{i=1}^n \sum_{k=1}^K \log(\pi_k) I(\mu_i = u_k) \quad (7.3)$$

where  $I(\cdot)$  is an indicator function.

To maximize the log likelihood function when  $I(\mu_i = u_k), i = 1, \dots, n, k = 1, \dots, K$  is not observed, we resort to the Expectation Maximization algorithm by alternating through E-step and M-step defined as follows.

**E-step** Compute

$$\hat{p}_{ik} = \mathbb{E}(I(\mu_i = u_k) | X, \hat{u}, \hat{\alpha}, \hat{\sigma}, \hat{\pi}) \quad (7.4)$$

$$= \mathbb{P}(\mu_i = u_k | X, \hat{u}, \hat{\alpha}, \hat{\sigma}, \hat{\pi}) \quad (7.5)$$

$$\text{Set } \hat{p}_{ik} = \hat{\pi}_k \prod_{j=1}^p \exp \left( -(X_{ij} - \hat{u}_k \hat{\alpha}_j)^2 / (2\hat{\sigma}_j^2) \right) / \hat{\sigma}_j \quad (7.6)$$

$$\hat{p}_{ik} = \hat{p}_{ik} / \sum_{k=1}^K \hat{p}_{ik} \quad (7.7)$$

Let  $cl(u, \alpha, \sigma, \pi) = \mathbb{E} (l(u, \alpha, \sigma, \pi) | X, \hat{u}, \hat{\alpha}, \hat{\sigma}, \hat{\pi})$ .

**M-step**

$$\text{Maximize}_{u,\alpha,\sigma,\pi} cl(u, \alpha, \sigma, \pi) \text{ s.t. } \alpha_j \geq 0, \sum_j \alpha_j = 1. \quad (7.8)$$

$$\text{Maximize}_{u,\alpha,\sigma,\pi} \sum_{i=1}^n \sum_{j=1}^p \sum_{k=1}^K \left[ -\frac{(x_{ij} - u_k \alpha_j)^2}{2\sigma_j^2} - \log(\sigma_j) \right] \hat{p}_{ik} + \sum_{i=1}^n \sum_{k=1}^K \log(\pi_k) \hat{p}_{ik} \quad (7.9)$$

$$\text{s.t. } \alpha_j \geq 0, \sum_j \alpha_j = 1. \quad (7.10)$$

where  $\hat{p}_{ik}, i = 1, \dots, n, k = 1, \dots, K$  are computed from the the E-step which depends on the estimated parameter  $\hat{u}, \hat{\alpha}, \hat{\sigma}, \hat{\pi}$ .

$l(u, \alpha, \sigma, \pi)$  is not a convex function but by recognizing that  $\max_u l(u, \alpha, \sigma, \pi)$ ,  $\max_\alpha l(u, \alpha, \sigma, \pi)$ ,  $\max_\sigma l(u, \alpha, \sigma, \pi)$ ,  $\max_\pi l(u, \alpha, \sigma, \pi)$  all have very simple and efficient solution, we can alternate among these individual optimization problems and the algorithm is summarized as below.

We stop when  $\|u(k) - u(k-1)\|/\|u(k-1)\| + \|\sigma(k) - \sigma(k-1)\|/\|\sigma(k-1)\| + \|\alpha(k) - \alpha(k-1)\|/\|\alpha(k-1)\| < \text{TOL}$ . The update step for  $\alpha(t)$  is a typical quadratic programming problem with linear constraints and a variety of methods are commonly used including interior point, augmented Lagrangian, conjugate gradient and gradient projection (Bonnans et al. (2006), Calamal and More (1987)). We use gradient projection method for updating  $\alpha(t)$  as it is easy to implement and has good convergence rate. Its convergence conditions are shown in Bertsekas (1976).

For initial value of Algorithm 3, we set  $\pi_k(0) = \frac{1}{K}, k = 1, \dots, K$ . Let  $\mathcal{PC} \in \mathbb{R}^n$  be the first principal component of  $X$ . We apply k-means clustering on  $\mathcal{PC}$  and set  $u(0)$  to be the ordered estimated cluster means for  $K$  clusters. Let  $\mathcal{U} \in \mathbb{R}^n$  and  $\mathcal{U}_i$  be the mean of the assigned cluster for subject  $i$ . Regress rows of  $X$  on  $\mathcal{U}$  and take the estimated coefficients as  $\alpha(0) \in \mathbb{R}^p$ . If  $\alpha(0) \neq \mathbf{0}$ , standardize  $\alpha(0)$  to be positive and sum up to 1. Otherwise, set  $\alpha_j(0) = \frac{1}{p}, j = 1, \dots, p$ . Take the sample standard deviations of the residuals from the regression across SNPs as  $\sigma(0)$ .

In reality, we don't know the number of possible CNVs and so we use the following BIC criterion to select  $K$ :

---

**Algorithm 3** EM based CNV genotyping

---

**Require:**  $X$ ,  $K$ , initialize  $\text{TOL} = 10^{-4}$ ,  $\text{Max.Iter} = 50$  and  $u(0), \alpha(0), \sigma(0), \pi(0)$  $t \leftarrow 0$ **while** maximum relative error larger than  $\text{TOL}$  and  $t < \text{Max.Iter}$  **do**

compute

$$p_{ik}(t+1) = \pi_k(t) \prod_{j=1}^p \exp\left(\frac{(X_{ij} - u_k(t)\alpha_j(t))^2}{2\sigma_j(t)^2}\right) / \sigma_j(t)$$

$$p_{ik}(t+1) = p_{ik}(t+1) / \sum_{k=1}^K p_{ik}(t+1)$$

compute

$$\pi(t+1) = \underset{\pi}{\operatorname{argmax}} cl(u(t), \alpha(t), \sigma(t), \pi)$$

i.e.  $\pi_k(t+1) = \frac{\sum_{i=1}^n p_{ik}(t+1)}{n}$

  compute  $\alpha(t+1) = \underset{\alpha}{\operatorname{argmax}} cl(u(t), \alpha, \sigma(t), \pi(t+1))$ , s.t.  $\alpha_j \geq 0, \sum_j \alpha_j = 1$ 

compute

$$u(t+1) = \underset{u}{\operatorname{argmax}} cl(u, \alpha(t+1), \sigma(t), \pi(t+1))$$

i.e.  $u_k(t+1) = \frac{\sum_{i=1}^n \sum_{j=1}^p \frac{X_{ij}\alpha_j(t+1)p_{ik}(t+1)}{\sigma_j(t)^2}}{\sum_{i=1}^n \sum_{j=1}^p \frac{\alpha_j(t+1)^2 p_{ik}(t+1)}{\sigma_j(t)^2}}$

compute

$$\sigma(t+1) = \underset{\sigma}{\operatorname{argmax}} cl(u(t+1), \alpha(t+1), \sigma, \pi(t+1))$$

i.e.  $\sigma_j(t+1) = \sqrt{\frac{\sum_{i=1}^n \sum_{k=1}^K (X_{ij} - u_k(t+1)\alpha_j(t+1))^2 p_{ik}(t+1)}{n}}$

 $t \leftarrow t + 1$ **end while**

---

$$BIC(k) = -2\hat{l} + (2p + 2k - 2) \log(n).$$

## 7.2 Performance on Synthetic Data

We simulate log intensity values with sample size  $n = 200$ , copy number changes  $K = 2$  and response rate for  $j$ th SNP  $\alpha_j$  i.i.d. exponential with mean 0.5,  $j = 1, \dots, p$ , true log intensity  $u_k = \log(k/2)$ ,  $k = 1, 2$ , that is, samples can either have a deletion or have normal copy number 2 in this segment. The probability to have a deletion is  $\pi_1 = 0.25$ , the probability to have 2 copies is  $\pi_2 = 0.75$ . With estimated posterior  $\{\hat{p}_{il}\}_{i=1, \dots, n, l=1, \dots, K}$ , we assign subject  $i$  to level  $u_k$  only if  $k = \arg \max_l \hat{p}_{il}$ . Our precision measure is the proportion of labels assigned correctly. The simulation is run 100 times. We compare our method with that of CNVtools which essentially is soft clustering on the first principal component. For both methods, we know that the number of clusters is  $K = 2$ .

| $\sigma_j$                   | p  | CNVtools | EM   |
|------------------------------|----|----------|------|
| $0.3 + 0.5I(\alpha_j > 0.5)$ | 5  | 0.66     | 0.82 |
|                              | 50 | 0.83     | 0.99 |
| $0.1 + 0.5I(\alpha_j > 0.5)$ | 5  | 0.70     | 0.96 |
|                              | 50 | 0.89     | 0.99 |

Table 7.1: Comparison of our method(EM) and CNVtools in percentage of correct identification.

Table 7.1 shows that the EM method has better precision than CNVtools. And our improvement is more pronounced when the number of SNPs is small. To see why EM method works better, we generate an artificial Gaussian mixture of two equal size clusters in  $\mathbb{R}^2$  with  $\sigma_1 = 0.1$  for the x coordinate and  $\sigma_2 = 0.6$  for the y coordinate as shown in the Figure 7.1. The difference of noise levels in x,y coordinates makes CNVtools more error prone while the EM based clustering splits the data neatly by estimating variance structure across features. As shown in the figure, CNVtools has more mislabeling than our method.

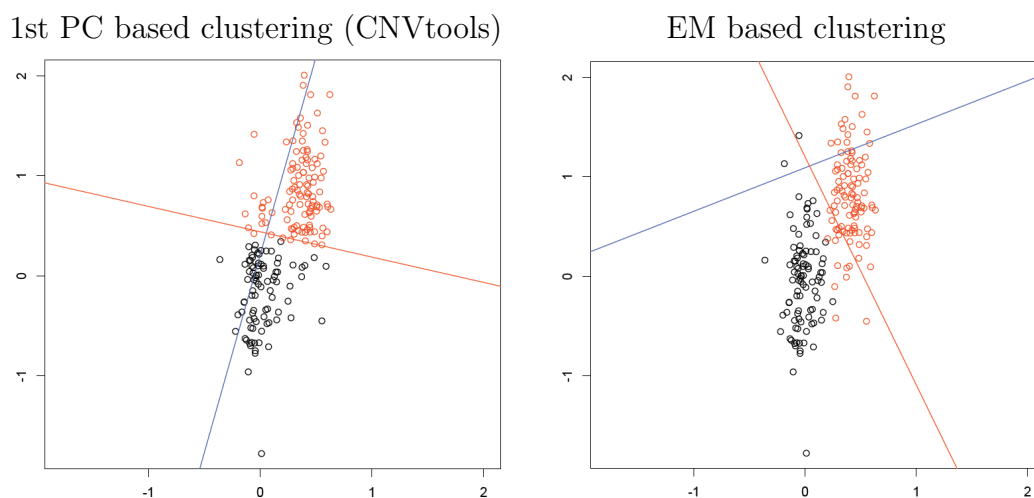


Figure 7.1: This figure shows the clustering performance between CNVtools and our EM based method. In the left panel, blue line is the principal component direction and red line is orthogonal to the blue line. In the right panel, the blue line is the optimal projection direction and the red line is the direction that splits two clusters found by EM based method. Detected clusters are colored in red and black respectively for CNVtools and our method.

The simulation result is similar when  $K$  is unknown and we can use BIC criterion to choose the number of clusters for EM and CNVtools respectively.

### 7.3 Real Data

We validate the model (7.1) with Levinson 302 data which contain 151 pairs of replicate samples, i.e., the same sample assayed twice in repeated experiments. This data set was generated for quality control within the Stanford psychiatric department. Let's call these data sets  $D_1$  and  $D_2$ . Then, true copy number variation region should be concordant across replicates. Also, the true underlying genotypes (i.e. copy number level) should match. Any scanning and genotyping algorithm can be evaluated based on the concordance rate. So, we can compare our method versus CNVtools in Barnes et al. (2008) on this data set.

We apply CNV segmentation as in Zhang et al. (2010) to data sets  $D_1$  and  $D_2$

respectively. We define an intersection segment be the index range  $[a, b]$  such that  $a = \max(a_1, a_2), b = \min(b_1, b_2)$  where  $[a_i, b_i]$  is a segment found in  $D_i, i = 1, 2$  and  $[a_2, b_2]$  is the closest segment found in  $D_2$  to segment  $[a_1, b_1]$  and  $[a_1, b_1]$  overlaps  $[a_2, b_2]$ . The distance between two intervals is defined as the distance between the middle points between two intervals. We find 55 such intersection segments between  $D_1, D_2$ . We then compare our method versus CNVtools in Barnes et al. (2008) on those intersection segments of  $D_1, D_2$  respectively. Let  $l_{ij}^1$  be the cluster label estimated for  $i$ th segment and  $j$ th subject in data set  $D_1$  and  $l_{ij}^2$  for that of data set  $D_2$ . For a particular segment  $i$ , we define classification error rate  $r_i$  as  $\frac{\sum_{j=1}^{151} I(l_{ij}^A \neq l_{ij}^B)}{151}$ . To simplify the comparison, we combine levels of copy number and only consider three levels for copy number changes: deletion ( $< 2$ ), normal ( $= 2$ ) and insertion ( $> 2$ ). And for both methods, we use BIC to select number of mixtures (number of types of possible copy number changes). The pair of estimators from data  $D_1$  and  $D_2$  agree with each other if both fall into the same category. This error rate measures the disagreement of the results from the replicate data sets. As the plot shown below, in general, our method gives a much smaller classification error rate and it indicates better concordance than CNVtools. We noticed that there are a few cases that CNVtools gave exactly zero classification error but our method gave small but positive errors and they are the cases when CNVtools detected only one cluster for both data sets  $D_1, D_2$  while our method identifies 2 or more clusters for at least one data set. If one method is extremely conservative and can not detect more clusters other than the normal two copy group, then our classification error will be exactly zero all the times no matter what might be the underlying truth. Hence it is not always better if we get exactly zero classification error. On the other hand, when our method had zero classification error but CNVtools did not, our methods was able to identify more than one clusters in both data sets  $D_1, D_2$  and the labels for the samples agreed exactly.

We define another measure of concordance of results from the repeated data sets  $D_1$  and  $D_2$  is the sum of correlation of scores defined as:

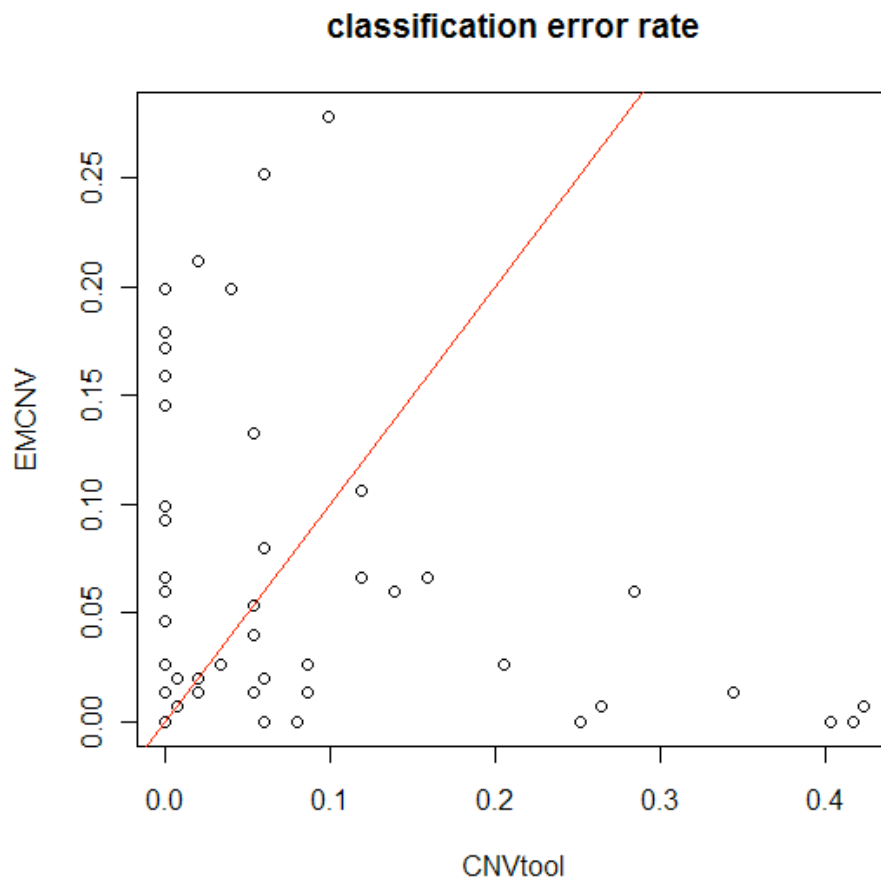


Figure 7.2: Classification Error: x axis is the error rate for CNVtools and y axis is the error rate for our EM based method

$$\sum_{i=1}^{55} cor(s_{1i}, s_{2i})$$

where  $s_{1i}, s_{2i} \in \mathbb{R}^{151}$  are the estimated putative CNVs for  $i$ th overlapped segments in data sets  $D_1, D_2$  respectively.

The sum of correlation of scores computed by using CNVtools is 25.61 and that of our method is 40.4. This result again suggests that our method has better concordance.

In the real probe intensity data set we have, a significant number of segments have

varying noise levels across SNPs. In the following Figure 7.3, there are two SNPs in the detected CNV region and our EM clustering method discovered the third hidden cluster with the BIC criterion while the CNVtools was not able to.

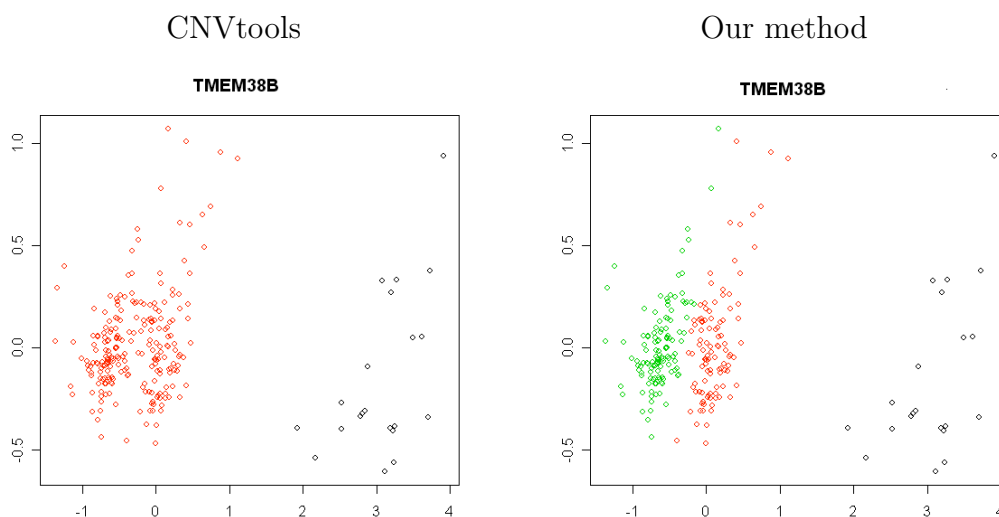


Figure 7.3: x coordinate: probe intensity of SNP 1; y coordinate: probe intensity of SNP 2. Clusters detected are colored. CNVtools detected 2 clusters (red, black) and our method detected 3 clusters (green, red, black).

## 7.4 Discussion

We introduced a general framework for CNV genotyping in a detected CNV segment where the boundaries of CNV are shared across samples. Our method is based on Expectation-Maximization method which will guarantee convergence but may end up in the local minima. A good initial value is important and we suggest using multistart. Because of the normality assumption of our method, it is sensitive to outliers and can lead to misleading result if outliers are not filtered in the first place. The simulation and the real data example show that our method performs better than the existing method CNVtools Barnes et al. (2008).

## Appendix

Our proposal splits the data set into two pieces. We then apply two different methods to those pieces. Those methods are described here.

### A.1 $\Theta$ -IPOD

Here we describe the  $\Theta$ -IPOD algorithm of She and Owen (2011). We use the standard regression notation, so that, for example,  $Y$  has a usual regression response meaning in this appendix, not the matrix version we use in the body of the article.

The additive outlier model is

$$Y = X\beta + \gamma + \varepsilon,$$

where  $Y \in \mathbb{R}^n$ ,  $X \in \mathbb{R}^{n \times p}$ ,  $\beta \in \mathbb{R}^p$ ,  $\gamma \in \mathbb{R}^n$  and  $\varepsilon_i$  are IID mean zero random variables. In terms of vector indices, the model is

$$Y_i = X_i^\top \beta + \gamma_i + \varepsilon_i.$$

The parameter  $\gamma_i$  is the coefficient of a dummy variable that eliminates a potential outlying  $i$ 'th observation. This formulation was earlier used by Gannaz (2006)

There are  $n + p$  parameters. Those in  $\gamma$  need to be strongly regularized to prevent fitting a saturated model.

The first criterion they consider is

$$\|y - X\beta - \gamma\|_2 + \sum_{i=1}^n \lambda_i |\gamma_i|$$

with  $\lambda_i = \lambda \sqrt{1 - h_i}$ , where  $h_i$  is the  $i$ 'th diagonal element of the hat matrix  $H = X(X^\top X)^{-1}X^\top$  and  $\lambda \geq 0$  must be determined.

Given  $\gamma$  we get  $\beta$  by regression of  $Y - \gamma$  on  $X$ . Given  $\beta$  we get  $\gamma$  by thresholding:

$\gamma_i = \Theta(y_i - X_i\beta, \lambda_i)$ . An  $L_1$  penalty corresponds to

$$\Theta(z, \lambda) = \Theta_{\text{soft}}(z, \lambda) = \begin{cases} 0 & |z| \leq \lambda \\ z - \text{sgn}(z)\lambda & |z| > \lambda. \end{cases}$$

An  $L_1$  penalty corresponding to soft thresholding does not produce robust results. Better results come from hard thresholding:

$$\Theta_{\text{hard}}(z, \lambda) = \begin{cases} 0 & |z| \leq \lambda \\ z & |z| > \lambda. \end{cases}$$

The choice of  $\lambda$  is based on a modified BIC from Chen and Chen (2008).

To handle unequal error variances, we write

$$Y_i = X_i^T \beta + \gamma_i + \sigma_i \varepsilon_i,$$

for  $\sigma_i > 0$ . Making the replacements  $\tilde{Y}_i = Y_i/\sigma_i$ ,  $\tilde{X}_i = X_i/\sigma_i$  and  $\tilde{\gamma}_i = \gamma_i/\sigma_i$  we get

$$\tilde{Y}_i = \tilde{X}_i^T \beta + \tilde{\gamma}_i + \varepsilon_i.$$

We apply the original IPOD algorithm to the reweighted points and then multiply the estimated  $\gamma_i$  by  $\sigma_i$ .

## A.2 Criss-cross regressions

Gabriel and Zamir (1979) studied how to fit regression models of the form

$$Y = X\beta^T + \delta Z^T + UV^T + E$$

where  $X$  are measured features of the rows with coefficients in  $\beta$ ,  $Z$  are measured features of the columns with coefficients in  $\delta$  and  $UV^T$  are latent variables (both unknown) and  $E$  is an error matrix.

They estimate  $\beta$ , then  $\delta$  then  $U$  and  $V$  sequentially. First

$$\hat{\beta} = Y^T X (X^T X)^{-1},$$

then

$$\hat{\delta} = (Y - X\hat{\beta})Z(Z^T Z)^{-1}$$

and finally setting  $R = Y - X\hat{\beta}^T - \hat{\delta}Z^T$  he computes the SVD  $R = U_R D_R V_R^T$  and picks  $\hat{U}$  and  $\hat{V}$  such that  $\hat{U}\hat{V}^T = U_R D_R V_R^T$ . For example the singular values  $D_R$  can be absorbed into either the left or right factors, or  $D_R^{1/2}$  can be absorbed into each. In our analysis we suppose that  $\hat{V}$  is normalized to be a unit vector.

When the rows have unequal variances we may replace  $E$  by  $\Sigma E$ . Our approach to criss-cross regression is then to alternate between fitting criss-cross regression to  $\hat{\Sigma}^{-1}E$  and estimating  $\Sigma$  by

$$\hat{\Sigma}^2 = \frac{1}{n} \text{diag} \left( (Y - X\hat{\beta}^T - \hat{\delta}Z^T - \hat{U}\hat{V}^T)(Y - X\hat{\beta}^T - \hat{\delta}Z^T - \hat{U}\hat{V}^T)^T \right).$$

Our setting is somewhat simpler than the general case. We do not have  $\delta Z^T$  term. We followed the steps above but with  $\delta Z^T = 0$ .

# Bibliography

- Allen, G. and Tibshirani, R. (2010). Inference with transposable data: modeling the effects of row and column correlations. Technical report, Stanford University, Department of Statistics.
- Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.
- Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.
- Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221.
- Balding, D. and Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identify and paternity. *Genetica*, 96:3–12.
- Barnes, C., Plagnol, V., and Fitzgerald, T. (2008). A robust statistical method for case-control association testing with copy number variation. *Nature Genetics*, 40:1245–1252.
- Benjamini, Y. and 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*, 57:289–300.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The annals of Statistics*, 29(4):1165–1188.

- Bertsekas, D. (1976). On the Goldstein-Levitin-Polyak gradient projection method. *Automatic Control, IEEE transactions*, 21:174–184.
- Bonnans, J., Gilbert, J., Lemarechal, C., and Sagastizabal, C. (2006). *Numerical optimization: theoretical and practical aspects*. Springer, New York, second edition.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *foundations and trends in machine learning*, 3(1):1–122.
- Broder, A. (1997). On the resemblance and containment of documents. *In compression and complexity of sequences*, 1:21–29.
- Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540.
- Calamal, P. and More, J. (1987). Projected gradient methods for linearly constrained problems. *Mathematical Programming*, 39:93–116.
- Candes, E., Li, X., Ma, Y., and Wright, J. (2009). Robust principal component analysis? *Journal of ACM*, 58(1):1–37.
- Candes, E. J. and Randall, P. A. (2006). Highly robust error correction by convex programming. *IEEE Transactions on Information Theory*, 54:2829–2840.
- Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2008). High dimensional sparse factor modeling: applications in gene expression genomics. *Journal of American Statistical Association*, 103(484):1438–1455.
- Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion. *Biometrika*, 94:759–771.
- Colella, S., Yau, C., Taylor, J., Mirza, G., Butler, H., Clouston, P., Bassett, A., Seller, A., Holmes, C., and Ragoussis, J. (2007). Quantisnp: an objective Bayes hidden Markov model to detect and accurately map copy number variation SNP genotyping data. *Nucleic Acids Research*, 35.

- Cooper, G. and Zerr, T. (2008). Systematic assessment of copy number variant detection via genome-wide snp genotyping. *Nature Genetics*, 40:1199–1203.
- Diskin, J. S., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J., and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole genome SNP genotyping platforms. *Nucleic Acids Research*, 36.
- Dudoit, S., van der Laan, M., and Pollard, K. (2004). Multiple testing. part I. single-step procedures for control of general type I error rates. *Statistical applications in genetics and molecular biology*, 3(1).
- Dudoit, S. and van der Laan, M. J. (2008). *Multiple testing procedures with applications to genetics*. Springer-Verlag, New York.
- Efron, B. (2005). Local false discovery rate. Technical report, Stanford University, Department of Statistics.
- Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102:93–103.
- Efron, B. (2008). Microarrays, empirical Bayes and two-groups model. *Statistical Science*, 23(1):1–22.
- Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing and prediction*. Cambridge University Press, Cambridge.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.
- Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160.
- Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–322.

- Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104:1406–1415.
- Gabriel, K. R. and Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21:489–498.
- Gannaz, I. (2006). Robust estimation and wavelet thresholding in partial linear models. Technical report, University Joseph Fourier.
- Gordon, A., Glazko, G., Qiu, X., and Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, 1(1):179–190.
- Guha, S., Li, Y., and Neuberg, D. (2006). Bayesian hidden Markov modeling of array CGH data. *Harvard University Biostatistics Working Paper Series*.
- Harding, M. C. (2009). Structural estimation of high-dimensional factor model. Technical report, Stanford University, Economics.
- Hedenfalk, I. (2001). Gene expression profiles in hereditary breast cancer. *New Engl. Jour. Medicine*, 344:539–548.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–803.
- Hsu, D., Kakade, S., and Zhang, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Trans. Info. Th.*
- Hsu, L., Self, S., Grove, D., Randolph, T., Wang, K., Delrow, J., Loo, L., and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6:211–226.
- Iafrate, J., Feuk, L., Rivera, M., Listewnik, M., Donahoe, P., Qi, Y., Scherer, S., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, 36:949–951.

- James, B., James, K., and Siegmund, D. (1987). Tests for a change-point. *Biometrika*, 74:71–83.
- Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327.
- Lai, T., Xing, H., and Zhang, N. (2007). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics*, 9:290–307.
- Leek, J. T., Scharpf, R. B., Corrada-Bravo, H., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerley, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739.
- Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Science*, 105:18718–18723.
- Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.
- Lucas, E., Kung, H., and Chi, A. J. (2010). Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers. *PLOS Computational Biology*, 6.
- McCarroll, S. and Altshuler, D. (2007). Copy-number variation and association studies of human disease. *Nature Genetics*, 39:S37–S42.
- Newton, M., Gould, M., Reznikoff, C., and Haag, J. (1998). On the statistical analysis of allelic-loss data. *Statistics in Medicine*, 17:1425–1445.
- Olshen, A., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5:557–572.
- Onatski, A. (2009). Asymptotics of the principal components estimator of large factor models with weak factors. Technical report, Columbia University, Department of Economics.

- Owen, A. and Perry, P. (2009). Bi-cross-validation of the svd and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3(2):564–594.
- Patterson, N. J., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12):2074–2093.
- Peiffer, D., Le, J., Steemers, F., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C., Belmont, J., Cheung, S., Shen, R., Barker, D., and Gunderson, K. (2006). High resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research*, 16:11360–1148.
- Perry, P. O. (2009). *Cross-validation for unsupervised learning*. PhD thesis, Stanford University.
- Perry, P. O. and Owen, A. B. (2010). A rotation test to verify latent structure. *Journal of Machine Learning Research*, 11:603–624.
- Pollard, K. and van der Laan, M. (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of statistical planning and inference*, 125:85–100.
- Price, A. L., Patterson, N. J., Plengt, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38:904–909.
- Rom, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77:663 – 665.
- Romano, J. and Wolf, M. (2005). Control of generalized error rates in multiple testing. *Annals of Statistics*, 35(4):1378–1408.
- Rouveirol, C., Stransky, N., Hupe, P., La Rosa, P., Viara, E., Barillot, E., and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioninformatics*, 22:849–856.
- Schiffman, J., Wang, Y., Mcpherson, L., Welch, K., Zhang, N., Davis, R., Lacayo, N., Dahl, G., Faham, M., and Ford, J. (2009). Molecular inversion probes reveal

- patterns of 9p21 deletion and copy number aberrations in childhood leukemia. *Cancer Genetics and Cytogenetics*, 193:9–18.
- She, Y. and Owen, A. B. (2011). Outlier identification using nonconvex penalized regression. *Journal of the American Statistical Association*.
- Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754.
- Stock, J. and Watson, M. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44:293–335.
- Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *PNAS*, 100(16):9440–9445.
- Storey, J. D., Akey, J. M., and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, 3(8):1380–1390.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J.Royal.Statist.Soc.B*, 58:267–288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2004). Sparsity and smoothness via the fused lasso. *J.Royal.Statist.Soc.B*, 67:91–108.
- Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9:18–29.
- Tracy, C. and Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159(1):295–327.
- Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.
- van der Laan, M., Birkner, M., and Hubbard, A. (2006). Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Statistical applications in genetics and molecular biology*, 4(1).

- van der Laan, M., Dudoit, S., and Pollard, K. (2004a). Augmentation procedures for control of a generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical application in genetics and molecular biology*, 3(1).
- van der Laan, M., Dudoit, S., and Pollard, K. (2004b). Multiple testing, part III. step down procedures for control of the family-wise error rate. *Statistical applications in genetics and molecular biology*, 3(1).
- Venkatraman, E. and Olshen, A. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23:657–663.
- Westfall, P. and Young, S. (1993). *Resampling based multiple testing*. Wiley-Interscience, New York, first edition.
- Xing, B., Greenwood, C., and Bull, S. (2007). Joint estimation of DNA copy number from multiple platforms. Technical report, Stanford University, Department of Statistics.
- Xu, H. and Hsu, J. (2007). Using the partitioning principle to control the generalized family error rate. *Biometrical journal*, 49(1):52–67.
- Yifan, H., Xu, H., Calian, V., and Hsu, J. (2006). To permute or not to permute. *Bioinformatics*, 22:2244–2248.
- Zahn, J., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeratna, A. T., Taub, D. D., Gorospe, M., Mazan-Mamczarz, K., Lakatta, E. G., Boheler, K. R., Xu, X., Mattson, M. P., Falco, G., Ko, M. S. H., Schlessinger, D., Firman, J., Kummerfeld, S. K., III, W. H. W., Zonderman, A. B., Kim, S. K., and Becker, K. G. (2007). AGEMAP: A gene expression database for aging in mice. *PLoS Genetics*, 3(11):2326–2337.
- Zhang, N. (2010). *DNA copy number profiling in normal and tumor genomes: Frontiers in Computational and Systems Biology*. Springer - Verlag:London.

- Zhang, N. and Siegmund, D. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63:22–32.
- Zhang, N., Siegmund, D., Ji, H., and Li, J. (2010). Detecting simultaneous change-points in multiple sequences. *Biometrika*, 97:631–645.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.