Adjustment Procedure to Permutation Tests in Epigenomic Differential Analysis

by

Dan Jiang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August, 2017

Abstract

In the analysis of genomic data, t-statistics are widely used to detect differential signals between different groups of samples. In many studies, each group has only a small number of replicate samples, making the variance estimation unstable. Small sample variances due to chance can create large t-statistics for genes or genomic loci that are not differential. In order to mitigate this problem, shrinkage estimators are now widely used for variance estimation. One example is moderated t-statistics. For statistical inference, null distributions need to be constructed for test-statistics. Permutation is a natural option to construct null distributions when they cannot be derived using a parametric model due to violations of parametric assumptions. When variance shrinkage estimators are involved, naive permutation can be misleading. This is because for a differential gene or locus, permuting measurements between two groups will inflate the variance estimate which in turn will influence the variance shrinkage estimator. This thesis investigates this issue and proposes a solution to this problem by permuting residuals. This approach is applied and evaluated in genomic applications that involve comparisons of one or multiple data types between two

ABSTRACT

biological conditions.

Advisor: Dr. Hongkai Ji

Acknowledgments

First, I would like to express my sincere gratitude to my advisor Dr. Hongkai Ji for the continuous support of my Ph.D study and related research, for his patience, motivation and immense knowledge. He has always been there to talk to, provide suggestions, and give encouragement. His guidance helped me in all the time of research and writing of this thesis. I am deeply grateful to him for supporting me through the Ph.D program.

I would like to thank Dr. Jiang Qian, Dr. Zhibin Wang, and Dr. Jeffrey Leek for serving as members on my oral exam committee and thesis defense committee and for giving me valuable comments on my proposal and thesis. I would like to thank Dr. Michael Beer for serving on my oral exam committee. I also would like to thank Dr. Kasper Hansen and Dr. Jiou Wang for serving as my alternative thesis defense and oral exam committee member.

I also want to express my gratitude to Weiqiang Zhou, Zhicheng Ji, and Bing He for their encouragement and valuable advice during my Ph.D study. Many other friends, especially my roommates Xindong Song, Yunching Chen, and Yunke Song

have helped me through setbacks in last five years.

Finally I want to thank my parents, who have always been standing by my side to support me spiritually throughout writing this thesis and my life in general.

Dedication

This thesis is dedicated to my family.

Contents

List of Tables

List of Figures

Chapter 1

Overview

This dissertation presents a method to adjust permutation tests in differential analysis of genomic data. It is organized as follows: Chapter 1 will provide the overview and related literature review. Chapter 2 presents the basic idea and the method through simulation. Chapter 3 discusses the application of this method in the context of comparing one data type between two biological conditions. Chapter 4 discusses the application of this method to comparing multiple data types jointly. Chapter 5 is the conclusion chapter that summarizes our analysis.

1.1 Introduction

With the rapid development of next-generation sequencing (NGS) technology, multiple epigenomic assays, including chromatin immunoprecipitation followed by

sequencing (ChIP-seq),[1] sequencing of DNase I hypersensitive sites (DNase-seq),[2] and Formaldehyde-Assisted Isolation of Regulatory Elements coupled with sequencing (FAIRE-seq),[3] are available to biologist. Differential analysis of epigenomic signals is an important analysis. It can provide insights on how gene activities vary across different cell types and biological condition. Because different epigenomic signals are correlated with each other, comparing multiple types of epigenomic assays jointly has also become a crucial question in differential analysis.[4567] Different analysis often involves two component: (1) rank the prioritize genomic loci for downstream analysis and studies; (2) determine which differences are statistically significant rather than being random noises.

1.2 Ranking and Variance Shrinkage

One challenge in characterizing and ranking differential genomic loci is that a typical dataset often involve a large number of loci but only a small number of replicate samples in each condition. This can create complications in constructing test-statistics for ranking. For example, t-statistics are commonly used to prioritize differential genes or genomic loci. For simplicity, we call both genes and genomic loci as genomic loci in following chapters. To construct a t-statistic, the estimation of variance is required. However, with the small number of replicate samples available and the large number of genomic loci involved, the estimation of variance could be un-

stable. Small sample variances due to chance can result in extremely large t-statistics and hence producing many false positives.

One solution to this problem is to borrow information from all genomic loci to stabilize the variance estimates for individual genomic loci. One commonly used approach is based on Bayesian or empirical Bayes methods. In an empirical Bayes approach, the sample variances are shrunk toward a shared variance estimate, which results in more stable variance estimates. In 2002, Lnnstedt and Speed used an empirical Bayes method to analyze replicated two-color microarray data.[8] They combine the data of all genomic loci to estimate the parameters of a prior distribution. Then for each specific genomic lous, they combine these prior parameter estimates with local means and standard deviations to form a posterior statistic, which is used to decide if a gene is differential or not.

Later, Smyth proposed the Limma model,[9] which can deal with microarray experiments with arbitrary numbers of treatments and RNA samples. They also proposed a moderated t-statistic, which improve the conventional t-statistic by using variance shrinkage estimators. They show that after incorporating variance shrinking, the moderated t-statistic improves the gene ranking compared to the conventional t-statistics. Variance shrinkage estimator has also been used in other data types and applications.[10]

1.3 Inference and Permutation

For inference, one needs to construct null distribution of the test statistics. Smyth[9] derived the null distribution for the moderated t-statistics under a hierarchical model with normality assumptions. When the model assumptions do not hold true, such a null distribution can be misleading. In that situation, a non-parametric procedure such as permutation can be used to construct the null distribution.

Permutation test is a commonly used non-parametric test, and has been widely used in differential analysis[1112131415 16].[17] For example, Significance Analysis of Microarrays (SAM)[11] is one of the earliest application of permutation tests in differential analysis. SAM assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. The p-value for each gene is calculated from permutations of repeated experiments. They assumed that for a given level of expression, the random fluctuations were locus-specific. Therefore, they designed a "relative difference" d_g in gene expression, which is the ratio of change in gene expression to the standard deviation for that gene. Then they compare d_g across all genes. For genes with d_g that exceeds a certain threshold, SAM uses a permutations of the repeated measurements to estimate the percentage of genes identified by chance, which is false discovery rate (FDR). SAM was used to analyze the transcriptional response of lymphoblastoid cells to ionizing radiation. Permutation test has also been applied in many other settings such as Gene Set Enrichment Analysis (GSEA),[12] QTL detection,[15] allelic association analysis,[16] and modeling ChIP

sequencing.[17]

However, when a variance shrinkage estimator is used, permutation-based differential analysis can create problems. The number of false positives can increase due to the inflated estimation of variance for non-differential genomic loci. After each permutation, the sample variances for differential genomic loci increase enormously. Their average, which incorporates these inflated sample variances would also be inflated. To construct posterior variance estimate for individual genomic locus, sample variances are shrunk toward the average variance. Therefore, the posterior variance estimates would also be inflated. For differential genomic loci, this inflation would not cause a big problem. However, for non-differential genomic loci, this inflation causes a smaller null t-statistic. Therefore, non-differential loci may be incorrectly identified as differential due to a more centered and light-tailed permutation null distribution.

1.4 Analysis of Multiple Data Types

Different epigenomic signals are correlated with each other. For example, different transcription factors may work together to regulate genes. Therefore, comparing multiple data types jointly has also become an important question in differential analysis. Previously, multiple methods have been developed for jointly analyzing multiple data types.

HHMM[4] is one method to analyze epigenomic data from ChIP-seq and ChIP-chip

experiments jointly by a hierarchical hidden Markov model. Here, a hidden Markov model was employed to infer the hidden states of genomic regions. The hidden states correspond to whether the signal intensity in a genomic region is different between two conditions or not. To analyze ChIP-seq and ChIP-chip data jointly, this hierarchical model consists of two levels of hidden Markov models. First, two individual-level hidden Markov models are constructed on ChIP-seq and ChIP-chip data separately. Then a master-level HMM is constructed to infer the true hidden state variable. The result of HHMM is compared with the result using ChIP-seq or ChIP-chip data alone. Analyzing two types of data jointly outperforms employing only one type of data.

Model-based Meta-analysis of ChIP data (MM-ChIP)[5] is another approach to combine multiple ChIP-seq and ChIP-chip datasets. MM-ChIP mainly focuses on dealing with variation across ChIP data samples due to different platform designs and laboratories. MM-ChIP proposed a two step process. The advantage of MM-ChIP is its employment of Stouffer's method. This method treats sources differently according to their quality. This is an improvement comparing to HHMM approach, which treats data from ChIP-seq and ChIP-chip datasets equally.

ChromaSig[6] is an unsupervised method to explore co-working mechanism of histone modifications by discovering histone modification patterns. Before ChromaSig was proposed, supervised classification method were also proposed to identify histone modification marks at known functional sites. However, since in some cases prior knowledge of relationship between functional sites in genome are not available, an

unsupervised learning method is necessary.

Spatial clustering is a qualitative approach of combining multiple genomic and epigenomic datasets. This is an unsupervised method based on learning an HMM model and inferring the most likely genomic layout. K-spatial clustering partitions the underlying genomic regions into disjoint and contiguous intervals, and each interval is tagged with cluster k. Then the goal of K-spatial clustering is to seek a way of tagging the intervals, so that the maximal score is achieved. K-spatial clustering has advantages over conventional clustering method. First, the result of K-spatial clustering can provide insights on relationship of functional genomic elements. For example, transcription start sites (TSS) are usually followed by the transcribed regions. This relationship would correspond to the coupling of two clusters. Second, spatial clustering takes in account the correlation of adjacent loci. The adjacent loci are likely to be assigned to the same cluster which corresponds to our common knowledge. However, one potential problem for this algorithm is that it requires considerable computing resources. As we can expect, if the number of tracks increases, this requirement would also increase exponentially. This is common combinatorial problem if we compare multiple data types qualitatively.

The above methods do not focus on analyzing differential signals. The first systematic approach for analyzing differential signals in multiple epigenomic data types jointly is differential principal component analysis (dPCA).[7] dPCA is proposed to deal with this combinatorial problem caused by qualitatively comparing multiple ChIP-seq

datasets. dPCA is an unsupervised learning method to discover covariation patterns of different chromatin marks. Besides ChIP-seq, dPCA can also handle DNase-seq, FAIRE-seq and other types of epigenomic data.

dPCA compares multiple epigenomic datasets quantitatively. There are other important features of dPCA. First, instead of applying PCA on observed data directly, it estimates the covariance matrix of true differences of multiple datasets under two conditions, and applies PCA on the estimated true differences. This helps to improve the efficiency of dimension deduction. Comparing to PCA, dPCA requires a smaller number of principal components (PCs) to explain the same amount of variance in the datasets.

Second, dPCA also integrates genomic loci ranking and statistical tests. Performing dPCA on estimated true difference matrix outputs principal components that show covariation patterns of these datasets. Each differential principal component (dPC) can be explained as a functional module of chromatin marks. If one chromatin mark corresponds to a large number in dPC, then it shows that this mark plays an important role in driving differences between two biological conditions. The major patterns discovered by dPCA are the differential patterns shared by many loci. Meanwhile, for each genomic locus, it also outputs principal component scores to indicate if a locus is differential with respect to that principal component. The principal component score for each candidate genomic locus can be used to rank these genomic loci, this is a prominent improvement comparing to qualitative differential analysis

method. This score is calculated by performing a t-test on each genomic locus. Inference based on t-test relies on the normality assumption of the noise. In reality, the noise may not be normal. Therefore, permutation test may be more appropriate. However, due to variance shrinkage, permutation again can create problems.

1.5 Our Approach

As described above, when we incorporate shrinkage estimator in a permuatation-based statistical test, the inflated sample variances in permutation may cause a overly centered null distribution. To solve this problem, this thesis investigates a revised permutation procedure. Suppose we want to compare samples from two different biological conditions, we calculate the mean expression level under each condition, and remove the mean difference between conditions. After this adjustment, the sample variances for differential genomic loci will be on the same scale with the sample variances for non-differential loci. This will avoid the overly centered permutation null distribution, and thus providing a more conservative decision in terms of differential signal detection.

Chapter 2

Basic Method and Illustration through Simulation

2.1 Method

We assume two biological conditions A and B. For each condition, we have $k = 1, ..., K$ samples. We use Y_{Akg} and Y_{Bkg} to denote signal intensity level of genomic locus g in the kth replicate under condition A and B respectively. We assume that the intensities are appropriately normalized and transformed (e.g. log2 transformation). The first step of our analysis involves calculating moderated t-statistics based on the data.

$$T_g = \frac{\bar{Y}_{Ag} - \bar{Y}_{Bg}}{\tilde{S}_g \sqrt{\frac{1}{K_A} + \frac{1}{K_B}}} \tag{2.1}$$

where

$$\tilde{S}_g^2 = \frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g} \tag{2.2}$$

$$S_g = \sqrt{\frac{(K_A - 1)S_A^2 + (K_B - 1)S_B^2}{K_A + K_B - 2}} \tag{2.3}$$

Here, $\bar{Y}_{Ag}$ and $\bar{Y}_{Bg}$ are the sample means of the expression for genomic locus g, under condition A and condition B. S_A^2 and S_B^2 are the sample variances. S_g^2 is the pooled sample variance. $\tilde{S}_g^2$ is the posterior variance estimators, which is a weighted average of a prior estimator S_0^2 and the pooled sample variance S_g^2. S_0^2 and d_0 are prior estimation of variance and prior degree of freedom, which is estimated through a Newton iteration as described in Limma.[14] $d_g = K_A + K_B - 2$ is the degree of freedom of S_g^2.

To construct the null distribution for T_g, we use permutations. Before permuting the data, we first force the data under two biological conditions to have the same mean by subtracting the mean from samples in each condition. The adjustment is as following:

$$\begin{aligned} \tilde{Y}_{Akg} &= Y_{Akg} - \bar{Y}_{Ag} \\ \tilde{Y}_{Bkg} &= Y_{Bkg} - \bar{Y}_{Bg} \end{aligned} \tag{2.4}$$

Where Y_{Akg} and Y_{Bkg} are the signal intensity level of the kth replicate, gth genomic locus, under biological condition A and B.

Without this adjustment procedure, differential genomic loci will have large sample variances after permutation. For example, consider a locus with three measure-

ments in each condition: $A = (0.1, 0.2, 0.3)$ and $B = (10.1, 10.2, 10.3)$. The sample variances before permutation are $S_A^2 = 0.01$ and $S_B^2 = 0.01$. After permutation, suppose the data becomes $A = (0.1, 10.2, 0.3)$ and $B = (10.1, 0.2, 10.3)$. The sample variances after permutation become $S_A^2 = 33.34$ and $S_B^2 = 33.34$. When constructing the shrinkage estimator, S_0^2 is estimated by looking at the average behaving of S_g^2 from all genes, therefore it would be inflated by these large sample variances of differential loci. As a result, for non-differential genomic loci, the posterior estimation of variances would be inflated because the calculation of posterior estimation involves pulling sample variance S_g^2 toward S_0^2. By subtracting the mean from samples in each group, differences are removed. As a result, permutation will not inflate the sample variances of differential loci.

After this adjustment, we permute the adjusted data by randomly rearranging the condition labels of samples. For each permutation, we follow the same procedure to calculate a null moderated t-statistic t_0. Then by performing the permutations for a large number of times, these t_0s form a permutation null distribution for moderated t-statistic t_g. Next, we use this permutation null distribution to calculate p-values and false discovery rates (FDR)[18] for each genomic locus.

2.2 Illustration through Simulations

In simulation data, we assume there are two biological conditions A and B. Under each condition, there are three replicates, $K_A = K_B = 3$. We assume there are $N = 15000$ genomic loci. Among all genomic loci, p of them are differential. We conduct simulation under different settings of p. Below we use $p = 0.3$ to illustrate the results, and results for other values of p are summarized at the end of this chapter. The data is simulated in the following way:

$$\begin{aligned} &\text{Condition A: } Y_{Akg} = \beta_{Ag} + \epsilon_{Akg} \\ &\text{Condition B: } Y_{Bkg} = \beta_{Bg} + \epsilon_{Bkg} \end{aligned} \tag{2.5}$$

We assume $\beta_{Bg} = 0$ for all genomic loci. For non-differential genomic loci, we set $\beta_{Ag} = 0$. For differential genomic loci, we assume $\beta_{Ag} \sim N(0, \sigma_\beta^2)$. For the noise ϵ we assume a hierarchical model, where $\epsilon_{kg} \sim N(0, \sigma_g^2)$, $\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 S_0^2}\chi_{d_0}^2$. We assume $d_0 = 4$, $S_0^2 = 2$, $\sigma_\beta^2 = 10$. This assumption follows the procedure from G.K.Smyths paper.[9] The empirical Bayes approach in Limma allows us to borrow information across all genomic loci to estimate the variance. Under the hierarchical model above, the posterior estimator of variance is calculated by formula (2.2) and the moderated t-statistic is calculated by formula (2.1). These statistics are calculated based on the simulated data.

Next, we permute the data by randomly assigning the label of biological condition

A or B to each sample. There are 20 possible distinct permutations. For each permutation, we follow the same procedure to calculate a null moderated t-statistic, t_0. After a number of permutations, these t-statistics form the naive permutation null distribution. Since the number of distinct permutations is small, we pooled t-statistics from all genomic loci to form the permutation null distributions before adjustment. According to our previous analysis, we expect that this null distribution should more centered around zero than it should be.

Then, we apply the adjustment procedure: centering the original data under two conditions by subtracting the mean as we described in formula (2.4). After this adjustment, we conduct the permutation again. This will generate another permutation null distribution, which is the null distribution based on our approach. We also generate a null distribution based on the parametric model, which is a moderated t-distribution with degree of freedom $df = d_0 + d_g$[14] under the hierarchical model above and the normality assumption of the noise term.

We compare the two permutation null distributions (before adjustment and after adjustment) and the moderated t-distribution. The comparison of three null distributions is presented in Figure 2.1. Permutation null distribution before adjustment (green curve) is the most centered one, while the permutation null after adjustment (red curve) is the least centered among three null distributions. This suggests that the inference based on the permutation after adjustment is the most conservative one, and the permutation before adjustment is the most optimistic one.

We calculated p-values of genomic loci based on the moderated t-statistics t_gs we have for each genomic locus with respect to two permutation null distributions and the parametric null distribution. The empirical CDF of p-values is presented in Figure 2.2. The red curve after adjustment has less p-values concentrated at 0, while the p-values based on the permutation null before adjustment are most concentrated near 0.

Next, we calculated true false discovery rates (FDR)[18] and estimated FDRs based on three null distributions. The true false discovery rates are calculated by counting the number of true non-differential genomic loci (with $\beta_{Ag} = 0$ in our simulated data) that have p-values below a certain threshold (false discoveries) divided by the total number of the genomic loci that have p-values below that threshold (all discoveries). The estimated false discovery rates are calculated by *qvalue* function in R package developed by Storey JD.[19] Given a list of p-values, it outputs corresponding estimated FDRs. Then we plot the estimated FDRs against the true FDRs, for three approaches respectively. The results are presented in Figure 2.3

One can see from this plot that the green curve of naive permutation is below the diagonal, indicating that the estimated FDR is smaller than the true FDR. This shows that before adjustment, our decisions for differential genomic loci are too optimistic. For revised permutation, the red curve is above the reference black diagonal. The estimated FDRs are larger than the true FDRs. Our adjustment would correct this optimism and make a more conservative decision for those non-differential genomic

loci. In this example, the estimated FDR based on the parametric null distribution matched the true FDRs relatively well. This is consistent with our expectation, since the normality assumption for error term holds true in this simulation.

Figure 2.4 shows results for $p = 0.05, 0.1, 0.3, 0.5$. Similar conclusions are reached. We also noticed that when p is small, parametric and naive permutations both work well. Comparing to naive permutation, our approach becomes more crucial when larger p is involved.

Next, we modified our simulation so that ϵ_g no longer follow a normal distribution. We assume ϵ_g follows a t-distribution with degree of freedom 2. As a result, the null distribution $t_{d_0+d_g}$ based on the parametric model assumptions in the *Limma* is no longer an appropriate null distribution. Figure 2.5 compares the null distribution derived using moderated t-distribution $t_{d_0+d_g}$, and permutation null distributions before and after adjustment. Now null distributions of naive permutation and parametric model are more close to each other, while the revised permutation approach has the least centered null. Figure 2.6 is the empirical CDF of p-values. The revised permutation has less p-values concentrated near 0. One can see from Figure 2.7 that $t_{d_0+d_g}$ no longer works: it was too optimistic. Permutation before adjustment also was too optimistic. However, permutation after our adjustment corrected the overly centered null distribution. We also investigate our method under different parameter settings, the results are presented in Figure 2.8. Similar conclusions are reached.

We conclude that if the error term does not follow a normal distribution, the null

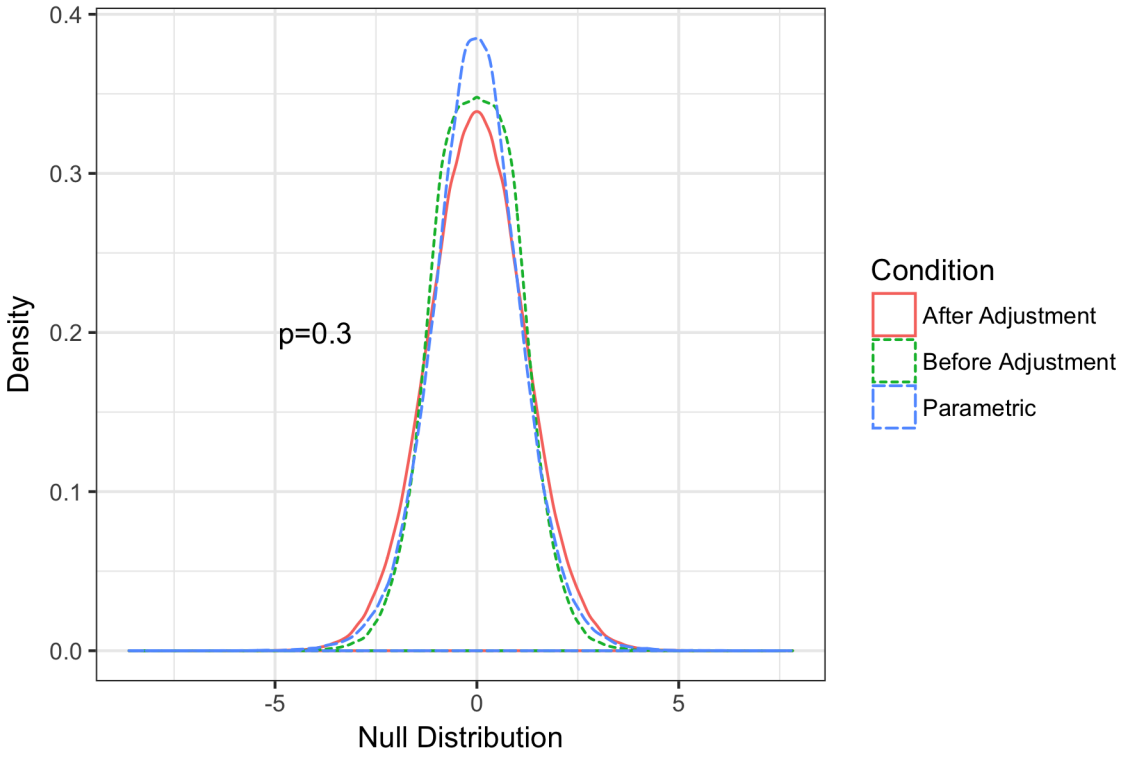

Figure 2.1: Comparison of Permutation Null Distributions before and after adjustment, and the Parametric Null Distribution, Simulation Data, $\epsilon_g \sim N(0, \sigma_g^2)$

distributions based on the parametric model and the permutation before adjustment are overly centered, hence would result in optimistic decisions for non-differential genomic loci. In this situation, our adjustment approach helps to correct the overly centered null distribution, and hence providing a more conservative decision for differential genomic loci.

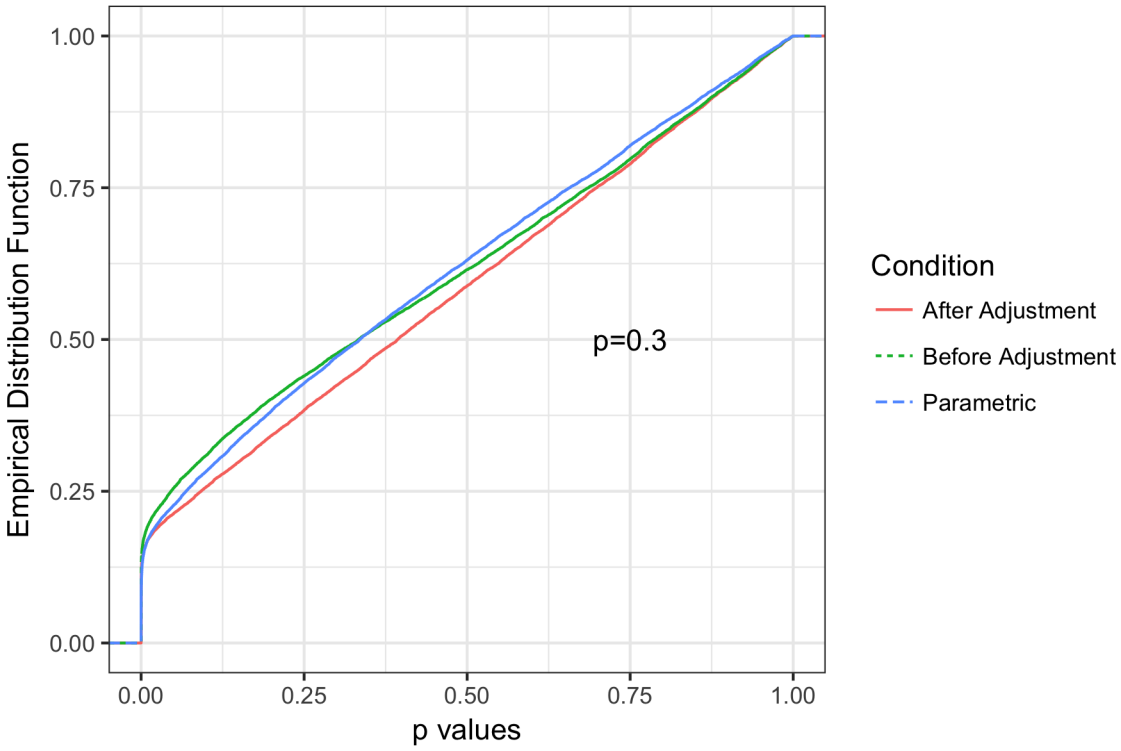

Figure 2.2: Empirical CDF of p-values, Simulation Data, $\epsilon_g \sim N(0, \sigma_g^2)$

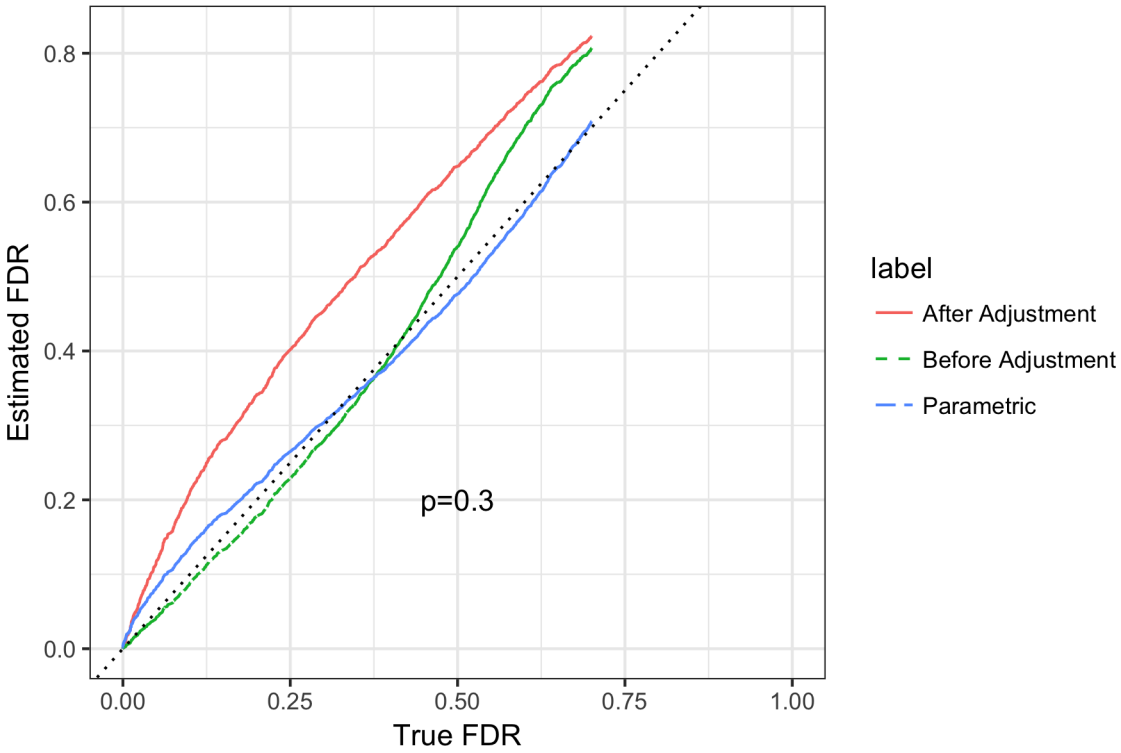

Figure 2.3: Estimated FDR v.s. True FDR, Simulation Data, $\epsilon_g \sim N(0, \sigma_g^2)$

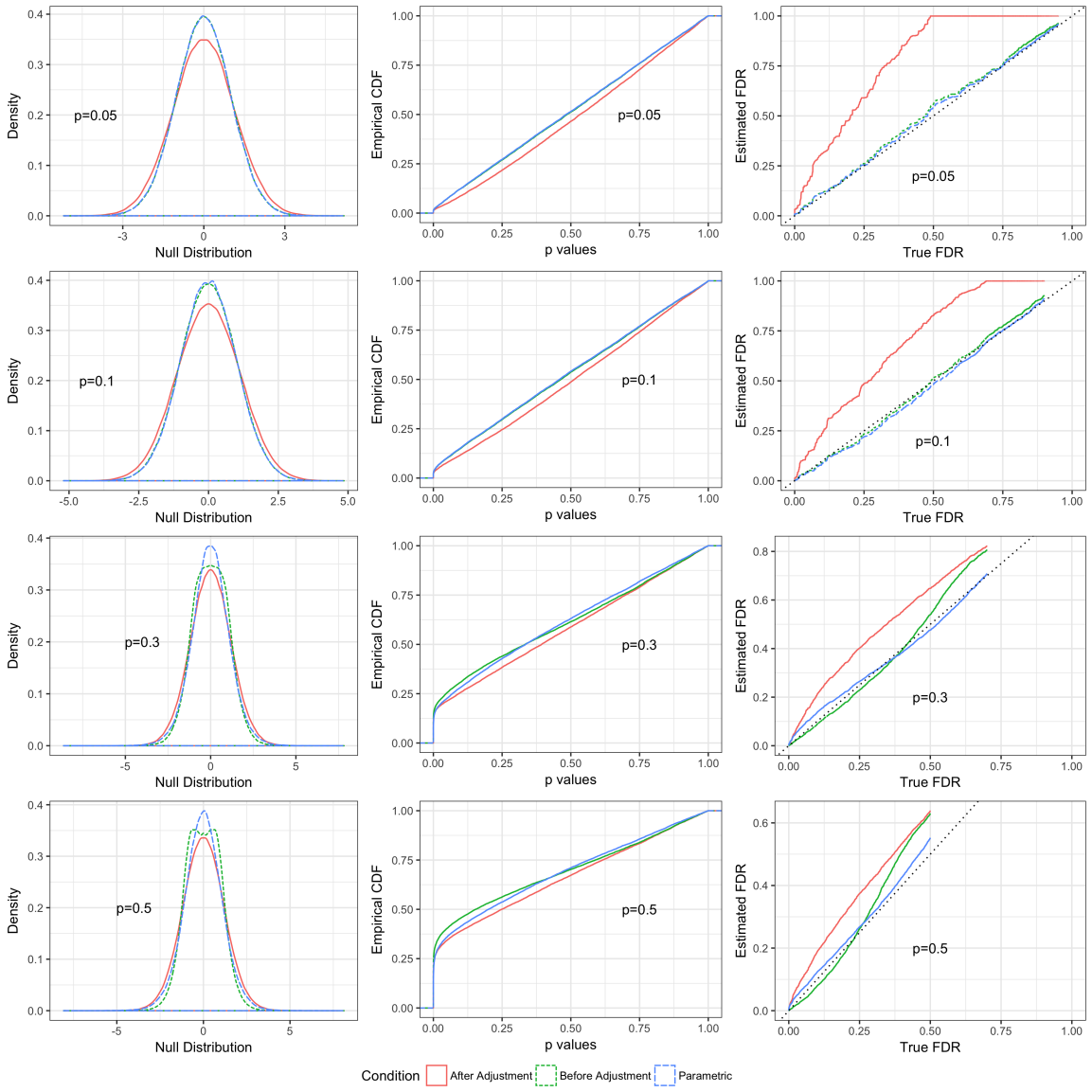

Figure 2.4: Results under Different Parameter Settings, Simulation Data, $\epsilon_g \sim N(0, \sigma_g^2)$

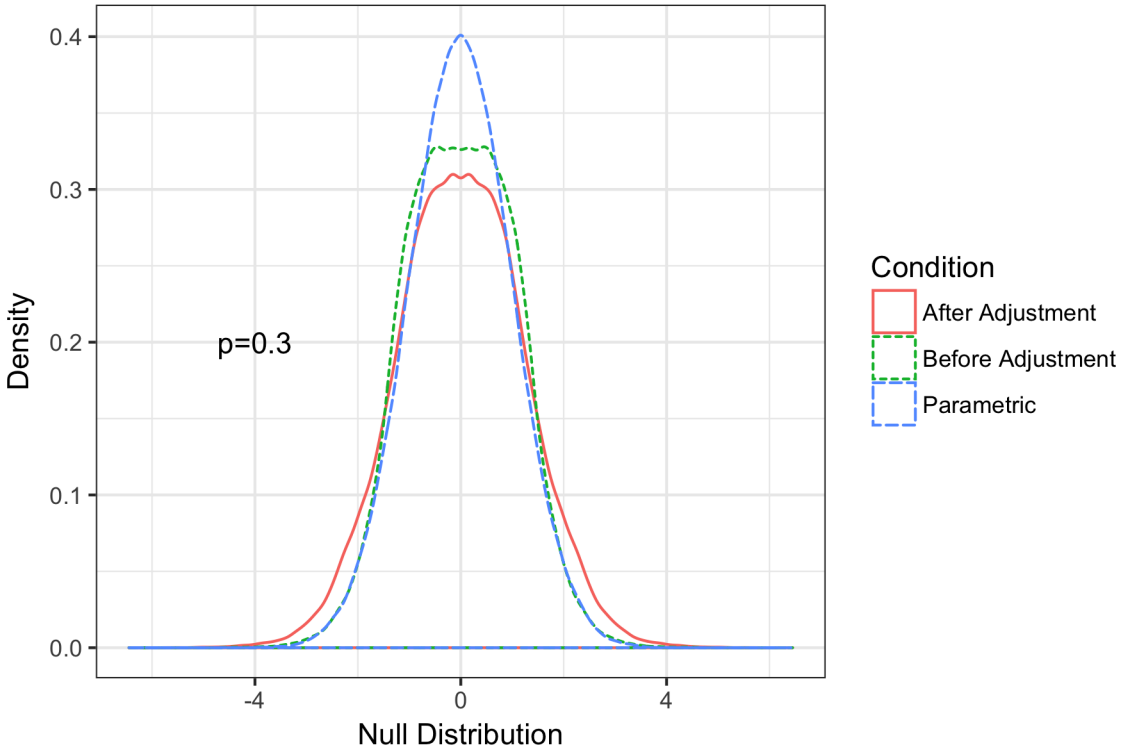

Figure 2.5: Comparison of Permutation Null Distributions before and after adjustment, and the Parametric Null Distribution, Simulation Data, $\epsilon_g \sim t_2$

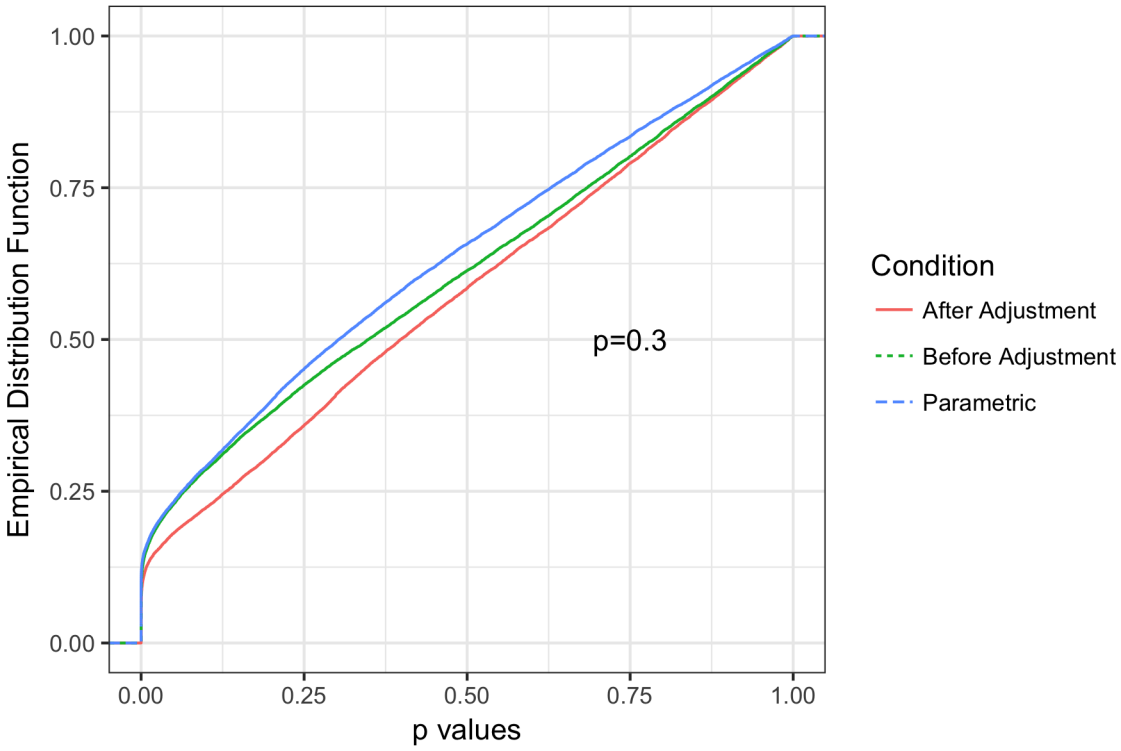

Figure 2.6: Empirical CDF of p-values, Simulation Data, $\epsilon_g \sim t_2$

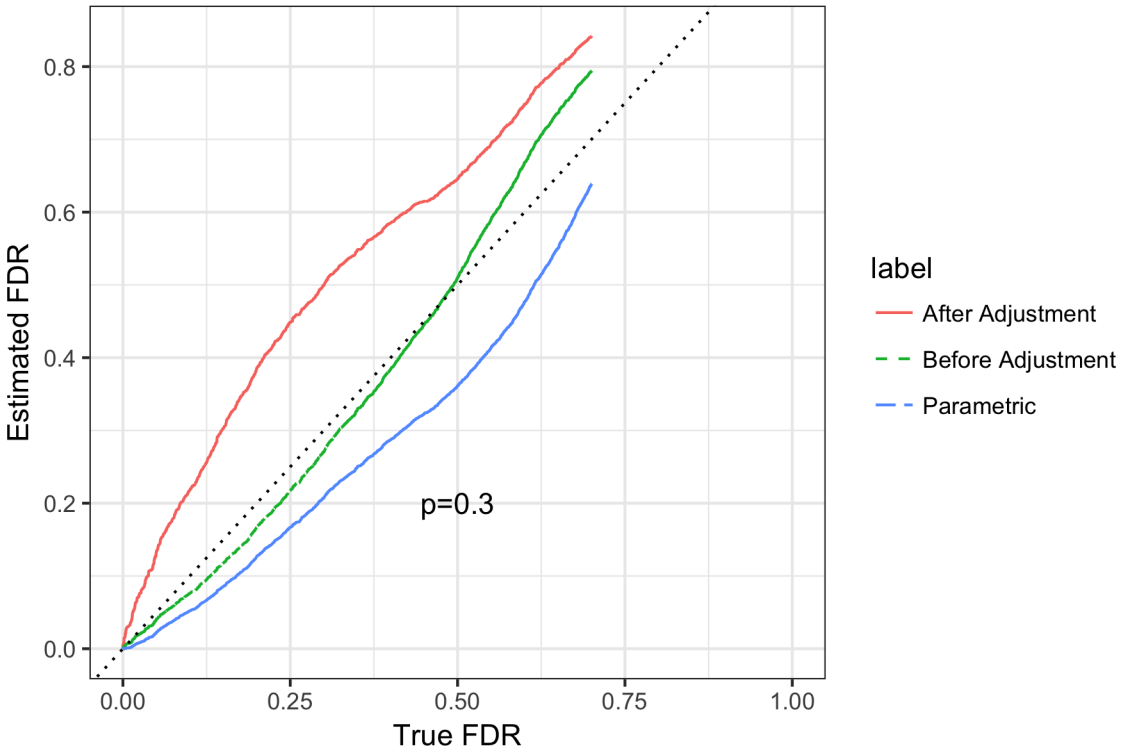

Figure 2.7: Estimated FDR v.s. True FDR, Simulation Data, $\epsilon_g \sim t_2$

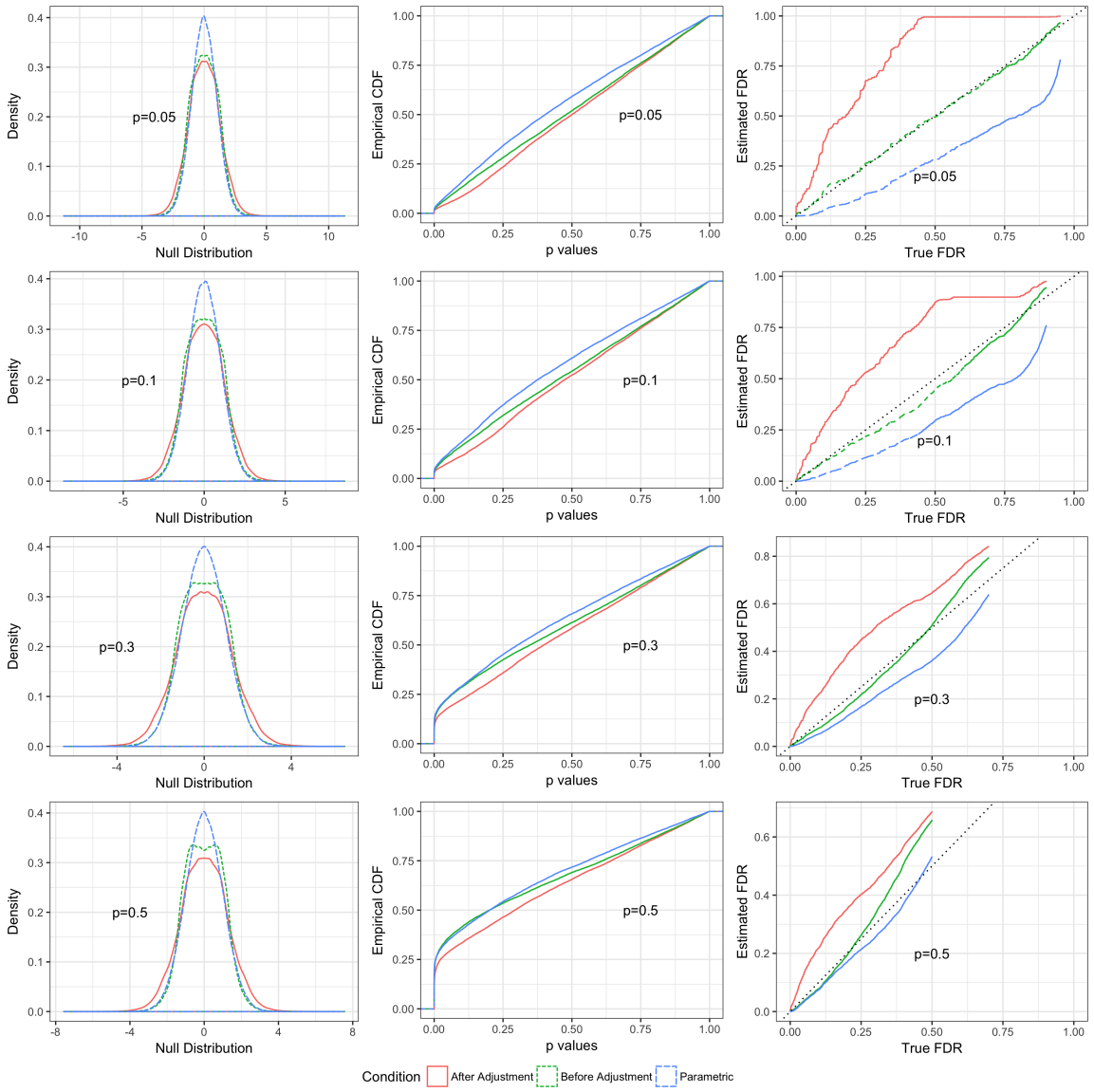

Figure 2.8: Results under Different Parameter Settings, Simulation Data, $\epsilon_g \sim t_2$

Chapter 3

Application including One Data Type

3.1 Brief Introduction

In last chapter, we discussed our method in simulation data. In this and next chapters, we investigate the behavior of our method in real data. The goal of analysis in this chapter is to detect differential signals by comparing one data type between two biological conditions. We use gene expression microarray data as test examples. We first conduct a realistic simulation on microarray data. Next, we apply our method directly to two real microarray datasets.

3.2 Realistic Simulation

The data used here come from a chronic lymphocytic leukemia (CLL) gene expression microarray dataset.[20] It has 24 samples in total with 12625 genomic loci from 24 CLL patients. The samples are categorized into two biological conditions according to the states of the disease of the patients: stable condition and progressive condition. This dataset has been preprocessed and normalized, and was provided as part of CLL package in R.

For simulation, we selected 6 samples randomly from stable states and divided them into two biological conditions. These two groups of samples should not have meaningful biological difference. We then added artificial differential signals in the following way. We randomly picked up p genomic loci as differential genomic loci, and we added artificial differences that follow a normal distribution $N(0, \sigma^2)$ to these genomic loci in one condition. Different choices of $p = (0.1, 0.3, 0.5)$ and $\sigma^2 = (1, 3, 5)$ are tested. The results from different combinations of parameters are similar. Here we present the results under $p = 0.3$ and $\sigma^2 = 3$ as an example, and results for other parameter settings are provided at the end of this chapter.

After the simulation, we have 6 samples under two biological conditions. We know which genomic loci are differential. Next, we used the Limma package[21] to analyze the simulated data and to calculate the moderated t-statistics. These moderated t-statistics are calculated by incorporating the shrinkage estimators of variance.

Next, we permuted the data by randomly assigning three samples as biological

condition A and the other three samples as biological condition B. There are 20 possible distinct permutations under our design, which is the same as the simulation study in Chapter 2. For each permutation, we applied the Limma package[21] again to get the moderated null t-statistic t_0. Since the number of distinct permutations is small, we pooled moderated null t-statistics from all genomic loci to form the permutation null distributions before adjustment.

Next, we applied our adjustment procedure. We force the data from two biological conditions to have the same mean for each genomic locus through mean centering. Then we permuted the data again and calculated moderated null t-statistics for each permutation. These t-statistics form the permutation null distribution after adjustment. Now, for each genomic locus, we have two permutation null distributions: before and after adjustment. We also generate the parametric null distribution for moderated t-statistics which is a t-distribution with augmented degrees of freedom $df = d_0 + d_g$ based on the hierarchical model of Limma.

Figure 3.1 compares the permutation null distributions before and after adjustment and the null distribution based on the parametric model. After adjustment, the permutation null distribution is less concentrated around zero than before adjustment. This is consistent with our expectation.

From Limma, we can get estimated S_0^2 for each permutation. We also compared the S_0^2 estimates before and after adjustment. According to our analysis above, we expect that before adjustment, because of the inflated sample variances of differential

sites, S_0^2 would be inflated. After adjustment, S_0^2 should be smaller. As we can see from Figure 3.2 the result is consistent with our expectation. The solid red curve is the distribution of S_0^2 after adjustment, which is smaller than the distribution of S_0^2 before adjustment.

Next, we calculated p-values for each genomic locus based on the moderated t-statistics with respect to the permutation null distributions before and after adjustment and the null distribution of parametric model. We compared these p-values. The empirical CDF of the p-values is shown in Figure 3.3. We can see from this figure that before adjustment, there are more genomic loci with smaller p-values. Therefore it is more likely to misclassify non-differential genomic loci as differential genomic loci before adjustment.

In simulation study, we know if one gene is differential or not. Therefore, we can calculate true false discovery rates (FDR). The estimated false discovery rates are calculated by *qvalue* function, as the same with last chapter. Then we plot the estimated FDRs against the true FDRs. The results are presented in Figure 3.4. We can see from this plot that estimated FDRs are smaller than the true FDRs for the permutation before adjustment and the parametric model. Therefore, differential sites based on these two null distributions may result in a large proportion of false discoveries. After adjustment, estimated FDRs are large than true FDRs, which will report less differential loci and make a more conservative decision for detecting differential signals.

We also conduct realistic simulation under different parameter settings, the results are presented in Figure 3.5 with $\sigma^2 = 1$ and $p = (0.1, 0.3, 0.5)$, Figure 3.6 with $\sigma^2 = 3$ and $p = (0.1, 0.3, 0.5)$, and Figure 3.7 with $\sigma^2 = 5$ and $p = (0.1, 0.3, 0.5)$. Our results are similar under different parameter settings. We also noticed that our adjustment approach is more necessary when a larger proportion of differential loci p or a larger difference magnitudes between two biological conditions is involved.

3.3 Microarray Data Result

In this section, we applied our method on two microarray datasets. The first microarray dataset is the same CLL dataset[20] as in last section. Here, we picked three samples from progressive state and three samples from stable state. Now we have three replicates in each biological condition.

Next, we follow the same procedure as above to calculate the moderated t-statistics for each genomic locus by applying limma.[21] Then naive permutation and revised permutation are conducted.

Figure 3.8 compares the null distribution based on parametric model and permutation null distributions before and after adjustment. After adjustment, the permutation null distribution is less concentrated near 0.

Similar to last section, we also plot the empirical CDF of S_0^2 in Figure 3.9, and the empirical CDF of p-values are in Figure 3.10. We can see from the plot that,

after our adjustment procedure, the S_0^2 are smaller, and the p-values are larger.

We then compared the number of differential sites at different FDR thresholds in Figure 3.11. We can see from the plot that less genomic loci are classified as differential after our adjustment. This is also consistent with what we expected. At $FDR = 0.05$, naive permutation reports 4523 differential genomic loci among all 12625 genomic loci. The parametric model reports 676 differential genomic loci. After adjustment, permutation reports 196 differential genomic loci. At $FDR = 0.1$, naive permutation reports 7922 differential genomic loci among all 12625 genomic loci. The parametric model reports 7088 differential genomic loci. After adjustment, permutation reports 5356 differential genomic loci. We can see that there is a large decrease in the number of differential sites between the permutation after adjustment and the other two null distributions. Therefore, our adjustment approach can have a substantial influence on the results of differential gene detection.

We also applied our method on another microarray dataset: a BRAFV600E A375 melanoma dataset.[22] This dataset was downloaded from GEO website, the data was normalized using GEOquery package[23] and affydata package[24] in R. In this dataset, the melanoma cells are treated with either vehicle or vemurafenib. Vemurafenib is a BRAF inhibitor, which suppresses the proliferation of BRAF mutant human melanoma cells. We have 6 samples in total, two treatments: vehicle and vemurafenib are considered to be two biological conditions. Three of the samples are treated with vehicle, and the other three are treated with vemurafenib. There are

32321 genomic loci in total. We applied the same procedure as we did in CLL dataset. The results are shown in Figure 3.12, Figure 3.13, Figure 3.14, and Figure 3.15. These figures showed the similar results with CLL dataset. After our adjustment procedure, the permutation null distribution is less centered, S_0^2s from permutations are smaller, thus p-values after adjustment are more uniformly distributed and less centered at 0. At the same FDR threshold, permutation after adjustment reports less differential loci comparing to the other two approaches. At $FDR = 0.05$, naive permutation reports 15854 differential genomic loci, parametric model reports 13812 differential loci, and the permutation after adjustment reports 11283 differential genomic loci. At $FDR = 0.1$, naive permutation reports 19313 differential genomic loci, parametric model reports 17556 differential loci, and the permutation after adjustment reports 14461 differential genomic loci. Therefore, under the same FDR threshold, less genomic loci are classified as differential after adjustment.

In summary, we can see that there is a decrease in number of differential loci between the revised permutation and the other two approaches. Under revised permutation, the null distribution is less centered, the p-values are less concentrated near 0, and the estimated FDRs are larger than the true FDRs. Moreover, according to our realistic simulation result, the parametric null distribution for moderated t-statistic is closer to the permutation null distribution before adjustment, suggesting that we can not rule out the possibility that the normality assumption of the error term does not hold true in real data. In this situation, our adjustment approach is necessary to

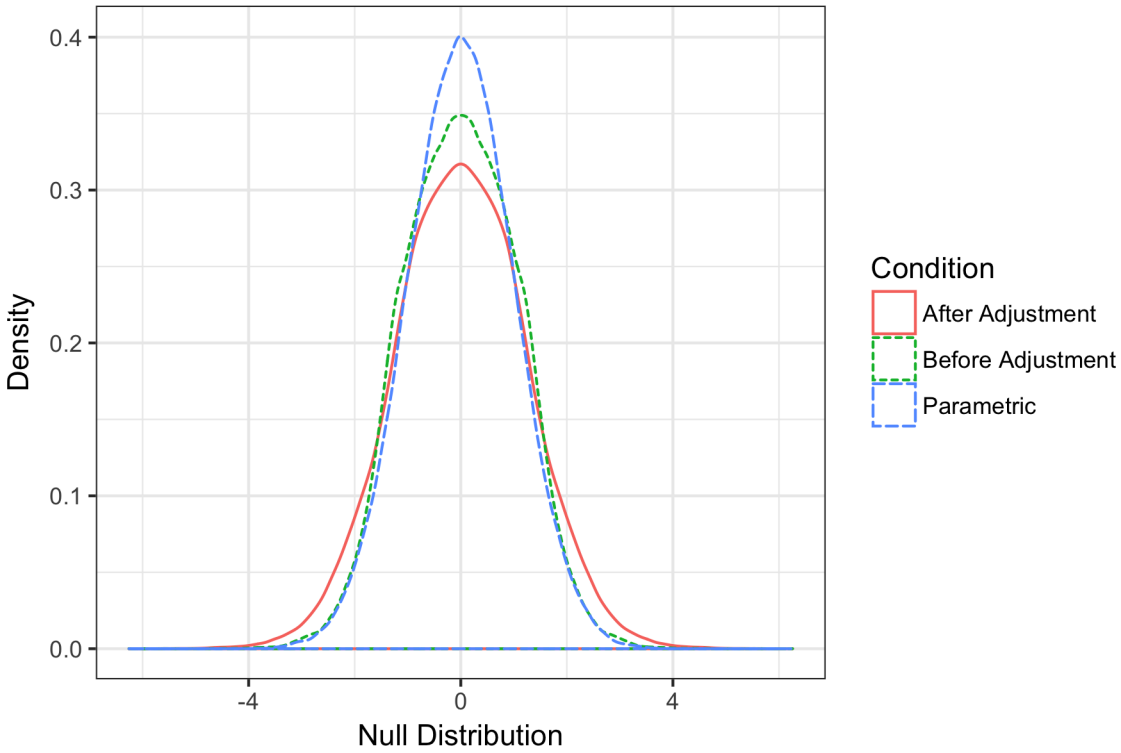

Figure 3.1: Comparison of Permutation Null Distributions before and after Adjustment, and Parametric Null Distribution, Realistic Simulation of CLL data

provide a conservative decision for differential genomic loci.

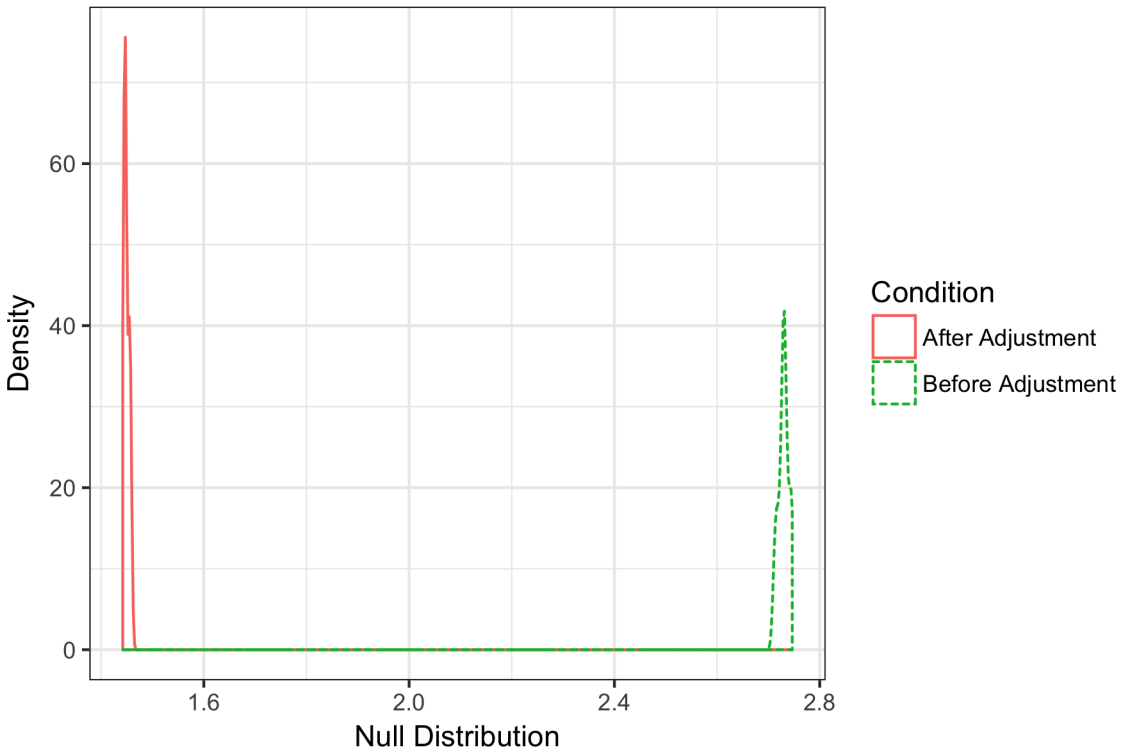

Figure 3.2: Comparison of S_0^2 before and after Adjustment, Realistic Simulation of CLL data

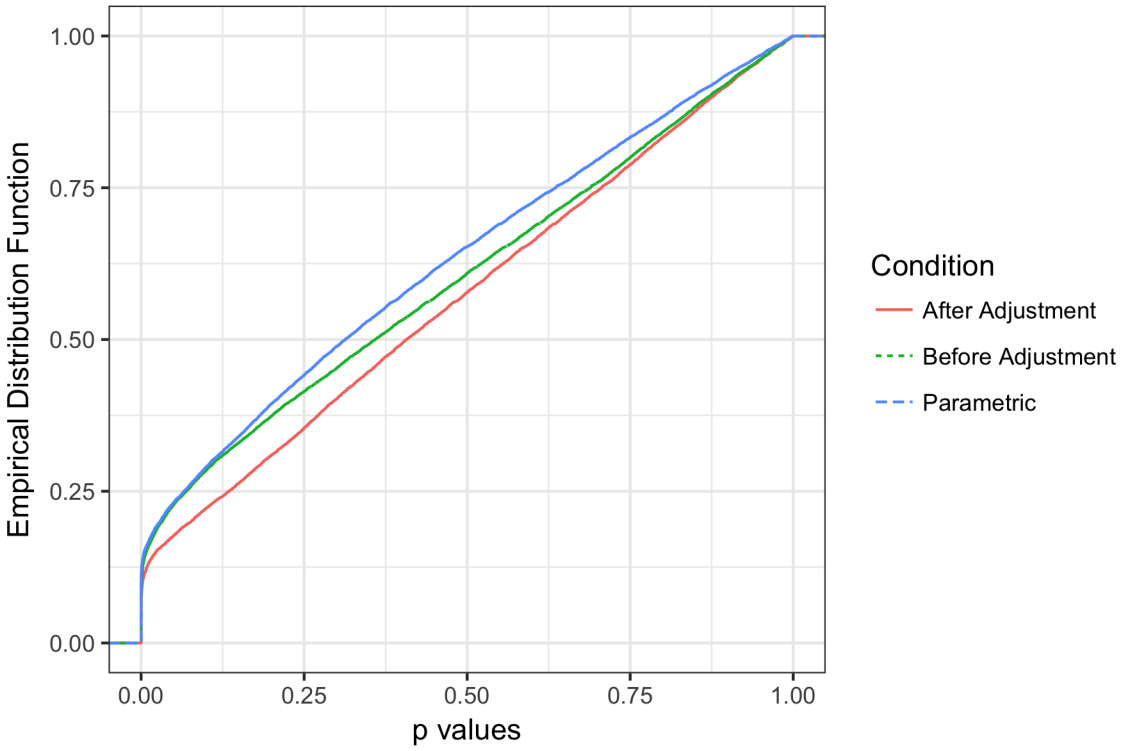

Figure 3.3: Empirical CDF of p-values, Realistic Simulation of CLL data

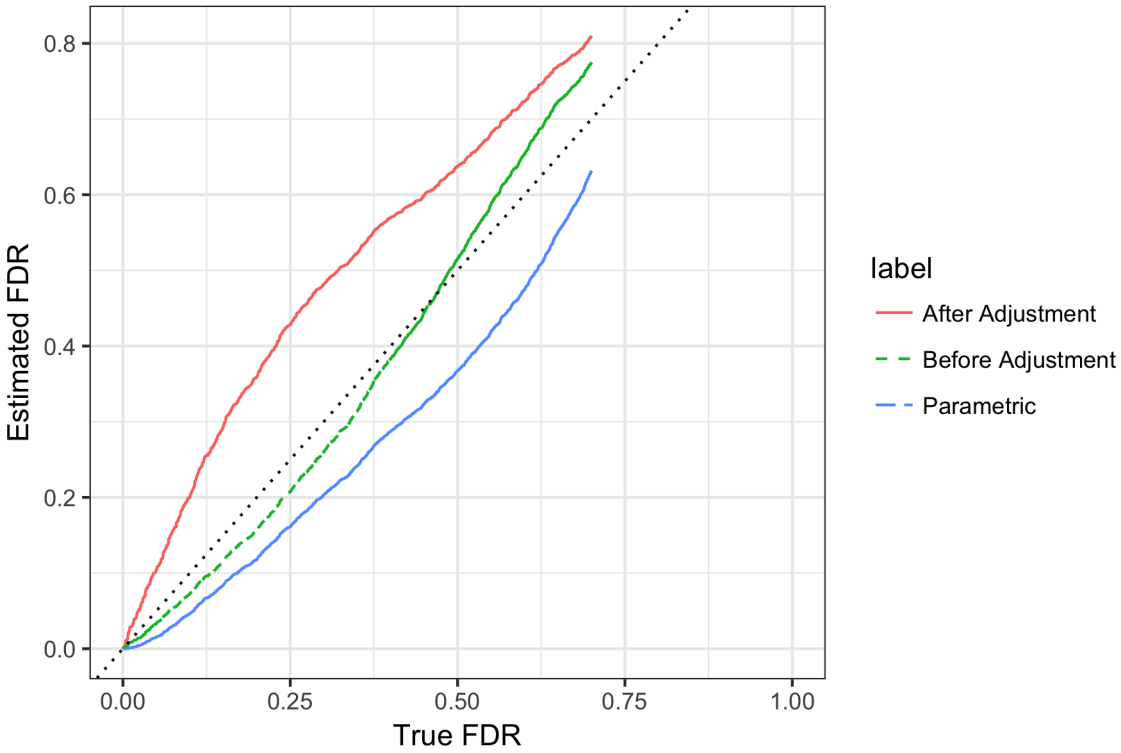

Figure 3.4: Estimated FDR v.s. True FDR, Realistic Simulation of CLL data

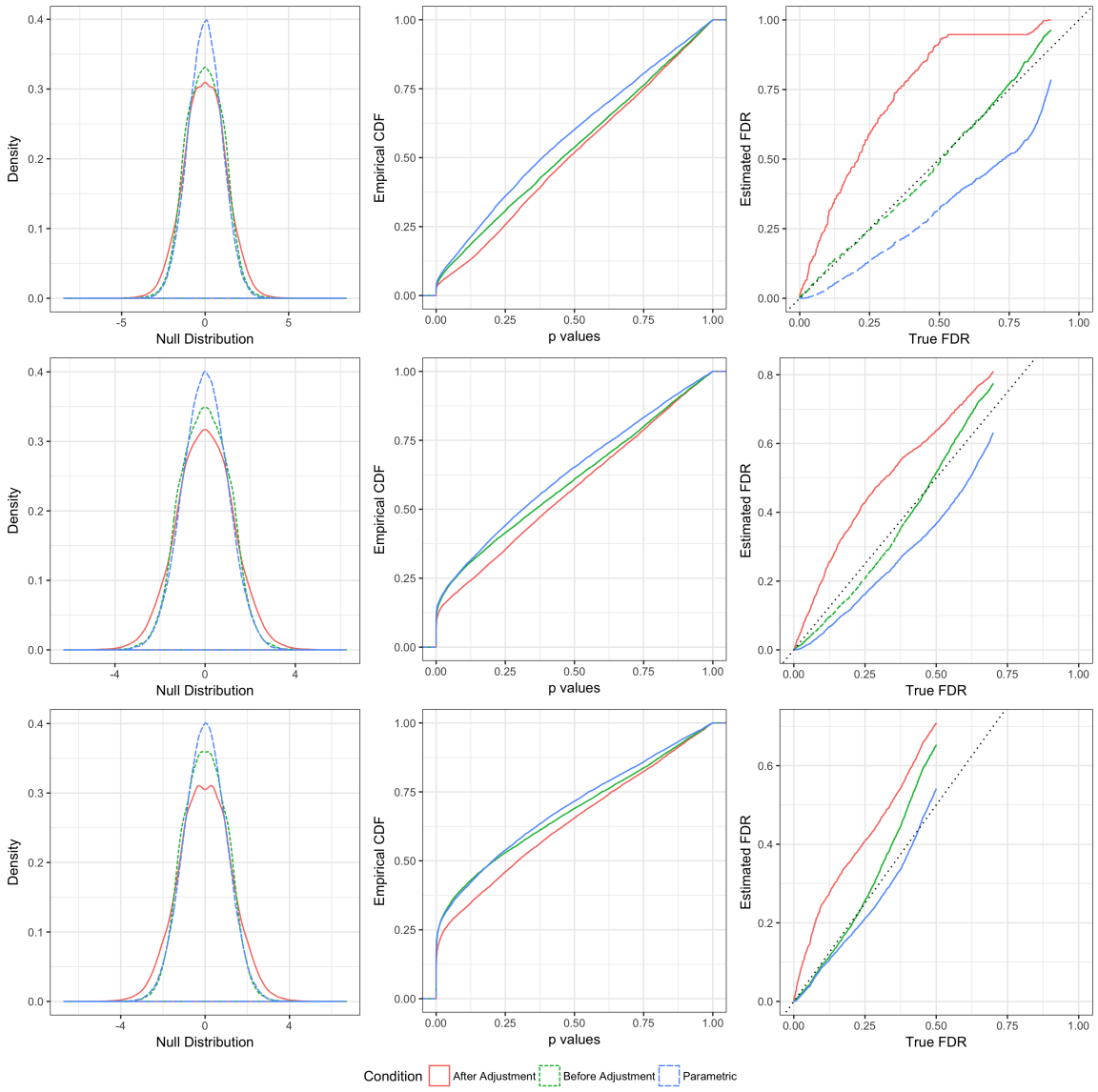

Figure 3.5: Realistic Simulation of CLL Data, p=(0.1,0.3,0.5), $\sigma^2 = 1$

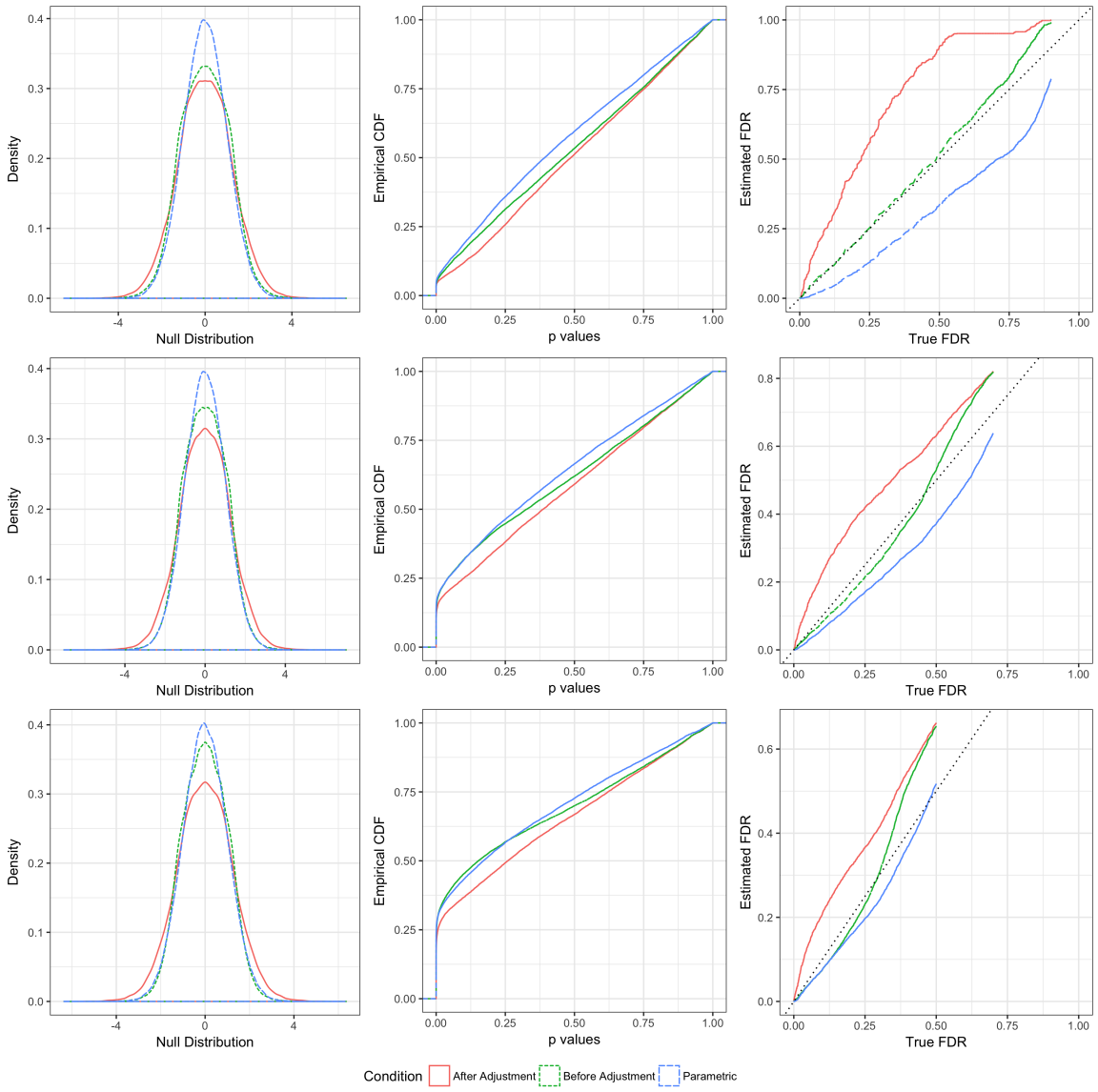

Figure 3.6: Realistic Simulation of CLL Data, p=(0.1,0.3,0.5), $\sigma^2 = 3$

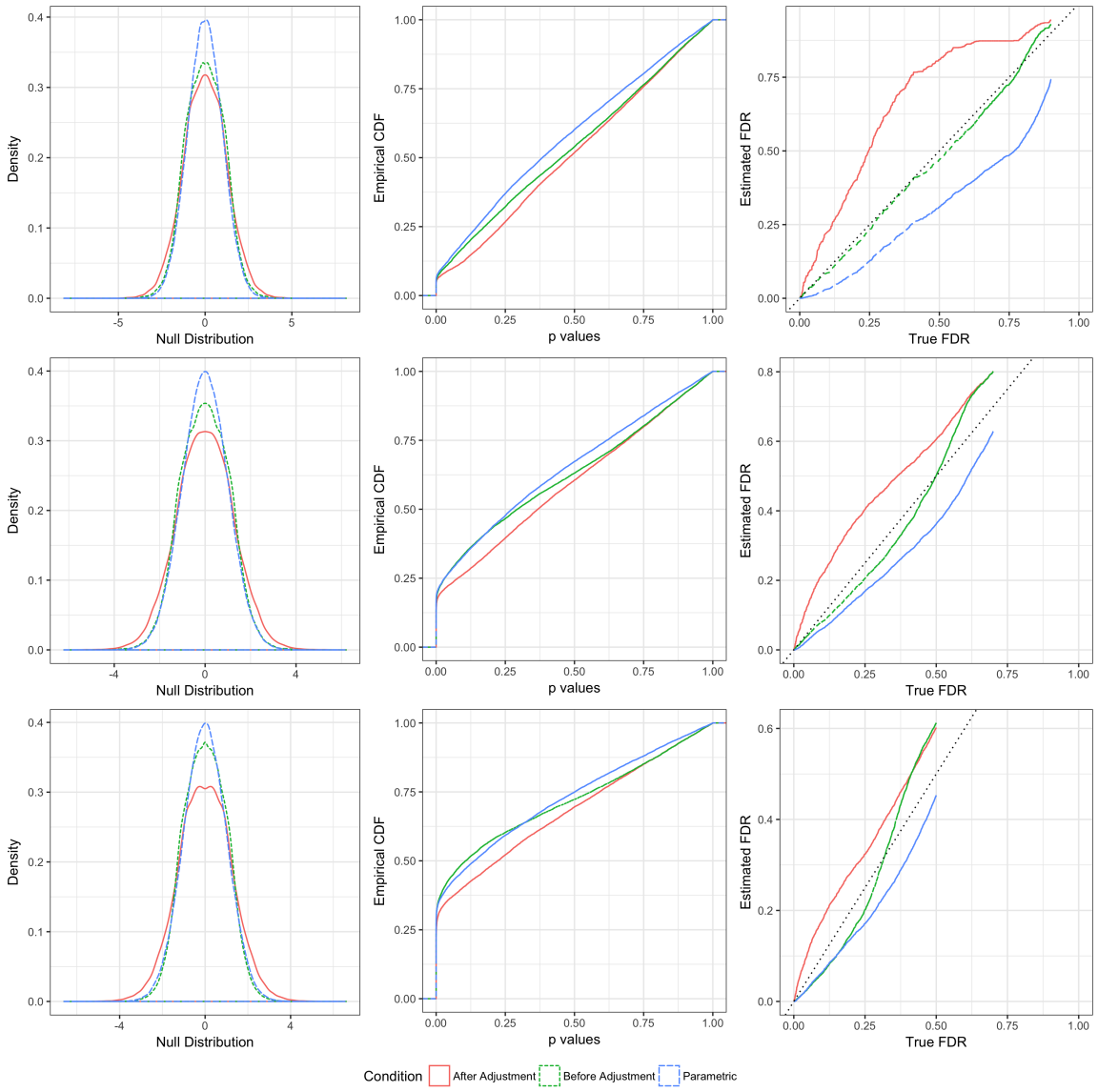

Figure 3.7: Realistic Simulation of CLL Data, p=(0.1,0.3,0.5), $\sigma^2 = 5$

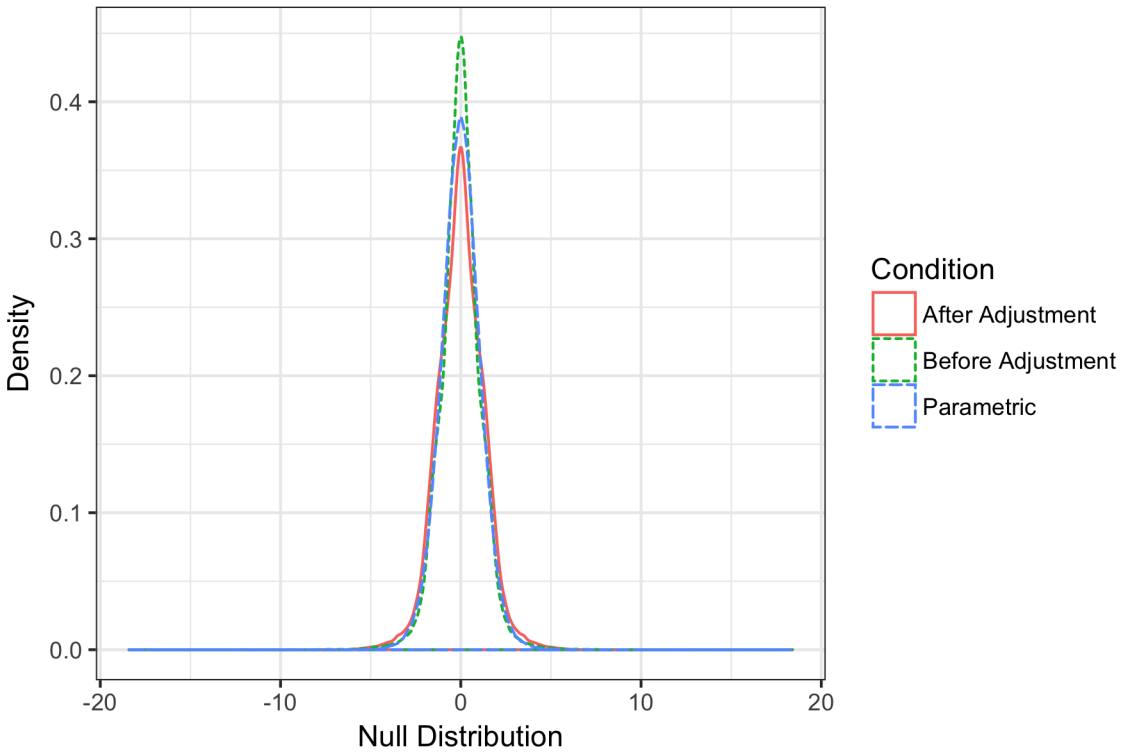

Figure 3.8: Comparison of Permutation Null Distributions before and after Adjustment and Parametric Null Distribution, CLL data

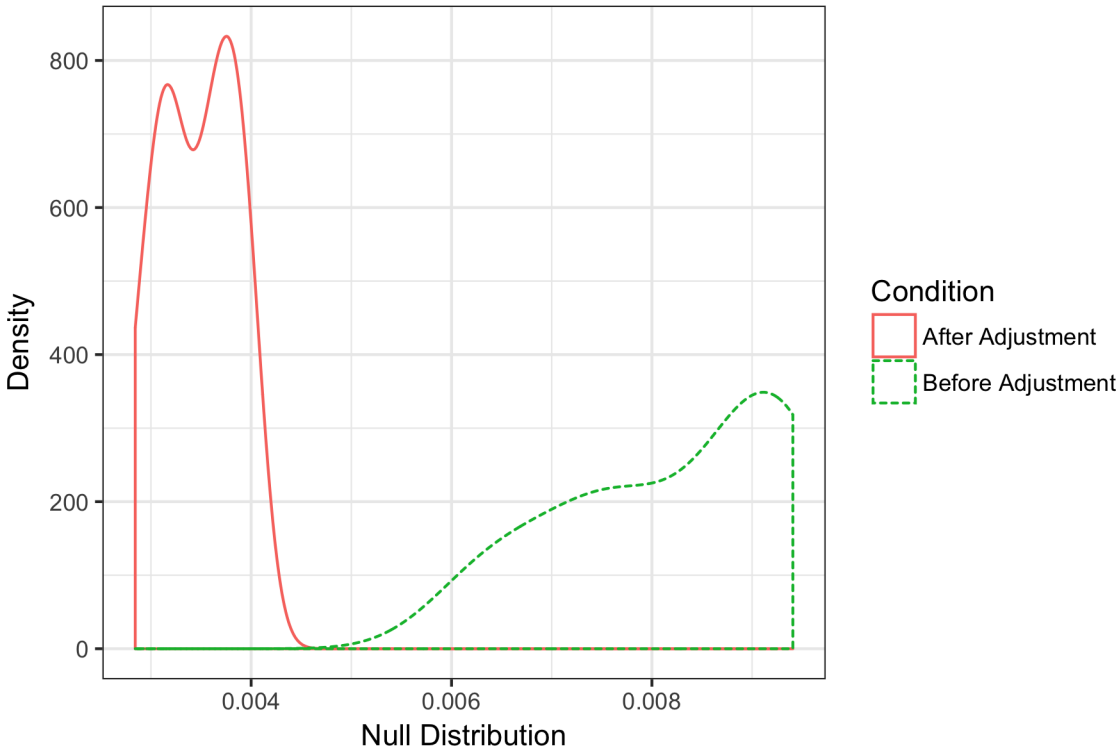

Figure 3.9: Comparison of S_0^2 before and after Adjustment, CLL data

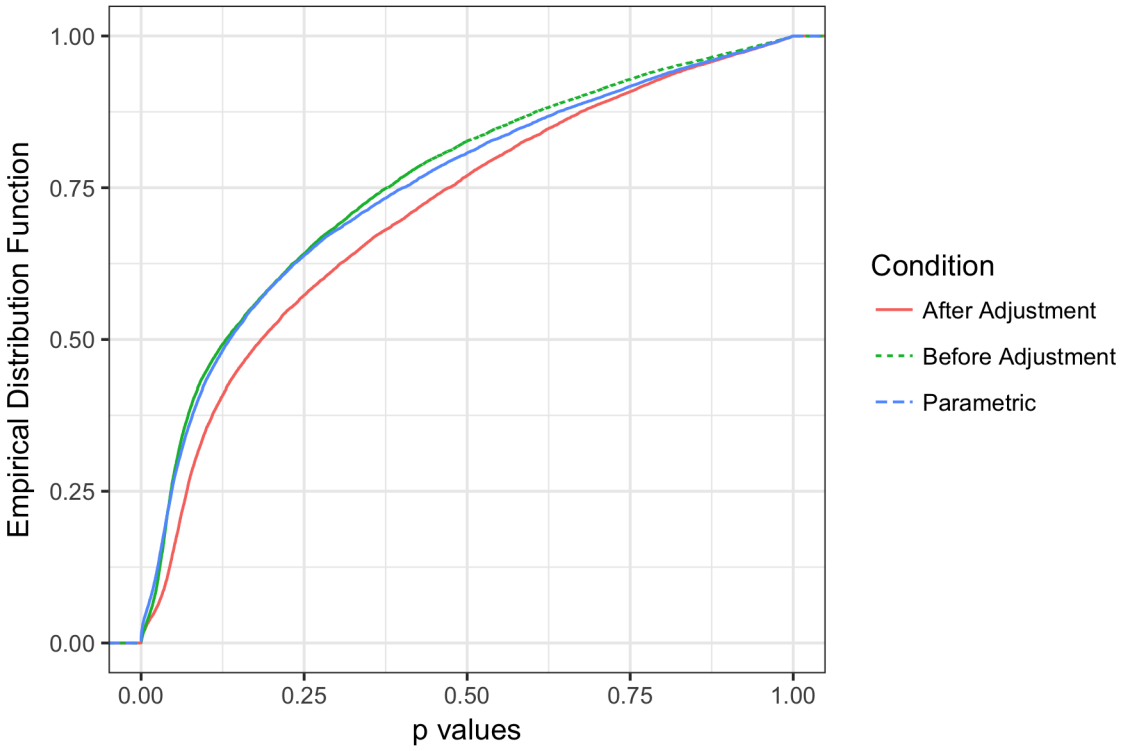

Figure 3.10: Empirical CDF of p-values, CLL data

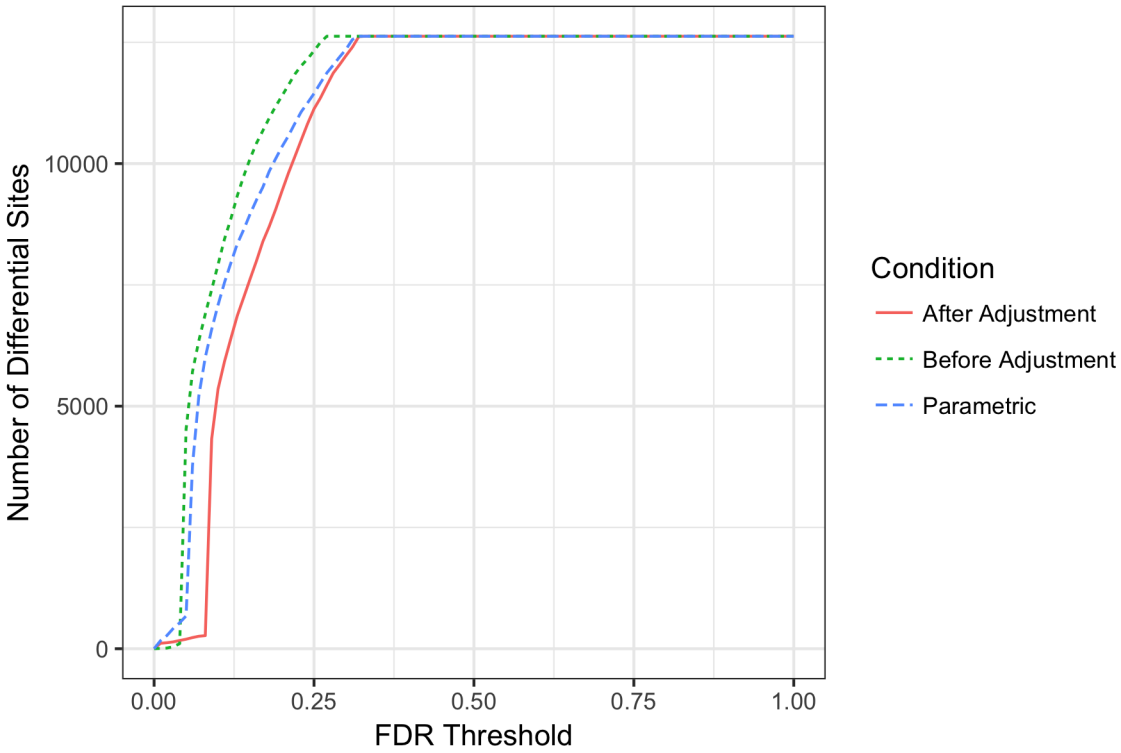

Figure 3.11: Comparison of Number of Differential Sites, CLL data

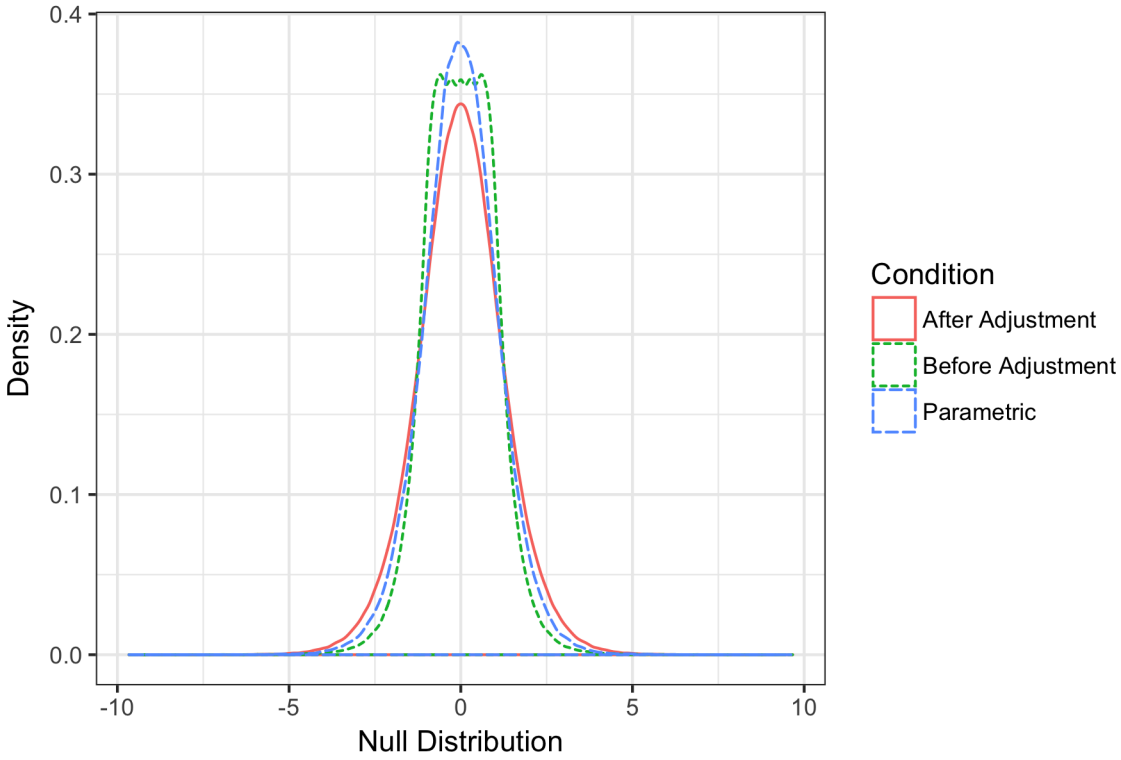

Figure 3.12: Comparison of Permutation Null Distributions before and after Adjustment and Parametric Null Distribution, Melanoma data

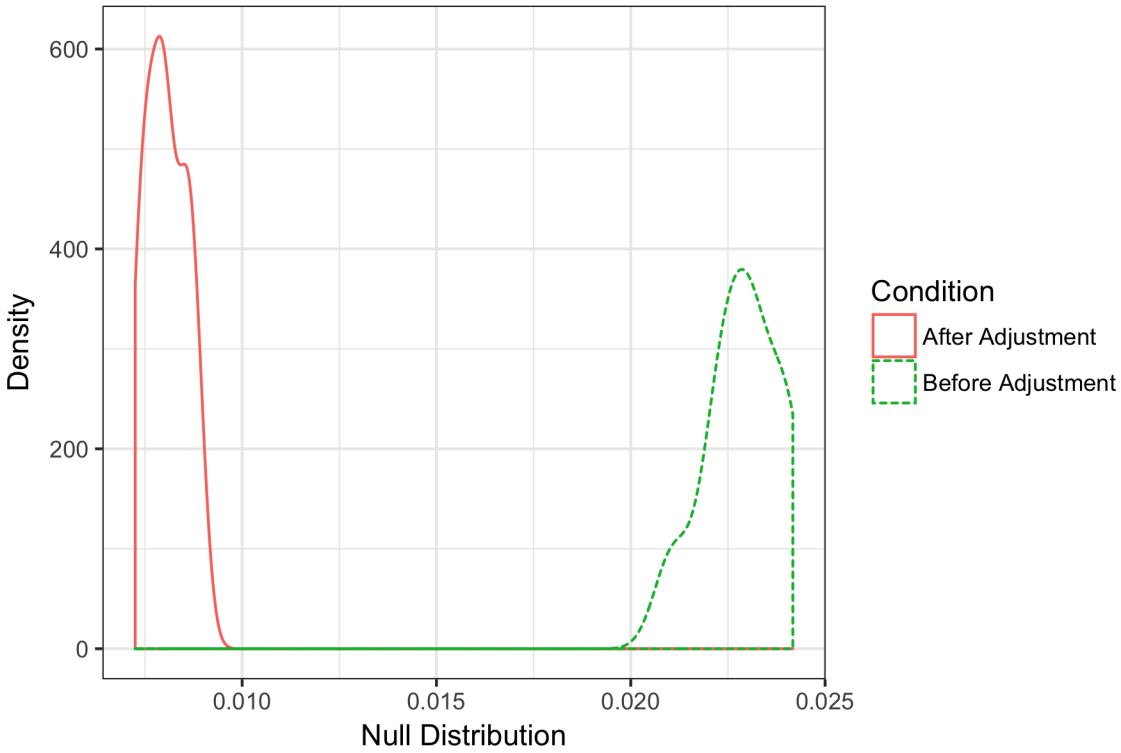

Figure 3.13: Comparison of S_0^2, Melanoma data

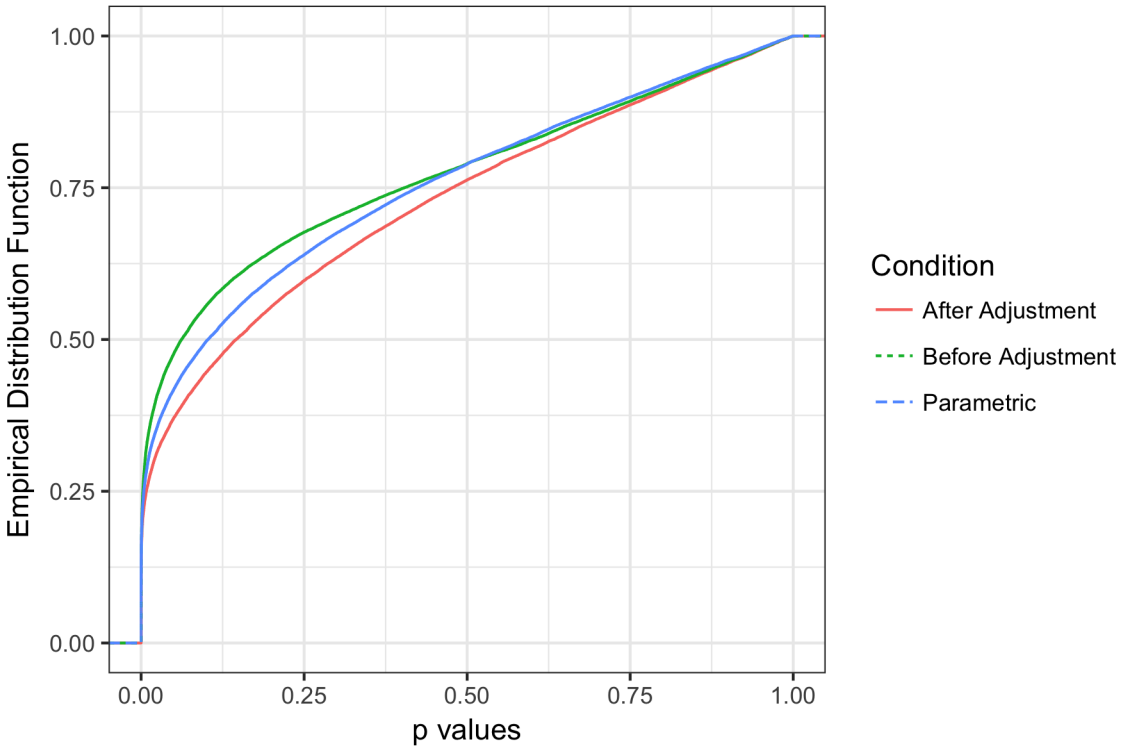

Figure 3.14: Empirical CDF of p-values, Melanoma data

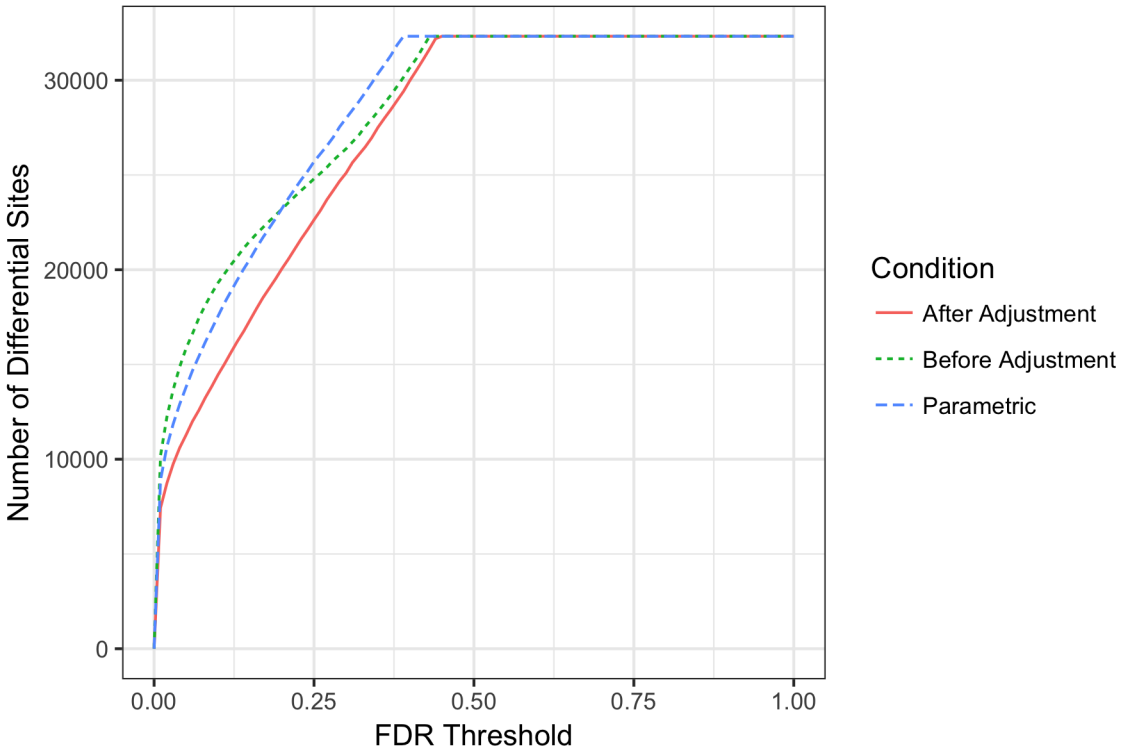

Figure 3.15: Comparison of Number of Differential Sites, Melanoma data

Chapter 4

Applications Including Multiple Data Types

4.1 Brief Introduction

In this chapter, we consider multiple types of epigenomic data jointly. We apply permutation to differential principal component analysis (dPCA).[7] In the analysis of one data type, the number of unique permutations is limited by the number of replicates available. For example, if there are 3 replicates in each biological condition, the number of distinct permutations is limited to 20. Therefore, when we construct permutation null distributions, we chose to pool moderated null t-statistics t_0s of all genomic loci together to form a shared permutation null distribution, instead of allowing each genomic locus to have its own permutation null distribution. When

multiple data types are analyzed jointly as in dPCA, each data type can be permuted. This allows one to create a large number of distinct permutations when all data types are analyzed together. Therefore, one can construct locus-specific permutation null distributions.

4.2 dPCA

We first briefly review dPCA using the data from the MYC example in dPCA paper.[7] This dataset consists of 18 different epigenomic data types, including ChIP-seq data, DNase-seq and FAIRE-seq data. dPCA is a method to analyze multiple data types jointly to identify differential protein-DNA interactions (PDI) between two biological conditions. dPCA outputs principal components as major covariation patterns of differential chromatin marks. Moreover, for each genomic locus, it outputs principal component score β to prioritize the differential genomic loci with respect to that principal component.

In our example, dPCA is used to discover differential patterns at MYC motif sites. We use $i = 1, 2$ to denote two cell types: K562 and Huvec. There are $M = 18$ different data types in total, including ChIP-seq data sets (e.g. H3k4me1, H3k4me2), DNase-seq data sets, and FAIRE-seq data sets. We have $N = 68$ samples and $G = 66364$ genomic loci in total. For each data type, we have $K = 1 \sim 3$ replicates in each cell type. The genomic loci were obtained by mapping MYC motif to human genome

using the CisGenome software.[25] After normalization and log2 transformation, the protein-DNA intensity for biological condition (cell type) i, data type m, replicate k, on genomic locus g is summarized into one number x_{gimk}. x_{gimk} can be decomposed to the true binding levels μ_{gim} and noises ϵ_{gimk}. In the original dPCA, ϵ_{gimk} is assumed to follow a normal distribution $N(0, \sigma^2)$. We calculate the mean of PDI intensity over replicates: $\bar{x}_{gim} = \sum_k x_{gimk}/K_{im}$, here K_{im} is the number of replicates of biological condition i and data type m. Then we calculate the difference of the intensity under two biological conditions: $d_{gm} = \bar{x}_{g1m} - \bar{x}_{g2m}$. This is our observed difference matrix D, the dimension of this matrix is $G \times M$. Each row of D indicates a genomic locus, and each column indicates a data type. dPCA decomposes the observed difference matrix D into two matrices: the unobserved truth Δ, which is the true difference between two biological conditions for each data type on each genomic locus; and the random sampling noise E. Based on the normality assumption, the elements in E, $e_{gm} \sim N(0, \sigma^2(\frac{1}{K_{1m}} + \frac{1}{K_{2m}}))$. The original dPCA method assumes equal variance for all genomic loci, that is $\sigma_g^2 = \sigma^2$. It is estimated by $\hat{\sigma}^2 = s^2 = \sum_g \sum_i \sum_m \sum_k \frac{(x_{gimk} - \bar{x}_{gim})^2}{\eta}$, where $\eta = G \times \sum_i \sum_m (K_{im} - 1)$. dPCA characterize Δ by principal components (PCs). There are M PCs: $v_1, v_2, ..., v_M$, and each PC represents a differential pattern, and PCs are orthogonal to each other. The true difference between two biological conditions for one genomic locus δ_g is the gth row of true differential matrix Δ, and it can be represented as a linear combination of these differential patterns. That is: $\delta_g = V\beta_g = \sum_j \beta_{gj} v_j$. Here the coefficient β_{gj}

is the principal component score, indicating if locus g is differential with respect to differential pattern j. β_{gj} will be used to rank genomic loci based on each dPC. In the original dPCA, whether β_{gj} is different from zero or not is tested using a t-distribution (or a normal distribution when the degree of freedom is large). However, in real data, noise may not be normal, therefore, we investigate the use of permutation in this context. We perform permutation tests on β_{gj}s to see if locus g is differential for pattern j or not. We have $H_0 : \beta_{gj} = 0$ and $H_1 : \beta_{gj} \neq 0$. We construct a t-statistic $T_{gj} = v_j^T d_g / \sqrt{\hat{\sigma}^2 v_j^T \Omega v_j}$ where d_g is the gth row of difference matrix D and $\Omega = diag((1/K_{11} + 1/K_{21}), ..., (1/K_{1M} + 1/K_{2M}))$ is a diagonal matrix. Under normality assumption of noise, the null distribution of t-statistics is a t-distribution, with degree of freedom $df = \sum_i \sum_m (K_{im} - 1)$. For simplicity, we use PC to indicate differential principal component (dPC) in following text.

4.3 dPCA Simulation Result

dPCA is a generative model, we can use principal components as basis to generate simulation data. First, we use PCs from original dPCA to do simulation on dPCA data. In this simulation, PCs are derived from real data. Principal component scores and random noise are simulated. The data simulation process is as following. We first pick PCs that have a signal-to-noise ratio $SNR_j = Var(v_j^T d_g)/Var(v_j^T e_g) > 5$ as recommended by dPCA.[25] In this case, top three PCs passed this threshold. For each

PC v_j, we set a proportion of differential genomic loci p_j, with respect to this PC. Then we randomly assign 0 or 1 that follows a Bernoulli distribution with parameter p_j to genomic loci to indicate if a genomic locus is differential or not with respect to v_j. Then for non-differential loci, we set coefficient $\beta_{gj} = 0$. For differential loci, we generate β_{gj} by assuming that they follow a normal distribution $N(0, \sigma_j^2)$. Then the true differential matrix $\Delta = V\beta_g = \sum_j \beta_{gj} v_j$. We set biological condition $i = 2$ (Huvec cell type) as the base line. We now assume that the random noise follows a normal distribution $\epsilon_{gimk} \sim N(0, \sigma_0^2)$. Therefore $x_{g2mk} = 0 + N(0, \sigma_0^2)$. For biological condition $i = 1$, we have $x_{g1mk} = \delta_{gm} + N(0, \sigma_0^2)$.

In this simulation process, there are three parts of the parameters. First, p_js are the proportion of differential genomic loci for each PC. Second, σ_j^2s are differential magnitude for each PC. Third, σ_0^2 is the noise level. Several combinations of parameters are used to see if our adjustment approach is stable under different parameter settings. The combinations of parameters are listed in Table **??**. The results from different parameter settings are similar. We presented the results from the second parameter setting here. $p_1 = 0.2$, $p_2 = 0.12$, and $p_3 = 0.04$. $\sigma_1 = 5$, $\sigma_2 = 3$, $\sigma_3 = 1$, and $\sigma_0^2 = 0.1$. Results from other parameter settings are presented at the end of the chapter.

After conducting dPCA on simulated data, we do permutations. We switch the label of biological conditions $i = 1, 2$ for each data type randomly to create a permuted data matrix. For example, if we originally have samples $x_{1m1}, x_{1m2}, x_{2m1}, x_{2m2}, x_{2m3}$.

That is, for data type m, we have two replicates for biological condition 1, and three replicates for condition 2, then one possible permutation is $\tilde{x}_{1m1} = x_{1m1}, \tilde{x}_{1m2} = x_{2m1}, \tilde{x}_{2m1} = x_{2m2}, \tilde{x}_{2m2} = x_{1m2}, \tilde{x}_{2m3} = x_{2m3}$. We have $M = 18$ data types in total, and the permutations of different data types are independent with each other. Therefore, we are able to create a large number of distinct permutations by combining permutations of different data types together. Data from two biological conditions are mixed together after permutations. Therefore, when we follow the same approach to calculate the t-statistic of β_{gj} for each permutation, those t-statistics will form an permutation null distribution for t_{gj}. We use t_{0gj} to indicate them. We conduct $B = 2000$ permutations in total, yielding a locus-specific null distribution for each locus. P-values and false discovery rates (FDR) are calculated based on the permutation null distributions.

dPCA assumes equal variances for all genomic loci $\sigma_g^2 = \sigma^2$. In this case, even though we do not use variance shrinkage estimator, $\hat{\sigma}^2$ is estimated by pooling information from all genomic loci. If we perform naive permutation, the sample variance for permuted data would be inflated by the increased sample variances of differential genomic loci after permutation. If we want to relax the equal variance assumption and assume locus-specific variance, variance shrinkage estimator should be used to achieve a more stable variance estimation due to the small number of replicates. When the variance shrinkage estimator is used, we would expect that the permutation null distributions generated by naive permutations are more centered than the

true null distribution. Consequently, some of the non-differential genomic loci would be misclassified as differential due to the small p-values based on the permutation null distributions.

In our proposed approach, we first force the data under two conditions to have the same mean for each data type before permutation, by subtracting the means. After this adjustment, for each genomic locus and each data type, we then permute data between two conditions.

To illustrate the effects of adjustment, we pick several genomic loci from the example data to compare the permutation null distributions of t-statistic before and after adjustment in Figure 4.1. Two curves are the permutation null distributions before and after adjustment for principal component score of this locus with respect to the first PC. These genomic loci are non-differential. The red solid curve is permutation null distribution after adjustment. After adjustment, the permutation null distributions are less centered than the naive permutation (green curve). This is consistent with our expectation

We also pick 10 true differential genomic loci on Figure 4.2. For differential loci, the behavior of two null distributions are different. We can see that for differential genomic loci, the permutation null distributions after adjustment (red curve) are actually more centered than before adjustment (green curve). This is reasonable because after adjustment the difference d_{gm} in permuted data also decreases. This may cause a more centered null distribution. This is also good because for differential

genomic loci, a more centered null distribution would result in smaller p-values.

Next, we compared the distributions of p-values of all genomic loci for top three PCs. The empirical CDF are presented in Figure 4.3. We also calculated the p-values based on the t-distributions as the assumption in original dPCA. From this plot, we can see that after adjustment, the distribution of p-values is more uniform than it is before adjustment. For each PC, when p-value is near 0, the value of empirical CDF after adjustment is closer to the true proportion of differential genomic loci for this PC in our parameter setting, which is $p_1 = 0.2$, $p_2 = 0.12$, and $p_3 = 0.04$. The parametric p-values (blue curve) is also distributed in a uniform way.

We presented the true FDRs and estimated FDRs for top three PCs in Figure 4.4. The red curve is based on the permutation after adjustment, the green curve is based on the permutation before adjustment, and the blue curve is based on the parametric t-distribution. The black line in the plot is the diagonal line as a reference. Among three approaches, FDR curve based on naive permutation is below the diagonal line. Therefore, this estimation is too optimistic. The other two approaches are both conservative, while the parametric approach works better. This is also consistent with our expectation, because the normality assumption of error term holds true in this simulation.

Next we simulate the data assuming that the error term follows a t-distribution with degree of freedom $df = 2$. Now the normality assumption of error term does not hold true. We conduct permutations in the same way. The results are present in

Figure 4.5 and Figure 4.6. One can see that when the normality assumption does not hold true, the estimated FDRs based on parametric t-distribution are smaller than the true FDRs. Therefore, t distribution is no longer an appropriate null distribution to prevent us from too optimistic decision for differential loci, as the same with the permutation null distribution before adjustment. In this situation, permutation after adjustment is the only approach among those three to provide a conservative decision for differential loci.

We investigate our approach under different parameter settings. The parameter settings are listed in Table **??**. Under each parameter setting, we assume two different distributions of error term: a normal distribution or a t-distribution with degree of freedom $df = 2$. The results are presented in Figure 4.7 to Figure 4.14. Each figure corresponds to one parameter setting. The six plots in the first row of each figure are the results under normality assumption of error term. The six plots in the second row are the results under t-distribution with $df = 2$ assumption of error term. Under different parameter settings, the results are similar and draw the same conclusions.

4.4 dPCA Realistic Simulation Result

One drawback of the naive simulation in last section is that we do not maintain the true noise structure in dPCA data. We simply assume that the error term follows a normal distribution or a t-distribution. In this section, we conduct a realistic

simulation to investigate the behavior of our method under the true noise structure in dPCA data. We only simulate the principal component score for each genomic locus.

The data simulation process is as following. First, we calculate a residual data matrix R by removing the mean of each data type under each cell type. $r_{gimk} = x_{gimk} - \bar{x}_{gim}$. Then this residual matrix R contains the true noise structure in dPCA data. Then we simulate principal component score β_{gj} as the same with naive simulation in last section. The true difference matrix $\Delta = V\beta_g = \sum_j \beta_{gj} v_j$. We set biological condition $i = 2$ (Huvec cell type) as the base line. The simulation data for biological condition $i = 2$ is simply the residual data matrix. $x_{g2mk} = r_{g2mk}$. For biological condition $i = 1$, we add the residual matrix to the true difference matrix. That is $x_{g1mk} = r_{g1mk} + \delta_{gm}$.

In this simulation process, there are two parts of the parameters. First, p_js are the proportion of differential genomic loci for each PC. Second, σ_j^2s are differential magnitude for each PC. Since the result from last section under different parameter settings are similar, we picked the second parameter settings in this section. $p_1 = 0.2$, $p_2 = 0.12$, and $p_3 = 0.04$. $\sigma_1 = 5$, $\sigma_2 = 3$, and $\sigma_3 = 1$.

Next, we conduct permutations as the same with last section. Permutation null distributions for each genomic locus are formed. Then, we conduct the adjustment by forcing each data type under two biological conditions to have the same mean. Then, we conduct permutations again to generate permutation null distributions af-

ter adjustment. We pick 10 non-differential genomic loci from the example data to compare the permutation null distributions before and after adjustment in Figure 4.15. We also pick 10 differential genomic loci, and the result is present in Figure 4.16. From these two plots we can draw the same conclusion. For non-differential genomic loci, the permutation null distribution after adjustment is less centered than it before adjustment. Therefore, these non-differential genomic loci are less likely to be misclassified as differential ones than before. On the other hand, for differential genomic loci, the permutation null distribution after adjustment is more centered, which would result in smaller p-values.

Next, we compared the distributions of p-values of all genomic loci for top three PCs. The empirical CDF are presented in Figure 4.17. After adjustment, p-values are distributed in a more uniform way than it before adjustment. The p-values from parametric t-distribution is also distributed in a uniform way. For each PC, when p-value is near 0, the value of empirical CDF after adjustment is closer to the true proportion of differential genomic loci for this PC in our parameter setting. We then compared the estimated FDRs and the true FDRs for top three PCs. The result is present in Figure 4.18. We can see from this plot that for the permutation null distribution before adjustment and the parametric t-distribution, the estimated FDRs are smaller than the true FDRs. Therefore, based on these two distributions, we will make too optimistic decisions for differential loci. After adjustment, estimated FDRs are large than the true FDRs, which will provide conservative decisions for differential

loci. Our adjustment is effective and the result is consistent with our expectation.

4.5 dPCA Real Data Result

In this section, we conduct permutations on the real data. We switch the label of biological conditions for each data type randomly to create a permuted data matrix. Then a null t-statistic is calculated for each permutation. These null t-statistics form permutation null distributions for each genomic locus. Then we apply the proposed adjustment approach by forcing the data under two conditions to have the same mean. Then we conduct permutations on the adjusted data again, which form a permutation null distributions after adjustment.

To illustrate the effects of adjustment, we picked 10 genomic loci to compare the permutation null distributions of t-statistics before and after adjustment in Figure 4.19. These 10 genomic loci have small t-statistics, implying that they are likely to be non-differential with respect to the first principal component. We can draw the same conclusion as before. After adjustment, the permutation null distributions are less centered.

Next, we calculated empirical p-values with respect to permutation null distributions before and after adjustment and the parametric null distribution. We plot the empirical CDF in Figure 4.20. After adjustment, the p-values are distributed in a more uniform way than it before adjustment. Though we do not know the hidden

truth of differential sites, we can see that the behavior of the permutation null distributions and the p-values in real data are similar to our simulation results before. Therefore, it is reasonable to conclude that, without the adjustment, the permutation null distribution is overly centered. Our adjustment is necessary, if we prefer a conservative decision for differential genomic loci.

Next, we calculated estimated FDRs based on the p-values. Given the same FDR threshold $FDR = 0.05$, naive permutation reports 29159 differential loci among all 66364 genomic loci for the first PC. Parametric model reports 17780 differential genomic loci. Permutation after adjustment reports 13934 differential loci. For the second PC, at $FDR = 0.05$, naive permutation reports 9179 differential loci. Parametric model reports 12508 differential loci. Permutation after adjustment reports 6886 differential loci. For the third PC, naive permutation reports 12104 differential loci. Parametric model reports 12043 differential loci. Permutation after adjustment reports 4641 differential loci. There is a significant decrease of the number of differential genomic loci based on permutation after adjustment, comparing to the other two approaches. Our adjustment therefore has substantial influence on how many loci will be reported as differential.

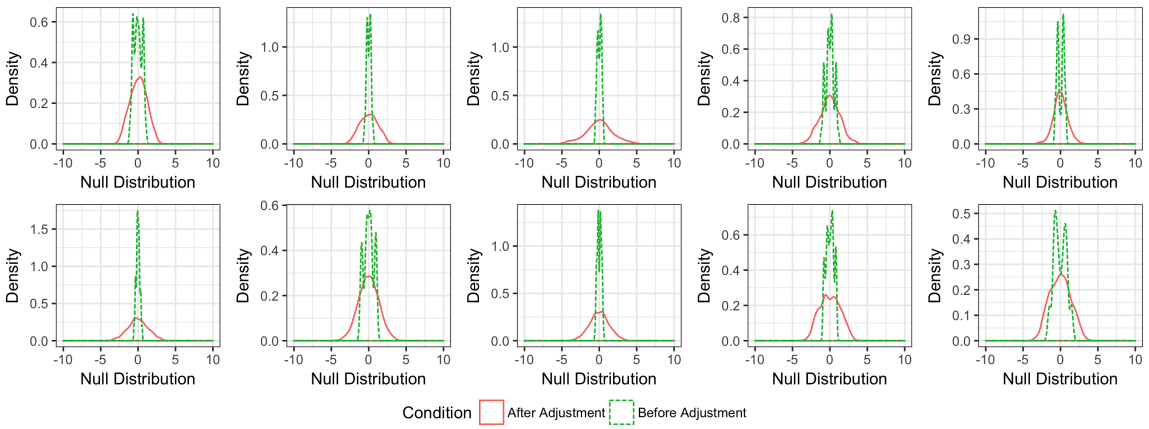

Figure 4.1: Comparison of Permutation Null Distributions before and after Adjustment, Simulation of dPCA Data, Non-differential Sites

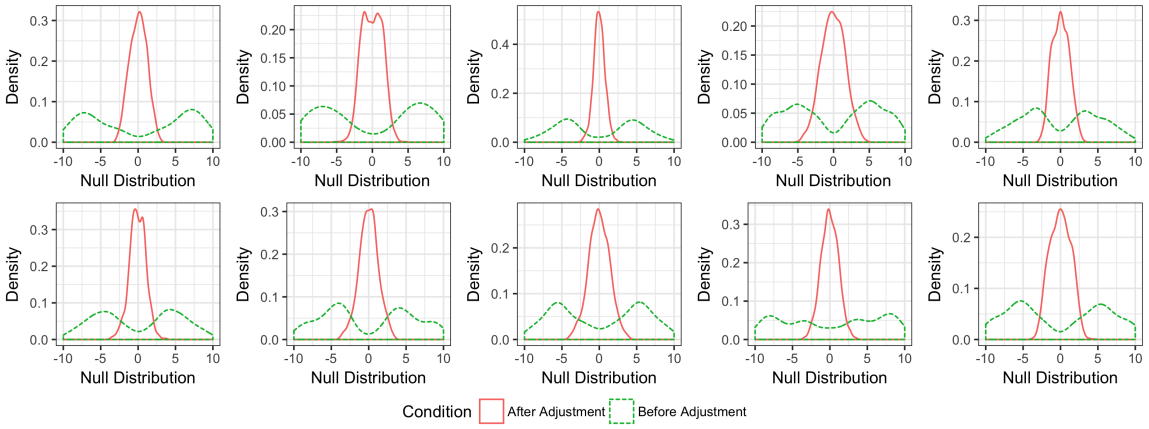

Figure 4.2: Comparison of Permutation Null Distributions before and after Adjustment, Simulation of dPCA Data, Simulation of dPCA data, Differential Sites

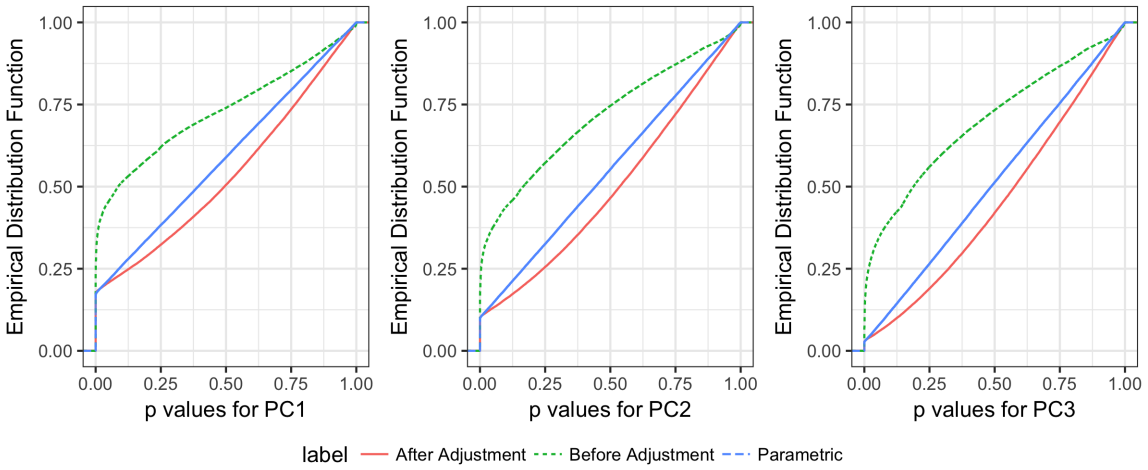

Figure 4.3: Empirical CDF of p-values for Top Three PCs, Simulation of dPCA Data, $\epsilon_{gimk} \sim N(0, \sigma_0^2)$

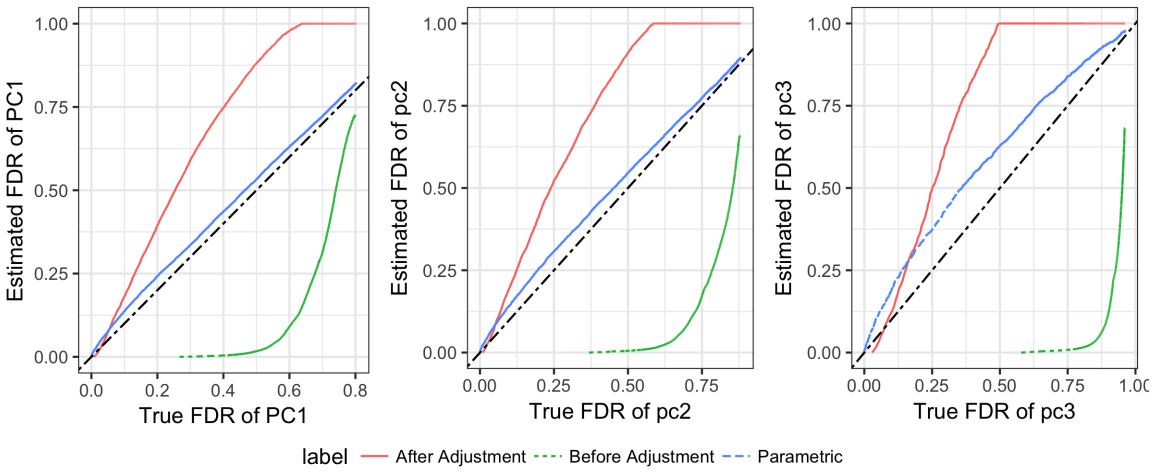

Figure 4.4: Estimated FDRs v.s. True FDRs for Top Three PCs, Simulation of dPCA Data, $\epsilon_{gimk} \sim N(0, \sigma_0^2)$

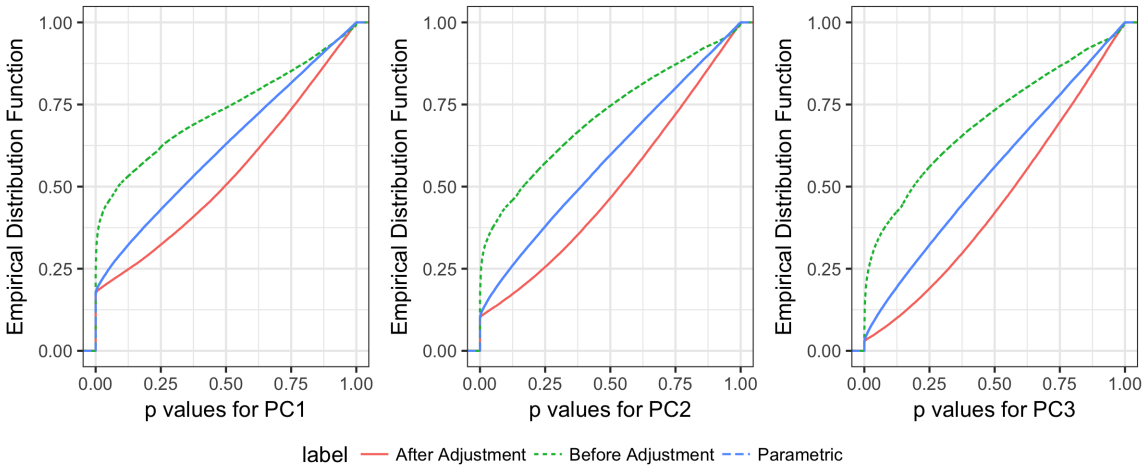

Figure 4.5: Empirical CDF of p-values for Top Three PCs, Simulation of dPCA Data, $\epsilon_{gimk} \sim t_2$

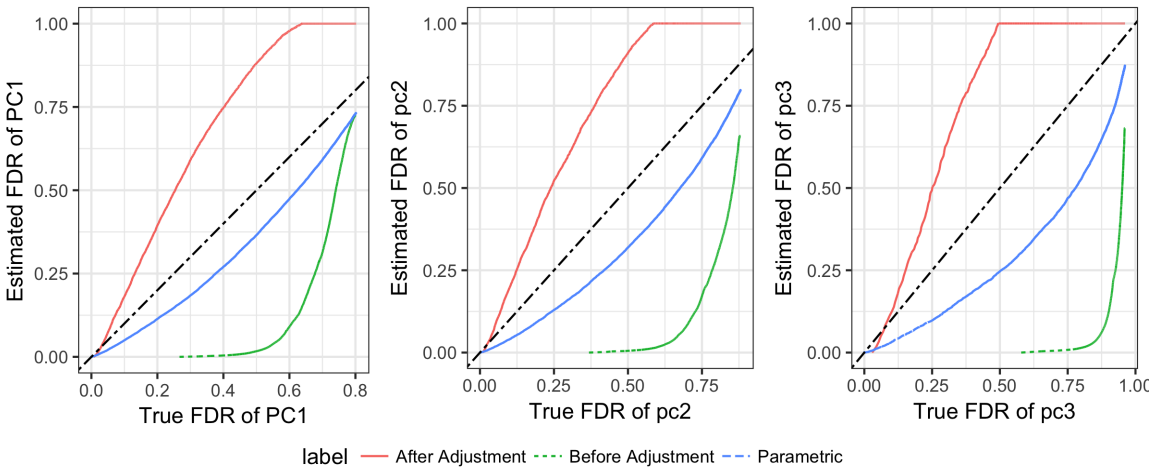

Figure 4.6: Estimated FDRs v.s. True FDRs for Top Three PCs, Simulation of dPCA Data, $\epsilon_{gimk} \sim t_2$

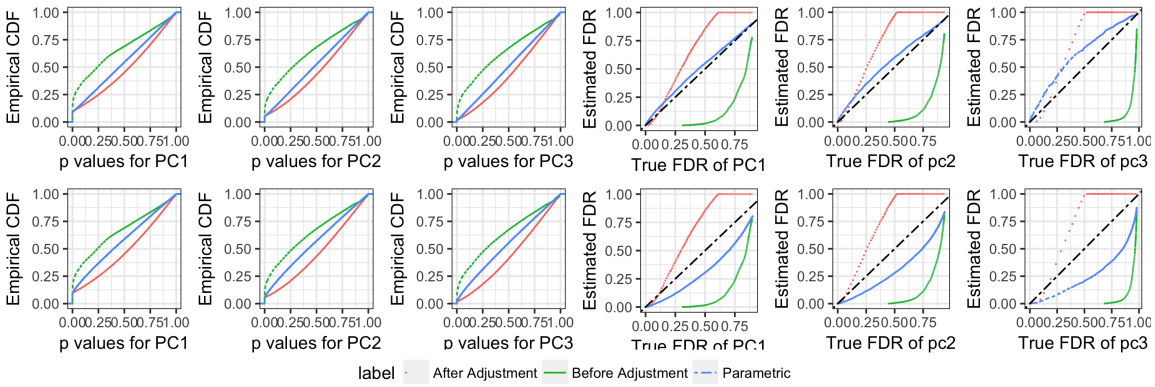

Figure 4.7: Comparison of p-values and FDRs for Top Three PCs, Simulation of dPCA Data, Parameter Setting 1

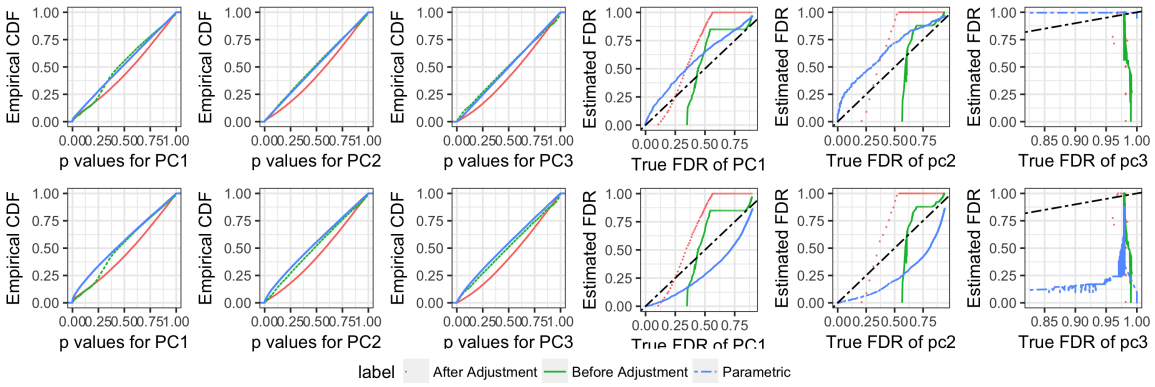

Figure 4.8: Comparison of p-values and FDRs for Top Three PCs, Simulation of dPCA Data, Parameter Setting 3

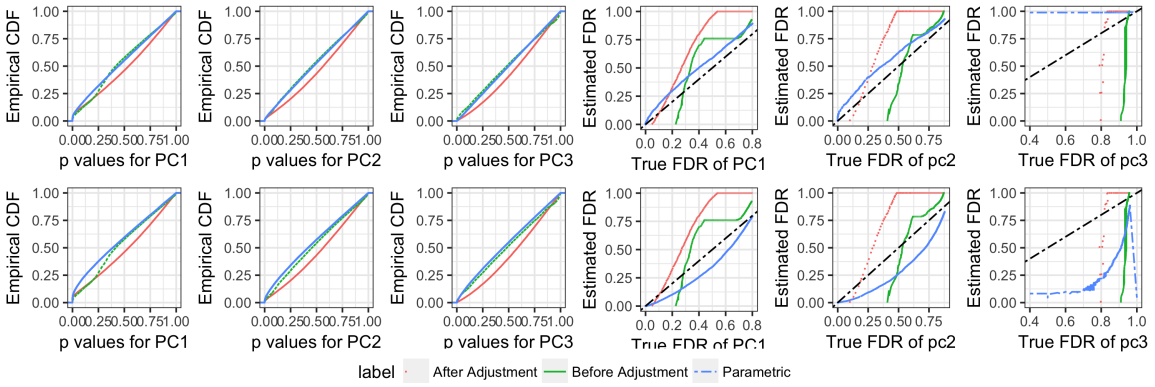

Figure 4.9: Comparison of p-values and FDRs for Top Three PCs, Simulation of dPCA Data, Parameter Setting 4

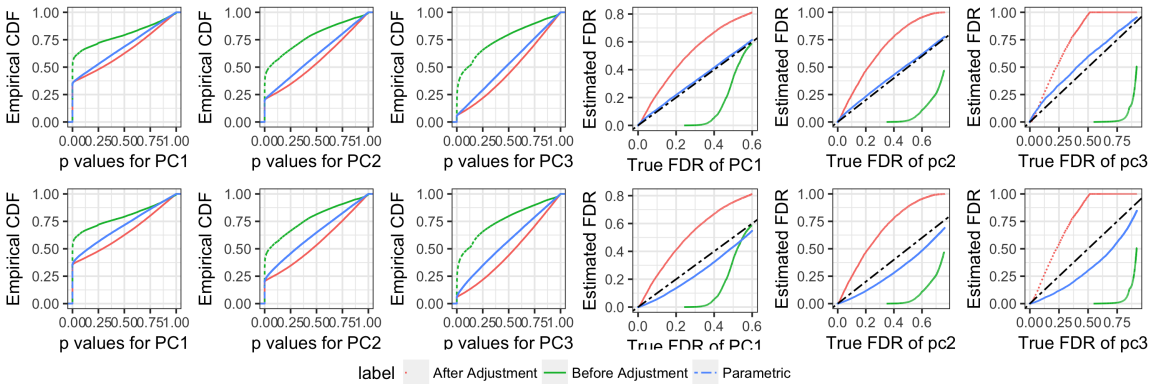

Figure 4.10: Comparison of p-values and FDRs for Top Three PCs, Simulation of dPCA Data, Parameter Setting 5

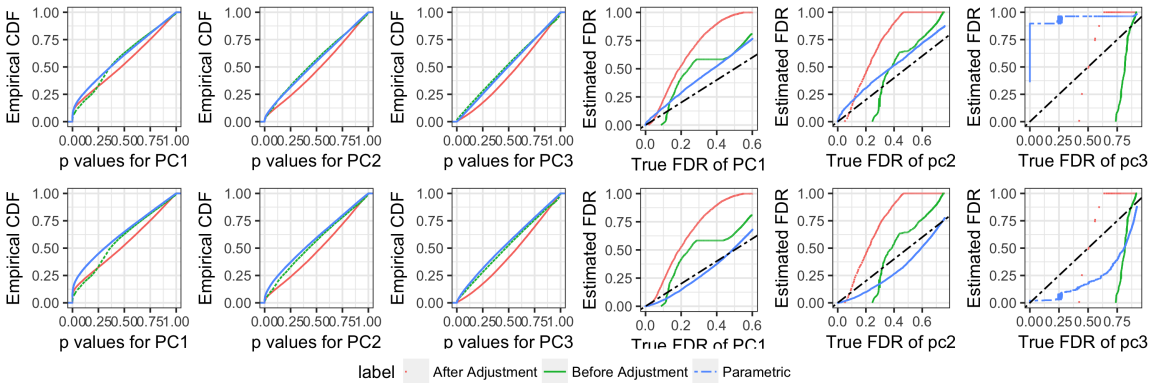

Figure 4.11: Comparison of p-values and FDRs for Top Three PCs, Simulation of dPCA Data, Parameter Setting 6

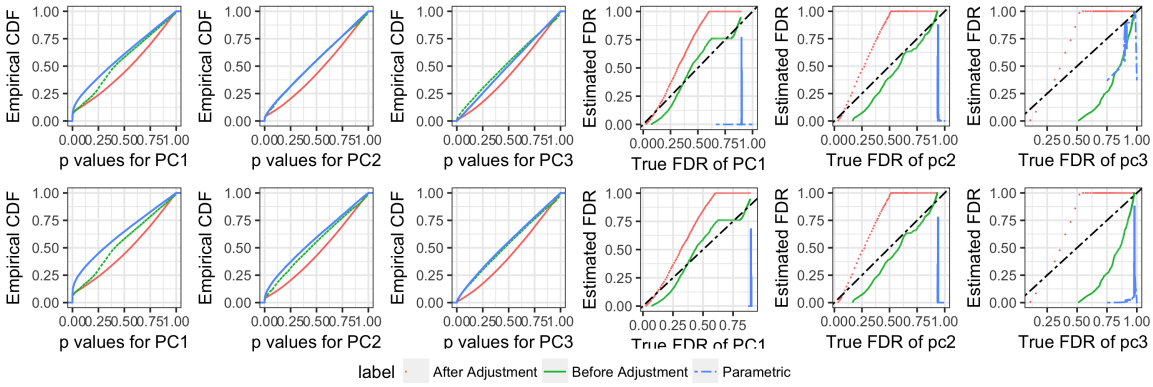

Figure 4.12: Comparison of p-values and FDRs for Top Three PCs, Simulation of dPCA Data, Parameter Setting 7

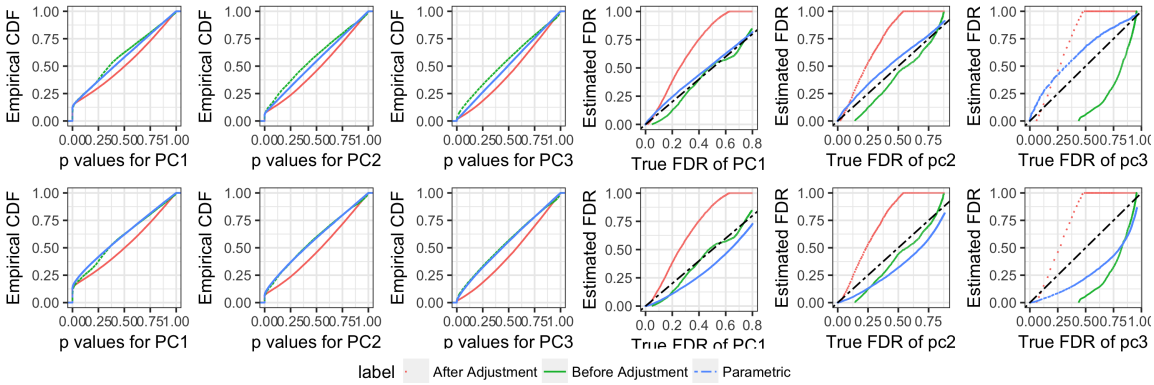

Figure 4.13: Comparison of p-values and FDRs for Top Three PCs, Simulation of dPCA Data, Parameter Setting 8

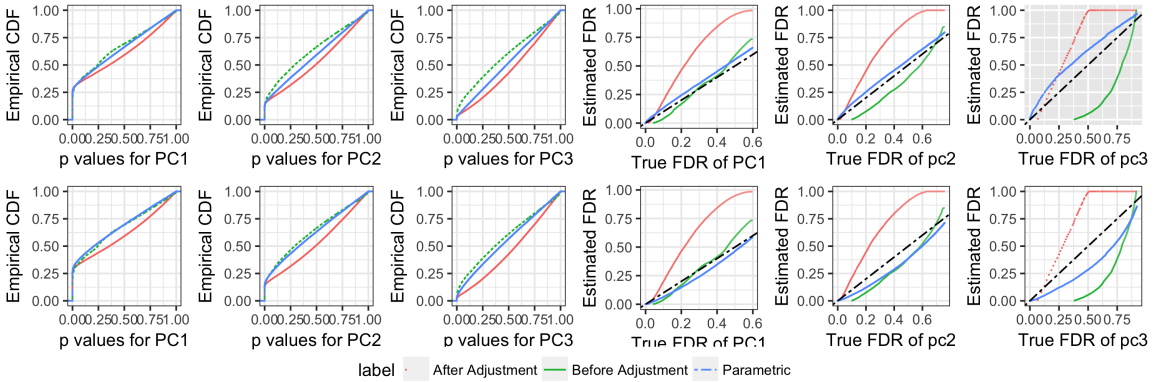

Figure 4.14: Comparison of p-values and FDRs for Top Three PCs, Simulation of dPCA Data, Parameter Setting 9

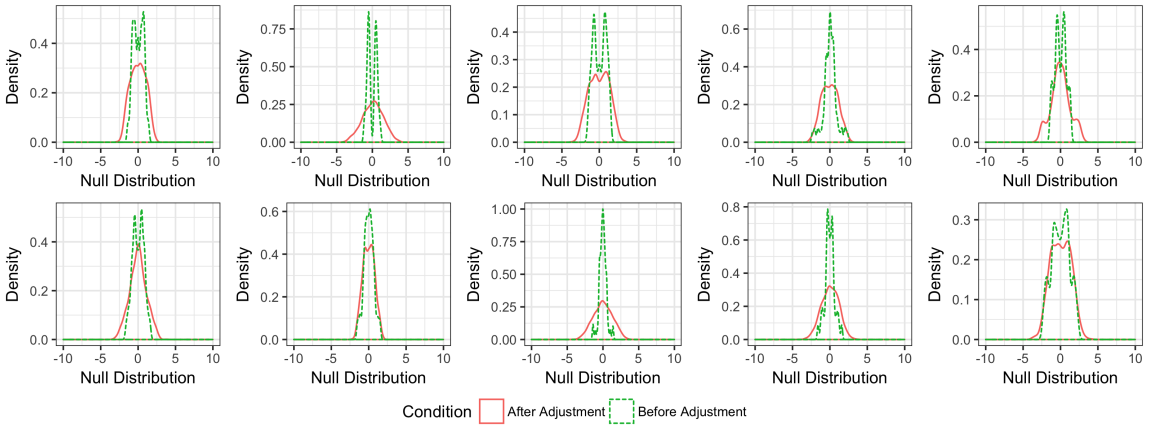

Figure 4.15: Comparison of Permutation Null Distributions before and after Adjustment, Realistic Simulation of dPCA Data, Non-differential Sites

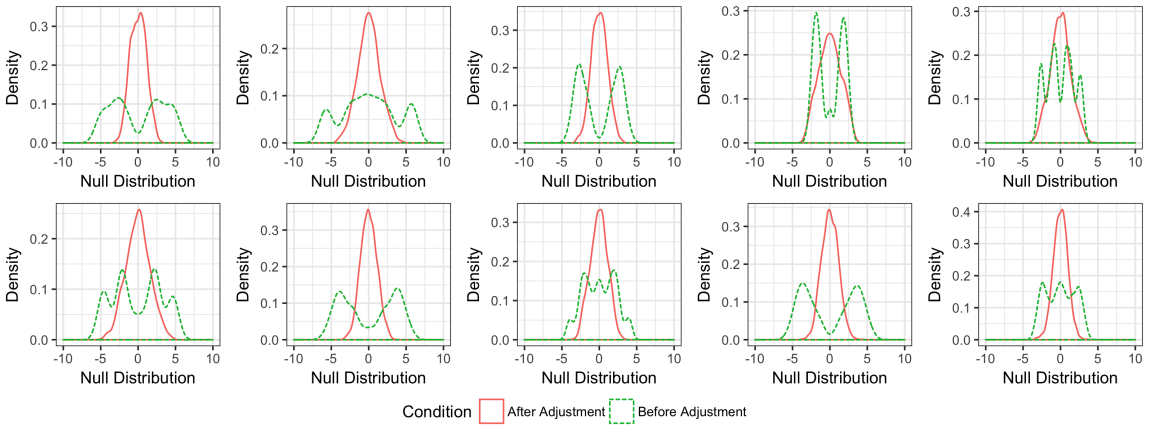

Figure 4.16: Comparison of Permutation Null Distributions before and after Adjustment, Realistic Simulation of dPCA Data, Differential Sites

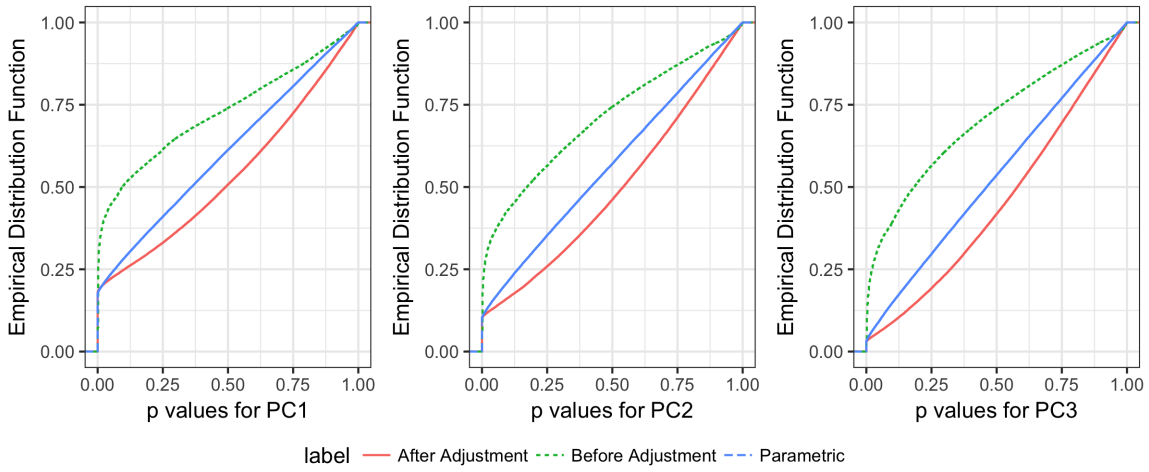

Figure 4.17: Empirical CDF of p-values for Top Three PCs, Realistic Simulation of dPCA Data

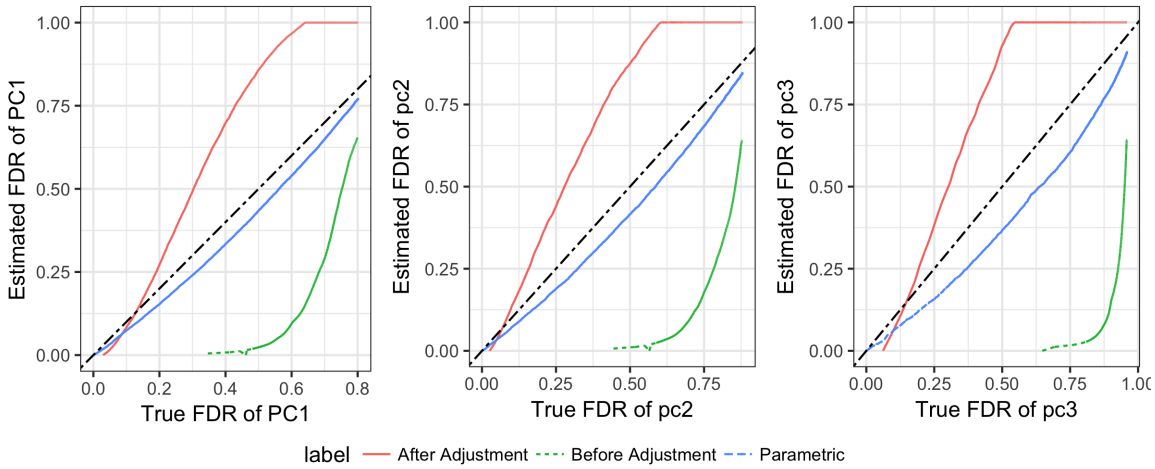

Figure 4.18: Estimated FDRs v.s. True FDRs for Top Three PCs, Realistic Simulation of dPCA Data

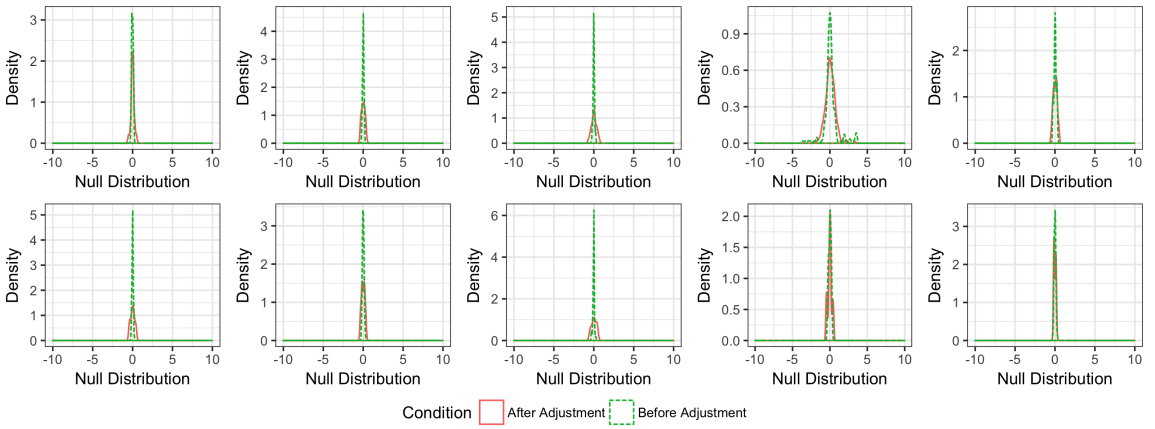

Figure 4.19: Comparison of Permutation Null Distributions before and after Adjustment, dPCA Data

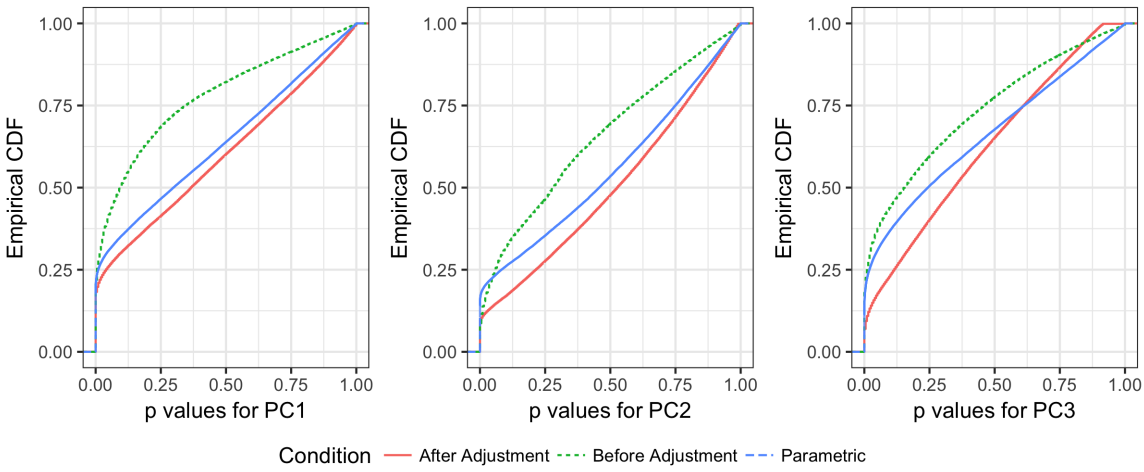

Figure 4.20: Empirical CDF of p-values for Top Three PCs, dPCA Data

Table 4.1: Parameter Settings for Simulation of dPCA data

Parameter Setting	dp_1	p_2	p_3	σ_1	σ_2	σ_3	σ_0
1	0.1	0.06	0.02	5	3	1	0.1
2	0.2	0.12	0.04	5	3	1	0.1
3	0.4	0.24	0.08	5	3	1	0.1
4	0.1	0.06	0.02	5	3	1	1
5	0.2	0.12	0.04	5	3	1	1
6	0.4	0.24	0.08	5	3	1	1
7	0.1	0.06	0.02	0.5	0.3	0.1	0.1
8	0.2	0.12	0.04	0.5	0.3	0.1	0.1
9	0.4	0.24	0.08	0.5	0.3	0.1	0.1

Chapter 5

Conclusion

In differential analysis of epigenomic data, our main interest is to detect differential signals between different biological conditions. To prioritize differential genomic loci, t-statistics are widely used. To calculate t-statistics, one needs to estimate variances. With small number of replicates in biology study, shrinkage estimator is commonly incorporated to stabilize the estimation of variance. Then moderated t-statistics are produced based on the shrinkage estimator. For statistical inference, null distributions need to be constructed for test statistics. When the assumption of normality of error term does not hold true, non-parametric tests, such as permutation test, are used instead. However, for differential genomic loci, permuting measurements between two groups will inflate the variance estimate, and thus form a misleading, overly centered null distribution. We proposed an adjustment approach by permuting residuals. We applied and evaluated this approach to simulation data, single type of microarray

data, and multiple types of epigenomic data.

In simulation studies, we assumed two different structures of noise: normal or t-distribution. Under the normality assumption, the parametric null distribution works well. However, when the normality assumption does not hold true, the parametric null distribution is no longer appropriate. In this case, permutation after adjustment is the only approach that provided a conservative decision about estimated FDRs. Naive permutation without adjustment was too optimistic under both noise structures.

In realistic simulation studies, the true noise structure in real data was maintained. The parametric null distribution was still overly centered. Based on parametric model, the estimated FDRs were smaller than the truth. This suggests that under true noise structure, the normality assumption of error term may be violated. Constructing null distribution in non-parametric way may be more appropriate than parametric model in real data. If we choose to conduct permutation, our adjustment procedure is necessary to prevent us from misleading, overly optimistic null distribution.

In real data, we observed a significant decrease in the number of differential genomic loci at the same FDR threshold when revised permutation is applied, comparing to the naive permutation or the parametric model. Combined with the previous realistic simulation results, we are confident that the permutation after adjustment provides a more conservative decision comparing to the other two approaches and decreases the number of false discoveries.

In simulation and realistic simulation studies, we investigated our method under

different parameter settings. We noticed that our adjustment procedure becomes more crucial when a large proportion of differential genomic loci are involved. In real data analysis, Melanoma dataset involves a large number of differential genomic loci. Our adjustment procedure therefore has a more profound impact comparing to the naive permutation. Therefore, our adjustment procedure is more suitable to apply to the datasets that involve a large number of differential loci.

Other similar non-parametric models such as bootstrapping can also be used to construct the null distribution. If other approaches are chosen, then we should apply a similar adjustment procedure to ensure that the estimation of variance of differential loci after permutation should be on the same scale with the estimation of variance of non-differential loci.

In summary, our adjustment procedure is valuable when incorporating shrinkage estimator of variance in permutation-based statistical test. In naive permutation, due to the inflation caused by shrinkage estimator of variance, the null distribution will be overly centered. Small p-values for non-differential loci based on this null distribution will result in large number of false positives. In our approach, we mitigate this inflation by removing the mean before permutation. This will help us to avoid the overly centered null distribution. Therefore, we will make a more conservative decision in terms of detecting differential signals between two biological conditions.

Bibliography

[1] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-dna interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[2] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg *et al.*, "Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss)," *Genome research*, vol. 16, no. 1, pp. 123–131, 2006.

[3] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, "Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin," *Genome research*, vol. 17, no. 6, pp. 877–885, 2007.

[4] H. Choi, A. I. Nesvizhskii, D. Ghosh, and Z. S. Qin, "Hierarchical hidden markov model with application to joint analysis of chip-chip and chip-seq data," *Bioinformatics*, vol. 25, no. 14, pp. 1715–1721, 2009.

[5] Y. Chen, C. A. Meyer, T. Liu, W. Li, J. S. Liu, and X. S. Liu, "Mm-chip enables integrative analysis of cross-platform and between-laboratory chip-chip or chip-seq data," *Genome biology*, vol. 12, no. 2, p. R11, 2011.

[6] G. Hon, B. Ren, and W. Wang, "Chromasig: a probabilistic approach to finding common chromatin signatures in the human genome," *PLoS Comput Biol*, vol. 4, no. 10, p. e1000201, 2008.

[7] H. Ji, X. Li, Q.-f. Wang, and Y. Ning, "Differential principal component analysis of chip-seq," *Proceedings of the National Academy of Sciences*, vol. 110, no. 17, pp. 6789–6794, 2013. [Online]. Available: http://www.pnas.org/content/110/17/6789.abstract

[8] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistica sinica*, pp. 31–46, 2002.

[9] G. Smyth, "Limma: linear models for microarray data," *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420, 2005.

[10] H. Ji and W. H. Wong, "Tilemap: create chromosomal map of tiling array hybridizations," *Bioinformatics*, vol. 21, no. 18, pp. 3629–3636, 2005.

[11] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National*

Academy of Sciences, vol. 98, no. 9, pp. 5116–5121, 2001. [Online]. Available: http://www.pnas.org/content/98/9/5116.abstract

[12] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005. [Online]. Available: http://www.pnas.org/content/102/43/15545.abstract

[13] R. Breitling, A. Amtmann, and P. Herzyk, "Iterative group analysis (iga): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments," *BMC Bioinformatics*, vol. 5, no. 1, p. 34, 2004. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-5-34

[14] G. K. Smyth *et al.*, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Stat Appl Genet Mol Biol*, vol. 3, no. 1, p. 3, 2004.

[15] R. W. Doerge and G. A. Churchill, "Permutation tests for multiple loci affecting a quantitative character," *Genetics*, vol. 142, no. 1, pp. 285–294, 1996.

[16] J. H. Zhao, D. Curtis, and P. C. Sham, "Model-free analysis and permutation tests for allelic associations," *Human heredity*, vol. 50, no. 2, pp. 133–139, 1999.

[17] Z. D. Zhang, J. Rozowsky, M. Snyder, J. Chang, and M. Gerstein, "Modeling chip sequencing in silico with applications," *PLoS Comput Biol*, vol. 4, no. 8, p. e1000158, 2008.

[18] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[19] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003. [Online]. Available: http://www.pnas.org/content/100/16/9440.abstract

[20] E. Whalen, *CLL: A Package for CLL Gene Expression Data*, 2017, r package version 1.16.0.

[21] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.

[22] T. J. Parmenter, M. Kleinschmidt, K. M. Kinross, S. T. Bond, J. Li, M. R. Kaadige, A. Rao, K. E. Sheppard, W. Hugo, G. M. Pupo *et al.*, "Response of braf-mutant melanoma to braf inhibition is mediated by a network of transcriptional regulators of glycolysis," *Cancer discovery*, vol. 4, no. 4, pp. 423–433, 2014.

[23] S. Davis and P. Meltzer, "Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor," *Bioinformatics*, vol. 14, pp. 1846–1847, 2007.

[24] L. Gautier, *affydata: Affymetrix Data for Demonstration Purpose*, 2016, r package version 1.22.0.

[25] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong, "An integrated software system for analyzing chip-chip and chip-seq data," *Nature biotechnology*, vol. 26, no. 11, pp. 1293–1300, 2008.

BIBLIOGRAPHY

Biographical Statement

Dan Jiang is a fifth year Ph.D. student in Bloomberg School of Public Health, Johns Hopkins University. She works with Dr. Hongkai Ji in Biostatistics department. Her research focuses on epigenomic differential analysis. Before graduation study, she earned her bachelor degree in Zhejiang University, Hangzhou China. She was majored in statistics.