Statistical Methods for Multi-Class Differential Gene Expression Detection

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xiting Cao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Baolin Wu Ph.D. and Wei Pan Ph.D.

November, 2011

Acknowledgements

I would like to thank all of those people who helped make this dissertation possible.

First, I wish to thank my advisor and mentor, Dr. Baolin Wu for all his guidance, encouragement, support, and patience. His sincere interests in statistics and education have been a great inspiration to me. Also, I would like to thank my co-advisor Dr. Wei Pan and the committee members Dr. Xiaotong Shen, and Dr. William Thomas for their very helpful insights, comments and suggestions at my proposal meeting. Additionally, I would like to acknowledge all the faculty members and staff at the division of Biostatistics to make my time here cherishable.

A special thanks to my parents, my fiancée, Junfeng Gao, and my friends for all their support during my time in graduate school.

Abstract

One of the major goals of microarray data analysis is to identify differentially expressed genes. In cancer studies, RNA is extracted from the tissue samples of cancer patients (case class) and healthy people (control class) to obtain the gene expression data and genes that are differentially expressed between case and control are identified to be candidate biomarkers which could undergo further studies. More often, we encounter situations where gene expression between more than two classes are being compared instead of the traditional case/control setup, e.g., multiple disease stages or different experimental conditions. In this dissertation, the problem of identifying differentially expressed genes in a multi-class comparison setting will be addressed.

To identify the differentially expressed genes, it is important to select a test statistic to rank the genes, and common approaches usually summarize each gene expression into a univariate test statistic and find a critical value for the ranking statistics to claim which gene is differentially expressed. In the dissertation, a univariate test statistic (the moderated F-statistics) is first used as a summary statistic and its distribution is empirically estimated using maximum likelihood. After that, A multivariate test statistic is proposed as a summary statistic for each gene and both parametric and non-parametric empirical Bayes approaches are adopted to rank the genes. The performances of the proposed methods are illustrated by extensive simulation studies and application to public microarray datasets. The results show that the proposed methods have better detection power than the commonly used approaches when controlling false discovery rates at the same level.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Microarray gene expression data has enabled researchers to simultaneously obtain quantitative information from hundreds of thousands of genes. One major task in microarray data analysis is detecting differentially expressed genes related to the change of conditions. For example, cancer studies have utilized microarray data, which measures the gene expression level derived from tumor tissues of a series of cancer patients, to select genes that are highly related to the disease. The expression patterns of truly differentially expressed genes are different among various tumor tissue types (e.g., tissues at various stages of development or disease states). Genes that show differential expressions are candidate biomarkers for further follow-up studies.

For significance testing of microarray data, many statistical methods have been proposed in the literature including the frequentist (1; 2; 3, e.g.), Bayesian (4; 5; 6; 7; 8, e.g.), and empirical Bayes approaches (9; 10; 11, e.g.) etc. Since the typical microarray experiment usually has small sample size for testing tens of thousands of genes simultaneously, false positives are very likely to happen just by chance. Many have proposed the shrunken estimation (1; 12; 13, e.g.) or empirical Bayes modeling approaches (14; 15, e.g.) to stabilize the variance

estimate and improve the gene selection power. Most existing methods have calculated a univariate summary statistic for each gene, which is then modeled for inference. For example, Dudoit *et al.* (2002) (2) proposed to use the t-statistic with a permutation test to rank genes. Tusher *et al.* (2001) (1) proposed the regularized t-statistic for gene ranking motivated by the commonly observed large variation associated with gene expressions. Efron (2003) (10) has proposed to model the distribution of ordinary or regularized t-statistics over all genes non-parametrically and make inference using the empirical Bayes approach.

Among the existing methods, the empirical Bayes approach (EB) has proven to be very useful for studying simultaneous significance testing problems commonly encountered in analyzing current large-scale microarray gene expression data and has been studied extensively. For example, Kendziorski *et al.* (2003) (11) and Smyth (2004) (15) have proposed the parametric EB approaches with different gene expression distribution assumptions. Efron (2001)(9) and Efron (2003, 2007) (10; 16) studied in detail the non-parametric EB approach, which models the univariate summary statistics (e.g., ordinary/regularized t/F-statistics) across genes, to borrow information for improving the individual gene inference and overall detection power.

One of the key components of EB is the information sharing across individual significance testing, which has proven to be quite useful and important for reducing the number of false positives. If we can increase the information shared across genes, we could further improve the detection power. It is shown that this information sharing sometimes could dramatically improve the overall detection power in the case of multi-class comparison problems, where gene expression of more than two groups are compared instead of the traditional case/control comparison. The intuitive idea is to model multivariate summary statistics instead of the commonly adopted approach of modeling univariate summary statistics. In this dissertation, several multivariate modeling methods are developed to implement

such an approach for large-scale significance testing of microarrays, which can also be applied to general large-scale simultaneous significance testing problems.

The dissertation is organized as follows. In Chapter 2, the gene expression data is summarized into a moderated F-statistic for each gene and maximum likelihood estimate is used to empirically estimate the null gene distribution. The performances of empirical null distribution and theoretical null distribution are compared through simulations and application to a lung study.

In Chapter 3, a multivariate t-statistic is proposed as a summary statistic for each gene and after a normal transformation, its distribution could be modeled with a multivariate normal mixture distribution. We adopt the empirical Bayes approach to rank the genes and compare its performance with the moderated F-statistics through simulations and application to a breast cancer data. A theoretical justification for the proposed method is also provided.

In Chapter 4, the same multivariate test statistic is summarized for each gene, and a non-parametric method is proposed to estimate both the null gene distribution and marginal distribution. Simulations and application to the breast cancer data is presented to show its performance compared with the moderated F-statistics.

The final chapter, Chapter 5 summarizes the approaches that have been proposed for detecting differentially expressed genes in the multi-class comparison situation. Strengths and weakness are discussed and some future improvements are outlined.

Chapter 2

Empirical null distribution modeling for multi-class differential gene expression detection

In detecting differentially expressed genes, the interaction between genes could influence the distribution of the test statistics for the genes and compromise the accuracy and efficiency of significant gene detection. For the traditional case/control comparison problems, Efron (16; 17; 18) has proposed to use empirical null distribution for modeling the two-sample t-statistic (and its variants, e.g., normally transformed) to partially incorporate the gene dependence (implemented in R package *locfdr* (18)).

In this chapter, the empirical null modeling approach is extended to multi-class differential gene expression detection. For significant gene selection, both empirical null distribution based p-value ranking and empirical Bayes ranking will be studied. Their performances will be illustrated through simulations and

application to public microarray data.

2.1 Statistical methods

Given a multi-class microarray data, denote x_{ijg} as the observed expression value for gene $i = 1, \cdots, m$ from sample $j = 1, \cdots, n_g$ of class $g = 1, \cdots, G$. Let $n = \sum_{g=1}^{G} n_g$ be the total number of samples. Most commonly used methods for testing differential expressions are based on variants of the F-statistics. For gene i, it is defined as

$$t_i = \frac{\sum_{g=1}^{G} n_g(\bar{x}_{ig} - \bar{x}_i)^2/(G-1)}{\hat{\sigma}_i^2},$$

where

$$\bar{x}_{ig} = \frac{\sum_{j=1}^{n_g} x_{ijg}}{n_g}, \quad g = 1, \cdots, G, \quad \bar{x}_i = \frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g} x_{ijg}}{n},$$

and $\hat{\sigma}_i^2$ is the estimated expression variance. A commonly used variance estimate is the sample variance $s_i^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}(x_{ijg} - \bar{x}_{ig})^2/(n-G)$. It has been widely accepted that for most microarray data, using some moderated/stabilized variance estimate instead of the sample variance estimates can often improve the differential gene expression detection power due to the relative small sample size compared to the large number of genes. Following (15), the expression variance for each gene is estimated with an empirical Bayes approach, which models the individual gene variance with an inverse chi-square prior distribution,

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_0}^2,$$

The posterior estimate of $\sigma_i^{-2}|s_i^2$ is the posterior mean, $\hat{\sigma}_i^{-2} = (n - G + d_0)/((n - G)s_i^2 + d_0 s_0^2)$, whose inverse is a weighted average of the prior information and the sample variance.

In the following discussion, the moderated F-statistics will be used as the test statistic for detecting differentially expressed genes. Here the prior degree of freedom d_0 and scale parameter s_0^2 are empirically estimated based on all genes.

Ideally when a gene is not differentially expressed, under normal distribution assumption, the moderated F-statistics will theoretically follow the F-distribution with $(G-1, n-G+d_0)$ degrees of freedom, which can be used to compute the p-value for each test statistic. Or a general permutation approach can also be used to derive p-values. Let $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(m)}$ be the ordered p-values. Benjamini and Hochberg (1995)(19) proposed the following false discovery rate (FDR) control procedure. For a given significance level $0 < \alpha < 1$, define

$$k_\alpha = \underset{i}{\operatorname{argmax}} \left\{ p_{(i)} \leqslant \frac{i}{m}\alpha \right\}.$$

When declaring all genes with p-values less than $p_{(k_\alpha)}$ as significant, FDR will be controlled at level α. An alternative approach is to directly estimate FDR (10). For example, when calling all genes with p-values less than $p_{(i)}$ as significant, the estimated FDR is

$$\widehat{\mathrm{FDR}} = m\hat{\theta}_0 \frac{p_{(i)}}{i},$$

where θ_0 is the proportion of true null genes, which needs to be estimated. With p-values available, the null gene proportion θ_0 can be estimated following the approach of (20), which fitted the p-value density with a convex decreasing density model. Comparing the FDR control and estimation approaches, the main difference lies in the use of the null gene proportion θ_0. In general, the estimation approach is more powerful than the control approach.

Since the commonly observed interactions among genes could influence the distribution of the test statistics across genes, the theoretical null distribution $F(G-1, n-G+d_0)$ in general does not provide an adequate fit for null genes. As shown in (17), the choice of null distribution is crucial in the estimation and control of FDR. With the large number of genes available, the null distribution can be empirically estimated to obtain a more accurate fit and provide appropriate FDR control.

2.1.1 Empirical modeling of null gene distribution

Figure 2.1 compares the theoretical and empirical null fitting of the moderated F-statistics for the gene expression data from a lung transplant study (21). The study measured the gene expressions of bronchoalveolar lavage cell samples from lung transplant recipients and the goal is to find the genes related to acute rejection. Biopsies were graded from 0 to 4 for "A" (perivascular inflammation) and "B" (lymphocytic bronchiolitis) scores, indicating an increase in the severity of acute rejection. The subjects are divided into three groups based on the sum of the A and B scores. In this case, the theoretical null distribution of the moderated F-statistics is the F-distribution with $(2, 22)$ degrees of freedom (estimated prior degree of freedom is $\hat{d}_0 = 1.0$). Assuming that those moderated F-statistics less than the sample median are coming from null genes, we could plot the histogram, quantile plot and empirical cumulative distribution function of the null moderated F-statistics, shown in Figure 2.1. In the left panel of Figure 2.1, the density of the theoretical null distribution shown in green solid curve has large deviation from the actual data, and provides a very poor fit. The red solid curve is the density of the estimated null distribution obtained using a maximum likelihood estimation approach which will be discussed later, and clearly fits the data much better. In the middle panel, the green solid curve is the quantile of theoretical null distribution against sample quantiles of the moderated F-statistics. The red solid curve is from empirical null distribution, and agrees with the diagonal line very well, indicating a better fit than the theoretical null distribution. In the right panel, we compare the empirical cumulative distribution function of moderated F-statistics against the c.d.f. of the theoretical null distribution (green solid line) and empirical null distribution (red solid line). Overall the empirical null distribution provides a much better fit to the data than the theoretical null distribution.

The maximum likelihood estimation of the empirical null distribution is based

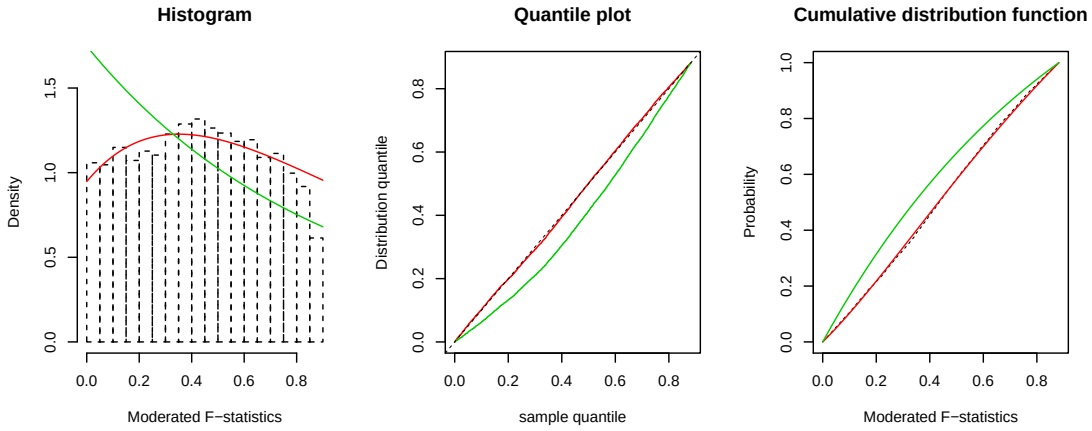

Figure 2.1: Theoretical and empirical null distribution comparison: black dotted lines are sample histogram, quantile plot, and empirical distribution function of moderated F-statistics less than the median; red lines are estimates from empirical null distribution fitting; and green lines are estimates from theoretical null distribution fitting.

on the assumption that all moderated F-statistics in a region near 0 come from the null genes, and the non-null density is supported outside that region. The assumption provides a reasonable approximation in most microarray data analysis, where researchers are only interested in detecting a relatively small number of differentially expressed genes, e.g. 10% among hundreds of thousands of candidate genes. Even if some truly significant genes have relatively small moderated F-statistics, they would not substantially affect the estimate of the null distribution.

The moderated F-statistics of the null genes are assumed to have the following density of a scaled F-distribution with $(G-1, n-G+d_0)$ degrees of freedom,

$$\frac{1}{\sigma_0} f(\frac{t}{\sigma_0}; G-1, n-G+d_0, \lambda_0),$$

where t is the moderated F-statistic, σ_0 is the scale parameter, and λ_0 is the non-centrality parameter.

Given a selected cutoff value C_0, define $\mathcal{I}_0 = \{i : t_i \leqslant C_0\}$ as the null gene set and let $N_0 = \sum_i I(t_i \leqslant C_0)$ be the corresponding number of null genes. Define,

$$H_0(\lambda_0, \sigma_0) = \Pr(t \leqslant C_0 | \text{null}) = F(\frac{C_0}{\sigma_0}; G-1, n-G, \lambda_0).$$

When assuming all moderated F-statistics less than C_0 are coming from null genes, we can compute the probability of a moderated F-statistic falls in the region of [0, C_0] as,

$$\theta = \Pr(t \leqslant C_0) = \theta_0 H_0(\lambda_0, \sigma_0).$$

Therefore the likelihood function can be written as follows, (non-null genes are those with moderated F-statistic larger than C_0 and are treated equally in a Binomial likelihood)

$$L(\lambda_0, \sigma_0, \theta_0) = \theta^{N_0}(1-\theta)^{m-N_0} \Big[\prod_{i \in \mathcal{I}_0} \frac{f(t_i/\sigma_0; G-1, n-G+d_0, \lambda_0)}{\sigma_0 H_0(\lambda_0, \sigma_0)} \Big]$$

For convenience of maximization, we did the following transformation,

$$\lambda_0 = \exp(\eta), \quad \sigma_0 = \exp(\tau), \quad \theta_0 = \frac{1}{1+\exp(-\phi)}.$$

With no constraint on (η, τ, ϕ), we can easily maximize the log likelihood numerically, which is done using the *optim* function in R.

For the lung transplant microarray data, choosing C_0 as the median of the moderated F-statistics, the parameters are estimated to be $\hat{\lambda}_0 = 3.40$, $\hat{\sigma}_0 = 0.39$ and $\hat{\theta}_0 = 0.99$. The estimated empirical null density fits the data very well as shown in Figure 2.1.

For large scale significant gene selection, a popular approach is using p-values to rank genes. P-values can be computed from the estimated empirical null distribution as follows:

$$p_i = 1 - F(t_i/\hat{\sigma}_0; G-1, n-G+d_0, \hat{\lambda}_0),$$

which can then be used to estimate FDR. We can similarly rank genes using p-values computed from the theoretical null distribution, and use the theoretical null based p-values to estimate FDR. Figure 2.2 (page 13) compares the estimated FDR using p-values computed from theoretical and empirical null distribution. When controlling FDR at 0.1, no gene is detected as significant if using theoretical null distribution, while 19 significant genes are identified using the empirical null distribution.

2.1.2 Empirical Bayes ranking with local false discovery rate

Alternatively we can use the empirical Bayes modeling approach to rank genes by local FDR (16). Suppose genes are either null or non-null with prior probability θ_0 and $\theta_1 = 1 - \theta_0$. Define h_0 as the density of the moderated F-statistics for null genes and h_1 for non-null genes. The marginal density of the moderated F-statistics is therefore $h = \theta_0 h_0 + \theta_1 h_1$. The local FDR is defined following (16; 18)

as:

$$fdr(t) = \Pr(\text{gene is null}|t) = \theta_0 \frac{h_0(t)}{h(t)},$$

which is the posterior probability of the gene being null given its moderated F-statistic according to the Bayes rule. h_0 can be estimated using empirical null distribution as described previously and h could be estimated non-parametrically given the large amount of genes. Following (16), Poisson regression model is adopted to estimate h based on all the moderated F-statistics $\{t_i\}_{i=1}^m$.

Divide the range of observations for t_i into K intervals with equal length (for moderated F-statistics, we typically choose the range as $[0, \max_i t_i]$), denote the sample count for each interval by $T_k = \#\{t_i \text{ in interval } k\}$, $k = 1, \cdots, K$ and fit T_k with a Poisson regression model,

$$T_k \sim Poisson(\lambda_k), \quad \log(\lambda_k) = \sum_{i=0}^{p} \beta_i b_i(x_k),$$

where x_k is the midpoint of interval k, $b_i(x_k)$ is the generated B-spline basis matrix for a natural cubic spline with p degrees of freedom, and λ_k is the expectation of T_k. The density at x_k can be estimated as

$$\hat{h}(x_k) = \frac{\lambda_k}{\sum_{j=1}^{K} \lambda_j} \frac{1}{\ell_k}, \quad \ell_k = \text{length of interval } k.$$

For general x, $\hat{h}(x)$ is estimated based on linear interpolation. The choice of K and p is not critical and the default value is $K = 120$ and $p = 7$ as implemented in R package *locfdr* (18).

The local FDR is estimated as

$$\widehat{fdr}(t) = \hat{\theta}_0 \frac{\hat{h}_0(t)}{\hat{h}(t)},$$

$\hat{\theta}_0$ and $\hat{h}_0$ are estimated based on the empirical null MLE fitting and $\hat{h}$ is the non-parametric estimate based on the Poisson regression model. Using local FDR as the ranking statistics, the overall FDR could be estimated using the Monte

Carlo approximation. Specifically we simulate a large set of null data based on the estimated empirical null distribution and then calculate the corresponding proportion of false positives to approximate the false positive rate in the observed data.

Given a cut off value c_0 for the local FDR, the overall FDR is estimated as

$$\widehat{\text{FDR}} = \frac{m\hat{\theta}_0 \sum_{b=1}^{B} I(\widehat{fdr}_b^0 < c_0)/B}{\sum_{i=1}^{m} I(\widehat{fdr}_i < c_0)}.$$

Here $\widehat{fdr}_b^0$ is the estimated local FDR for simulated null gene $b = 1, \cdots, B$. B is fixed at 10^6 in the simulation study and lung transplant data analysis.

2.2 Application to lung transplant data

Three methods discussed previously are applied to the lung transplant microarray data: (1) ranking genes with p-values derived from theoretical null distribution; (2) ranking genes with p-values derived from empirical null distribution and (3) ranking genes with local FDR. Figure 2.2 compares the estimated FDR for the three methods. Overall the empirical null distribution based p-value ranking performs the best. When controlling FDR at 0.1, both EB ranking and p-value ranking with empirical null identified the same set of 19 significant genes.

Table 2.1 shows the 19 genes sorted by their p-values computed from empirical null distribution. The last column is their ranking by local FDR. Some of the selected genes have been shown as associated with lung disease or human immune process, which is crucial in organ transplant allograft rejection. For example, EZR was found significantly associated with lung adenocarcinoma (22). STAT6 encodes a protein that plays a central role in exerting interleukin-4 (IL-4) responses and IL-4 is known associated with transplant allograft rejection (see 23; 24; 25, e.g.). PSMC3 regulates the class II transactivator, which is critical for

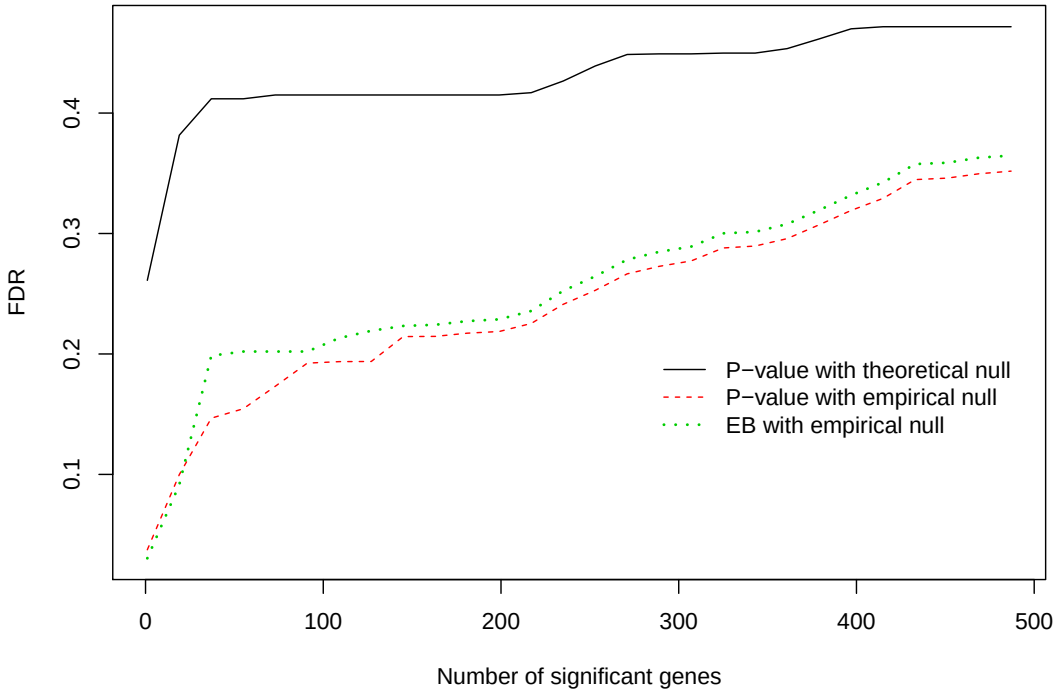

Figure 2.2: FDR estimates based on both empirical Bayes method with empirical null distribution and theoretical null distribution for lung transplant study.

initiation of adaptive immune responses (26). GSTA4 encodes an enzyme involved in cellular defense against toxic, carcinogenic, and pharmacologically active electrophilic compounds and was found significantly associated with regeneration of graft tissue after transplant (27). IGKC encodes a protein that combines with other proteins to produce an immunosuppressant for heart transplantation (28).

Gene	p-value	local FDR rank
EZR (ezrin)	2.1e-06	1
UBA52 (ubiquitin A-52 residue ribosomal protein fusion product 1)	3.8e-06	9
STAT6 (signal transducer and activator of transcription 6, interleukin-4 induced)	1.7e-05	16
NFS1 (NFS1 nitrogen fixation 1 homolog (S. cerevisiae))	2.1e-05	11
EIF3I (eukaryotic translation initiation factor 3, subunit I)	2.4e-05	8
RAD21 (RAD21 homolog (S. pombe))	2.6e-05	6
GINS1 (GINS complex subunit 1 (Psf1 homolog))	2.9e-05	2
RAE1 (RAE1 RNA export 1 homolog (S. pombe))	4.0e-05	3
PSMC3 (proteasome (prosome, macropain) 26S subunit, ATPase, 3)	4.1e-05	4
SELT (selenoprotein T)	4.2e-05	5
HCCS (holocytochrome c synthase (cytochrome c heme-lyase))	4.7e-05	7
NONO (non-POU domain containing, octamer-binding)	5.5e-05	10
UGT2B28 (UDP glucuronosyltransferase 2 family, polypeptide B28)	6.1e-05	12
CLINT1 (clathrin interactor 1)	6.7e-05	13
GSTA4 (glutathione S-transferase alpha 4)	7.4e-05	14
IGKC (immunoglobulin kappa constant)	7.6e-05	15
UBQLN2 (ubiquilin 2)	8.5e-05	17
CHRNA3 (cholinergic receptor, nicotinic, alpha 3)	9.0e-05	18
DAZAP2 (DAZ associated protein 2)	9.7e-05	19

Table 2.1: Top 19 genes ranked by p-value computed from empirical null distribution.

2.3 Simulation study

A simulation study is conducted to compare the performance of p-value ranking based on theoretical and empirical null distribution, and local FDR based ranking approach.

The F-statistics for $m = 10^4$ genes are simulated. 95% of them are assumed to be null genes, following a scaled non-central F-distribution density $\sigma_0^{-1} f(\sigma_0^{-1} t; 2, 40, \lambda_0)$, and the rest are differentially expressed genes, following $\sigma_0^{-1} f(\sigma_0^{-1} t; 2, 40, \lambda)$. The simulations are conducted for $\sigma_0 = (1, 2, 0.5)$, $\lambda_0 = (0, 0.5)$ and $\lambda = (5, 10)$ and results are averaged over 100 simulations.

Figure 2.3 compares the estimated FDR for all three methods, and their bias and variance are summarized in Table 2.2 and 2.3. When $\sigma_0 = 1$ and $\lambda_0 = 0$, the true null distribution is exactly the theoretical null distribution, using the theoretical null distribution performs best and estimate true FDR accurately. Because the other methods need to empirically estimate the null gene distribution, there are some sacrifices on the precision. However, when $\sigma_0 < 1$, indicating a deviance from the theoretical null distribution, using theoretical null distribution could significantly underestimate FDR, which gives very optimistic results, while for $\sigma_0 > 1$, the theoretical null yields very conservative estimates. The local FDR ranking performs comparable with p-value ranking based on empirical null distribution when $\lambda = 10$, indicating a larger expression difference between classes. With small λ, e.g. $\lambda = 5$, local FDR ranking could estimate true FDR better. However, because the marginal distribution needs to be non-parametrically estimated for the local FDR method, its true FDR is slightly larger than that of the p-value ranking with empirical null distribution.

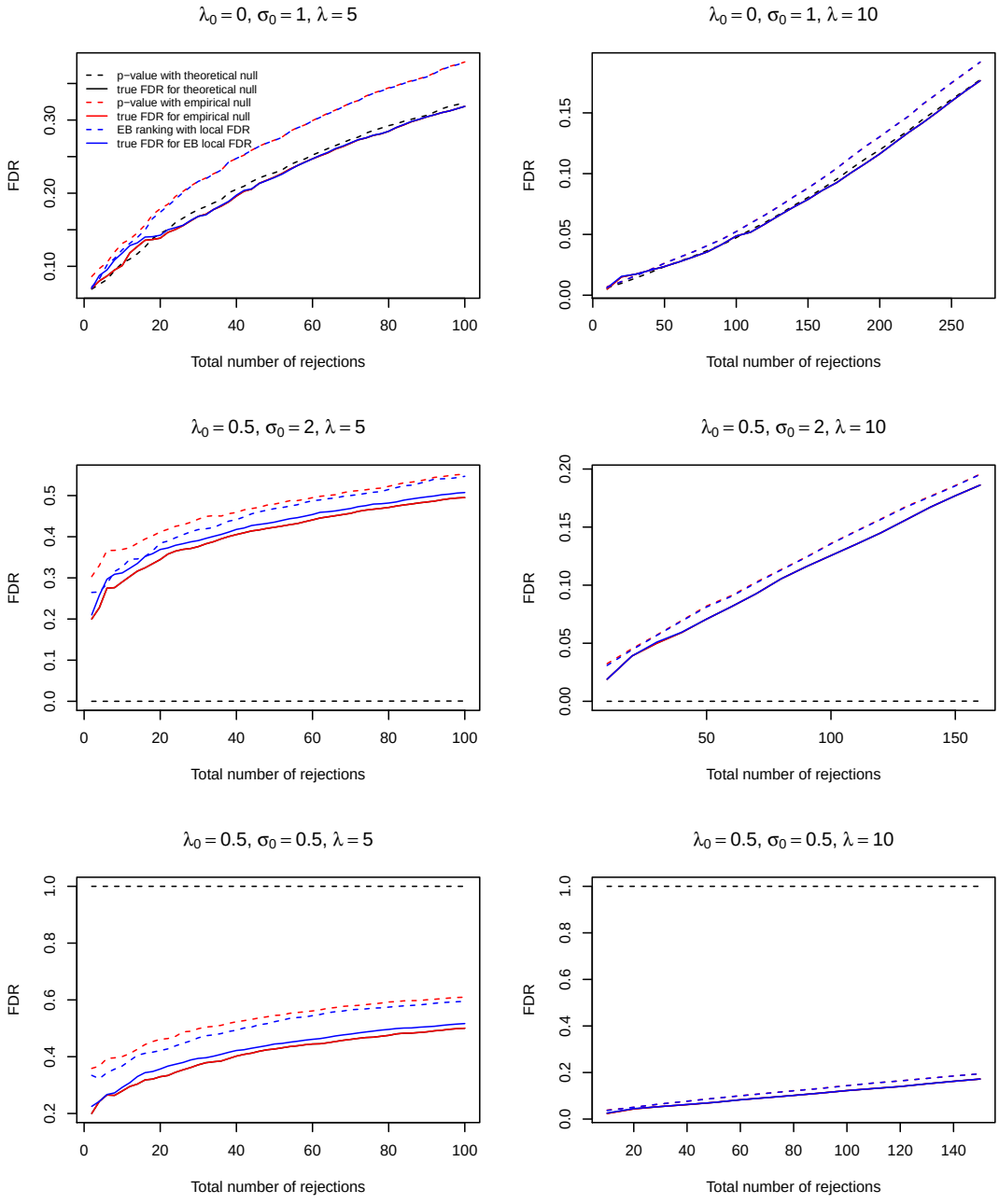

Figure 2.3: Estimated FDR for local FDR ranking, p-value ranking based on theoretical and empirical null distribution.

$\lambda_0 = 0$, $\sigma_0 = 1$, $\lambda = 5$									
	total rejection =20			total rejection=50			total rejection =100		
	true FDR	bias	variance	true FDR	bias	variance	true FDR	bias	variance
F-statistic with theoretical null	0.139	0.00644	0.00244	0.222	0.00591	0.00254	0.319	0.00557	0.00208
F-statistic with empirical null	0.139	0.0402	0.00497	0.222	0.0505	0.00474	0.319	0.0607	0.00456
local FDR with empirical null	0.143	0.0316	0.00475	0.221	0.0507	0.00474	0.319	0.0603	0.00458
$\lambda_0 = 0.5$, $\sigma_0 = 2$, $\lambda = 5$									
	total rejection =20			total rejection=50			total rejection =100		
	true FDR	bias	variance	true FDR	bias	variance	true FDR	bias	variance
F-statistic with theoretical null	0.344	-0.344	3.31e-09	0.423	-0.423	6.32e-09	0.495	-0.494	2.03e-08
F-statistic with empirical null	0.344	0.0672	0.0583	0.423	0.0563	0.0486	0.495	0.057	0.0461
local FDR with empirical null	0.369	0.0156	0.0487	0.435	0.0329	0.0456	0.507	0.0395	0.0457
$\lambda_0 = 0.5$, $\sigma_0 = 0.5$, $\lambda = 5$									
	total rejection =20			total rejection=50			total rejection =100		
	true FDR	bias	variance	true FDR	bias	variance	true FDR	bias	variance
F-statistic with theoretical null	0.33	0.67	0.802	0.427	0.573	0.203	0.5	0.5	0.0812
F-statistic with empirical null	0.33	0.132	0.0534	0.427	0.118	0.0429	0.5	0.109	0.0392
local FDR with empirical null	0.356	0.065	0.0409	0.445	0.0785	0.0369	0.517	0.0782	0.035

Table 2.2: Bias and variance of the estimated FDR for p-value ranking based on theoretical and empirical null distribution and local FDR ranking, $\lambda = 5$

$\lambda_0 = 0$, $\sigma_0 = 1$, $\lambda = 10$									
	total rejection =100			total rejection=150			total rejection =300		
	true FDR	bias	variance	true FDR	bias	variance	true FDR	bias	variance
F-statistic with theoretical null	0.0487	-0.00136	8.19e-05	0.0788	0.00162	0.000161	0.202	0.00183	0.00059
F-statistic with empirical null	0.0487	0.00368	0.000151	0.0788	0.00936	0.000318	0.202	0.0176	0.00132
local FDR with empirical null	0.0487	0.00364	0.000153	0.0788	0.00934	0.000311	0.202	0.0179	0.00132
$\lambda_0 = 0.5$, $\sigma_0 = 2$, $\lambda = 10$									
	total rejection =25			total rejection=70			total rejection =150		
	true FDR	bias	variance	true FDR	bias	variance	true FDR	bias	variance
F-statistic with theoretical null	0.0476	-0.0476	7.1e-12	0.0927	-0.0927	1.09e-10	0.177	-0.177	1.89e-09
F-statistic with empirical null	0.0476	0.00313	0.00109	0.0927	0.01	0.00303	0.177	0.00878	0.00672
local FDR with empirical null	0.0476	0.00275	0.00109	0.0927	0.00934	0.00303	0.177	0.00846	0.00668
$\lambda_0 = 0.5$, $\sigma_0 = 0.5$, $\lambda = 10$									
	total rejection =25			total rejection=70			total rejection =150		
	true FDR	bias	variance	true FDR	bias	variance	true FDR	bias	variance
F-statistic with theoretical null	0.0484	0.952	0.0734	0.0921	0.908	0.035	0.172	0.828	0.0291
F-statistic with empirical null	0.0484	0.00913	0.00158	0.0921	0.0186	0.00332	0.172	0.0224	0.00675
local FDR with empirical null	0.0472	0.011	0.00161	0.0921	0.0185	0.00337	0.172	0.0227	0.00675

Table 2.3: Bias and variance of the estimated FDR for p-value ranking based on theoretical and empirical null distribution and local FDR ranking, $\lambda = 10$

Chapter 3

Multi-class differential gene expression detection with finite multivariate normal mixture model

In the previous chapter, a univariate test statistic (moderated F-statistics) is used as the summary statistic for each gene, which is a common approach in multi-class differential gene detection situation. However, it is shown that summarizing gene expression data into some univariate test statistic leads to loss of information, so in this chapter, a multivariate test statistic is proposed for testing each hypothesis and a parametric multivariate modeling method that efficiently utilizes the information across multiple genes is developed to improve the overall significant gene detection power.

3.1 Statistical methods

Given a multi-class microarray data, denote x_{ijg} as the observed expression value (already preprocessed) for gene $i = 1, \cdots, m$ and sample $j = 1, \cdots, n_g$ of class $g = 1, \cdots, G$. Let $n = \sum_{g=1}^{G} n_g$. For gene i, assume expressions have class means μ_{ig} with overall mean μ_i and variance σ_i^2. Denote $\hat{\mu}_{ig}$, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ as their corresponding estimates, for example, sample means and variance, $\bar{x}_{ig} = \sum_j x_{ijg}/n_g$, $\bar{x}_i = \sum_g n_g \bar{x}_{ig}/n$ and $s_i^2 = \sum_g \sum_j (x_{ijg} - \bar{x}_{ig})^2/(n - G)$. It has long been recognized that variance estimation is often unstable across genes for microarray data due to small sample size, and it helps to use regularized variance estimate by borrowing information across different genes. Here the same hierarchical empirical Bayes modeling approach used in Chapter 2 is adopted to estimate a regularized variance (15).

Commonly used differential expression testing methods often construct a univariate statistic, denoted as t_i, based on $\hat{\sigma}_i, \hat{\mu}_{ig}, \hat{\mu}_i$ to summarize overall mean expression differences. Very often t_i is completely determined by the estimated variance $\hat{\sigma}_i^2$ and class mean differences $\hat{\mu}_{ig} - \hat{\mu}_i$, for example, $t_i = \hat{\sigma}_i^{-2} \sum_g n_g (\hat{\mu}_{ig} - \hat{\mu}_i)^2/(G-1)$. With the non-parametric empirical Bayes (EB) approach, the distribution of $\{t_i\}_{i=1}^m$ is modeled by a two-component mixture, $f(t) = \theta_0 f_0(t) + (1-\theta_0) f_1(t)$, where distributions f_0 and f_1 model those null and differentially expressed genes, and θ_0 is the proportion of null genes (see 10, e.g.). The marginal distribution f can be estimated based on $\{t_i\}_{i=1}^m$ using some non-parametric density estimation method, and the null distribution f_0 can be estimated based on the permutation approach. Here a parametric multivariate EB (MEB) method that directly models all class mean differences is developed and could potentially enable more information sharing across genes.

Define standardized class mean differences for gene i as

$$Z_i = \hat{\sigma}_i^{-1} (\hat{\mu}_{i1} - \hat{\mu}_i, \cdots, \hat{\mu}_{iG} - \hat{\mu}_i)^T, \tag{3.1}$$

where superscript T means matrix transpose. Here class mean differences are standardized by $\hat{\sigma}_i$ to make them more comparable across genes so that they could be modeled with some probability distributions. If $\hat{\mu}_{ig} - \hat{\mu}_i$ are linearly dependent (e.g., $\hat{\mu}_{ig}$ and $\hat{\mu}_i$ are sample mean estimates), those redundant components will be removed. The commonly used univariate test statistic t_i can often be completely determined by Z_i. Therefore in this sense Z_i contains more information than t_i and could also be used for testing differential expressions. We propose to model Z_i directly to borrow more information across genes and improve the overall detection power. Given tens of thousands of genes, the distribution of Z_i, $1 \leq i \leq m$, is modeled using the following two-component multivariate mixture:

$$h(Z) = \theta_0 h_0(Z) + (1 - \theta_0) h_1(Z), \quad 0 \leq \theta_0 \leq 1, \tag{3.2}$$

where distribution densities h_0 and h_1 model those null and differentially expressed genes, and θ_0 is interpreted as the proportion of null genes. The multivariate mixture model is motivated by the similar model used in previous univariate EB approach, and could be a very flexible modeling approach. Inference is drawn using the local false discovery rate (local FDR), $\theta_0 h_0(Z)/h(Z)$, (see 10; 18; 16; 19, e.g.). Note that for the purpose of ranking genes, it suffices to use the ratio $h_0(Z)/h(Z)$ only. In section 3.4, some theoretical justification is provided to show that this MEB approach of ranking genes with local FDR has some optimal properties.

3.1.1 Parametric MEB modeling with multivariate normal mixture distribution

Let us first consider one gene with expression values x_{jg} for sample $j = 1, \cdots, n_g$ of class $g = 1, \cdots, G$. Let $n = \sum_{g=1}^{G} n_g$. We assume that the gene has mean expression μ_g for class g and variance σ^2. Notice that the normal distribution

assumption $N(\mu_g, \sigma^2)$ is a special case. Now estimate μ_g by sample mean $\bar{x}_g$

$$\bar{x}_g = \sum_j x_{jg}/n_g,$$

we have $\mathbb{E}(\bar{x}_g) = \mu_g$ and $\mathrm{Var}(\bar{x}_g) = n_g^{-1}\sigma^2$. Denote $\hat{\sigma}^2$ as some (regularized) variance estimate. According to the central limit theorem, for a relatively large sample size n, the standardized class mean differences, $Z = \hat{\sigma}^{-1}(\bar{x}_2 - \bar{x}_1, \cdots, \bar{x}_G - \bar{x}_1)^T$, approximately follow a multivariate normal distribution. Since,

$$\mathrm{Var}(\bar{x}_g - \bar{x}_1) = \big(n_g^{-1} + n_1^{-1}\big)\sigma^2, \quad \mathrm{Cov}(\bar{x}_g - \bar{x}_1, \bar{x}_k - \bar{x}_1) = n_1^{-1}\sigma^2,$$

thus

$$\begin{aligned}
&\mathrm{Var}\big\{\hat{\sigma}^{-1}(\bar{x}_g - \bar{x}_1)\big\} \approx n_g^{-1} + n_1^{-1}, \\
&\mathrm{Cov}\big\{\hat{\sigma}^{-1}(\bar{x}_g - \bar{x}_1), \hat{\sigma}^{-1}(\bar{x}_k - \bar{x}_1)\big\} \approx n_1^{-1}, \quad g \neq 1, k \neq 1, k \neq g.
\end{aligned}$$

Therefore Z approximately has the (compound symmetry) covariance matrix

$$\Sigma = D^{-1} + n_1^{-1} J, \tag{3.3}$$

where $D = \mathrm{diag}\{n_2, \cdots, n_G\}$ is a $(G-1) \times (G-1)$ diagonal matrix, and J a $(G-1) \times (G-1)$ matrix with all elements equal to 1. It can be verified that

$$\Sigma^{-1} = D - n^{-1} DJD = D - n^{-1} N_G N_G^T,$$

where $N_G = (n_2, \cdots, n_G)^T$. Similar derivations apply to $\hat{\sigma}^{-1}(\bar{x}_2 - \bar{x}, \cdots, \bar{x}_G - \bar{x})^T$, with $\bar{x} = \sum_g n_g \bar{x}_g / n$ being the overall sample mean.

Consider the multi-class microarray data with observed expression values x_{ijg} for gene $i = 1, \cdots, m$ and sample $j = 1, \cdots, n_g$ from class $g = 1, \cdots, G$ and $n = \sum_{g=1}^{G} n_g$. Assume gene i follows a distribution with mean μ_{ig} for class g and variance σ_i^2. Denote the standardized class mean differences for gene i as

$$\Lambda_i = \sigma_i^{-1}(\mu_{i2} - \mu_{i1}, \cdots, \mu_{iG} - \mu_{i1})^T.$$

Given the observed values, Λ_i can be estimated by

$$Z_i = \hat{\sigma}_i^{-1}(\bar{x}_{i2} - \bar{x}_{i1}, \cdots, \bar{x}_{iG} - \bar{x}_{i1})^T.$$

When sample size is relatively small, the t-statistics usually have heavy tails. Every element of the Z statistics is normally transformed, and they still approximately have covariance matrix Σ and are denoted as Z in the following discussion for the convenience of notation.

The proposed method models Z_i using a hierarchical model approach to borrow information across genes by putting a probability distribution on Λ_i. Divide genes into two broad categories of null genes ($\|\Lambda_i\| = 0$) and differentially expressed genes ($\|\Lambda_i\| > 0$). The magnitude of differential expression might be different across genes, and thus Λ_i is modeled by a discrete distribution with one zero component and several nonzero components

$$\Pr(\Lambda_i = \Delta_k) = \theta_k, \quad \text{where } \Delta_k = (\delta_{k2}, \cdots, \delta_{kG})^T, \quad k = 1, \cdots, K,$$
$$\|\Delta_1\| = 0, \quad \|\Delta_{k>1}\| > 0, \quad \sum_{k=1}^{K} \theta_k = 1.$$

Here K is the total number of components of the mixture distribution. With relatively large sample size n, conditionally $Z_i|\Lambda_i$ approximately follows a multivariate normal distribution with mean Λ_i and variance Σ (3.3). Marginally we have the following finite multivariate normal mixture model

$$Z_i \sim \sum_{k=1}^{K} \theta_k N(\Delta_k, \Sigma),$$

which can be used to calculate the MEB statistic. Next an EM algorithm is developed to estimate the multivariate normal mixture model.

3.1.2 Computational algorithm development

Define indicators for gene i, $w_{ik} \in \{0, 1\}$, $k = 1, \cdots, K$, which are assumed to follow the multinomial distribution, $\Pr(w_{ik} = 1) = \theta_k$, $\sum_{k=1}^{K} w_{ik} = 1$. Conditionally

$Z_i|w_{ik} = 1 \sim N(\Delta_k, \Sigma)$. In the E-step, the conditional expectations can be shown to be

$$\theta_{ik} = \Pr(w_{ik} = 1|Z_i) = \frac{\theta_k \Phi(Z_i - \Delta_k, \Sigma)}{\sum_{l=1}^{K} \theta_l \Phi(Z_i - \Delta_l, \Sigma)}, \quad \text{where } \Phi(\Delta, \Sigma) = \frac{\exp(-\frac{1}{2}\Delta^T \Sigma^{-1} \Delta)}{(2\pi)^{(G-1)/2}|\Sigma|^{1/2}}.$$

In the M-step, the updates in each iteration are:

$$\theta_k = \frac{\sum_{i=1}^{m} \theta_{ik}}{m}, \quad k = 1, \cdots, K; \quad \Delta_k = \frac{\sum_{i=1}^{m} \theta_{ik} Z_i}{\sum_{i=1}^{m} \theta_{ik}}, \quad k > 1.$$

Iterating the E/M-steps, the maximum likelihood estimation can be obtained. The EM algorithm is repeated multiple times with random starting points to obtain good estimates and the number of components K can be inferred using BIC.

3.1.3 Empirical null distribution for dependence modeling

The gene dependence could compromise the accuracy and efficiency of the estimation procedure and change the distribution of summary statistics across genes. Efron (16; 17; 18) proposed to use empirical null distribution for one-dimensional statistics to partially incorporate the gene dependence (implemented in R package *locfdr* (18)). We propose to estimate empirical null distribution for the proposed parametric MEB based on central matching. The basic idea is to fit/approximate the center of the overall mixture of multivariate normal distribution empirically with one multivariate normal distribution.

Assuming the null distribution is a normal distribution with mean μ_0 and covariance matrix $\sigma_0^2\Sigma$,

$$\begin{aligned} \log(\theta_0 h_0(z)) &= \log\theta_0 - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\sigma_0^2\Sigma| - \frac{1}{2\sigma_0^2}(Z-\mu_0)^T\Sigma^{-1}(Z-\mu_0) \\ &= \log\theta_0 - \frac{1}{2}(n\log 2\pi + \log|\sigma_0^2\Sigma| - \frac{\mu_0^T\Sigma^{-1}\mu_0}{\sigma_0^2}) + \frac{\mu_0^T\Sigma^{-1}}{\sigma_0^2}Z - \frac{Z^T\Sigma^{-1}Z}{2\sigma_0^2}. \end{aligned}$$

By matching the coefficients of the linear and quadratic term in the above equation with the first and second derivative of $\log\hat{h}(Z)$ at $Z = 0$, $\hat{\sigma}_0$ and $\hat{\mu}_0$ can be solved.

Since the Hessian matrix of $\log h(Z)$ at $Z = 0$ is usually not proportional to Σ^{-1}, then σ_0 is estimated by minimizing the quadratic difference between $\sigma_0^{-2}\Sigma^{-1}$ and $-g_2$. Denote $g_1 = \frac{\partial}{\partial Z}\log h(Z)|_{Z=0}$ and $g_2 = \frac{\partial^2}{\partial Z \partial Z^T}\log h(Z)|_{Z=0}$, we have

$$\hat{\sigma}_0 = \text{argmin}_\sigma ||\sigma^{-2}\Sigma^{-1} + g_2||_2, \quad \hat{\mu}_0 = \hat{\sigma}_0^2 \Sigma g_1. \tag{3.4}$$

The estimated empirical null distribution is then a multivariate normal distribution, $\hat{h}_0(Z) = N(\hat{\mu}_0, \hat{\sigma}_0^2\Sigma)$ and θ_0 is estimated as $\hat{h}(0)/\hat{h}_0(0)$.

Local FDR is used as a ranking statistic and is estimated as,

$$\widehat{fdr}(Z_i) = \hat{\theta}_0 \frac{\hat{h}_0(Z_i)}{\hat{h}(Z_i)}.$$

If all genes with local FDR less than a cut off value c_0 are declared as significant, the overall FDR could be estimated through Monte Carlo approximation. Specifically B summary statistics from the estimated empirical null distribution $\hat{h}_0$ are simulated, and their corresponding local FDR are computed, denoted as $\widehat{fdr}_b^0$ for $b = 1, \cdots, B$. The overall FDR is then estimated as

$$\widehat{FDR} = \frac{\hat{\theta}_0 m \sum_{b=1}^{B} I(\widehat{fdr}_b^0 < c_0)/B}{\sum_{i=1}^{m} I(\widehat{fdr}_i < c_0)}.$$

In the simulation study and applications, B is chosen to be 10^6.

3.2 Simulation study

In the simulation study, a total number of 10^4 genes in three groups are generated. Each group has n observations. For each gene, the variance is generated from a χ^2 distribution with 10 degree of freedom. Assuming dependent data structure, the genes are divided into 50 blocks, each having 200 genes. Within each block, the genes are correlated according to a compound symmetry covariance matrix with correlation ρ. Suppose $\theta_0 \times 100\%$ genes are null genes. For the non-null genes, the

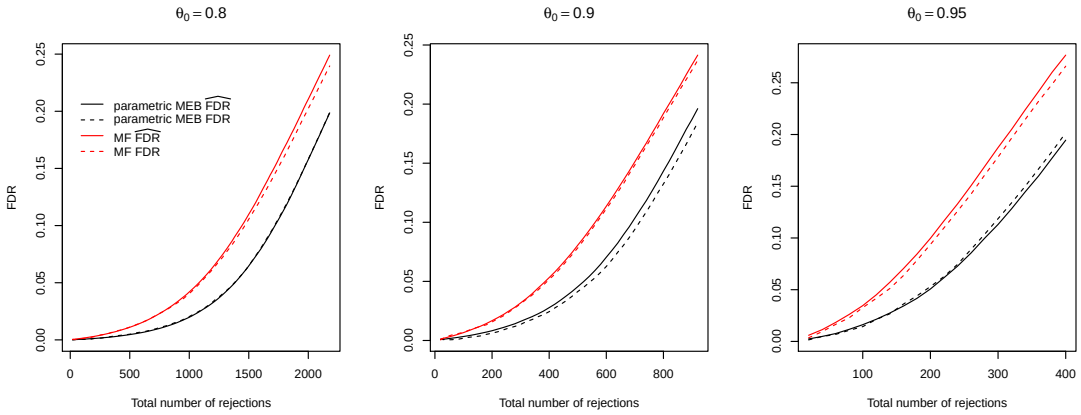

Figure 3.1: estimated FDR and true FDR based on EB parametric method and moderated F-statistic, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95 from left to right.

standardized expression for gene i, $(\mu_{1i}, \mu_{2i}, \mu_{3i})/\sigma_i$ is generated from a mixture of $(0, 0.5, 1)$ and $(0, -1, -0.5)$ with equal probability.

The simulations are done under settings $n = 10, 20$, $\rho = 0.25, 0.5, 0.75$ and $\theta_0 = 0.8, 0.9, 0.95$ and are repeated for 100 times. The proposed method (denoted as parametric MEB) is compared with the moderated F-statistics (denoted as MF). For each method, both the true FDR and estimated FDR are calculated to compare the performances. Figure 3.1 shows the estimated FDR and true FDR under $n = 20$, $\rho = 0.5$ by both methods. The estimated FDR could approximate the true FDR very well for both methods, but parametric MEB has higher power than MF at a certain FDR level. Under the same situation, Table 3.1 compares the average number of true positives detected by both methods at FDR level (0.01, 0.05, 0.1). Results under other simulation settings are included in the Appendix A.1.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	484	1045	1315	137	370	512	38	129	190
parametric MEB	753	1329	1555	237	506	639	80	198	269

Table 3.1: Number of true positives when fixing FDR by both moderated F-statistics and parametric MEB method, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8, 0.9, 0.95$.

3.3 Application to public microarray data

The breast cancer microarray data (29) is analyzed, which consists of 49 tumor samples with 7129 genes measured. The tumors are characterized by estrogen receptor (ER) status (negative/positive), and lymph node (LN) status (negative/positive), and are divided into 13 ER+/LN+ tumor samples, 11 ER-/LN+ tumor samples, 12 ER+/LN- tumor samples and 13 ER-/LN- tumor samples. The data is normalized using quantile normalization (30), and then log transformed for follow-up statistical analysis. The ER-/LN- group is set as the reference group. Both the proposed method (denoted as parametric MEB) and moderated F-statistic (denoted as MF) were applied to the data. P-values for MF statistics are derived from permutations. The proposed method yields an estimate of 0.90 for the null gene proportion θ_0 while MF based method gives 0.77. The marginal distribution of the Z statistics is estimated as a multivariate normal mixture model with 7 components, which is selected based on BIC criterion. Fixing FDR at different levels, Table 3.2 summarizes the number of significant genes detected by both methods. Figure 3.2 compares their estimated FDR and corresponding number of significant genes. Overall the proposed method detected more significant genes.

The identified significant genes (controlling FDR$\leq$0.05) by both methods are

FDR	0.001	0.01	0.05	0.1	0.2
parametric MEB	35	117	254	422	687
MF	16	50	183	275	517

Table 3.2: Number of detected significant genes at given FDR for breast cancer microarray data: the proposed method (denoted as parametric MEB) versus the moderated F-statistic (denoted as MF)

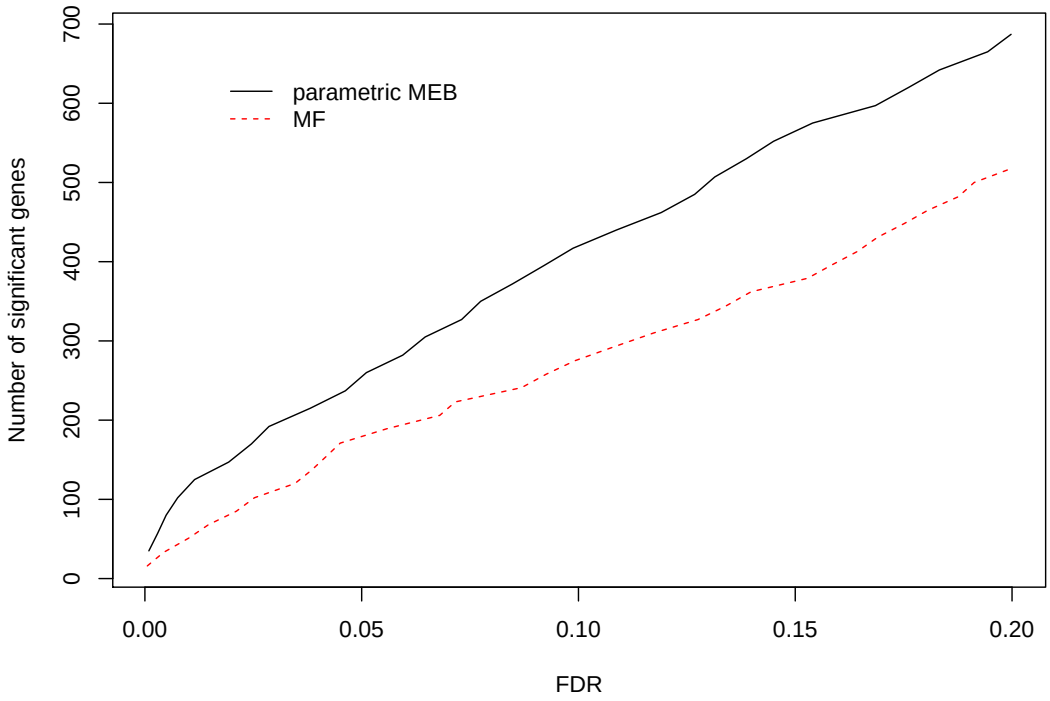

Figure 3.2: Detection power comparison of the proposed method (denoted as parametric MEB) versus the moderated F-statistic (denoted as MF) for the breast cancer microarray data: the y-axis is the number of detected significant genes, the x-axis is the estimated FDR.

then submitted to the online tool DAVID (Database for Annotation, Visualization and Integrated Discovery (31; 32)) to search for enriched Gene Ontology (GO) biological process (33). We compare the significant GO enrichment at FDR=0.01.

For the top 254 significant genes selected by the proposed parametric MEB, there are 14 significantly enriched biological processes: response to organic substance, response to hormone stimulus, response to endogenous stimulus, negative regulation of programmed cell death, negative regulation of cell death, negative regulation of apoptosis, response to steroid hormone stimulus, response to wounding, response to estrogen stimulus, immune response, defense response, regulation of cell proliferation, cell cycle, mitotic cell cycle. Many of them are closely related to cancer development. For example, apoptosis is the process of programmed cell death, and defective cell dealth processes have been implicated in an extensive variety of diseases including cancer (34). The cell cycle and cell proliferation are known to play important roles in cancer development. Cancer is a disease of inappropriate cell proliferation, and cell cycle machinery controls cell proliferation (35). The estrogen stimulus is closely associated with breast cancer: it causes the cells to divide and pushes the breast cancer forward. For the top 183 significant genes selected by MF, there is no significantly enriched biological process identified.

3.4 Theoretical justification for multivariate modeling approach

A theoretical justification is provided for the proposed MEB to show that it generally has more power than the univariate EB approach.

Optimality Lemma. Given the observed data vector $\boldsymbol{X}_j$ for hypothesis test $j = 1, \cdots, m$, suppose we can construct a multivariate statistic Z_j based on $\boldsymbol{X}_j$

for hypothesis j, and model $\{Z_j\}_{j=1}^m$ marginally with the following multivariate mixture distribution

$$h(Z) = \pi_0 h_0(Z) + (1-\pi_0)h_1(Z).$$

Here assuming both $h_0(Z)$ and $h_1(Z)$ are positive distribution density functions, they model those true null and truly significant hypotheses respectively, and π_0 is the overall proportion of true null hypotheses. Suppose that we want to construct statistics for testing each hypothesis based on the multivariate statistics $\{Z_j\}_{j=1}^m$. Given a test statistic $\Gamma(Z)$, where $\Gamma(\cdot)$ is some measurable function, without loss of generality, denote its rejection region as $\{Z : \Gamma(Z) \geq t_0\}$. Define the test power and Type I error as the average number of true/false positives (TPN/FPN)

$$\text{TPN} = \int_{\Gamma(Z)\geq t_0} m(1-\pi_0)h_1(Z)dZ, \quad \text{FPN} = \int_{\Gamma(Z)\geq t_0} m\pi_0 h_0(Z)dZ.$$

The multivariate empirical Bayes method (MEB) with its rejection region defined as

$$R = \left\{ Z : \frac{h_1(Z)}{h_0(Z)} \geq r_0 \right\},$$

has the maximum power among all testing methods with their rejection regions being of the form $\{Z : \Gamma(Z) \geq t_0\}$, and Type I errors equal to $\int_R m\pi_0 h_0(Z)dZ$.

Proof Given any test method with its rejection region being $A = \{Z : \Gamma(Z) \geq t_0\}$, assume

$$\int_A m\pi_0 h_0(Z)dZ = \int_R m\pi_0 h_0(Z)dZ.$$

Denote $A \cap R = B$, $A \setminus B = A_1$, $R \setminus B = R_1$. We then have

$$\int_{A_1} h_0(Z)dZ = \int_{R_1} h_0(Z)dZ.$$

According to the definition of R, we have

$$h_1(Z)/h_0(Z) < r_0, \quad \forall Z \in A_1; \quad h_1(Z)/h_0(Z) \geq r_0, \quad \forall Z \in R_1.$$

Therefore

$$\int_{A_1} h_1(Z)dZ \leq \int_{A_1} r_0 h_0(Z)dZ = \int_{R_1} r_0 h_0(Z)dZ \leq \int_{R_1} h_1(Z)dZ.$$

Since $A = B \cup A_1$ and $R = B \cup R_1$, we have

$$\int_{A} h_1(Z)dZ \leq \int_{R} h_1(Z)dZ.$$

The lemma is proven.

For very large-scale simultaneous significance testing problems, FDR approximately equals to FPN/(FPN+TPN). So the proposed MEB could also maximize TPN while controlling FDR. The previous optimality results are related to the Neyman-Pearson lemma (36), which says that for traditional single hypothesis testing, the likelihood ratio statistic generally has the optimal power for any given significance level (i.e., Type I error). The MEB rejection region R would correspond to the likelihood ratio statistic. Note that the optimality results do not require the independence of $\{Z_j\}_{j=1}^m$. Very often they are correlated due to information sharing across different tests. It is easily verified that the MEB rejection region could be equivalently based on ratios h_1/h or h_0/h.

Chapter 4

Non-parametric empirical Bayes modeling of multi-class differential gene expression detection

Chapter 3 discusses a parametric multivariate normal mixture model for estimating the marginal and empirical null density. However, there are situations where the data is not normally distributed or not continuous, where parametric approach might not be feasible. In this chapter, a general non-parametric multivariate empirical Bayes (MEB) approach is proposed to estimate the densities and local FDR for large-scale significance analysis.

4.1 Statistical methods

The same multivariate test statistic discussed in Chapter 3 is adopted again to summarize the expression differences for gene $i = 1, \cdots, m$:

$$T_i = (t_{i2}, \cdots, t_{iG})^T, \quad t_{ig} = \frac{\bar{x}_{ig} - \bar{x}_{i1}}{\hat{\sigma}_i}, \tag{4.1}$$

where $\bar{x}_{ig}$ and $\bar{x}_{i1}$ are the sample means in class g and class 1 for gene i, n_g and n_1 are the sample sizes in class g and class 1.

For inference, the local FDR (16; 18) is used to rank genes:

$$fdr(T) = \frac{\theta_0 f_0(T)}{f(T)},$$

where θ_0 is the proportion of null genes, f_0 models the distribution of the test statistics for null genes, and f is the marginal distribution of the test statistics.

The marginal distribution $f(T)$ will be non-parametrically estimated based on $\{T_i\}_{i=1}^m$, and the null distribution $f_0(T)$ will be estimated with the permutation approach. Using balanced permutation, the gene dependence is naturally taken into account in the null distribution estimation (18). In the following a Poisson regression model for non-parametric estimation and inference will be discussed.

4.1.1 Non-parametric multivariate density estimation with de-correlation and Poisson regression

First note that the individual components of T_i are correlated with each other. It can be verified that $\mathrm{Cov}(t_{ig}, t_{ik}) \approx n_1^{-1}$, $\forall k \neq g$. Generally it is not easy to accurately estimate the multivariate densities $f_0(T)$ and $f_1(T)$ non-parametrically due to the curse of dimensionality. The strategy is to first transform T_i into approximately independent components by de-correlation, and then estimate the density of each individual component separately. Since the gene dependence could

further change the dependence of the test statistics, the correlation structure is empirically estimated through permutation.

For each random permutation, the test statistics for all genes are recomputed each time (preserving the gene correlation structure). Denote the permutation statistics as $\{T_i^b\}$ for gene $i = 1, \cdots, m$ and permutation $b = 1, \cdots, B$. The covariance matrix is estimated as:

$$\hat{R} = \frac{\sum_{i=1}^{m}\sum_{b=1}^{B}(T_i^b - \bar{T}_0)(T_i^b - \bar{T}_0)^T}{mB - 1}, \quad \bar{T}_0 = \frac{\sum_{i=1}^{m}\sum_{b=1}^{B} T_i^b}{mB}.$$

Define $Y_i = (y_{i2}, \cdots, y_{iG})^T = \hat{R}^{-1/2}T_i$ and $Y_i^b = (y_{i2}^b, \cdots, y_{iG}^b)^T = \hat{R}^{-1/2}T_i^b$, so they have approximately independent components. Denote their densities as $h(Y)$ and $h_0(Y)$, which are the products of their individual components' densities:

$$h(Y_i) = \prod_{g=2}^{G} h_g(y_{ig}), \quad h_0(Y_i) = \prod_{g=2}^{G} h_{0g}(y_{ig}).$$

Therefore,

$$f_0(T) = |\hat{R}^{-1/2}|h_0(\hat{R}^{-1/2}T), \quad f(T) = |\hat{R}^{-1/2}|h(\hat{R}^{-1/2}T).$$

For gene i, the local FDR is computed as:

$$fdr(T_i) = fdr(Y_i) = \frac{\theta_0 \prod_{g=2}^{G} h_{0g}(y_{ig})}{\prod_{g=2}^{G} h_g(y_{ig})}. \tag{4.2}$$

In the following, the same Poisson regression model as described in Chapter 2 will be used to estimate individual component's densities h_g and h_{0g} based on Y_i and Y_i^b.

Divide the range of observations for y_{ig} into K intervals with equal length and denote the sample count for each interval by $N_k = \#\{y_{ig} \text{ in interval } k\}$, $k = 1, \cdots, K$, then fit N_k with a Poisson regression model,

$$N_k \sim Poisson(\lambda_k),$$
$$\log(\lambda_k) = \sum_{i=0}^{p} \beta_i b_i(x_k),$$

where x_k is the midpoint of interval k, $b_i(x_k)$ is the generated B-spline basis matrix for a natural cubic spline with p degrees of freedom, and λ_k is the expectation of N_k. The density at x_k can be estimated as

$$\hat{f}_g(x_k) = \frac{\lambda_k}{\sum_{j=1}^{K} \lambda_j} \frac{1}{\ell_k}, \quad \ell_k = \text{length of interval } k.$$

For general x, $\hat{h}_g(x)$ is estimated based on linear interpolation. Similarly $\hat{h}_{0g}(x)$ can be computed based on Y_i^b. As shown in (18), the choice of K and p is not critical and the default is $K = 120$ and $p = 7$ as implemented in R package *locfdr* (18).

4.1.2 Local FDR estimation

In order to compute local FDR, the null gene proportion θ_0 needs to be estimated. Here a central matching idea is adopted. Specifically, first identify the mode of the estimated null density $\hat{h}_{0g}(x)$, denoted as w_g, and then conservatively estimate null gene proportion as

$$\hat{\theta}_0 = \prod_{g=2}^{G} \frac{\hat{h}_g(w_g)}{\hat{h}_{0g}(w_g)}.$$

Then compute local FDR for all genes and permuted data as

$$fdr_i = \frac{\hat{\theta}_0 \prod_{g=2}^{G} \hat{h}_{0g}(y_{ig})}{\prod_{g=2}^{G} \hat{h}_g(y_{ig})}, \quad fdr_i^b = \frac{\hat{\theta}_0 \prod_{g=2}^{G} \hat{h}_{0g}(y_{ig}^b)}{\prod_{g=2}^{G} \hat{h}_g(y_{ig}^b)}.$$

When calling genes with $fdr_i \leq c_0$ as significant, the corresponding global FDR can be estimated as

$$\widehat{\text{FDR}} = \frac{\hat{\theta}_0 \sum_{b=1}^{B} \sum_{i=1}^{m} I(fdr_i^b \leq c_0)}{B \sum_{i=1}^{m} I(fdr_i \leq c_0)}.$$

In the next section, simulation studies are conducted to evaluate the performance of the proposed method.

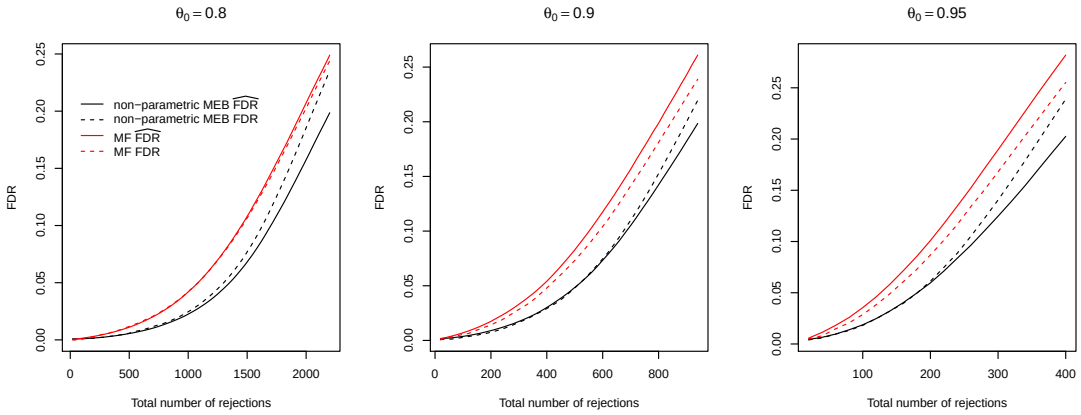

Figure 4.1: estimated FDR and true FDR based on non-parametric MEB method and moderated F statistic, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8, 0.9, 0.95$ from left to right.

4.2 Simulation study

In the simulation study, the proposed method (denoted as non-parametric MEB) is compared to a representative of traditional statistical approach, the moderated F-statistics (denoted as MF; 15). The same simulation settings as the previous chapter are used and here the simulation result for $n = 20$ and $\rho = 0.5$ is reported with 100 permutations. More simulation results are provided in Appendix A.2.

Figure 4.1 shows the total number of rejections versus the estimated FDR over 100 simulations for both methods. Overall we can see that the proposed approach performs reasonably well for estimating FDR and has larger power especially for small FDR value. Table 4.1 compares the detected average number of true positives at FDR=(0.01,0.05,0.1) for the non-parametric MEB and MF. In general we can see that the non-parametric MEB has more power than MF.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	471	1033	1312	137	369	507	40	126	185
non-parametric MEB	690	1281	1514	220	486	619	66	176	240

Table 4.1: Number of true positives when fixing FDR at 0.01, 0.05, 0.1 by both moderated F-statistics and non-parametric MEB method, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

Next the same public breast cancer microarray data is analyzed to empirically compare the detection power of the proposed method and the moderated F-statistics.

4.3 Application to breast cancer microarray data

The breast cancer microarray data studied in previous chapter (29) is revisited. Recall each sample has two binary outcomes describing the status of lymph node involvement in breast cancer (negative and positive, denoted as LN-/LN+), and estrogen receptor status (negative and positive, denoted as ER-/ER+). In total there are 49 samples divided into four groups: 13 ER+/LN+, 12 ER-/LN+, 12 ER+/LN- and 12 ER-/LN-.

Both the non-parametric MEB and moderated F-statistics are applied to detect differentially expressed genes among the four groups. P-values for MF statistics are derived from permutations. Table 4.2 compares their detected number of significant genes at FDR=(0.001,0.01,0.05,0.1,0.2). Figure 4.2 shows the complete results for FDR in the range of [0,0.2]. In general we can see that the proposed

FDR	0.001	0.01	0.05	0.1	0.2
non-parametric MEB	30	121	289	467	867
MF	16	50	183	275	517

Table 4.2: Number of detected significant genes at given FDR for breast cancer microarray data: the proposed method (denoted as non-parametric MEB) versus the moderated F-statistics (denoted as MF)

non-parametric MEB identified more significant genes than MF. The identified significant genes by both methods controlling FDR$\leq$0.05 are submitted to the online tool, DAVID (Database for Annotation, Visualization and Integrated Discovery, 31; 32), to search for enriched Gene Ontology (GO) biological process (33). We compare the significant GO enrichment at FDR=0.01. For the top 289 significant genes selected by the non-parametric MEB, there are 13 significantly enriched biological processes: negative regulation of apoptosis, negative regulation of programmed cell death, regulation of apoptosis, regulation of programmed cell death, mitotic cell cycle, response to stimulus, inflammatory response, response to wounding, response to chemical stimulus, cell cycle phase, response to stress, defense response, cell cycle. Many of them are closely related to cancer development. For example, cell cycle, cell death and apoptosis are all known to play important roles in cancer development. Cell cycle machinery controls cell proliferation while cancer is a disease of inappropriate cell proliferation (35). Apoptosis is the process of programmed cell death. Defective cell death processes have been implicated in cancer (37). For the top 183 significant genes selected by MF, there are no significantly enriched biological process identified. In addition, the top 289 genes ranked by MF are also submitted for analysis of the GO enrichment no significantly enriched biological process at FDR=0.01 is found either.

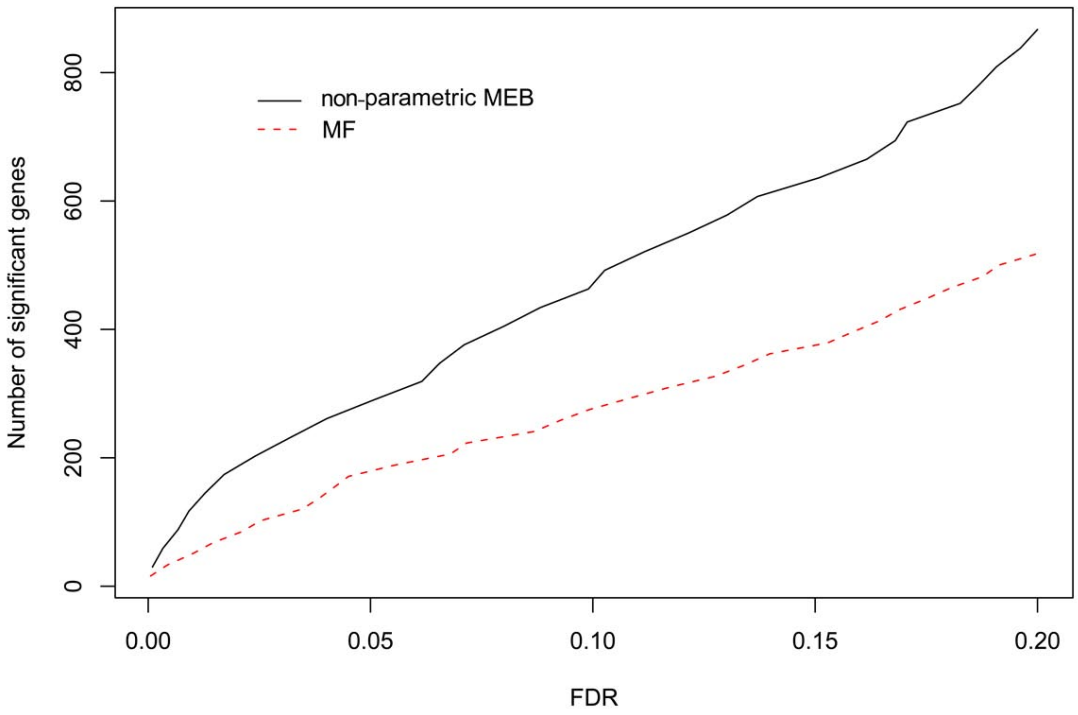

Figure 4.2: Detection power comparison of the proposed method (denoted as MEB) versus the moderated F-statistics (denoted as MF) for the breast cancer microarray data: the y-axis is the number of detected significant genes, the x-axis is the estimated FDR.

Chapter 5

Conclusion and discussion

This dissertation focuses on detecting differentially expressed genes comparing gene expression data from multiple classes. In this large-scale simultaneous testing situation, where thousands of hypothesis tests are performed at the same time, the empirical Bayes idea can often enable us to borrow information across different tests to improve the overall testing power.

In Chapter 2, the choice of empirical null distribution and theoretical null distribution is discussed. With theoretical null distribution, it is easy to calculate p-values associated with the test statistics. However, it may be clear to researchers that the theoretical null distribution is inappropriate as illustrated in Figure 2.1, which usually happens when gene expressions are correlated. MLE method to estimate the null distribution is discussed and yields some good results comparing with the theoretical null distribution through simulations and real data analysis. Although we might sacrifice some accuracy when estimating the empirical null distribution, it is sometimes necessary. When ranking genes, two approaches are discussed, ranking with p-values calculated from empirical null distribution and ranking with local FDR, which adopts the empirical Bayes method. Overall the two approaches work well and could estimate FDR with higher accuracy than the

theoretical null distribution.

A multivariate test statistic for multi-class differential expression detection is proposed in Chapter 3 and 4 instead of using the traditional univariate test statistics, in the sense that more information will be retained. First in Chapter 3, the marginal distribution of the multivariate test statistics is approximated by a mixture of normal distributions, where the parameters of the normal components are estimated through EM algorithm. The null gene distribution is also empirically estimated instead of using the theoretical null distribution through central matching method. The parametric MEB performs quite well as shown in the simulation and real data analysis as compared to the moderated F-statistics. Genes are ranked based on local FDR estimates, which is proportional to the likelihood ratio statistics. It is proved that this is the most powerful test with fixed type I error. However, when the sample size is small, the normal approximation might not be appropriate. One way to solve this is to use the more robust mixture t distribution to estimate the marginal distribution, which will be a future task for the method to improve or find an alternative non-parametric modeling approach.

Chapter 4 discusses the multivariate modeling approach using a non-parametric method, Poisson regression, to estimate both the null gene distribution and marginal distribution of the test statistics. Without parametric assumptions, the non-parametric method is more flexible in the case when normal assumptions are not appropriate. Similarly, the local FDR is adopted to rank the genes which enables information sharing across genes and improves the detection power.

Here, several approaches on multi-class differential gene expression detection incorporating the empirical Bayes idea are discussed. By summarizing the gene expression into multivariate test statistics, more information is retained. Both simulation results and real data analysis have shown improvement in detection power in that more relevant genes are detected when controlling FDR at a certain level. There are many statistical problems of interest for analyzing current

large-scale biomedical data. We expect that the general idea underlying the proposed multivariate modeling method would be generalizable to other statistical problems.

References

[1] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98:5116–5121, 2001.

[2] S. Dudoit, Y. H. Yang, T. P. Speed, and M. J. Callow. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, 12:111–139, 2002.

[3] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461, 2002.

[4] J. Ibrahim, M. H. Chen, and R. Gray. Bayesian models for gene expression with dna microarray data. *Journal of the American Statistical Association*, 97:88–99, 2002.

[5] J. Townsend and D. Hartl. Bayesian analysis of gene expression levels: statistical quantification of relative mrna level across multiple strains or treatments. *Genome Biology*, 3:research0071.1–16, 2002.

[6] H. Ishwaran and J. Rao. Detecting differentially expressed genes in microarrays using bayesian model selection. *Journal of the American Statistical Association*, 98:438–455, 2003.

[7] M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

[8] R. Gottardo, A. E. Raftery, K. Yee Yeung, and R. E. Bumgarner. Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, 62(1):10–18, 2006.

[9] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[10] B Efron. Robbins and empirical bayes and microarrays. *The Annals of Statistics*, 31:366–378, 2003.

[11] C. M. Kendziorski, M. A. Newton, H. Lan, and M. N. Gould. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22(24):3899–3914, 2003.

[12] B Wu. Differential gene expression detection using penalized linear regression models: the improved sam statistics. *Bioinformatics*, 21:1565–1571, 2005.

[13] X. Cui, J. T. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75, 2005.

[14] I. Lönnstedt and T. Speed. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

[15] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 3, 2004.

[16] B Efron. Size and power and and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377, 2007.

[17] B Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99:96–104, 2004.

[18] B Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102:93–103, 2007.

[19] 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)*, 57:289–300, 1995.

[20] M. Langaas, E. Ferkingstad, and B. Lindqvist. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society Series B*, 67:555–572, 2005.

[21] V. J. Gimino, J. D. Lande, T. R. Berryman, R. A. King, and M. I. Hertz. Gene expression profiling of bronchoalveolar lavage cells in acute lung rejection. *American Journal of Respiratory and Critical Care Medicne*, 168:1237–1242, 2003.

[22] M. Tokunou, T. Niki, Y. Saitoh, H. Imamura, M. Sakamoto, and S. Hirohashi. Altered expression of the erm proteins in lung adenocarcinoma. *Lab Invest*, 80(11):1643–1650, 2000.

[23] P. K. Manchanda and R. D. Mittal. Analysis of cytokine gene polymorphisms in recipient's matched with living donors on acute rejection after renal transplantation. *Molecular and Cellular Biochemistry*, 311(1).

[24] A. Pawlik, L. Domanski, J. Rozanski, B. Czerny, Z. Juzyszyn, G. Dutkiewicz, M. Myslak, M. Hałasa, M. Słojewski, and E. Dabrowska-Zamojcin. The association between cytokine gene polymorphisms and kidney allograft survival.

Annals of Transplantation: Quarterly of the Polish Transplantation Society, 13(2):54–58, 2008.

[25] P. K. Manchanda, A. Kumar, R. K. Sharma, H. Goel, and R. D. Mittal. Association of pro/anti-inflammatory cytokine gene variants in renal transplant patients with allograft outcome and cyclosporine immunosuppressant levels. *Biologics: Targets & Therapy*, 2(4):875–884, 2008.

[26] A. D. Truax, O. I. Koues, M. K. Mentel, and S. F. Greer. The 19s atpase s6a (s6'/tbp1) regulates the transcription initiation of class ii transactivator. *Journal of Molecular Biology*, 395(2):254–269, 2010.

[27] S. P. Bradley, M. Pahari, M. E. Uknis, C. Rastellini, and Cicalese L. Gene expression profiles characterize early graft response in living donor small bowel transplantation: a case report. *Transplant Proceedings*, 38(6):1742–1743, 2006.

[28] B. M. Meiser, C. Reiter, H. Reichenspurner, P. Uberfuhr, E. Kreuzer, E. P. Rieber, G. Riethmüller, and B. Reichart. Chimeric monoclonal cd4 antibody-a novel immunosuppressant for clinical heart transplantation. *Transplantation*, 58(4):419–423, 1994.

[29] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, 98:11462–11467, 2001.

[30] B. Bolstad, R. Irizarry, M. Astrand, and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 9:185–193, 2003.

[31] G. Dennis, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, and R. Lempicki. David: database for annotation and visualization and and integrated discovery. *Genome Biology*, 4(9):R60, 2003.

[32] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[33] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[34] Wikipedia.org. Apoptosis. `http://en.wikipedia.org/wiki/apoptosis`.

[35] K. Collins, T. Jacks, and N. P. Pavletich. The cell cycle and cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 94(7):2776–2778, 1997.

[36] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London Series A and Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[37] G. Klein. Cancer, apoptosis, and nonimmune surveillance. *Cell Death Differ*, 11(1):13–17, 2003.

Appendix A

Detailed simulation results

A.1 Supplementary simulation result for parametric multivariate modeling approach

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	6	110	276	1	17	54	0	3	10
parametric MEB	20	229	486	5	54	127	1	10	30

Table A.1: Number of true positives when fixing FDR by both moderated F-statistics and parametric MEB method, the correlation parameter $\rho = 0.25$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	8	119	308	1	19	56	0	3	9
parametric MEB	21	222	478	5	45	113	1	10	25

Table A.2: Number of true positives when fixing FDR by both moderated F-statistics and parametric MEB method, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	11	121	291	2	23	59	0	5	15
parametric MEB	29	221	466	6	40	93	3	17	35

Table A.3: Number of true positives when fixing FDR by both moderated F-statistics and parametric MEB method, the correlation parameter $\rho = 0.75$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	468	1027	1319	142	372	508	37	125	185
parametric MEB	756	1330	1568	257	526	662	86	202	271

Table A.4: Number of true positives when fixing FDR by both moderated F-statistics and parametric MEB method, the correlation parameter $\rho = 0.25$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	484	1045	1315	137	370	512	38	129	190
parametric MEB	753	1329	1555	237	506	639	80	198	269

Table A.5: Number of true positives when fixing FDR by both moderated F-statistics and parametric MEB method, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	493	1026	1297	138	382	545	44	133	190
parametric MEB	636	1169	1416	216	513	675	74	182	240

Table A.6: Number of true positives when fixing FDR by both moderated F-statistics and parametric MEB method, the correlation parameter $\rho = 0.75$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

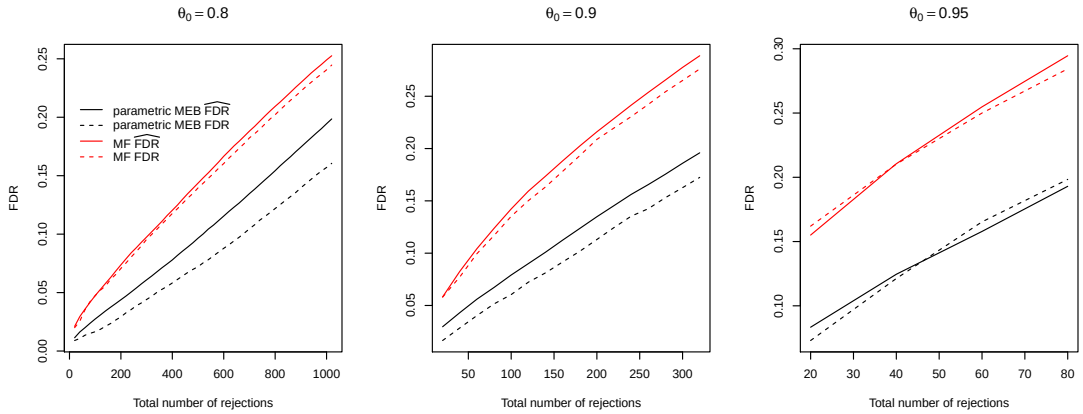

Figure A.1: estimated FDR and true FDR based on EB parametric method and moderated F-statistics, the correlation parameter $\rho = 0.25$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95 from left to right.

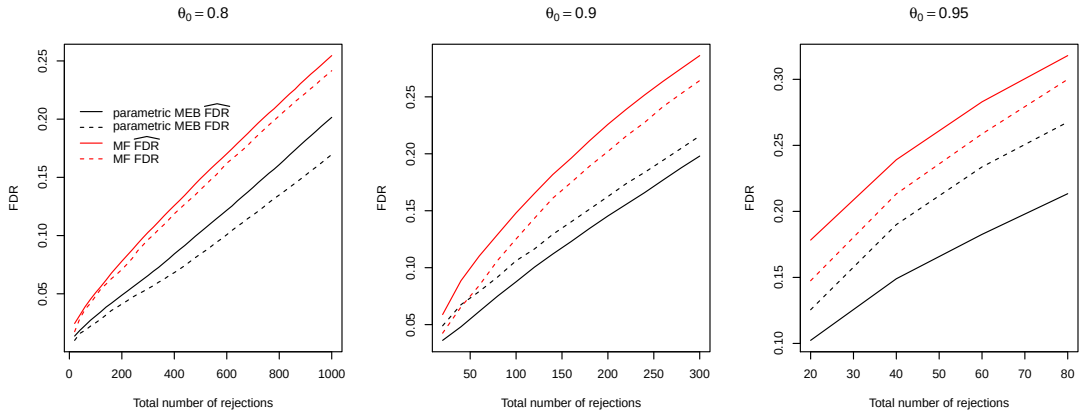

Figure A.2: estimated FDR and true FDR based on EB parametric method and moderated F-statistics, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95 from left to right.

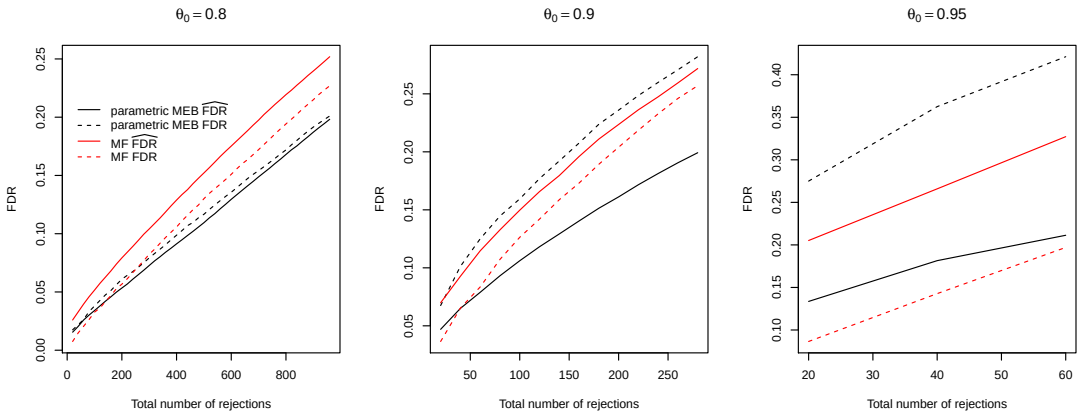

Figure A.3: estimated FDR and true FDR based on EB parametric method and moderated F-statistics, the correlation parameter $\rho = 0.75$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95 from left to right.

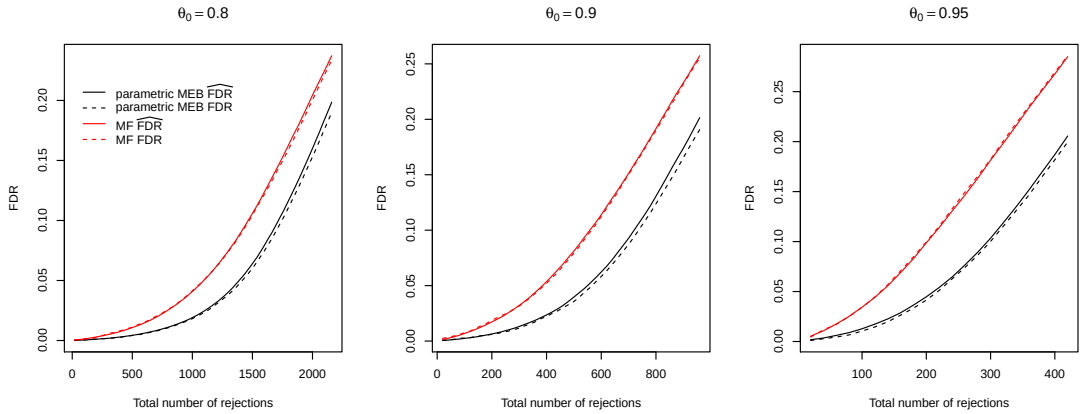

Figure A.4: estimated FDR and true FDR based on EB parametric method and moderated F-statistics, the correlation parameter $\rho = 0.25$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95 from left to right.

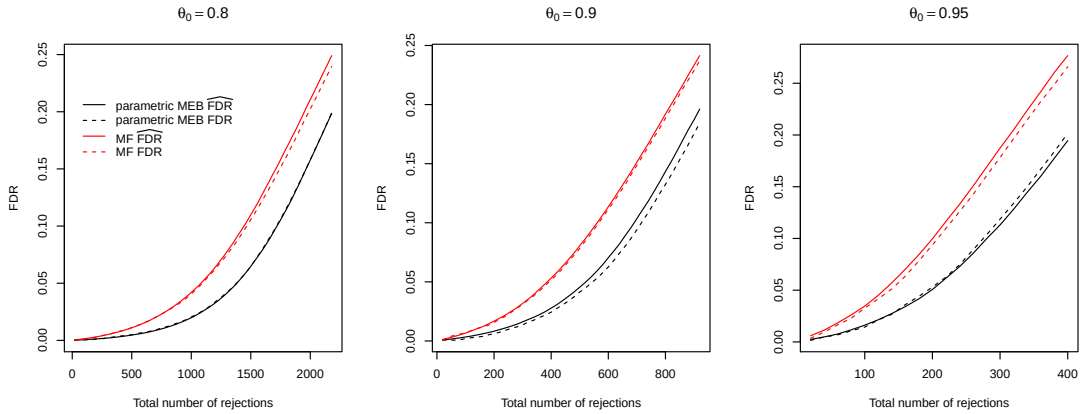

Figure A.5: estimated FDR and true FDR based on EB parametric method and moderated F-statistics, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95 from left to right.

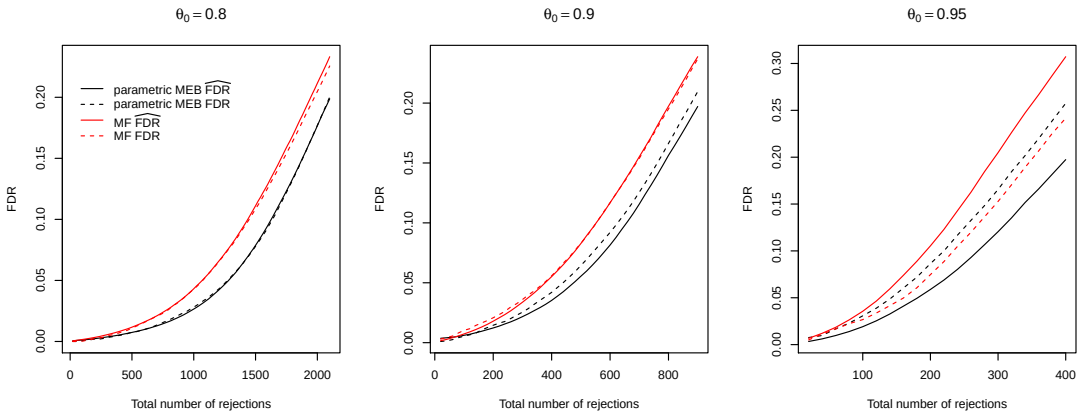

Figure A.6: estimated FDR and true FDR based on EB parametric method and moderated F-statistics, the correlation parameter $\rho = 0.75$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8, 0.9, 0.95$ from left to right.

A.2 Supplementary simulation result for non-parametric multivariate empirical Bayes modeling

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.05	0.1	0.15	0.05	0.1	0.15	0.1	0.15	0.2
MF	112	287	454	16	54	98	10	20	32
non-parametric MEB	236	508	735	40	112	186	21	43	66

Table A.7: Number of true positives when fixing FDR at 0.05, 0.1, 0.15 (0.1, 0.15, 0.2 for $\theta_0 = 0.95$) by both moderated F-statistics and non-parametric MEB method, the correlation parameter $\rho = 0.25$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.05	0.1	0.15	0.05	0.1	0.15	0.1	0.15	0.2
MF	109	280	447	18	53	97	10	20	33
non-parametric MEB	233	499	718	42	114	186	22	43	64

Table A.8: Number of true positives when fixing FDR at 0.05, 0.1, 0.15 (0.1, 0.15, 0.2 for $\theta_0 = 0.95$) by both moderated F-statistics and non-parametric MEB method, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.05	0.1	0.15	0.05	0.1	0.15	0.1	0.15	0.2
MF	105	270	439	21	57	101	11	20	32
non-parametric MEB	232	498	707	49	121	195	22	41	62

Table A.9: Number of true positives when fixing FDR at 0.05, 0.1, 0.15 (0.1, 0.15, 0.2 for $\theta_0 = 0.95$) by both moderated F-statistics and non-parametric MEB method, the correlation parameter $\rho = 0.75$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	472	1033	1313	141	370	506	38	124	183
non-parametric MEB	689	1283	1521	217	486	622	62	173	238

Table A.10: Number of true positives when fixing FDR at 0.01, 0.05, 0.1 by both moderated F-statistics and non-parametric MEB method, the correlation parameter $\rho = 0.25$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	471	1033	1312	137	369	507	40	126	185
non-parametric MEB	690	1281	1514	220	486	619	66	176	240

Table A.11: Number of true positives when fixing FDR at 0.01, 0.05, 0.1 by both moderated F-statistics and non-parametric MEB method, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95.

	$\theta_0 = 0.8$			$\theta_0 = 0.9$			$\theta_0 = 0.95$		
FDR	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
MF	482	1037	1317	146	375	511	41	127	185
non-parametric MEB	700	1276	1502	225	488	616	66	171	234

Table A.12: Number of true positives when fixing FDR at 0.01, 0.05, 0.1 by both moderated F-statistics and non-parametric MEB method, the correlation parameter $\rho = 0.75$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8, 0.9, 0.95$.

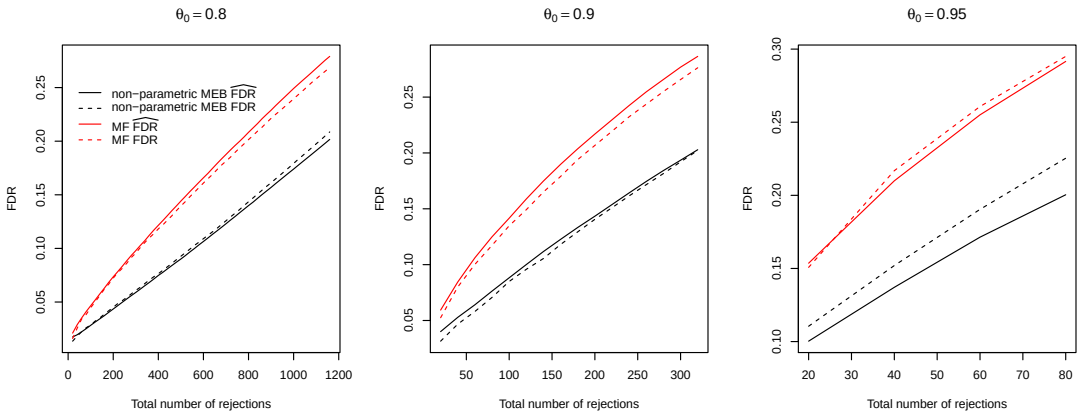

Figure A.7: estimated and true FDR based on non-parametric MEB method and moderated F-statistics, the correlation parameter $\rho = 0.25$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8, 0.9, 0.95$ from left to right.

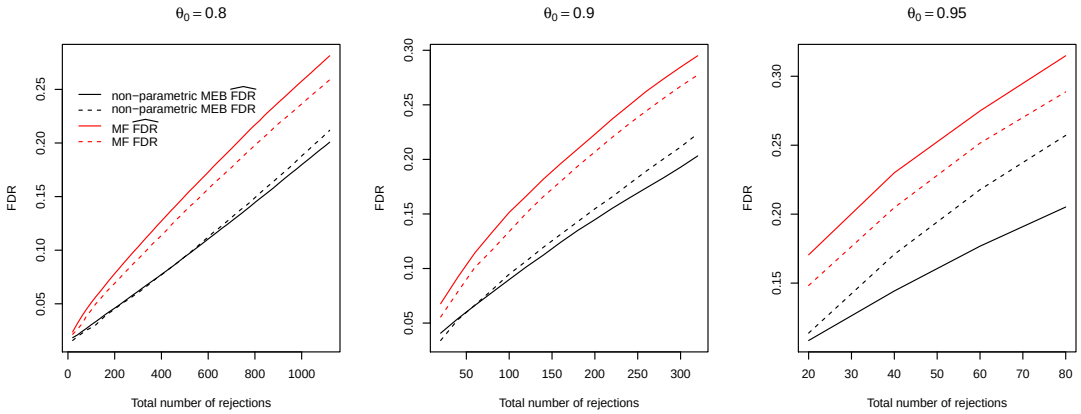

Figure A.8: estimated and true FDR based on non-parametric MEB method and moderated F-statistics, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95 from left to right.

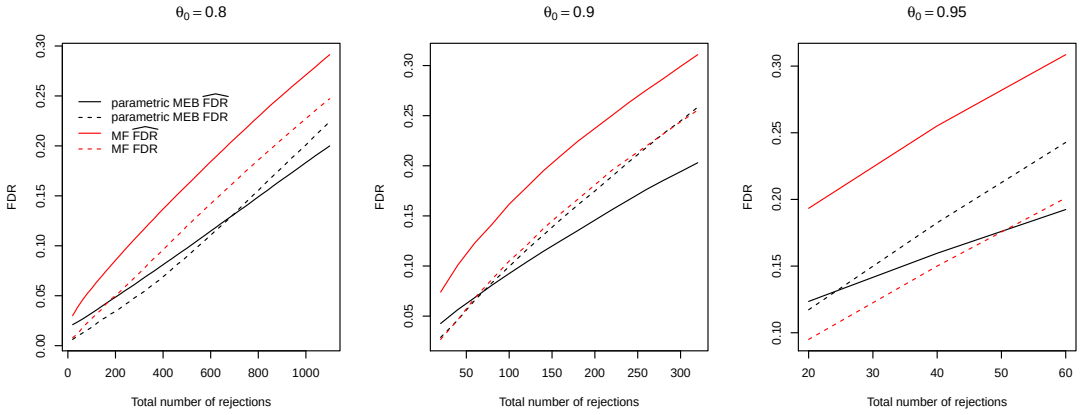

Figure A.9: estimated and true FDR based on non-parametric MEB method and moderated F-statistics, the correlation parameter $\rho = 0.75$, the sample size in each group $n = 10$, and the null gene probability $\theta_0 = 0.8$, 0.9, 0.95 from left to right.

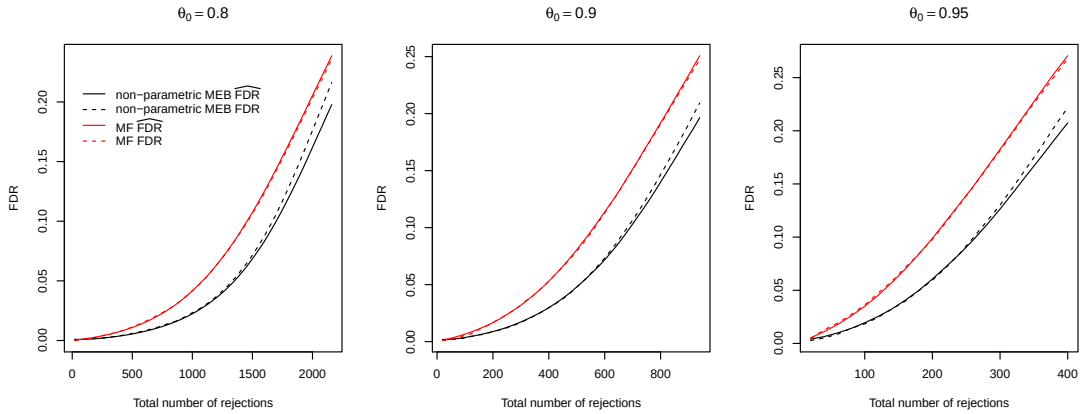

Figure A.10: estimated and true FDR based on non-parametric MEB method and moderated F-statistics, the correlation parameter $\rho = 0.25$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8, 0.9, 0.95$ from left to right.

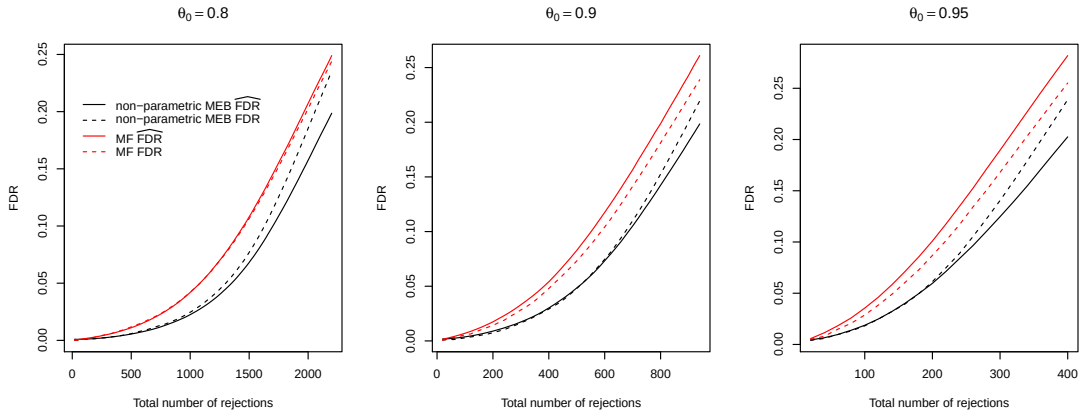

Figure A.11: estimated and true FDR based on non-parametric MEB method and moderated F-statistics, the correlation parameter $\rho = 0.5$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8, 0.9, 0.95$ from left to right.

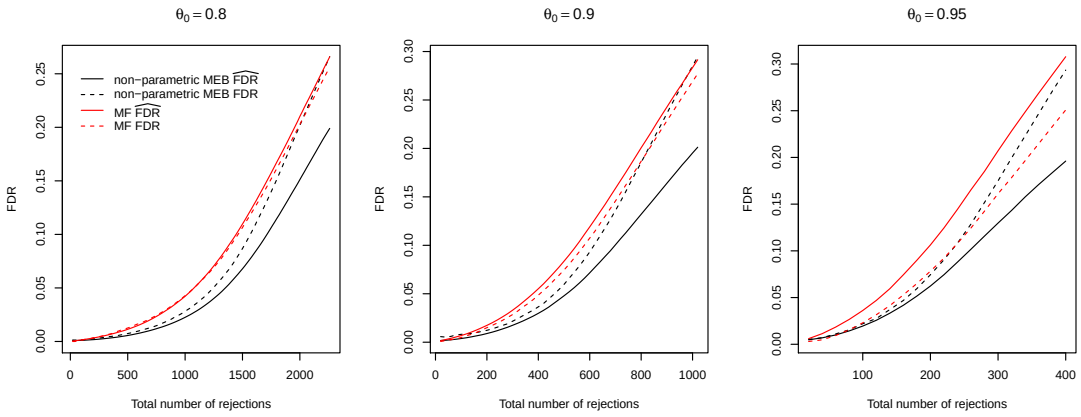

Figure A.12: estimated and true FDR based on non-parametric MEB method and moderated F-statistics, the correlation parameter $\rho = 0.75$, the sample size in each group $n = 20$, and the null gene probability $\theta_0 = 0.8, 0.9, 0.95$ from left to right.