Multiple Testing in Genome-Wide Studies

by
Moonsu Kang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2007

Approved by:

Dr.Pranab K.Sen, Advisor
Dr.Chirayath Suchindran, Reader
Dr.Fei Zou, Reader
Dr.Ethan Lange, Reader
Dr.Mayetri Gupta, Reader

ABSTRACT

MOONSU KANG: Multiple Testing in Genome-Wide Studies.
(Under the direction of Dr.Pranab K.Sen.)

DNA microarray technologies allow us to monitor expression levels of thousands of genes simultaneously. A basic task in analyzing microarray data is the identification of differentially expressed genes under different experimental conditions. The null hypothsis is no association between the expression levels and explanatory variables or covariates. Family-wise error rate (FWER), although very conservative, controls type I error. False Discovery Rate (FDR) is a less stringent approach which aims to control the expected proportion of Type I errors among the rejected hypotheses. Since there are thousands of genes tested simultaneously, FDR may be enhanced. High correlation between tested genes, attributed to co-regulations and dependency in the measurement errors, further complicates the problem. Most of the current FDR procedures assume independence or rather restrictive dependence structures, resulting in being less reliable.

In this work, we address these very large multiplicity problems by adopting a two-stage FDR controlling procedure under suitable dependence structures and based on Poisson distributional approximation, which eliminates the need to assume restricted dependence structures. We compare the performance of the proposed FDR procedure with that of other FDR controlling procedures, with illustration of the leukemia microarray study of Golub et al. (1999) and simulated data. In these studies, the proposed FDR procedure has greater power without much elevation of FDR.

Current FDR procedures have not been used extensively in genomic sequences involving count or discrete, or purely qualitative responses, confronted with high-dimensional

low sample size constraints. Using the 2002-03 SARS epidemic model, it is shown that proposed FDR procedure along with an appropriate test statistic based on a pseudo-marginal approach with Hamming distance performs better.

Finally, for classfication of genes of dependent genes with heterogeneity amidst a small sample, standard robust inference may not work out. This issue involves setting up a hypothesis when parameters of interest are subject to inequality restrictions. Usual (restricted) likelihood based statistical inference procedures may not be computationally intensive. Roy's union-intersection principle may be a viable alternative. The breast cancer study of Lobenhofer et al. is included for numerical illustration.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr.Pranab K. Sen. I could not have imagined having a better advisor and mentor for my PhD, and without his help, knowledge, encouragement, perceptiveness and suggestions I would never have finished. I want to express my gratitude to Dr.Suchindran, Melissa hobgood, Veronica Stalling, and Dr.Gary Koch for helping me to overcome various hardships in my life. I would also like to thank my committee member, other Bios staffs, and Dr.Qaqish for their interest and valuable comments and suggestions. Finally, I would like to thank my family with my sister getting married on August, even though I may not attend her wedding and friends for their advice and conversation.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

List Of Abbreviations

- **MTP$_2$** : **X** has MTP_2 property if for all **x** and **y**, $f(\mathbf{x}) \cdot f(\mathbf{y}) \leq f(min(\mathbf{x}, \mathbf{y})) \cdot f(max(\mathbf{x}, \mathbf{y}))$, where f is the joint density and the minimum and maximum are evaluated componentwise.

- **TP$_2$** : A nonnegative bivariate function f (x , y) is said to be TP_2 in (x , y) if the following condition holds: f (x , y) f (x', y') 2 f (x , y')f (x', Y) for all $x < x'$ and $y < y'$.

- **PRDS** : A multivariate distribution is said to have positive regression dependency (PRDS) if for any increasing set $D, P(\mathbf{X} \in D | X_1 = x_1, \ldots, X_i = x_i)$ is nondecreasing in $(x_1, \cdots, x_i)$.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

The Human Genome Project announced the completion of a map of the human genome in 2003. DNA microarrays are used to measure the level of expression of genes under different enviromental setups by hybridizing a labeled cRNA representation of the mRNA to cDNA sequences (cDNA microarrays) or by hybridizing a labeled cRNA representation of the mRNA to short specific segments (synthetic oligonucleotide microarrays).

These new developments have to analyze genomic data. The technology creates an abundance of complex and enormously large dimensional data models, resulting in the high-dimension (K) low sample size (n) environments. Genes tend to be heavily correlated for co-regulations on genomic locations and gene expression biases based on the effects of aneuploidy, resulting in complicated dependency structures. One of the important aims of this study is the identification of differentially expressed genes. This issue can be restated as a problem in multiple hypothesis testing: the simultaneous mul-

tiple test for each gene of no association between the expression levels and explanatory variables or covariates. Thus, we are faced with large multiplicity problems generated in such studies. Two types of errors are involved: a false positive, Type I error is committed when a gene is declared to be differentially expressed when it is not, and a false negative, Type II error, is committed when a gene is not declared to be differentially expressed when it is. The traditional approach to the multiplicity is control of the familywise error rate (FWER) which adjusts the p-value so that it reflects the chance of at least 1 false positive being found in the list. The FWER methods are unduly conservative when there are thousands of hypotheses (or genes) tested and thus a different approach to this problem is needed. Benjamini and Hochberg (1995) defined False discovery rate (FDR) as the expected proportion of Type I error among the number of rejections. Compared to FWER, the FDR is a better way to deal with uncertainty in large screening data sets, where a small number of false positives is acceptable. It is said that merely controlling the FDR could lose power, considered as false nonrejection. This false nonrejection rate (FNR) is defined as the expected rate of false acceptance against the number of total acceptances. An approach with the balance between the FDR and the FNR would be better than one purely controlling the FDR. Current FDR controlling procedures do not take into account complex dependency structures among the genes, resulting in loss of power and unreliable estimation. They control the FDR only when the p-values meet some regularity conditions under which central limit theorems apply. In reality, it's not easy to find suitable mixing conditions for central limit theorems, under complex dependence structures of the genes. Under fairly mild regularity conditions about the dependence of genes, we adopt a new false discovery rate controlling procedure. For these problems, two-stage FDR and FNR are proposed using alternative limit theorems for dependent genes by the Chen-Stein methods. These procedures attain both more power and exact estimation. We apply proposed FDR

procedure along with an appropriate test statistics to microarray experiment as well as categorical genomic sequences in Chapters 2 and 3.

The high-dimension (K) low sample size (n) environments make it hard to classfiy thousands of genes. These problems make it unreasonable to adopt standard models where the number of parameters outnumber the sample size. Studies such as dose-response microarray experiments or time-course data mainly involves order-restricted inference. In these enviroments, Roy's (1953) union-intersection principle have some advantanges (Silvapulle and Sen 2004, Tsai and Sen 2005). Based on the Union-Intersection principle, robust M-statistics , insensitive to outlier arrays, and linear rank statistics, a locally most powerful test, is proposed in Chapter 4. The real microarray datasets, real genomic sequence and simulation models are presented in Chapter 5 to evaluate proposed FDR and the corresponding test statistics. Overview of research work on this problems is summarized in section 1.3.

1.2 Literature Review

1.2.1 Multiple Testing And Adjusted p-values

Multiple hypothesis testing issues arise frequently in biomedical and genomic research. For example, a number of recent articles have addressed multiple testing in DNA microarrays, but the solutions proposed so far have not always been in the standard framework Dudoit et al. (2003). A key feature of this methodology is the general characterization and an explicit construction of a test statistics null distribution. We shall briefly review some of the existing methodologies and also describe some recent developements in this field. The adjusted p-values are one of the useful tools to describe some multiple testing procedure. We shall also address it.

1.2.1.1 Multiple Testing In DNA Microarray Experiments

Define multiple hypotheis testing procedure in microarray experiment. An $m \times n$ matrix $\mathbf{X} = (x_{ji}) = (X_1, \ldots, X_m)$ represents the gene expression level data with rows corresponding to genes and columns corresponding to individual microarry experiments. The expression measures x_{ji} are in general highly preprocessed data. We use the sample data $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$ formed by the expression profiles $\mathbf{x}_i$ and response or covariates y_i in order to test hypotheses regarding the joint distrubution of the expression measures $\mathbf{X} = (X_1, \ldots, X_m)$ and response or covariate Y. A standard approach to the multiple testing problem includes two aspects:

- computing an appropriate test statsistic T_j for each gene j,
- applying a multiple testing procedure to determine which hypotheses are rejected while controlling a suitably defined Type I error rate.

1.2.1.2 Type I Error Rates

A multiple testing procedure controls a particular Type I error rate at level α if this error rate is less than or equal to α when the given procedure is applied to a set of rejected hypotheses.

Consider the problem of simultaneous testing m null hypotheses, $H_j, j = 1, \ldots, m$ which are assumed to be known, of which m_0 are true and unknown. The corresponding p-values are $P_1, \ldots, P_m$. This situation can be expressed by the table below. R is the number of hypotheses rejected, which is an observable random variable. U, V, S, and T are unobservable random variable.

The focus is on the proportion of false positives V with respect to the number of rejected hypotheses R. When multiple testing procedure is applied to high-dimensional genomic data, one may wish to bear some false positives as long as their number is small.

TABLE I: Number of errors committed when testing m null hypotheses

	Number not rejected	Number rejected	Total
Non-differentially expressed	U	V	m_0
Differentially expressed	T	S	$m - m_0$
	$m - R$	R	m

In the microarray setting, there is a null hypothesis H_i for each gene i and rejection of H_i corresponds to declaring that gene i is differentially expressed. In general, we'd like to minimize the number V corresponding to Type I error and the number T corresponding to Type II error. When testing a single hypothesis H, the probability of Type I error is controlled at prespecified level α. This may be achieved by choosing a critical value c_α so that $Pr(|T| \geq c_\alpha|H) \leq \alpha$ and rejecting the null hypothesis when $|T| > c_\alpha$. The Type I error rates shown below are the most standard ones Shaffer (1995).

- The per-comparison error rate (PCER):the expected value of the number of Type I errors divided by the number of hypotheses, that is, $PCER = E(V)/m$.
- The per-family error rate (PFER):the expected number of Type I errors, $E(V)$.
- The family-wise error rate (FWER):the probability of at least one Type I error, that is, $FWER = Pr(V \geq 1)$.
- The false discovery rate (FDR) of Benjamini and Hochberg (1995):the expected proportion of Type I errors among the rejected hypotheses.

It is easy to prove that $PCER \leq FDR \leq FWER \leq PFER$. Note that the error rates are defined under the true and typically unknown data generating distribution for gene expression data $\mathbf{X} = (x_{ji}) = (X_1, \ldots, X_m)$ where a gene expression profile is $\mathbf{x_i} = (x_{1i}, \ldots, x_{mi})$. In particular, they depend on which specific subset $\Lambda_0 \subset \{1, \ldots, m\}$

of null hypotheses is true for this distribution. Weak control refers to control of Type I error rate when all the null hypotheses are true. In the microarray setting, it seems more appropriate to have strong control of the Type I error rate, that is, control under any combination of true and false null hypothesis.

1.2.1.3 Adjusted p-values

The multiple testing procedure may be defined in terms of unadjusted p-values or adjusted p-values. Unadjusted p-value gives the probability of obtaining a value of a test statistic that is at least as unfavorable to H_0 as the observed one, that is, $p_j = Pr(|T_j| \geq |t_j||H_j)$ for hypothesis H_j. The adjusted p-value for H_j is defined as the nominal level of the entire test procedure at which H_j to be rejected, provided that the values of all test statistics are given. For FWER controlling procedure, $\tilde{p}_j$ is defined as $\inf\{\alpha \in [0,1] : H_j \text{ is rejected at nominal } FWER = \alpha\}$. For FDR controlling procedure, $\tilde{p}_j$ is defined as $\inf\{\alpha \in [0,1] : H_j \text{ is rejected at nominal } FDR = \alpha\}$(Yekutieli and Benjamini, 1999). An advantage of reporting adjusted p-values is that the level of the test does not have to be determined in advance.

1.2.2 Simes Inequality And MTP_2 Property

To test the overall null hypothesis $H_0 = \bigcap_{i=1}^{n} H_i$ with their corresponding P values at a prespecified significance level α is a common problem in practice. For example, when identifying differentially expressed genes, multiple studies are often performed: the simultaneous multiple test for each gene. Type I error should be controlled at preassigned level in multiple testing procedure. Simes method is one of the methods to control type I error. We intend to identify the correlation structure among the genes. It is said that MTP_2 property may characterize a general class of positive dependence structures among the genes. We introduce this concept along with positive regression dependence.

The classical and well-known Bonferroni method rejects H_0 IF $P_i \leq \alpha/m$ for at least one i. But this method is very conservative, particulary when the dependence among the test statistics is very high. Simes (1986) proposed modified Bonferroni methods. Let $P_{(1)} \leq, \ldots, \leq P_{(m)}$ be the ordered P values. Simes suggested the test procedure to reject H_0 if $P_i \geq \frac{i\alpha}{m}$ at least one i. Under Simes inequality, this method controls the type I error rate for the test statistics having the folllowing distributions.

The null distributions of test statistics, $X_1, \ldots, X_m$, have probability densities of the form

$$\int \prod_{i=1}^{m} f(x_i, z) g(z) dz \qquad (1)$$

for some probability densities $f(x, z)$ and $g(z)$, where $f(x, z)$ is TP_2 in (x, z). Statistics whose distributions has the form (1) are called positively dependent. The Simes conjecture holds only for positively dependent test statistics. Equicorrelated Multivariate normal with nonnegative correlation, absolute-valued equicorrelated multivariate normal, absolute-valued central multivariate t, central multivariate F, and Bayes methods including frailty model when a parameter z is random and other multivariate distribu-

tions have densities of the form (1). The following MTP_2 property Karlin and Rinott (1980) characterizes a class of positive dependent distribution. A multivariate distribution is said to have positive regression dependency (PRDS) if for any increasing set $D, P(\mathbf{X} \in D|X_1 = x_1, \ldots, X_i = x_i)$ is nondecreasing in $(x_1, \cdots, x_i)$. A stricter condition, that is, positive regression dependency, is multivariate total positivity of order 2, MTP_2 if for all $\mathbf{x}$ and $\mathbf{y}$, $f(\mathbf{x}) \cdot f(\mathbf{y}) \leq f(\min(\mathbf{x}, \mathbf{y}))\cdot$ f$(\max(\mathbf{x}, \mathbf{y}))$ where f is either the joint density or the joint probability function, and the minimum and maximum are evaluated componentwise.

Let $X_{(1)}, \ldots, X_{(m)}$ be the ordered values of a set of MTP_2 random variables $X_1, \ldots, X_m$ with a marginal F. Then $Pr(X_{(j)} \leq a_j, j = 1, \ldots, m) \leq 1 - \alpha$. If $\{a_j\}$ are such that $F(a_j) = \frac{j\alpha}{m}$, with the equality holding when when $\{X_i\}$ are independent. The Simes inequality holds in general for all MTP_2 distributions. However, Karlin and Rinott (1980) considered the strongly multivariate reverse rule of order two $(S - MRR_2)$ condition characterizing negatively dependent multvariate distributions. They proved that the Simes conjecture is not true in general for such distributions.

1.2.3 Control Of FWER

The common approach to the multiplicity problem is to control the FWER at preassigned level. The FWER is said to be controlled at level α by a particular multiple testing procedure if FWER $\leq \alpha$. We shall introduce the existing FWER methodologies here.

- Single-step procedures: Strong control of FWER is provided based on Boole's inequality.

  $$FWER = Pr(V \geq 1) = Pr(\bigcup_{j=1}^{m_0}\{\tilde{P}_j \leq \alpha\}) \leq \sum_{j=1}^{m_0} Pr(\tilde{P}_j \leq \alpha) = \sum_{j=1}^{m_0} Pr(P_j \leq \frac{\alpha}{m}) \leq \frac{m_0\alpha}{m}.$$

  Single-step Bonferroni adjusted p-values are given by $\tilde{p}_j = min(mp_j, 1)$ The fol-

lowing Šidák's procedure controls for FWER for test statistics that satisfy the Šidák's inequality.

$$Pr(|T_1| \leq c_1, \ldots, |T_m| \leq c_m) \leq \prod_{j=1}^{m} Pr(|T_j| \leq c_j).$$

The single-step Šidák' adjusted p-values are given by $\tilde{p}_j = 1 - (1 - p_j)^m$.

Westfall and Young (1993) proposed adjusted p-values for less conservative procedures which take into account the dependence structure among test statistics. The single-step min P adjusted p-values are given by

$$\tilde{p}_j = Pr(min_{1 \leq l \leq m} P_l \leq p_j | H_0^C).$$

The single-step max T adjusted p-values are given by

$$\tilde{p}_j = Pr(max_{1 \leq l \leq m} T_l \leq t_j | H_0^C).$$

- Step-down procedures: Step-down FWER procedures achieve higher power rather than by single-step procedures. Let $p_{r_1} \leq p_{r_2} \leq \cdots \leq p_{r_m}$ denote the observed ordered unadjusted p-values and $H_{r_1}, H_{r_2}, \ldots, H_{r_m}$ denote the corresponding null hypotheses. The Holm (1979) procedure operates in the following manner. Define $j^* = min\{j : p_{r_j} > \alpha/(m-j+1)\}$ and reject hypotheses H_{r_j}, for $j = 1, \ldots, j^* - 1$. If no such j^* exists, reject all hypotheses. Similarly, the step-down Holm adjusted p-values are given by $\tilde{p}_{r_j} = max_{k=1,\ldots,j}\{min((m-k+1)p_{r_k}, 1)\}$. The step-down Šidák adjusted p-values are defined as $\tilde{p}_{r_j} = max_{k=1,\ldots,j}\{1 - (1 - p_{r_k})^{(m-k+1)}\}$. The Westfall and Young (1993) step-down min P adjusted p-values are defined by

$$\tilde{p}_{r_j} = max_{k=1,\ldots,j}\{Pr(min_{l \in \{r_k,\ldots,r_m\}} P_l \leq p_{r_k} | H_0^C)\}.$$

And the step-down max T adjusted p-values are defined by

$$P\tilde{p}_{r_j} = max_{k=1,\ldots,j}\{Pr(max_{l\in\{r_k,\ldots,r_m\}}|T_l| \geq |t_{r_k}||H_0^C)\}.$$

where $|t_{r_1}| \geq |t_{r_2}| \geq \cdots \geq |t_{r_m}|$ denote the observed ordered test statistics.

- Step-up procedures: Under the complete null hypothesis H_0^C and for independent test statistics, the ordered unadjusted p-values $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ satisfy $Pr(P_{(j)} \geq \frac{j\alpha}{m}, \quad \forall j = 1,\ldots,m|H_0^C) \geq 1-\alpha$ with equality in the continuous case. Step-up procedures start with this Simes inequality (1986). Hochberg (1988) applied the Simes inequality to derive the following FWER controlling procedure. Let $j^* = max\{j : p_{r_j} \leq \alpha/(m-j+1)\}$ and reject hypotheses H_{r_j}, for $j = 1,\ldots,j^*$. If no such j^* exists, reject no hypotheses. The step-up Hochberg adjusted p-values are given by $\tilde{p}_{r_j} = min_{k=j,\ldots,m}\{min(m-k+1)p_{r_k}, 1)\}$. Related procedure are those of Hommel (1988) and Rom (1990). All procedures based upon the Simes inequality have the assumption that the result derived under independence is a conservative procedure for dependent tests.

However, Benjamini and Hochberg (1995) argued that this FWER approach has the following limitations.

- Much of the methodology of FWER controlling procedures is concerned with comparisons of multiple treatments and families whose test statstics have multivariate normal (or t).

- Strong control of the FWER tends to be less powerful than the per comparison procedure of the same levels.

- The control of the FWER is not quite often needed.

In many situations, control of the FWER is too restrictive at the expense of substantially lower power in detecting false hypotheses. One may wish tolerate some Type I errors, provided their number is small in comparison to the number of rejected hypotheses.

1.2.4 Control Of FDR

As we have seen before, control of the FWER is too conservative when there are many hypotheses such as in microarray experiments. The number of erroneous rejections should be considered in many multiplicity problems. At the same time, the seriousness of the loss by erroneous rejections is related to the number of rejected hypotheses.

Benjamini and Hochberg (1995) introduced the concept of the false discovery rate (FDR) in order to control for the conservativeness of the FWER. Let the unobserved random variable Q= $\frac{V}{V+S}$-the proportion of the rejected null hypothses which are erroneously rejected. $Q = 0$ when $V + S$=0. We define the FDR Q_e to be the expectation of Q, $Q_e = E(Q) = E\{\frac{V}{V+S}\} = E\{\frac{V}{R}\}$. Under the complete null hypotheses, control of the FDR implies control of the FWER in the weak sense. When $m_0 < m$, the FDR is smaller than or equal to the FWER.

Benjamini and Hochberg (1995) proposed the following step-up FDR controlling procedure. Consider testing $H_1, H_2, \ldots, H_m$ with the corresponding p-values $P_1, P_2, \ldots, P_m$. Let $P_{(1)} \leq P_{(2)} \leq, \ldots, \leq P_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. Define the following Bonferrroni type multiple-testing procedure :

Let k be the largest i for which $P_{(i)} \leq \frac{i}{m}q$; then reject all $H_{(i)}, i = 1, 2, \ldots, k$. If no such i exists, reject no hypothesis Benjamini and Hochberg (1995).

Benjamini and Liu (1999) derived a new step-down procedure when test statistics are independent. Define the m critical values by

$$\delta_i \equiv 1 - [1 - min(1, \frac{m}{(m-i+1)}q)]^{\frac{1}{(m-i+1)}} \qquad , 1 \leq i \leq m.$$

The step-down procedure then operates as follow . Let k be the smallest i for which $P_{(i)} > \delta_i$. Reject $H_{(1)}, \ldots, H_{(k-1)}$. This procedure controls the FDR at level q Benjamini and Liu (1999). They proved that their procedure neither dominates nor is dominated by the step-up procedure.

Benjamini and Yekutieli (2001) showed that if the joint distribution of the test statistics is PRDS on the subset of test statistics corresponding to true null hypotheses, the Benjamini-Hochberg procedure controls the FDR at less than or equal to $\frac{m_0}{m}q$. They also introduced a simple conservative modification of the procedure which controls the FDR for arbitrary dependence structures. Adjusted p-values for this modified step-up procedures are $\tilde{p_{r_j}} = min_{k=j,\ldots,m}\{min(\frac{m\sum_{j=1}^{m} 1/j}{k} p_{r_k}, 1)\}$.

Choosing the critical values $c_1 \leq \cdots \leq c_m$ subject to a preassigned level α, is equivalent to finding constants $a_1 \leq \cdots \leq a_m$ satisfying the following set of inequalities Sarkar (2000):$P(X_{1:k} \leq a_1, \ldots, X_{k:k} \leq a_k) \geq 1 - \alpha$ where $X_{1:k} \leq \ldots X_{k:k}$ denote the ordered components of $(X_1, \ldots, X_k)$. Finner and Roters (1998) illustrated this, which provided that the critical values $(a_1 \leq a_2 \leq a_3)$ satisfying this inequality involving an equicorrelated trivariate standard normal distribution is in fact monotone when the common correlation is positive and at most $(z_{\alpha/2}^2 - z_{1/4}^2)/(z_{\alpha/2}^2 + z_{1/4}^2)$, where $z_{\alpha/2}$ is the upper 100α percent point of $N(0,1)$. This example is generalized to the fact that the desired monotonicity property holds in general for MTP_2 test statistics.

The work of Benjamini and Yekutieli (2001) and the Benjamini-Liu step-down procedure can be extended to a more general stepwise procedure. Sarkar(2002) obtained an

explicit expression of the FDR of a generalized step-up-step-down procedure of order r, in terms of right-tailed test based on $X_1, \ldots, X_m$ with the corresponding critical values $c_1, \ldots, c_m$. Starting with the formula $FDR = \sum_{i \in I_0} \sum_{j=0}^{m-1} \frac{1}{m-j} P[\{X_i \geq c_{(j+1)}\}] \cap B_{j,m}^r$, where $\{H_i, i \in I_0\}$ is the list of true null hypotheses.

$$\begin{aligned} FDR &= \frac{1}{m-r+1} \sum_{i \in I_0} P\{X_i \geq c_{(r)}\} \\ &+ \sum_{i \in I_0} \sum_{j=1}^{r-1} E[\phi_{j,m}^r(X_i)(\frac{I(X_i \geq c_{(j)})}{m-j+1} - \frac{I(X_i \geq c_{(j+1)})}{m-j})] \\ &+ \sum_{i \in I_0} \sum_{j=r}^{m-1} E[\Psi_{j,m}^r(X_i)(\frac{I(X_i \geq c_{(j+1)})}{m-j} - \frac{I(X_i \geq c_{(j)})}{m-j+1}) \end{aligned}$$

with $\phi_{j,m}^r(X_i) = P\{X_{(j)} \geq c_{(j)}, \ldots, X_{(r)} \geq c_{(r)} | X_i\}$

and $\Psi_{j,m}^r(X_i) = P(X_{(r)} < c_{(r)}, \cdots, X_{(j)} < c_{(j)} | X_i)$. Sarkar (2004)

Under a variety of distributional settings of the $X_i's$, the $c_{(i)}^{'}s$ can be obtained such that the FDR in above formula is controlled at less than or equal to $m_0\alpha/m$, and hence less than or equal to α. For example, when the $X_i's$ are stochastically independent, or have a multivariate distribution exhibiting a positive dependence propoerty in the sense that the $X_i's$ are PRDS on the subset $\{X_i, i \in I_0\}$, $c_{(i)}^{'}s$ satisfying $F(c_{(i)}) = 1 - (m-i+1)\alpha/m, i = 1, \ldots, m$ with $F(\cdot)$ being the common marginal null cumulative distribution function of the $X_i's$, provide a control of the FDR at α (Sarkar,2002). These are the Simes (1986) critical values used in the Benjamini-Hochberg step-up test with independent test statistics. It is important to note that the step-up test with these critical values in the independent case is actaully exactly equal to $m_0\alpha/m$ (Benjamini and Yekutieli 2001; Finner and Roters 2001; Sarkar 2002). The FDR-controlling property of other step-up-step-down procedures for these types of null hypotheses is not clear. There is another step-down procedure suggested by Benjamini and Liu (1999) that controls the FDR at level α. The critical values $c_{(i)}^{'}s$ of this

step-down procedure are such that $F(c_{(i)}) = \left[1 - min(1, \frac{m}{i}\alpha)\right]^{1/i}, \qquad i = 1, \ldots, m.$ This step-down procedure controls the FDR at α for the independent statistics. Sarkar (2002) has strengthened this fact by proving that the FDR-controlling property still holds when the test statistics are positively dependent in the sense of being MTP_2 under any alternatives, and exchangeable under the null distribution.

The proportion of false negatives among the accepted null hypotheses is defined as $N = T/A$ if $A > 0$ and $= 0$ if $A = 0$, and then we define the false negatives rate (FNR) by $E(N)$. The FNR of a generalized step-up-step-down procedure of order r is given by

$$\begin{aligned} FNR &= \frac{1}{r}\sum_{i\in I_1} P\{X_i \le c_{(r)}\} \\ &+ \sum_{i\in I_1}\sum_{j=2}^{r} E[\phi^r_{j,m}(X_i)\{\frac{I(X_i \le c_{(j)})}{j-1} - \frac{I(X_i \le c_{(j)})}{j}\}] \\ &+ \sum_{i\in I_1}\sum_{j=r+1}^{n} E[\Psi^r_{j,m}(X_i)\{\frac{I(X_i \le c_{(j)})}{j} - \frac{I(X_i \ge c_{(j-1)})}{j-1}\}] \end{aligned}$$

Sarkar (2004) A step-down procedure can be used to control the FNR under certain conditions, for example, independence or PRDS, on the test statistics.

The difference $1 - (FDR + FNR)$ indicates the strength of unbiasedness as well as a measure of power of a multiple testing procedure. Between two procedures, the one with higher value of this difference is more powerful, in that it maintains either a higher proportion of corretly accepted null hypotheses or a low proportion of falsely rejected null hypotheses. Based on simulation data from equi-correlated multivariate normals, the performance of the Benjamini-Hochberg test performs better than any other step-up-step-down test.

Sarkar (2003) also proposed single-step FDR and FNR testing procedures. First, under fixed configuration of true and false null hypotheses, inequalities are obtained repre-

senting how the results show FDR-or-FNR-controlling single-step procedure, like Bonferroni or Sidak procedure, can be improved by borrowing information about m_0 or m_1 in the sprit of Benjamini and Hochberg (2000), Benjamini, Krieger, and Yekutieli (2002), Storey (2002) and Storey, Taylor, and Siegmund (2004). Storey, Taylor, and Siegmund (2004) provided procedures modifying the BH procedure using estimates of m_0 and proved that they control the FDR under independence. Two families of procedures ,one modifying the FDR-controlling and the other modifying the FNR-controlling Sidak procedures are proposed. These control FDR or FNR under independence less conservatively than the corresponding families modifying the FDR-or FNR-controlling Bonferroni procedure by using the estimates of m_0 considered in Storey, Taylor, and Siegmund(2004). Sarkar extends Storey's (2002, 2003) result to dependent case by considering a mixture model where different configurations of true and false null hypotheses are assumed to have certain probabilities Sarkar (2004).

However, it was shown that most of all FDR controlling procedures were shown to control the FDR in cases of restricted dependency but they were not designed to make use of the dependency structure to gain more power when possible. Denote the true null hypotheses by $\{H_{01}, \ldots, H_{0m_0}\}$ and the false null hypotheses by $\{H_{11}, \ldots, H_{1m_1}\}$. The corresponding vectors of p-values are $\mathbf{P_0}$ and $\mathbf{P_1}$, respectively. Knowing how $\mathbf{P_0}$ is distrubuted, we can construct more powerful MCPs. Resampling-based FDR controlling procedure along the line of Westfall and Young(1993) for FWE control, use p-value resampling to simulate $\mathbf{P_0}$ and utilize the dependency structure of the data so as to construct more powerful MCPs. p-value resampling is conducted under the complete null hypothesis. The resampling procedure makes use of these simulated sets of p-values. Based on p-value resampling, the FDR of the generic MCP, the FDR local estimators is estimated.

The BH FDR local estimator is defined as $Q_{est}^{BH}(p) = \begin{cases} m \cdot p/r(p) & \text{if } r(p) \geq 1 \\ 0 & \text{otherwise} \end{cases}$

Based on the resampling-based distribution R^*, Benjamini and Yekutieli also introduced two resampling-based estimators differing in their treatment of $s(p)$: point estimator and an upper limit. The first estimator is $r(p) - mp$. Using this downward biased estimator, the resampling-based FDR local estimator is given by

$$Q^*(p) = \begin{cases} E_{R^*} \frac{R^*(p)}{R^*(p)+r(p)-p \cdot m} & \text{if } r(p) - r^*_\beta(p) \geq p \cdot m \\ Pr_{R^*}\{R^*(p) \geq 1\} & \text{otherwise} \end{cases}$$

The second estimator is $r(p)$-$r^*_\beta(p)$, assuming subset pivotality conditioning on $S(p) = s(p)$, The resampling based $1 - \beta$ FDR upper limit is defined as

$Q^*_\beta(p) = sup_{x \in [0,p]} \begin{cases} E_{R^*} \frac{R^*(p)}{R^*(p)+r(p)-r^*_\beta(p)} & \text{if } r(p) - r^*_\beta(p) \geq 0 \\ Pr_{R^*} R^*(p) \geq 1 & \text{otherwise} \end{cases}$

Based on $\hat{Q}$, the FDR local estimator computed, the size q MCP based on the FDR local estimator is :if $k_q = max_k\{\hat{Q}(p_{(k)} \leq q\}$, reject $H^0_{(1)}, \ldots, H^0_{(k_q)}$ Yekutieli and Benjamini (1999).

The traditional FDR controlling procedures involve sequential p-value rejection methods. For example, Benjamini and Hochberg (1995) provided a sequential p-value method. A sequential p-value method gives us an estimate $\hat{k}$ that leads to reject $p_{(1)}, p_{(2)}, \ldots, p_{(\hat{k})}$, where $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(\hat{k})}$ are the ordered observed p-values. However, $\hat{k}$ may not be reliable case by case. Secondly, the method controls the error rate for all possible values of m_0, which is the number of true null hypotheses simultaneously without any information about m_0. Instead of fixing the error rate and then estimating k, we propose the opposite approach to fix the rejection region and then estimate α.

Storey (2002) suggested an updated version of FDR called pFDR, which is defined as

the conditional FDR given that there is at least one rejection. pFDR protects false discovery better than FDR since it is conditioned on the occurence of discovery. Let $FDR(t)$ denote the FDR when rejecting all null hypotheses with $p_i \leq t$ for $i = 1, \ldots, m$. For $t \in [0, 1]$, let $V(t), S(t)$, and $R(t)$ denote the number of $\{\text{null } p_i : p_i \leq t\}$, the number of $\{\text{alternative } p_i : p_i \leq t\}$, and $V(t)+S(t)$, respectively. In terms of these empirical processes, $FDR(t) = E[\frac{V(t)}{R(t)\vee 1}]$. Similarly, $pFDR = E(\frac{V(t)}{R(t)}|R(t) > 0)$.

Storey proposed a bootstrap-based algorithm to control FDR and $pFDR$. Similarly, $pFDR(\lambda)$ is estimated by $\hat{pFDR}_(\lambda)(t) = \frac{\hat{\pi}_0(\lambda)t}{\hat{Pr}(P\leq t)}\{1-(1-t)^m\}$. For B bootstrap samples of $p_1, \ldots, p_m$, caculate the bootstrap esimates $\hat{pFDR}^{*b}_{\lambda}(b = 1, \ldots, B)$.

Form a $1-\alpha$ upper confidence interval for $pFDR(t)$ by taking the $1-\alpha$ quantile of the $\hat{pFDR}^{*b}_{\lambda}(t)$ as the upper confidence bound. Since FDR is not conditioned on at least one rejection occuring, we can set $\hat{FDR}_{\lambda}(t) = \frac{\pi_0(\lambda)t}{\hat{Pr}(P\leq t)}$. Tibshirani et al (2001) develop the software package SAM applying this approach.

Let's look at the finite sample setting of m.

Theorem 1.2.1 *If the p-values for the true null hypotheses are independent and have uniform distribution,* $E\{\hat{pFDR}_{\lambda}(t)\} \geq pFDR(t)$ *and* $E\{\hat{FDR}_{\lambda}(t)\} \geq FDR(t)$ *for all* t *and* π_0.

$\hat{FDR}_{\lambda}(t)$ *and* $\hat{pFDR}_{\lambda}(t)$ *is a conservative point esimate of* $FDR_{\lambda}(t)$ *and* $pFDR_{\lambda}(t)$, *respectively.*

Storey (2002)

Theorem 1.2.2 *If the p-values corresponding to the true null hypotheses are independent, then, for* $\lambda > 0$, $FDR\{(t_{\alpha}(\hat{FDR}^{*}_{\lambda}\} \leq (1-\lambda^{\pi_0 m})\alpha \leq \alpha$.

Storey et al. (2004)

Hence, the thresholding procedure using $\hat{FDR}_{\lambda}(t)$, which is a family of conservatively biased estimate of $FDR(t)$, controls the FDR at prespecified level α in the strong

sense. So, the goals of the BH procedure and this procedure can be met with this one family of estimates.

Now, let's look at the case when m is large. For large m, the assumption of independence can be weakened to "weak dependence". The following three assumptions are needed for large m results.

$$\begin{aligned} lim_{m\to\infty}\{\tfrac{V(t)}{m_0}\} &= G_0(t), lim_{m\to\infty}\{\frac{S(t)}{m_1}\} = G_1(t) \quad a.s.t \in (0,1] \qquad (1), \\ 0 < G_0(t) \leq t &\quad , t \in (0,1]; \qquad (2), \\ lim_{m\to\infty}\tfrac{m_0}{m} &\equiv \pi_0 \qquad (3). \end{aligned}$$

where G_0 and G_1 are continous functions.

$$F\hat{D}R^{\infty}_{\lambda}(t) = \{\frac{1-G_0(t)}{1-\lambda}\pi_0 + \frac{1-G_1(t)}{1-\lambda}\pi_1\}G_0(t)/\{\pi_0 G_0(t) + \pi_1 G_1(t)\}.$$

This is the pointwise limit of $F\hat{D}R_{\lambda}(t)$ under the assumptions (1)-(3)Storey et al. (2004).

Theorem 1.2.3 *Suppose that the convergence assumptions of equations* $(1)-(3)$ *hold. For each* $\delta > 0$,
$lim_{m\to\infty} inf_{t\geq\delta}\{F\hat{D}R_{\lambda} - FDR_{\lambda}(t))\} \geq 0$ *and* $lim_{m\to\infty} inf_{t\geq\delta}\{F\hat{D}R_{\lambda} - \frac{V(t)}{R(t)\vee 1}\} \geq 0$ *with probability 1.*

Storey et al. (2004)

Hence, an estimate of $FDR(t)$ proposed in Storey (2002) a conservative estimate of the error rate over all significance regions simultaneously in the asymptotic setting. Thus, the goals of the traditional sequential p-value method and a new method are equivalent.

Let's investigate the statistical properties of the pFDR. First, under the assumption

that the test statistics have a random mixture of the null and alternative distributions, the pFDR can be restated as a simple Bayesian posterior probability as shown below. Second, these properties remain asymtotically under general conditions, even under certain form of dependence in that the realizd V/R, the FDR, and the pFDR all converge to the Bayesian form of the pFDR simultaneously over all significance regions. Third, the pFDR can be used to define the q-value, a natural pFDR analogue to the p-value.

For an observed statistic $T = t$, the q-value of t is defined as $q(t) = inf_{\{\Gamma_\alpha :\, t \in \Gamma_\alpha\}}\{pFDR(\Gamma_\alpha)\}$. In words, the q-value is a measure of the strenth of an observed statistic with respect to pFDR. The q-value is "posterior Baysian p-value"-the minimum posterior probability $H = 0$ over all significance containing the statistic. Fourth, the pFDR has a connection to classification theory, and the set of Bayes rule can be used to minimize $(1 - w) \cdot pFDR + w \cdot pFNR$, where the pFNR is the natural counterpart to the pFDR,where $pFNR = E[\frac{T}{W}|W > 0]$.

We have shown that in both finite sample and asymptotic settings, the goals of two approaches are equivalent. Using this new approach, we reject a greater number of hypotheses while controlling the same error rate as the Benjamini and Hochberg (1995) method, which leads to higher power. If the number of tests is large, it's appropriate to tolerate more than one false rejection provided the number of such cases is controlled, therefore increasing the ability of the procedure to detect false null hypotheses. E.L.Lehmann and J.P.Romano (2005) derived single-step and stepdown k-FWER procedures, controlling the probability of k or more false rejections, without any assumptions about the dependence structure of the p-values Lehmann and Romano (2005).

Theorem 1.2.4 *For testing* $H_i : P \in w_i, \qquad i = 1, \ldots, m$, *suppose* $\hat{p}_i$ *satisfies the following:* $P\{\hat{p}_i \leq u\} \leq u$ *for any* $u \in (0, 1)$ *and any* $P \in w_i$. *Consider the procedure*

that rejects any H_i for which $\hat{p}_i \leq k\alpha/s$. This procedure controls the k-FWER, so that $P\{\hat{p}_i \leq u\} \geq P\{X \in S_i(u)\}$. holds. Equivalently, if each of the hypotheses is tested at level $k\alpha/s$, then the k-FWER is controlled.

Theorem 1.2.5 *(i) Let the α_i be given below.*

$$\alpha_i = \begin{cases} \frac{k\alpha}{m} & i \leq k \\ \frac{k\alpha}{m+k-i} & i > k \end{cases}$$

For any $i \geq k$ there exists a joint distribution for $\hat{p}_1, \ldots, \hat{p}_s$ such that $m + k - i$ of the $\hat{p}_i$ are uniformly distributed on (0,1) and the following holds.

$$P\{\hat{p}_{(1)} \leq \alpha_1, \hat{p}_{(2)} \leq \alpha_2, \ldots, \hat{p}_{(i-1)} \leq \alpha_{i-1}, \hat{p}_{(i)} \leq \alpha_i\} = \alpha.$$

(ii) For testing $H_i : P \in w_i, i = 1, \ldots, m$, suppose $\hat{p}_i$ satisfies the following: $P\{\hat{p}_i \leq u\} \leq u$ for any $u \in (0, 1)$. Let $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_m$ be constants. If $\hat{p}_{(1)} > \alpha_1$, reject no null hypotheses. If $hatp_{(1)} \leq \alpha_1, \ldots, \hat{p}_{(r)} > \alpha_r$, reject hypotheses $H_{(1)}, \ldots, H_{(r)}$. For this stepdown procedure with α_i , one cannot increase even one of the constants α_i (for $i \geq k$) without violating the k-FWER.

Lehmann and Romano (2005)

Lehmann and Romano (2005) proposed one stepdown procedure to control the FDP under mild conditions on the dependence structure of p-values. $\hat{p}_1, \ldots, \hat{p}_s$ denotes the p-values of the individual tests. Also let $\hat{q}_1, \ldots, \hat{q}_{|I|}$ denote the p-values corresponding to the $|I| = |I(P)|$ture null hypotheses. So $q_i = p_{j_i}$, where $j_1, \ldots, j_{|I|}$ correspond to the indices of the true null hypotheses. Also let $\hat{r}_1, \ldots, \hat{r}_{s-|I|}$ denote the p-values of the false null hypotheses.

Theorem 1.2.6 *Assume the following condition:* $P\{\hat{q}_i \leq u|\hat{r}_1, \ldots, \hat{r}_{s-|I|}\} \leq u$, *the stepdown procedure with* α_i *given by* $\alpha_i = \frac{(\lfloor \gamma i \rfloor + 1)\alpha}{s + \lfloor \gamma i \rfloor + 1 - i}$ *controls the FDP in the sense that* $P\{FDP > \gamma\} \leq \alpha$. *They also proposed more conservative stepdown methods without any dependence assumptions.*

Lehmann and Romano (2005)

Lehmann and Romano constructed stepdown procedures to control the FDR with a dependence assumptions on the joint distribution of the p-values.

Theorem 1.2.7 *For testing* $H_i : P \in w_i, i = 1, \ldots, s$, *suppose* $\hat{p}_i$ *satisfies* $P\{\hat{p}_i \leq u\} \leq u$ *for any* $u \in (0, 1)$.*Consider the stepdown procedure with constants* $\alpha_i^* = min\{\frac{s\alpha}{(s-i+1)^2}, 1\}$ *and assume the condition* $P\{\hat{q}_i \leq u|\hat{r}_1, \ldots, \hat{r}_{s-|I|}\} \leq u$. *Then* $FDR \leq \alpha$.

Lehmann and Romano (2005)

1.2.5 Recent Proposals For DNA Microarray Experiments

Let us review the recent proposals for DNA Microarray Experiments. Golub et al. (1999) proposed neighborhood analysis for identifying genes that are differentially expressed in patients with two types of leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The authors computed a test statistic t_j for each gene, $t_j = \frac{\overline{x}_{1j} - \overline{x}_{2j}}{\overline{s}_{1j} + \overline{s}_{2j}}$ where $\overline{x}_{kj}$ and s_{kj} denote the average and standard deviation of the expression measures of gene j in the class $k = 1, 2$ samples,respectively. Golub et al.used the term *neighborhood* to refer to sets of genes with test statistics T_j greater in absolute value than a given critical value $c > 0$, sets of rejected hypotheses $\{j : T_j \geq c\}$ or $\{j : T_j \leq -c\}$. The ALL/AML labels were permuted $B = 400$ times to estimate the complete null distribution of the numbers $R(c) = V(c) = \sum_{j=1}^{m} I(T_j \geq c)$ of false positives for different critical values c. However, there are some limitations in this approach. Golub et al. did not provide

further guidelines for selecting the critical value c or discussion of the Type I error control of the procedure. The error rate controlled by this analysis is in fact a p-value for the number of rejected hypotheses under the complete null, $G(c) = Pr(R(c) \geq r(c)|H_0^C)$. A critical value c is selected to control this unusual error at a preassigned nominal level α. $G(c)$ is not, in fact, decreasing overall and there may be several values of c with $G(c) = \alpha$. Dudoit, Shaffer, and Boldrick (2002)considered a step-down and a step-up version of neighborhood analysis in order to handle the monotonicity of $G(c)$ and they derived corresponding adjusted p-values. Since neighborhoood analysis is based on the distribution of order statistics under the complete null, this analysis controls the Type I error rate in the weak sense. The step-down version controls the FWER weakly, whereas the step-up analysis does not control any error rate.

We consider the Significance Analysis of Microarrays. The earlier version of SAM procedure (Efron et al.,2000) and Tusher, Tibshirani and Chu (2001) version of SAM procedure seems very similar. SAM procedure from Tusher, Tibshirani and Chu(2001).

1. compute a test statistic t_j for each gene j and define order statistics $t_{(j)}$ such that $t_{(1)} \geq t_{(2)} \cdots \geq t_{(m)}$.

2. Perform B permutations of the response/covariates $y_1, \ldots, y_n$. For each permutation b compute the test statistic $t_{j,b}$ and the corresponding order statistics $t_{(1),b} \geq t(2), b \geq \cdots t(m), b$.

3. From the B permutations, estimate the expected value(under the complete null) of the order statistics by $\bar{t}_{(j)} = (1/B) \sum_b t_{(j),b}$.

4. Form a quantile-quantile plot of the observed $t_{(j)}$ versus the expected $\bar{t}_{(j)}$.

5. For a fixed threshold Δ, let

$j_0 = max\{j : \bar{t}_{(j)} \geq 0\}, j_1 = max\{j \leq j_0 : t_{(j)} - \bar{t}_{(j)} \geq \Delta\}$ and

$j_2 = min\{j > j_0 : t_{(j)} - \bar{t}_{(j)} \leq -\Delta\}$. All genes with $j \leq j_1$ are called *significant positive* and all genes with $j \geq j_2$ are called *significant negative*. Define the uppper cut point, $cut_{up}(\Delta) = min\{t_{(j)} : j \leq j_1\} = t_{(j_1)}$, and the lower cut point, $cut_{low}(\Delta) = max\{t_{(j)} : j \geq j_2\} = t_{j_2)}$. If no such $j_1(j_2)$ exists, set $cut_{up}(\Delta) = \infty(cut_{low}(\Delta) = -\infty)$.

6. For a given threshold Δ, the expected number of false positives, PFER, is estimated by computing for each of the B permutations the number of genes with $t_{j,b}$ above $cut_{up}(\Delta)$ or below $cut_{low}(\Delta)$, and averaging this number over permutations.

7. A threshold Δ is chosen to control the expected number of false positives, PFER, under the complete null, at an acceptable nominal level.

The only difference between the latter version of SAM and standard procedures which rejects the null H_j for $|t_j| \geq c$ is in the use of asymmetric critical values chosen from a Q-Q plot. Otherwise, SAM does not provide any new definition of Type I error rate nor any new procedure for controlling this error rate. However, there are number of problems linked to the implementation of the Tusher, Tibshirani and Chu (2001) SAM procedure.

1.2.6 Classification Of Genes

There are various clustering techniques of the genes which has their own issues. First, various clustering algorithms produce different sets of clusters. There is not a standard criterion or algorithm for choosing a cutoff point for a dendrogram. Second, a more fundamental issue is which samples will be clustered in the first place, and on which genes (for example, whether or not to include control samples). Third, the difficulties are inherent in not only assessing cluster reliability, but also determining the number of clusters. As a typical statistical clustering method, k-means method are not flexible: It is not effective for handling different within-variations (variations

within each cluster) and for finding outliers. For these problems, a model-based clustering method using a normal mixture model and a well-conceived penalized likelihood was proposed Fujisawa et al. (2004). Ridge regression, Principal components regression, and Partial least squares regression which are regularized regression models were proposed to deal with classification problems in gene expression studies Ghosh (2003). These regression procedures were used to classify the genes with the optimal scoring algorithm. A combination of the results across several microarray experiments helps to gain significant increases in power of identifying differentially expressed genes.

1.2.7 The Chen-Stein Method

The following method may be a useful tool to approximate a distribution to the Poisson distribution. Arratia, Goldstein and Gordon (1989) verify the following theorem. Write $\mathcal{L}(Y)$ for the law of Y.
One may write $||\mathcal{L}(Y_0) - \mathcal{L}(Y_1)||$=2sup$_A|P(Y_0 \in A) - P(Y_1 \in A)|$=2minP$(Y_0 \neq Y_1)$. For each $\alpha \in I$, let X_α be a Bernoulli random variable with p_α=P$(X_\alpha = 1)> 0$. Let $W = \sum_{\alpha \in I} X_\alpha$ and $\lambda = EW$. Z is denoted as a Poisson random variable with the same mean as W. For each $\alpha \in I$, we choose $B_\alpha \subset I$ with $\alpha \in B_\alpha$. Define

$$\begin{aligned} b_1 &= \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta, \\ b_2 &= \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} p_{\alpha\beta}, \\ \textit{and} \quad b_3 &= \sum_{\alpha \in I} E|E\{X_\alpha - p_\alpha|\sigma(X_\beta : \beta \notin B_\alpha)|, \end{aligned}$$

where $p_{\alpha\beta} = E[X_\alpha X_\beta]$. In applications where X_α is independent of the collection $|X_\beta : \beta \notin B_\alpha|$, the term b_3=0.

Theorem 1.2.8 $||\mathcal{L}(W) - \mathcal{L}(Z)|| \leq 2(b_1 + b_2 + b_3)$

Arratia et al. (1990)

1.3 Overview of Research

This dissertation was motivated by high-dimension low-sample size perspective for identifying differentially expressed genes among thousands of genes. It involves defining an appropriate multiple testing procedure with the associated test statistics for each gene. The traditional approach to the multiplicity is familywise error rate (FWER), but it is known to be unduly conservative. As stated before, false discovery rate (FDR) is the better procedure in the microarray setting and genomic sequence, in that we are interested in detecting as many differentially expressed genes as possible. Main concerns are to estimate the underlying null distribution of test statistics, that is, genes. Many researchers tried to oversimplify this distribution under independence or restrictive dependence structure among the genes. Or they have exploited unfeasible conditions under which central limit theorems apply, resulting in too much restrictive mathematical assumptions. It is natural that we don't know the real dependence structures among the genes. Besides, the assumption among the genes has been a practical issue and checking this distributional assumption in a whole genomic study may not be easy. In these sense, we use false discovery rate procedure to overcome these difficulties. The Chen-Stein method addressed in section 1.2.7 plays a fundamental role in deriving an appropriate false discover rate. This theorem presents alternative limit theorems and its ramification wherein Poisson approximation for more general dependent sequences. This is attainable under mild regularity conditions regarding the dependence of the genes: the classification into two subsets of non-differentially expressed genes and

differentially expressed genes crucial to sort plausible dependence patterns out. A suitable false discover rate procedure must provide the exact estimation of true FDR and attain better power than other procedures.

Developing this false discovery rate in the high dimension low sample size data in miroarray experiment is discussed in Chapter 2. First, the first-stage FDR procedure is derived and then we take another testing procedure to this procedure. This procedure is designed to minimize V and maximize S in FDR. We do not estimate a smaller false discovery rate than truly exists. In Chapter 3, we address the complexity of high-dimension categorical genomic models that the full multisample, multi-dimensional multinomial law may not be reasonable. The categories are not even ordered, and a stochastic ordering may not be applicable. Diversity measures such as the Hamming distance can have stochastic ordering. But individual statistics, even coordinatewise ones, based on Hamming distance do not have a known null hypothesis distribution. In these sense, we use jackknife variance estimation and permutation distribution to construct some permutation tests. A pseudo-marginal approach based on these facts is used to construct an appropriate marginal test statistics. For the problem of small sample size along with discrete p-values, we simulate the permutation distribution of this marginal test statistics and use the exact permutation theory. In Chapter 4, we develop two nonstandard robust methods to classify genes in order-restricted inference and small size perspective, without assuming a linear or any specific nonlinear form that other researchers have used. One method is to construct a locally most powerful test statistics using a suitable rank scores, instead of deriving a uniformly most powerful test statistics which is not feasible in our studies. Gene expression data usually has many outliers, and is highly probable to be noisy. The small sample sizes results in unreliable estimation of variance. Because of the large number of genes and small number of arrays, and

higher signal-noise ratio in microarray data, traditional approaches do not work properly. For these reasons, we propose the test statistics for each gene based on robust M-estimator insensive to outlier arrays. Union-Intersection principle is used to construct these test statistics. However, such tests are in general conservative. The locally smoothed Kendall's tau statistics is also illustrated in microarray study with continuous responses. In Chapter 5, numerical studies are conducted assessing proposed false discovery rate and the associated test statistics. Most of the researchers have evaluated the performance of their FDR procedures only, not comparing with other procedures. They have neglected a numerical study with application to real data example. In this chapter, our numerical studies present simulated data as well as three real data examples, by comparing with other conventional procedures.

CHAPTER 2

FALSE DISCOVERY RATE IN MICROARRAY STUDIES

2.1 Dependence structures among tested genes

2.1.1 Introduction

DNA microarrays have been used for monitoring expression levels of thousands of genes simultaneously and ultimately, selecting differentially expressed genes: Multiple hypothesis testing of thousands of gene expression levels under different experimental conditions. As stated before, one wish to find as many differentially expressed genes as possible. In these senese, FDR has been mostly used in microarray studies rather than FWER. However, high correlation structures between test genes due to the gene coregulation patterns and dependency in the measurement errors has been a great concern in developing an appropriate FDR controlling procedure. Ignoring complexed correlation structures among tested genes results in increase of the variability of the FDR estimate, which inevitably misses a large portion of the informations produced by a microarray experiment.

Mostly, current FDR controlling procedures developed thus far control the FDR under independence or positive regression dependence using MTP_2 property or do not exploit the joint distribution of the test statistics, resulting in unduly conservativeness. This motivates us to find out another approach to account for more general dependence structures.

Most of the FDR controlling procedures in microarray data focused on estimating an underlying null distribution of genes (or test statistics). This required a rather restricted dependence assumption among thousands of genes. In fact, correlation structures among tested genes (or p-values) is still unknown. This motivates us to take into account that test genes might have more general dependence structures. They have assumed some regularity conditions under which central limit theorems may work. Unfortunately, without some knowledge of any positional ordering of the genes, it was hard to find these conditions. For these reason, we propose a new approach to false discovery rates, which directly estimate the distributions of V and R, accounting for more general dependence structures among tested genes. One fundamental property underlying the analysis of microarray data is that p-values from non-differentially expressed genes, the null hypotheses are uniformly distributed on (0,1) (Casella and Berger, 1990). There are another two assumptions behind our model: the classification into two subsets of non-differentially expressed genes and differentially expressed genes. One important assumption is that any correlation between a non-differentially expressed gene (a null hypothesis) and a differentially expressed gene (an alternative hypothesis) appears to be negligible. The other important assumption is that any correlation between non-differentially expressed genes is small. Non-differentially expressed gene expression levels having stochastically small expression levels may not have significant interaction among themselves as well as differentially expressed genes. On the other hand, stochastic

dependence among differentially expressed genes may be significant. Incorporating these fairly milder regularity conditions, this problem motivates us to utilize alternative limit theorems by the Chen-Stein theorem where Poission approximations for more general dependent sequences are allowed.

2.1.2 Model

Consider a DNA microarray expression data on large m genes, of which m_0 is the number of non-differentially expressed genes and m_1 is the number of differentially expressed genes. $\frac{m_1}{m}$ is assumed to be close to 0 ,that is, m_1 may not be small. Numerous false positives are due to the large number of non-differentially expressed genes. There is a null hypothesis H_i for each gene i and rejection of H_i corresponds to declaring that a gene i is differentially expressed. For each hypothesis H_i, a test statistic T_i is calculated with the corresponding $P_i = Pr(|T_i| \geq t_i)$. Let V_{m_0} denote the number of genes among the m_0 genes erroneously rejected, S_{m_1} the number of genes among the m_1 genes, declared to be differentially expressed and $R_{m_0} = V_{m_0} + S_{m_1}$ be the number of genes rejected by a procedure. Let α_m denote Pr(a non-differentially expressed genes will be errorneously rejected), for example, $\alpha_m = \frac{\alpha}{m}$. Let α_m^* denote Pr(a differentially expressed gene will be declared to be differentially expressed), for example, $\alpha_m^* = 1 - (1 - \alpha_m)^\lambda$. α_m^* is assumed to be greater than α_m.

2.1.3 Distributions

Theorem 2.1.1 *V_{m_0}, S_{m_1}, and R_{m_0} follow Poisson distribution with rates μ_{m_0}, λ_{m_1}, and $\mu_{m_0}^*$, respectively, where $\mu_{m_0} = m_0 \cdot \alpha_m, \lambda_{m_1} = m_1 \cdot \alpha_m^*$, and $\mu_{m_0}^* = m_0 \cdot \alpha_m + m_1 \cdot \alpha_m^*$.*

The theorem follows from the Chen-Stein methods, whose proof is given in Appendix A. For this theorem to hold we must use the two assumptions described above. In fact. if two variables are Poisson variables, it is obvious that given the sum of two variables, one variable has Binomial distribution. The following elementary corollaries, whose proof is omitted, incorporate the distributions we need for deriving the FDR.

Corollary 2.1.2 *V_{m_0}, given $R_{m_0} = r$, follows Binomial distribution with r and* $\frac{\mu_{m_0}}{\mu^*_{m_0}} = \frac{m_0\alpha_m}{m_0\alpha_m+(m-m_0)\alpha^*_m}$.

Corollary 2.1.3 *S_{m_1} given $R_{m_0} = r$ follow Binomial distribution with r and* $1 - \frac{\mu_{m_0}}{\mu^*_{m_0}} = \frac{(m-m_0)\alpha^*_m}{m_0\alpha_m+(m-m_0)\alpha^*_m}$.

2.1.3.1 FDR

Using the distribution results above, we prove the following theorem giving an explicit expression of FDR. We consider the large m case in that data of interest is high-dimension (large m) genomic data.

Theorem 2.1.4 $FDR = \frac{1}{1+\frac{m_1}{m_0}(\frac{\alpha^*_m}{\alpha_m})}(1 - exp(-(m\alpha_m + m_1(\alpha^*_m - \alpha_m)))$

Proof.
[Proof of the Main Theorem]

$$
\begin{aligned}
FDR &= E(\frac{V_{m_0}}{R_{m_0}}|R_{m_0} > 0) \cdot Pr(R_{m_0} > 0) \\
&= \sum_{r=1}^{m} E[\frac{V_{m_0}}{R_{m_0}}|R_{m_0} = r] \cdot Pr(R_{m_0} = r|R_{m_0} > 0) \cdot Pr(R_{m_0} > 0) \\
&= \sum_{r=1}^{m} \frac{1}{r} \cdot r \cdot \frac{\mu_{m_0}}{\mu^*_{m_0}} \cdot \frac{exp(-\mu^*_{m_0})(\mu^*_{m_0})^r}{r!} / Pr(R_{m_0} > 0) \cdot Pr(R_{m_0} > 0) \\
&= \frac{\mu_{m_0}}{\mu^*_{m_0}}(1 - exp(-\mu^*_{m_0})) \\
&= \frac{m_0\alpha_m}{m_0\alpha_m + (m - m_0)\alpha^*_m} \cdot (1 - exp(-(m_0\alpha_m + (m - m_0)\alpha^*_m))) \\
&= \frac{1}{1 + \frac{m_1}{m_0}(\frac{\alpha^*_m}{\alpha_m})}(1 - exp(-(m\alpha_m + m_1(\alpha^*_m - \alpha_m)))
\end{aligned}
$$

In fact, as m goes to infinity, $m_1(\alpha_m^* - \alpha_m)$ becomes small but $m\alpha_m$ becomes large. $exp(-(m\alpha_m + m_1(\alpha_m^* - \alpha_m))$ goes to 0. The ratio $\frac{m_1}{m_0} \cdot \frac{\alpha_m^*}{\alpha_m} (= \frac{\lambda_{m_1}}{\mu_{m_0}})$ is much larger than 1. Since the dominating term $\frac{1}{1+\frac{m_1}{m_0}(\frac{\alpha_m^*}{\alpha_m})}$ becomes much smaller than 1, the FDR becomes much smaller. Hence, FDR is controlled at prespecified level α , where $0 < \alpha < 1$.

2.1.4 Stochastic ordering

In this section, we prove some elementary stochastic ordering results for FDR and FNR. These results are used to prove the monotonicities of FDR and FNR. Let V_{m_0-1}, S_{m_1+1}, and R_{m_0-1} be the number of genes among the $m_0 - 1$ genes erroneously rejected, the number of genes among the $m_1 + 1$ genes, declared to be differentially expressed and the number of genes rejected by a procedure, respectively after a non-differentially expressed gene becomes infected to a differentially expressed gene. Let $F_{V_{m_0}}$, $F_{S_{m_1}}$, and $F_{R_{m_0}}$denote the distribution functions of V_{m_0}, S_{m_1},and R_{m_0} respectively. Likewise, $F_{V_{m_0-1}}$, $F_{S_{m_1+1}}$, $F_{R_{m_0-1}}$denote the distribution functions of V_{m_0-1}, S_{m_1+1}, and R_{m_0-1}, respectively. In fact, the relationship between V_{m_0-1} and V_{m_0} is defined based on the probability α_m.

$$Pr\{V_{m_0-1} = v - 1 | V_{m_0} = v\} = Pr\{P_i < c_\alpha\} = \alpha_m \quad i = 1, 2, \ldots, m_0.$$

$$Pr\{V_{m_0-1} = v | V_{m_0} = v\} = Pr\{P_i \geq c_\alpha\} = 1 - \alpha_m \quad i = 1, 2, \ldots, m_0.$$

Theorem 2.1.5 $V_{m_0} >^{st} V_{m_0-1}$.

Proof.

[Proof of the Main Theorem]

$$\begin{aligned} F_{V_{m_0-1}} &= Pr(V_{m_0-1} \le v) \\ &= E(I(V_{m_0-1} \le v)) \\ &= E(E(I(V_{m_0-1} \le v)|V_{m_0} = v_{m_0})) \\ &= \alpha_m E(I(V_{m_0} - 1 \le v)) + (1-\alpha_m)E(I(V_{m_0} \le v))) \\ &= \alpha_m F_{V_{m_0}}(v+1) + (1-\alpha_m)F_{V_{m_0}}(v) \end{aligned}$$

$$\begin{aligned} F_{V_{m_0-1}}(v) - F_{V_{m_0}}(v) &= \alpha_m F_{V_{m_0}}(v+1) + (1-\alpha_m)F_{V_{m_0-1}}(v) - F_{V_{m_0}}(v) \\ &= \alpha_m (F_{V_{m_0}}(v+1) - F_{V_{m_0}}(v)) \\ &= \alpha_m \cdot Pr(V_{m_0} = v) \ge 0, \qquad v \ge 0. \end{aligned}$$

By the relationship between cumulative distribution function and stochastic ordering, we can prove

$$\overline{F}_{V_{m_0}} \ge \overline{F}_{V_{m_0-1}} \Leftrightarrow V_{m_0} >^{st} V_{m_0-1}.$$

Similarly, there is a relationship between S_{m_1+1} and S_{m_1} is defined based on the probability α_m^*.

$$Pr\{S_{m_1+1} = s+1 | S_{m_1} = s\} = Pr\{P_i < c_\alpha\} = \alpha_m^* \quad i = m_0+1, \ldots, m$$

$$Pr\{S_{m_1+1} = s | S_{m_1} = s\} = Pr\{P_i \ge c_\alpha\} = 1-\alpha_m^* \quad i = m_0+1, \ldots, m.$$

Theorem 2.1.6 $S_{m_1+1} >^{st} S_{m_1}$

Proof.

[Proof of the main theorem]

$$\begin{aligned} F_{S_{m_1+1}} &= Pr(S_{m_1+1} \leq s) \\ &= E(I(S_{m_1+1} \leq s)) \\ &= E(E(I(S_{m_1+1} \leq s)|S_{m_1} = s_{m_1})) \\ &= \alpha_m^* E(I(S_{m_1} + 1 \leq s)) + (1 - \alpha_m^*)E(I(S_{m_1} \leq s))) \\ &= \alpha_m^* F_{S_{m_1}}(s-1) + (1 - \alpha_m^*)F_{S_{m_1}}(s) \end{aligned}$$

$$\begin{aligned} \overline{F}_{S_{m_1+1}}(s) - \overline{F}_{S_{m_1}}(s) &= F_{S_{m_1}}(s) - \alpha_m^* F_{S_{m_1}}(s-1) - (1 - \alpha_m^*)F_{S_{m_1}}(s) \\ &= \alpha_m^*(F_{S_{m_1}}(s) - F_{S_{m_1}}(s-1)) \\ &= \alpha_m^* \cdot Pr(S_{m_1} = s-1) \geq 0, \quad s \geq 1 \end{aligned}$$

The corollary 2.17 directly comes from theorem 2.16, whose proof is omitted.

Corollary 2.1.7 $S_{m_1+1} <^{st} S_{m_1} + 1$

Proof.

$$\begin{aligned} \overline{F}_{1+S_{m_1}}(s) - \overline{F}_{S_{m_1+1}}(s) &= 1 - F_{S_{m_1}}(s-1) - (1 - \alpha_m^* F_{S_{m_1}}(s-1) - (1 - \alpha_m^*)F_{S_{m_1}}(s)) \\ &= -F_{S_{m_1}}(s) + \alpha_m^*(F_{S_{m_1}}(s-1) + (1 - \alpha_m^*)F_{S_{m_1}}(s)) \\ &= (\alpha_m^* - 1) \cdot F_{S_{m_1}}(s-1) + (1 - \alpha_m^*)F_{S_{m_1}}(s)) \\ &= (1 - \alpha_m^*)Pr(S_{m_1} = s-1) \geq 0, \quad s \geq 1 \end{aligned}$$

The relationship between R_{m_0-1} and R_{m_0} is defined in terms of both α_m and α_m^* in

the following theorem.

$$
\begin{aligned}
Pr \quad & \{R_{m_0-1} = r | R_{m_0} = r\} \\
= \quad & Pr(V_{m_0-1} = v, S_{m_1+1} = s | V_{m_0} = v, S_{m_1} = s) \\
+ \quad & Pr(V_{m_0-1} = v-1, S_{m_1+1} = s+1 | V_{m_0} = v, S_{m_1} = s) \\
= \quad & Pr(S_{m_1+1} = s | V_{m_0-1} = v, V_{m_0} = v, S_{m_1} = s) \cdot Pr(V_{m_0-1} = v | V_{m_0} = v, S_{m_1} = s) \\
+ \quad & Pr(S_{m_1+1} = s+1 | V_{m_0-1} = v-1, V_{m_0} = v, S_{m_1} = s) \cdot Pr(V_{m_0-1} = v-1 | V_{m_0} = v, S_{m_1} = s) \\
= \quad & (1-\alpha_m^*) \cdot (1-\alpha_m) + \alpha_m^* \cdot \alpha_m \\
= \quad & 1 - \alpha_m - \alpha_m^* + 2\alpha_m^* \cdot \alpha_m
\end{aligned}
$$

$$
\begin{aligned}
Pr \quad & \{R_{m_0-1} = r+1 | R_{m_0} = r\} \\
= \quad & Pr(V_{m_0-1} = v, S_{m_1+1} = s+1 | V_{m_0} = v, S_{m_1} = s) \\
= \quad & Pr(S_{m_1+1} = s+1 | V_{m_0-1} = v, V_{m_0} = v, S_{m_1} = s) \cdot Pr(V_{m_0-1} = v | V_{m_0} = v, S_{m_1} = s) \\
= \quad & (1-\alpha_m) \cdot \alpha_m^*
\end{aligned}
$$

Theorem 2.1.8 $R_{m_0-1} >^{st} R_{m_0}$

Proof.
[Proof of the main theorem]

$$
\begin{aligned}
\overline{F}_{R_{m_0-1}}(r) \quad & -\overline{F}_{R_{m_0}}(r) \\
= \quad & 1 - (1 - \alpha_m - \alpha_m^* + 2\alpha_m^* \cdot \alpha_m) F_{R_{m_0}}(r) - (1-\alpha_m) \cdot \alpha_m^* F_{R_{m_0}}(r-1) - 1 + F_{R_{m_0}}(r) \\
= \quad & (1-\alpha_m) \cdot \alpha_m^* (F_{R_{m_0}}(r) - F_{R_{m_0}}(r-1)) + (1-\alpha_m^*) \cdot \alpha_m F_{R_{m_0}}(r) \geq 0, \quad r \geq 1
\end{aligned}
$$

Like V_{m_0}, S_{m_1}, and R_{m_0}, V_{m_0-1}, S_{m_1+1}, and R_{m_0-1} follow Poisson distribution with rates $(m_0-1)\alpha_m, (m_1+1)\alpha_m^*$, and $(m_0-1)\alpha_m + (m_1+1)\alpha_m^*$, respectively.

2.1.5 Monotonicity property of FDR

By using stochastic ordering, we can prove the following theorem.

Theorem 2.1.9 *FDR is a monotone decreasing function of* m_1.

Proof.

Given $R_{m_0} > 0$ and $R_{m_0-1} > 0$,

$$\begin{aligned} \frac{V_{m_0-1}}{R_{m_0-1}} &= \frac{V_{m_0} + (V_{m_0-1} - V_{m_0})}{R_{m_0} + (R_{m_0-1} - R_{m_0})} \\ &<^{st} \frac{V_{m_0}}{R_{m_0}} \end{aligned}$$

$$E(\frac{V_{m_0-1}}{R_{m_0-1}} | R_{m_0-1} > 0) < E(\frac{V_{m_0}}{R_{m_0}} | R_{m_0} > 0)$$

$E(\frac{V_{m_0}}{R_{m_0}} | R_{m_0} > 0)$ is a nonincreasing function of m_1. Since $\alpha_m^* \approx \alpha_m$,

$$Pr(R_{m_0-1} > 0) - Pr(R_{m_0} > 0) = exp(-\mu_{m_0}^*) \cdot [1 - exp(-(\alpha_m^* - \alpha_m))] \approx 0$$

.

2.1.6 FNR

We will derive in this section an explicit expression of FNR analogous to that of FDR.

$$\begin{aligned} FNR &= E(\frac{T_{m_1}}{A_{m_0}} | A_{m_0} > 0) \cdot Pr(A_{m_0} > 0) \\ &= E(\frac{m_1 - S_{m_1}}{m - R_{m_0}} | R_{m_0} < m) \cdot Pr(R_{m_0} < m) \\ &= \sum_{r=0}^{m-1} E[\frac{m_1 - S_{m_1}}{R_{m_0}} | R_{m_0} = r] \cdot \frac{Pr(R_{m_0} = r)}{Pr(R_{m_0} < m)} \cdot Pr(R_{m_0} < m) \\ &= \sum_{r=1}^{m-1} \frac{1}{m-r} \cdot [m_1 - E(S_{m_1} | R_{m_0} = r)] \cdot \frac{exp(-\mu_{m_0}^*)(\mu_{m_0}^*)^r}{r!} \\ &= \sum_{r=1}^{m-1} \frac{1}{m-r} \cdot [m_1 - r \cdot \frac{(m-m_0)\alpha_m^*}{m_0\alpha_m + (m-m_0)\alpha_m^*}] \cdot \frac{exp(-\mu_{m_0}^*)(\mu_{m_0}^*)^r}{r!} \end{aligned}$$

2.1.7 Monotonicity property of FNR

Making the similar arguments as we made before the monotonicity property of the FDR, we notice the relationship between FNR and m_1.

Theorem 2.1.10 *FNR is a monotone increasing function of* m_1.

Proof.

Given $m - R_{m_0} > 0$ and $m - R_{m_0-1} > 0$

$$\frac{m_1 + 1 - S_{m_1+1}}{m - R_{m_0-1}} = \frac{m_1 + 1 - S_{m_1} + (1 + S_{m_1} - S_{m_1+1})}{m - R_{m_0} - (R_{m_0-1} - R_{m_0})}$$

where $R_{m_0-1} >^{st} R_{m_0}$ and $S_{m_1+1} <^{st} 1 + S_{m_1}$.

$$\frac{m_1 + 1 - S_{m_1+1}}{m - R_{m_0-1}} >^{st} \frac{m_1 - S_{m_1}}{m - R_{m_0}}$$

$$E(\frac{m_1 + 1 - S_{m_1+1}}{m - R_{m_0-1}} | R_{m_0-1} < m) > E(\frac{m_1 - S_{m_1}}{m - R_{m_0}} | R_{m_0} < m)$$

$$\begin{aligned} Pr(R_{m_0} < m) &= 1 - Pr(R_{m_0} = m) \\ &= 1 - \frac{exp(-m_0 \cdot \alpha_m - m_1 \cdot \alpha_m^*) \cdot (m_0 \cdot \alpha_m + m_1 \cdot \alpha_m^*)^m}{m!} \end{aligned}$$

$$\begin{aligned} Pr(R_{m_0-1} < m) & -Pr(R_{m_0} < m) \\ = & \frac{exp(-m_0 \cdot \alpha_m - m_1 \cdot \alpha_m^*) \cdot (m_0 \cdot \alpha_m + m_1 \cdot \alpha_m^*)^m}{m!} \\ - & \frac{exp(-(m_0 - 1) \cdot \alpha_m - (m_1 + 1) \cdot \alpha_m^*) \cdot ((m_0 - 1) \cdot \alpha_m + (m_1 + 1) \cdot \alpha_m^*)^m}{m!} \\ \leq & \frac{exp(-m_0 \cdot \alpha_m - m_1 \cdot \alpha_m^*)(1 - exp(\alpha_m - \alpha_m^*)) \cdot ((m_0 - 1) \cdot \alpha_m + (m_1 + 1) \cdot \alpha_m^*)^m}{m!} \\ \approx & 0 \end{aligned}$$

2.2 Two stage FDR procedure

2.2.1 Introduction

In last section, we find out the strategy to allow for more general dependence structures among tested genes. Researchers are aware of high rate of false positives in microarray data studies. Many small microarray studies has reported the large FDR problem. We propose two-stage procedure to add the first stage FDR derived in last section to another testing procedure. This leads to not only minimize the FDR level but also increase power. We still have two great concerns involved in developing an appropriate FDR procedure. For FDR estimation purpose, we don't want to report a smaller false discovery rate than truly exists. On the other hand, FDR procedure must be controlled at preassigned level α. Optimal FDR procedure maximizes the expected number of true positives (S) for each fixed level of expected false positives (V), which ideally corresponds to better estimate of false-discovery rates (estimation) and minimized false positives and false negatives (Power) Storey (2007). We develop proposed FDR procedure to achieve these goals. We will show stochastic ordering thoroughly in this section. The number of rejected hypotheses, R, the number of accepted hypotheses in favor of the alternative, the number of true null hypotheses m_1 turn out to be stochastic in nature. It is feasible only when data are continous. However, for the categorical models, another techiniques will be needed. We will investigate this case in details in Chapter 3.

2.2.2 Model

Consider the following two-stage FDR procedure. There is a null hypothesis H_i for each gene i , with the corresponding alternative hypothesis H_i^c and rejection of H_i corresponds to declaring that gene i is differentially expressed. For each hypothesis H_i, a test statistic X_i is caculated with the corresponding $P_i = Pr(X_i \geq x_i)$ for a right-tailed test.

Stage 1 : $H_0 : \bigcap_{i=1}^{m} H_i$ vs $H_1 : \bigcup_{i=1}^{m} H_i^c$.

Let $\alpha_{1m} = Pr(P_i < C_\alpha, \quad i = 1, \ldots, m_0)$ denote Pr(a underexpressed gene will be errorneously rejected at the first stage), for example, $\alpha_{1m} = \frac{\alpha}{m}$. Let $\alpha^*_{1m} = Pr(P_i < C_\alpha, \quad i = m_0 + 1, \ldots, m)$ denote Pr(an overexpressed gene will be declared to be differentially expressed at the first stage), for example, $\alpha^*_{1m} = 1 - (1 - \alpha_{1m})^\lambda$. α^*_{1m} is assumed to be greater than α_{1m}. Assume that the $P_i's$ are uniformly distributed among the m_0 genes. Let $V_{1(m_0)}$ denote the number of genes among the m_0 genes erroneously rejected, $S_{1(m_1)}$ the number of genes among the m_1 genes, declared to be differentially expressed and $R_{1(m_0)} = V_{1(m_0)} + S_{1(m_1)}$ be the number of genes rejected by a procedure. There is a Type I error that inactive genes are declared to be active genes.

Stage 2 : Among the set of genes not rejected at the first stage,$m_0 - R_{1(m_0)}$ genes, repeat performing the same testing procedure with different critical values. Let $\alpha_{2m} = Pr(P_i < C^*_\alpha | P_i > C_\alpha, \quad i = 1, \ldots, m_0)$ denote Pr(a underexpressed gene will be errorneously rejected at the second stage), for example, $\alpha_{2m} = \frac{\alpha}{m}$. Let $\alpha^*_{2m} = Pr(P_i < C^*_\alpha | P_i > C_\alpha, \quad i = m_0 + 1, \ldots, m)$ denote Pr(an overexpressed gene will be declared to be differentially expressed at the second stage), for example, $\alpha^*_{2m} = 1 - (1 - \alpha_{2m})^\lambda$. α^*_{2m} is assumed to be greater than α_{2m}. Assume that the $P_i's$ are uniformly distributed among the $m_0 - R_{1(m_0)}$ genes. Let $V_{2(m_0)}$ denote the number of genes among the $m_0 - V_{1(m_0)}$ genes erroneously rejected, S_{2m_1} the number of genes among the $m_1 - S_{1(m_1)}$ genes, declared to be differentially expressed and $R_{2(m_0)} = V_{2(m_0)} + S_{2(m_1)}$ be the number of genes rejected by a procedure.

2.2.3 Distributions

Making the same arguments in the previous section, using the Chen-Sten method, we derive V, S, and R at both stages.

Theorem 2.2.1 *$V_{1(m_0)}, S_{1(m_1)}$, and $R_{1(m_0)}$ follow Poisson distribution with rates $\mu_{1(m_0)}, \lambda_{1(m_1)}$, and $\mu^*_{1(m_0)}$, respectively, where $\mu_{1(m_0)}(= m_0 \cdot \alpha_{1m})$, $\lambda_{1(m_1)}(= m_1 \cdot \alpha^*_{1m})$, and $\mu^*_{1(m_0)}(= m_0 \cdot \alpha_{1m} + m_1 \cdot \alpha^*_{1m})$.*
Similarly, $V_{2(m_0)}, S_{2(m_1)}$, and $R_{2(m_0)}$, given $V_{1(m_0)}, S_{1(m_1)}$, and $R_{1(m_0)}$,follow Poisson distribution with rates $\mu_{2(m_0)}, \lambda_{2(m_1)}$, and $\mu^*_{2(m_0)}$, respectively, where
$\mu_{2(m_0)}(= (m_0 - V_{1(m_0)}\alpha_{2m})$, $\lambda_{2(m_1)}(= (m_1 - S_{1(m_1)})\alpha^*_{2m})$, and
$\mu^*_{2(m_0)}(= (m_0 - V_{1(m_0)})\alpha_{2m} + (m_1 - S_{1(m_1)})\alpha^*_{2m})$.

The corollaries are directly proven by the theorem above.

Corollary 2.2.2 *$V_{1(m_0)}$, given $R_{1(m_0)} = r_1$, follows Binomial distribution with r_1 and $\frac{\mu_{1(m_0)}}{\mu^*_{1(m_0)}} = \frac{m_0\alpha_{1m}}{m_0\alpha_{1m}+(m-m_0)\alpha^*_{1m}}$.*

Corollary 2.2.3 *$S_{1(m_1)}$ given $R_{1(m_0)} = r_1$ follow Binomial distribution with r_1 and $1 - \frac{\mu_{1(m_0)}}{\mu^*_{1(m_0)}} = \frac{(m-m_0)\alpha^*_{1m}}{m_0\alpha_{1m}+(m-m_0)\alpha^*_{1m}}$.*

2.2.4 $FDR^{(2)}$

Using the distibutional settings of V, S, and R at both stages, we get an explicit form of the FDR. It is important to note that $FDR^{(2)}$ is a little bit smaller than the FDR in section 2.1.2.1. Analogous to the FDR, we derive two stage pFDR in the theorem.

Theorem 2.2.4 $FDR^{(2)} = \frac{1}{1+\frac{m_1(\alpha^*_{1m}\cdot e^{\alpha^*_{2m}+\alpha^*_{2m})}}{m_0(\alpha_{1m}\cdot e^{\alpha_{2m}+\alpha_{2m})}}}$

Proof.

[Proof of the Main Theorem]

$$
\begin{aligned}
FDR^{(2)} &= E(E(\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}}|V_{1(m_0)}=v_1, S_{1(m_1)}=s_1)|R_{1(m_0)}>0)\cdot Pr(R_{1(m_0)}>0)\\
&= E(\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}}|R_{1(m_0)}>0)\cdot Pr(R_{1(m_0)}>0)\\
&= \sum_{r_1=1}^{m} E(\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}}|R_{1(m_0)}=r_1)\cdot Pr(R_{1(m_0)}=r_1|R_{1(m_0)}>0)\cdot Pr(R_{1(m_0)}>0)\\
&= \sum_{r_1=1}^{m} E(\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}}|R_{1(m_0)}=r_1)\cdot Pr(R_{1(m_0)}=r_1>0)\\
&= \sum_{r_1=1}^{m}\sum_{v_1=0}^{m_0}\sum_{v_2=0}^{m_0}\sum_{r_2=0}^{m}\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}}Pr(V_{1(m_0)}=v_1, V_{2(m_0)}=V_2, R_{2(m_0)}=r_2|R_{1(m_0)}=r_1)\\
&\times \frac{Pr(R_{1(m_0)}=r_1)}{Pr(R_{1(m_0)}>0)}\cdot Pr(R_{1(m_0)}>0)\\
&= E(\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}})
\end{aligned}
$$

In fact, $R = R_{1(m_0)} + R_{2(m_0)}$ is always positive in this FDR formula due to two stage rejection procedures. The distribution of $V_{1(m_0)} + V_{2(m_0)}$ is as below.

$$
\begin{aligned}
Pr(V_{1(m_0)}+V_{2(m_0)}=v) &= E(Pr(V_{2(m_0)}=v-v_1|V_{1(m_0)}=v_1))\\
&= \sum_{v_1=0}^{v} Pr_{V_{1(m_0)}}(v_1)Pr_{V_{2(m_0)}|V_{1(m_0)}}(v-v_1)\\
&= \sum_{v_1=0}^{v} e^{-m_0\cdot\alpha_{1m}}\cdot\frac{(m_0\cdot\alpha_{1m})^{v_1}}{v_1!} * e^{-(m_0-v_1)\alpha_{2m}}\cdot\frac{(m_0-v_1)\cdot\alpha_{2m})^{v-v_1}}{(v-v_1)!}\\
&= \sum_{v_1=0}^{v} e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}\cdot e^{\alpha_{2m}\cdot v_1}\cdot\frac{m_0^{v_1}\cdot\alpha_{1m}^{v_1}}{v_1!}\alpha_{2m}^{v-v_1}\frac{(m_0-v_1)^{v-v_1}}{(v-v_1)!}\\
&= m_0^v\cdot\frac{e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}}{v!}\cdot\sum_{v_1=0}^{v}\binom{v}{v_1}\alpha_{2m}^{v-v_1}(\alpha_{1m}e^{\alpha_{2m}})^{v_1}(1-\frac{v_1}{m_0})^{v-v_1}\\
&= m_0^v\cdot\frac{e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}}{v!}\cdot\sum_{v_1=0}^{v}\binom{v}{v_1}\alpha_{2m}^{v-v_1}(\alpha_{1m}e^{\alpha_{2m}})^{v_1}e^{(v-v_1)ln(1-\frac{v_1}{m_0})}\\
&= m_0^v\cdot\frac{e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}}{v!}\cdot\sum_{v_1=0}^{v}\binom{v}{v_1}\alpha_{2m}^{v-v_1}(\alpha_{1m}e^{\alpha_{2m}})^{v_1}e^{(v-v_1)g(v_1)}\\
&= m_0^v\cdot\frac{e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}}{v!}\cdot\sum_{v_1=0}^{v}\binom{v}{v_1}(\alpha_{2m}\cdot e^{g(v_1)})^{v-v_1}(\alpha_{1m}e^{\alpha_{2m}})^{v_1}\\
&= m_0^v\cdot\frac{e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}}{v!}\cdot\sum_{v_1=0}^{v}\binom{v}{v_1}\cdot e^{g(v_1)}\theta^{v-v_1}(1-\theta)^{v_1}\cdot(\alpha_{2m}+\alpha_{1m}e^{\alpha_{2m}})^{v}\\
&= m_0^v\cdot\frac{e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}}{v!}\cdot E(e^{g(v_1)})\cdot(\alpha_{2m}+\alpha_{1m}e^{\alpha_{2m}})^{v}\\
&= m_0^v\cdot\frac{e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}}{v!}\cdot(1-\frac{1}{m_0}\cdot\frac{\alpha_{1m}e^{\alpha_{2m}}}{\alpha_{2m}+\alpha_{1m}e^{\alpha_{2m}}})\cdot(\alpha_{2m}+\alpha_{1m}e^{\alpha_{2m}})^{v}\\
&= \frac{e^{-m_0\cdot(\alpha_{1m}+\alpha_{2m})}}{v!}\cdot(1-\frac{1}{m_0}\cdot\frac{\alpha_{1m}e^{\alpha_{2m}}}{\alpha_{2m}+\alpha_{1m}e^{\alpha_{2m}}})\cdot(m_0(\alpha_{2m}+\alpha_{1m}e^{\alpha_{2m}}))^{v}
\end{aligned}
$$

where $\theta = \frac{\alpha_{2m}}{\alpha_{2m}+\alpha_{1m}e^{\alpha_{2m}}}$. The distribution has the form of Poisson distibution with rate $m_0\cdot(\alpha_{1m}+\alpha_{2m})$, except for the second term $(1-\frac{1}{m_0}\cdot\frac{\alpha_{1m}e^{\alpha_{2m}}}{\alpha_{2m}+\alpha_{1m}e^{\alpha_{2m}}})$.

The distribution of $S_{1(m_1)} + S_{2(m_1)}$ is as below.

$$
\begin{aligned}
Pr(S_{1(m_1)} + S_{2(m_1)} = s) &= E(Pr(S_{2(m_1)} = s - s_1 | S_{1(m_1)} = s_1)) \\
&= \sum_{s_1=0}^{s} Pr_{S_{1(m_1)}}(s_1) Pr_{S_{2(m_1)}|S_{1(m_1)}}(s - s_1) \\
&= \sum_{s_1=0}^{s} e^{-m_0 \cdot \alpha^*_{1m}} \cdot \frac{(m_1 \cdot \alpha^*_{1m})^{s_1}}{s_1!} * e^{-(m_0 - s_1)\alpha^*_{2m}} \cdot \frac{(m_1 - s_1) \cdot \alpha^*_{2m})^{s - s_1}}{(s - s_1)!} \\
&= \sum_{s_1=0}^{s} e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})} \cdot e^{\alpha^*_{2m} \cdot s_1} \cdot \frac{m_1^{s_1} \cdot \alpha^{*s_1}_{1m}}{s_1!} \alpha^{*(s-s_1)}_{2m} \frac{(m_1 - s_1)^{s - s_1}}{(s - s_1)!} \\
&= m_1^s \cdot \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{s!} \cdot \sum_{s_1=0}^{s} \binom{s}{s_1} \alpha^{*(v - v_1)}_{2m} (\alpha^*_{1m} e^{\alpha^*_{2m}})^{s_1} (1 - \frac{s_1}{m_1})^{s - s_1} \\
&= m_1^s \cdot \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{s!} \cdot \sum_{s_1=0}^{s} \binom{s}{s_1} \alpha^{*(s - s_1)}_{2m} (\alpha^*_{1m} e^{\alpha^*_{2m}})^{s_1} e^{(s - s_1) ln(1 - \frac{s_1}{m_1})} \\
&= m_1^s \cdot \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{s!} \cdot \sum_{s_1=0}^{s} \binom{s}{s_1} \alpha^{*(s - s_1)}_{2m} (\alpha^*_{1m} e^{\alpha^*_{2m}})^{s_1} e^{(s - s_1) g(s_1)} \\
&= m_1^s \cdot \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{s!} \cdot \sum_{s_1=0}^{s} \binom{s}{s_1} (\alpha^*_{2m} \cdot e^{g(s_1)})^{s - s_1} (\alpha^*_{1m} e^{\alpha^*_{2m}})^{s_1} \\
&= m_1^s \cdot \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{s!} \cdot \sum_{s_1=0}^{s} \binom{s}{s_1} \cdot e^{g(s_1)} \theta^{s - s_1} (1 - \theta)^{s_1} \cdot (\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}})^s \\
&= m_1^s \cdot \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{s!} \cdot E(e^{g(s_1)}) \cdot (\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}})^s \\
&= m_1^s \cdot \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{s!} \cdot (1 - \frac{1}{m_1} \cdot \frac{\alpha^*_{1m} e^{\alpha^*_{2m}}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}}) \cdot (\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}})^s \\
&= \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{s!} \cdot (1 - \frac{1}{m_1} \cdot \frac{\alpha^*_{1m} e^{\alpha^*_{2m}}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}}) \cdot (m_1(\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}))^s
\end{aligned}
$$

where $\theta = \frac{\alpha^*_{2m}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}}$.

The distribution of $S_{1(m_1)} + S_{2(m_1)}$ has the form of Poisson distibution with rate $m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})$, except for the second term $(1 - \frac{1}{m_1} \cdot \frac{\alpha^*_{1m} e^{\alpha^*_{2m}}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}})$.
For convenience of notation, let $V_{1(m_0)} + V_{2(m_0)}$be V_{m_0} and $S_{1(m_1)} + S_{2(m_1)}$ be S_{m_1}.

$$
\begin{aligned}
Pr & (V_{m_0} + S_{m_1} = r) \\
= & \sum_{v=0}^{r} \frac{e^{-m_0 \cdot (\alpha_{1m} + \alpha_{2m})}}{v!} \cdot (m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^v \cdot \frac{e^{-m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})}}{(r - v)!} \cdot (m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m})^{r - v} \\
\times & (1 - \frac{1}{m_0} \cdot \frac{\alpha_{1m} e^{\alpha_{2m}}}{\alpha_{2m} + \alpha_{1m} e^{\alpha_{2m}}}) \cdot (1 - \frac{1}{m_1} \cdot \frac{\alpha^*_{1m} e^{\alpha^*_{2m}}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}}) \\
= & \frac{exp(-m_0 \cdot (\alpha_{1m} + \alpha_{2m}) - m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m}))}{r!} \\
\times & \sum_{v=0}^{r} rCv \cdot (m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^v \cdot (m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m})^{r - v} \\
\times & (1 - \frac{1}{m_0} \cdot \frac{\alpha_{1m} e^{\alpha_{2m}}}{\alpha_{2m} + \alpha_{1m} e^{\alpha_{2m}}}) \cdot (1 - \frac{1}{m_1} \cdot \frac{\alpha^*_{1m} e^{\alpha^*_{2m}}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}}) \\
= & \frac{exp(-m_0 \cdot (\alpha_{1m} + \alpha_{2m}) - m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m}))}{r!} \cdot (m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^r \\
\times & (1 - \frac{1}{m_0} \cdot \frac{\alpha_{1m} e^{\alpha_{2m}}}{\alpha_{2m} + \alpha_{1m} e^{\alpha_{2m}}}) \cdot (1 - \frac{1}{m_1} \cdot \frac{\alpha^*_{1m} e^{\alpha^*_{2m}}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}})
\end{aligned}
$$

The distributions of V_{m_0} and S_{m_1} do not have the exact forms of Poisson distibution. However, based on the distribution of V_{m_0} and S_{m_1}, we can derive the conditional distribution of V_{m_0}, given $V_{m_0} + S_{m_1} = r$ and prove that it is a Binomial distribution with r and p.

$$
\begin{aligned}
Pr \quad & (V_{m_0} = v | V_{m_0} + S_{m_1} = r) \\
= \quad & \frac{Pr(V_{m_0} = v, S_{m_1} = r - v)}{Pr(V_{m_0} + S_{m_1} = r)} \\
= \quad & \frac{\frac{exp(-m_0 \cdot (\alpha_{1m} + \alpha_{2m}) - m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m}))}{r!(r-v)!}}{\frac{exp(-m_0 \cdot (\alpha_{1m} + \alpha_{2m}) - m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m}))}{r!}} \cdot \frac{(m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^v (m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}))^{r-v}}{(m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^r} \\
\times \quad & \frac{(1 - \frac{1}{m_0} \cdot \frac{\alpha_{1m} e^{\alpha_{2m}}}{\alpha_{2m} + \alpha_{1m} e^{\alpha_{2m}}}) \cdot (1 - \frac{1}{m_1} \cdot \frac{\alpha^*_{1m} e^{\alpha^*_{2m}}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}})}{(1 - \frac{1}{m_0} \cdot \frac{\alpha_{1m} e^{\alpha_{2m}}}{\alpha_{2m} + \alpha_{1m} e^{\alpha_{2m}}} \cdot (1 - \frac{1}{m_1} \cdot \frac{\alpha^*_{1m} e^{\alpha^*_{2m}}}{\alpha^*_{2m} + \alpha^*_{1m} e^{\alpha^*_{2m}}})} \\
= \quad & \binom{r}{v} p^v (1-p)^{r-v}
\end{aligned}
$$

where p=$\frac{m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}{m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}$
Based on the conditional distribution shown, we can derive two stage FDR as below.

$$
\begin{aligned}
E(\frac{V_{1(m_0)} + V_{2(m_0)}}{R_{1(m_0)} + R_{2(m_0)}}) \quad & = \quad E(\frac{V_{m_0}}{V_{m_0} + S_{m_1}}) \\
& = \quad E(E(\frac{V_{m_0}}{V_{m_0} + S_{m_1}} | V_{m_0} + S_{m_1})) \\
& = \quad E(\frac{V_{m_0} + S_{m_1}}{V_{m_0} + S_{m_1}} \cdot p) \\
& = \quad \frac{m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}{m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})} \\
& = \quad \frac{1}{1 + \frac{m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m})}{m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}}
\end{aligned}
$$

In fact, α_{1m} and α^*_{1m} are estimated by max($V_{1(m_0)}/R_{1(m_0)}$,c_α) and $S_{1(m_1)}/R_{1(m_0)}$, respectively.
Similary, Two stage pFDR is defined as follow.

$$
\begin{aligned}
pFDR \quad & = \quad E(\frac{V_{1(m_0)} + V_{2(m_0)}}{R_{1(m_0)} + R_{2(m_0)}} | R_{1(m_0)} > 0) \\
& = \quad \frac{1}{1 + \frac{m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m})}{m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}} / (1 - Pr(R_{1(m_0)} = 0)) \\
& = \quad \frac{1}{1 + \frac{m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m})}{m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}} / (1 - exp(-(m_0 \cdot \alpha_{1m} + m_1 \cdot \alpha^*_{1m})))
\end{aligned}
$$

2.2.5 Stochastic ordering

We can make the same arguments as done in section 2.1.3. Let $V_{1(m_0-1)}, S_{1(m_1+1)}$, and $R_{1(m_0-1)}$ be the number of genes among the $m_0 - 1$ genes erroneously rejected, the number of genes among the $m_1 + 1$ genes, declared to be differentially expressed and the number of genes rejected by the first stage procedure, respectively after a underexpressed gene becomes infected to an overexpressed gene. Let $V_{2(m_0-1)}, S_{2(m_1+1)}$, and $R_{2(m_0-1)}$ be the number of genes among the $m_0 - 1$ genes erroneously rejected, the number of genes among the $m_1 + 1$ genes, declared to be differentially expressed and the number of genes rejected by the second stage procedure, respectively after a underexpressed gene becomes infected to an overexpressed gene. This stochastic ordering property supports the fact that the increase in the proportion of the true null hypotheses (π_0), the greater FDRs are and the smaller FNRs are.

Let $F_{V_{1(m_0)}}$, $F_{S_{1(m_1)}}$, and $F_{R_{1(m_0)}}$denote the distribution functions of $V_{1(m_0)}, S_{1(m_1)}$,and $R_{1(m_0)}$ respectively. Likewise, $F_{V_{1(m_0-1)}}$, $F_{S_{1(m_1+1)}}$, $F_{R_{1(m_0-1)}}$denote the distribution functions of $V_{1(m_0-1)}, S_{1(m_1+1)}$, and $R_{1(m_0-1)}$, respectively. Similarly, Let $F_{V_{2(m_0)}}$, $F_{S_{2(m_1)}}$, and $F_{R_{2(m_0)}}$denote the distribution functions of $V_{2(m_0)}, S_{2(m_1)}$,and $R_{2(m_0)}$ respectively. Likewise, $F_{V_{2(m_0-1)}}$, $F_{S_{2(m_1+1)}}$, $F_{R_{2(m_0-1)}}$denote the distribution functions of $V_{2(m_0-1)}, S_{2(m_1+1)}$, and $R_{2(m_0-1)}$, respectively. We define $\alpha_{1m}, \alpha^*_{1m}, \alpha_{2m}$,and α^*_{2m} in terms of C_α and C^*_α.

$$\begin{aligned} Pr(P_i < C_\alpha) &= \alpha_{1m} \qquad i = 1, 2, \ldots, m_0 \\ &= \alpha^*_{1m} \qquad i = m_0 + 1, 2, \ldots, m \end{aligned}$$

$$\begin{aligned} Pr(P_i < C_\alpha^* | P_i > C_\alpha) &= \alpha_{2m} \qquad i = 1, 2, \ldots, m_0 \\ &= \alpha_{2m}^* \qquad i = m_0 + 1, 2, \ldots, m \end{aligned}$$

where C_α and C_α^* are cutoffpoints of the first stage and the second stage, respectively. In fact, the relationship between $V_{1(m_0-1)}$ and $V_{1(m_0)}$ is explained by α_{1m} and the relationship between $V_{2(m_0-1)}$ and $V_{2(m_0)}$ is explained by α_{2m}.

$$\begin{aligned} Pr\{V_{1(m_0-1)} = v_1 - 1 | V_{1(m_0)} = v_1\} &= \alpha_{1m} \\ Pr\{V_{1(m_0-1)} = v_1 | V_{1(m_0)} = v_1\} &= 1 - \alpha_{1m} \\ Pr\{V_{2(m_0-1)} = v_2 - 1 | V_{2(m_0)} = v_2\} &= \alpha_{2m} \\ Pr\{V_{2(m_0-1)} = v_2 | V_{2(m_0)} = v_2\} &= 1 - \alpha_{2m} \end{aligned}$$

Similarly, we can find out the relationship between $S_{1(m_1+1)}$ and $S_{1(m_1)}$ and the relationship between $S_{2(m_1+1)}$ and $S_{2(m_1)}$.

$$\begin{aligned} Pr\{S_{1(m_1+1)} = s_1 + 1 | S_{1(m_1)} = s_1\} &= \alpha_{1m}^* \\ Pr\{S_{1(m_1+1)} = s_1 | S_{1(m_1)} = s_1\} &= 1 - \alpha_{1m}^* \\ Pr\{S_{2(m_1+1)} = s_2 + 1 | S_{2(m_1)} = s_2\} &= \alpha_{2m}^* \\ Pr\{S_{2(m_1+1)} = s_2 | S_{2(m_1)} = s_2\} &= 1 - \alpha_{2m}^* \end{aligned}$$

Based on these relationships, we can derive the following stochastic orderings.

Theorem 2.2.5 *$V_{1(m_0)} >^{st} V_{1(m_0-1)}$, $S_{1(m_1+1)} >^{st} S_{1(m_1)}$, and $R_{1(m_0-1)} >^{st} R_{1(m_0)}$, where $R_{1(m_0)} = V_{1(m_0)} + S_{1(m_1)}$*

Proof.

[Proof of the main theorem]

$$\begin{aligned} F_{V_{1(m_0-1)}}(v_1) - F_{V_{1(m_0)}}(v_1) &= F_{V_{1(m_0)}}(v_1) - \alpha_{1m} F_{V_{1(m_0)}}(v_1+1) - (1-\alpha_{1m}) F_{V_{1(m_0)}}(v_1) \\ &= -\alpha_{1m} \cdot (F_{V_{1(m_0)}}(v_1+1) - F_{V_{1(m_0)}}(v_1)) \\ &= -\alpha_{1m} \cdot Pr(V_{1(m_0)} = v_1) \le 0, \qquad v_1 \ge 0. \end{aligned}$$

By the relationship between cumulative distribution function and stochastic ordering, we can prove

$$\overline{F}_{V_{1(m_0)}} \ge \overline{F}_{V_{1(m_0-1)}} \Leftrightarrow V_{1(m_0)} >^{st} V_{1(m_0-1)}.$$

As for $S_{1(m_1)}$,

$$\begin{aligned} F_{S_{1(m_1+1)}}(s_1) - F_{S_{1(m_1)}}(s_1) &= F_{S_{1(m_1)}}(s_1) - \alpha^*_{1m} F_{S_{1(m_1)}}(s_1-1) - (1-\alpha^*_{1m}) F_{S_{1(m_1)}}(s_1) \\ &= \alpha^*_{1m} \cdot (F_{S_{1(m_1)}}(s_1) - F_{S_{1(m_1)}}(s_1-1)) \\ &= \alpha^*_{1m} \cdot Pr(S_{1(m_1)} = s_1 - 1) \le 0, \qquad s_1 \ge 1. \end{aligned}$$

$$\overline{F}_{S_{1(m_1)}} \le \overline{F}_{S_{1(m_1+1)}} \Leftrightarrow S_{1(m_1+1)} >^{st} S_{1(m_1)}.$$

Stochastic ordering of $R_{1(m_0)}$ can be defined in terms of $V_{1(m_0)}$ and $S_{1(m_1)}$.

$$\begin{aligned} P\{R_{1(m_0-1)} = r_1 | R_{1(m_0)} = r_1\} &= P(V_{1(m_0-1)} = v_1 - 1, S_{1(m_1+1)} = s_1 + 1 | V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\ &+ P(V_{1(m_0-1)} = v_1 - 1, S_{1(m_1+1)} = s_1 + 1 | V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\ &= P(S_{1(m_1+1)} = s_1 | V_{1(m_0-1)} = v_1, V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\ &\times P(V_{1(m_0-1)} = v_1 | V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\ &+ P(S_{1(m_1+1)} = s_1 + 1 | V_{1(m_0-1)} = v_1 - 1, V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\ &\times P(V_{1(m_0-1)} = v_1 - 1 | V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\ &= (1-\alpha^*_{1m}) \cdot (1-\alpha_{1m}) + \alpha^*_{1m} \cdot \alpha_{1m} \\ &= 1 - \alpha_{1m} - \alpha^*_{1m} + 2\alpha^*_{1m} \cdot \alpha_{1m} \end{aligned}$$

$$
\begin{aligned}
P\{R_{1(m_0-1)} = r_1 + 1 | R_{1(m_0)} = r_1\} &= P(V_{1(m_0-1)} = v_1, S_{1(m_1+1)} = s_1 + 1 | V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\
&= P(S_{1(m_1+1)} = s_1 + 1 | V_{1(m_0-1)} = v_1, V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\
&\times \quad P(V_{1(m_0-1)} = v_1 | V_{1(m_0)} = v_1, S_{1(m_1)} = s_1) \\
&= (1 - \alpha_{1m}) \cdot \alpha^*_{1m}
\end{aligned}
$$

$$
\begin{aligned}
\overline{F}_{R_{1(m_0-1)}}(r_1) - \overline{F}_{R_{1(m_0)}}(r_1) &= 1 - (1 - \alpha_{1m} - \alpha^*_{1m} + 2\alpha^*_{1m} \cdot \alpha_{1m}) F_{R_{1(m_0)}}(r_1) \\
&- (1 - \alpha_{1m}) \cdot \alpha^*_{1m} F_{R_{1(m_0)}}(r_1 - 1) - 1 + F_{R_{1(m_0)}}(r_1) \\
&= (1 - \alpha_{1m}) \cdot \alpha^*_{1m} (F_{R_{1(m_0)}}(r_1) - F_{R_{1(m_0)}}(r_1 - 1)) \\
&+ (1 - \alpha^*_{1m}) \cdot \alpha_{1m} F_{R_{1(m_0)}}(r_1) \geq 0, \quad r_1 \geq 1
\end{aligned}
$$

Thus, $R_{1(m_0-1)} >^{st} R_{1(m_0)}$.

Like $V_{1(m_0)}, S_{1(m_1)}$, and $R_{1(m_0)}$, $V_{1(m_0-1)}, S_{1(m_1+1)}$, and $R_{1(m_0-1)}$ follow Poisson distribution with rates $(m_0 - 1)\alpha_{1m}, (m_1 + 1)\alpha^*_{1m}$, and $(m_0 - 1)\alpha_{1m} + (m_1 + 1)\alpha^*_{1m}$, respectively.

Theorem 2.2.6 *$V_{2(m_0)} >^{st} V_{2(m_0-1)}$, $S_{2(m_1+1)} >^{st} S_{2(m_1)}$, and $R_{2(m_0-1)} >^{st} R_{2(m_0)}$, where $R_{2(m_0)} = V_{2(m_0)} + S_{2(m_1)}$*

Proof.

$$
\begin{aligned}
F_{V_{2(m_0-1)}}(v_2) - F_{V_{2(m_0)}}(v_2) &= F_{V_{2(m_0)}}(v_2) - \alpha_{2m} F_{V_{2(m_0)}}(v_2 + 1) - (1 - \alpha_{2m}) F_{V_{2(m_0)}}(v_2) \\
&= -\alpha_{2m} \cdot (F_{V_{2(m_0)}}(v_2 + 1) - F_{V_{2(m_0)}}(v_2)) \\
&= -\alpha_{2m} \cdot Pr(V_{2(m_0)} = v_2) \leq 0, \qquad v_2 \geq 0.
\end{aligned}
$$

$$
\overline{F}_{V_{2(m_0)}} \geq \overline{F}_{V_{2(m_0-1)}} \Leftrightarrow V_{2(m_0)} >^{st} V_{2(m_0-1)}.
$$

$$
\begin{aligned}
F_{S_{2(m_1+1)}}(s_2) - F_{S_{2(m_1)}}(s_2) &= F_{S_{2(m_1)}}(s_2) - \alpha^*_{2m} F_{S_{2(m_1)}}(s_2 - 1) - (1 - \alpha^*_{2m}) F_{S_{2(m_1)}}(s_2) \\
&= \alpha^*_{2m} \cdot (F_{S_{2(m_1)}}(s_2) - F_{S_{2(m_1)}}(s_2 - 1)) \\
&= \alpha^*_{2m} \cdot Pr(S_{2(m_1)} = s_2 - 1) \leq 0, \qquad s_2 \geq 1.
\end{aligned}
$$

$$
\overline{F}_{S_{2(m_1)}} \leq \overline{F}_{S_{2(m_1+1)}} \Leftrightarrow S_{2(m_1+1)} >^{st} S_{2(m_1)}.
$$

$$
\begin{aligned}
P\{R_{2(m_0-1)} = r_2 | R_{2(m_0)} = r_2\} &= P(V_{2(m_0-1)} = v_2 - 1, S_{2(m_1+1)} = s_2 + 1 | V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&+ P(V_{2(m_0-1)} = v_2 - 1, S_{2(m_1+1)} = s_2 + 1 | V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&= P(S_{2(m_1+1)} = s_2 | V_{2(m_0-1)} = v_2, V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&\times P(V_{2(m_0-1)} = v_2 | V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&+ P(S_{2(m_1+1)} = s_2 + 1 | V_{2(m_0-1)} = v_2 - 1, V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&\times P(V_{2(m_0-1)} = v_2 - 1 | V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&= (1 - \alpha^*_{2m}) \cdot (1 - \alpha_{2m}) + \alpha^*_{2m} \cdot \alpha_{2m} \\
&= 1 - \alpha_{2m} - \alpha^*_{2m} + 2\alpha^*_{2m} \cdot \alpha_{2m}
\end{aligned}
$$

$$
\begin{aligned}
P\{R_{2(m_0-1)} = r_2 + 1 | R_{2(m_0)} = r_2\} &= P(V_{2(m_0-1)} = v_2, S_{2(m_1+1)} = s_2 + 1 | V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&= P(S_{2(m_1+1)} = s_2 + 1 | V_{2(m_0-1)} = v_2, V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&\times P(V_{2(m_0-1)} = v_2 | V_{2(m_0)} = v_2, S_{2(m_1)} = s_2) \\
&= (1 - \alpha_{2m}) \cdot \alpha^*_{2m}
\end{aligned}
$$

$$
\begin{aligned}
\overline{F}_{R_{2(m_0-1)}}(r_2) - \overline{F}_{R_{2(m_0)}}(r_2) &= 1 - (1 - \alpha_{2m} - \alpha^*_{2m} + 2\alpha^*_{2m} \cdot \alpha_{2m}) F_{R_{2(m_0)}}(r_2) \\
&- (1 - \alpha_{2m}) \cdot \alpha^*_{2m} F_{R_{2(m_0)}}(r_2 - 1) - 1 + F_{R_{2(m_0)}}(r_2) \\
&= (1 - \alpha_{2m}) \cdot \alpha^*_{2m} (F_{R_{2(m_0)}}(r_2) - F_{R_{2(m_0)}}(r_2 - 1)) \\
&+ (1 - \alpha^*_{2m}) \cdot \alpha_{2m} F_{R_{2(m_0)}}(r_2) \geq 0, \quad r_2 \geq 1
\end{aligned}
$$

Thus, $R_{2(m_0-1)} >^{st} R_{2(m_0)}$. Like $V_{2(m_0)}, S_{2(m_1)}$, and $R_{2(m_0)}$, $V_{2(m_0-1)}, S_{2(m_1+1)}$, and $R_{2(m_0-1)}$, given $V_{1(m_0-1)}, S_{1(m_1+1)}$, and $R_{1(m_0-1)}$, follow Poisson distribution with rates $(m_0 - V_{1(m_0)} - 1)\alpha_{2m}, (m_1 + 1 - S_{1(m_1)})\alpha^*_{2m}$, and $(m_0 - V_{1(m_0)} - 1)\alpha_{2m} + (m_1 + 1 - S_{1(m_1)})\alpha^*_{2m}$, respectively.

2.2.6 Monotonicity of $FDR^{(2)}$

By using stochastic ordering described above, we can prove the following theorem.

Theorem 2.2.7 *$FDR^{(2)}$ is a monotone decreasing function of m_1.*

Proof.

$$FDR^{(2)} = E(\frac{V_{1(m_0)} + V_{2(m_0)}}{R_{1(m_0)} + R_{2(m_0)}} | R_{1(m_0)} > 0) \cdot Pr(R_{1(m_0)} > 0)$$

Given $R_{1(m_0)} + R_{2(m_0)} > 0$, in particular, $R_{1(m_0)} > 0$, and $R_{1(m_0-1)} + R_{2(m_0-1)} > 0$, in particular, $R_{1(m_0-1)} > 0$, $\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}} < \frac{V_{1(m_0-1)}+V_{2(m_0-1)}}{R_{1(m_0-1)}+R_{2(m_0-1)}}$.
So, $E(\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}} | R_{1(m_0)} > 0)$ is a monotone nonincreasing function of m_1. Since $\alpha^*_{1m} \approx \alpha_{1m}$,

$$Pr(R_{1(m_0-1)} > 0) - Pr(R_{1(m_0)} > 0) = exp(-\mu^*_{1(m_0)}) \cdot [1 - exp(-(\alpha^*_{1m} - \alpha_{1m}))] \approx 0.$$

The dominating term $E(\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}} | R_{1(m_0)} > 0)$ is a monotone nonincreasing function of m_1. Hence, FDR is a monotone decreasing function of m_1.
In fact, since $\mu_{1(m_0)}(= m_0 \cdot \alpha_{1m}) < \lambda_{1(m_1)}(= m_1 \cdot \alpha^*_{2m})$, $V_{1(m_0)} <^{st} S_{1(m_1)}$. Likewise, $V_{2(m_0)} <^{st} S_{2(m_1)}$.
Given $R_{1(m_0)} + R_{2(m_0)} > 0$, in particular, $R_{1(m_0)} > 0$,

$$\begin{aligned}
\tfrac{V_{1(m_0)}}{R_{1(m_0)}} & - \frac{V_{1(m_0)} + V_{2(m_0)}}{R_{1(m_0)} + R_{2(m_0)}} \\
= & \frac{V_{1(m_0)} \cdot R_{2(m_0)} - V_{2(m_0)} \cdot R_{1(m_0)}}{R_{1(m_0)} \cdot (R_{1(m_0)} + R_{2(m_0)})} \\
= & \frac{V_{1(m_0)} \cdot (V_{2(m_0)} + S_{2(m_1)}) - V_{2(m_0)} \cdot R_{1(m_0)}}{R_{1(m_0)} \cdot (R_{1(m_0)} + R_{2(m_0)})} \\
= & \frac{V_{1(m_0)} \cdot S_{2(m_1)} - V_{2(m_0)} \cdot S_{1(m_1)}}{R_{1(m_0)} \cdot (R_{1(m_0)} + R_{2(m_0)})} \\
\geq & \frac{S_{1(m_1)} \cdot S_{2(m_1)} - V_{2(m_0)} \cdot S_{1(m_1)}}{R_{1(m_0)} \cdot (R_{1(m_0)} + R_{2(m_0)})} \\
= & \frac{S_{1(m_1)} \cdot (S_{2(m_1)} - V_{2(m_0)})}{R_{1(m_0)} \cdot (R_{1(m_0)} + R_{2(m_0)})} \\
\geq^{st} & \ 0.
\end{aligned}$$

Hence, The dominating term $E(\frac{V_{1(m_0)}+V_{2(m_0)}}{R_{1(m_0)}+R_{2(m_0)}} | R_{1(m_0)} > 0)$ of $FDR^{(2)}$ is smaller than

the dominating term $E(\frac{V_{1(m_0)}}{R_{1(m_0)}}|R_{1(m_0)} > 0)$ of FDR. $FDR^{(2)}$ is smaller(better) than the first-stage FDR, FDR.

2.2.7 $FNR^{(2)}$

We introduce the following two-stage FNR procedure. Let $T_{1(m_1)}$, and $A_{1(m_0)}$ be the number of genes among the $m - m_0$ genes not rejected, the number of genes not rejected by the first stage procedure. Let $T_{2(m_1)}$, and $A_{2(m_0)}$ be the number of genes among the $m - m_0$ genes not rejected, the number of genes not rejected by the second stage procedure.

$$
\begin{aligned}
FNR^{(2)} &= E(\frac{T_{1(m_1)} + T_{2(m_1)}}{A_{1(m_0)} + A_{2(m_0)}}|A_{1(m_0)} + A_{2(m_0)} > 0) \cdot Pr(A_{1(m_0)} + A_{2(m_0)} > 0) \\
&= E(\frac{m_1 - (S_{1(m_1)} + S_{2(m_1)})}{m - R_{1(m_0)} - R_{2(m_0)}}|R_{1(m_0)} + R_{2(m_0)} < m) \cdot Pr(R_{1(m_0)} + R_{2(m_0)} < m) \\
&= \sum_{r=0}^{m-1} (E[\frac{m_1 - (S_{1(m_1)} + S_{2(m_1)})}{m - R_{1(m_0)} - R_{2(m_0)}}|R_{1(m_0)} + R_{2(m_0)} = r] \\
&\times \frac{Pr(R_{1(m_0)} + R_{2(m_0)} = r)}{Pr(R_{1(m_0)} + R_{2(m_0)} < m)} \cdot Pr(R_{1(m_0)} + R_{2(m_0)} < m)) \\
&= \sum_{r=0}^{m-1} \frac{1}{m-r} \cdot [m_1 - E(S_{1(m_1)} + S_{2(m_1)}|R_{1(m_0)} + R_{2(m_0)} = r)] \cdot pr(R_{1(m_0)} + R_{2(m_0)} = r) \\
&= \sum_{r=0}^{m-1} \frac{1}{m-r} \cdot [m_1 - E(S_{1(m_1)} + S_{2(m_1)}|R_{1(m_0)} + R_{2(m_0)} = r)] \cdot pr(R_{1(m_0)} + R_{2(m_0)} = r \\
&= \sum_{r=0}^{m-1} \frac{1}{m-r} \cdot [m_1 - r(1-p)] \cdot \frac{exp(-m_0 \cdot (\alpha_{1m} + \alpha_{2m}) - m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m}))}{r!} \\
&\times (m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^v \cdot (m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^r \\
&= exp(-m_0 \cdot (\alpha_{1m} + \alpha_{2m}) - m_1 \cdot (\alpha^*_{1m} + \alpha^*_{2m})) \cdot \sum_{r=0}^{m-1} \frac{1}{m-r} \cdot [\frac{m_1 - r(1-p)}{r!}] \\
&\times (m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^v \cdot (m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^r
\end{aligned}
$$

where p=$\frac{m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}{m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}$

2.2.8 Monotonicity of $FNR^{(2)}$

By using stochastic ordering described above, we can prove the following theorem.

Theorem 2.2.8 *$FNR^{(2)}$ is a monotone increasing function of m_1.*

Proof.

$$
\begin{aligned}
FNR^{(2)} &= E(\frac{T_{1(m_1)} + T_{2(m_1)}}{A_{1(m_0)} + A_{2(m_0)}}|A_{1(m_0)} + A_{2(m_0)} > 0) \cdot Pr(A_{1(m_0)} + A_{2(m_0)} > 0) \\
&= E(\frac{m_1 - (S_{1(m_1)} + S_{2(m_1)})}{m - R_{1(m_0)} - R_{2(m_0)}}|R_{1(m_0)} + R_{2(m_0)} < m) \cdot Pr(R_{1(m_0)} + R_{2(m_0)} < m)
\end{aligned}
$$

$S_{1(m_1+1)} <^{st} 1 + S_{1(m_1)}$ and $1 - S_{1(m_1+1)} >^{st} 1 + S_{1(m_1)}$.
Given $R_{1(m_0-1)} + R_{2(m_0-1)} < m$ and $R_{1(m_0)} + R_{2(m_0)} < m$,

$$\frac{m_1 + 1 - S_{1(m_1+1)} - S_{2(m_1+1)}}{m - R_{1(m_0-1)} - R_{2(m_0-1)}} >^{st} \frac{m_1 - S_{1(m_1)} - S_{2(m_1)}}{m - R_{1(m_0)} - R_{2(m_0)}}$$

$E(\frac{m_1+1-S_{1(m_1+1)}-S_{2(m_1+1)}}{m-R_{1(m_0-1)}-R_{2(m_0-1)}}|R_{1(m_0-1)} + R_{2(m_0-1)} < m) > E(\frac{m_1-S_{1(m_1)}-S_{2(m_1)}}{m-R_{1(m_0)}-R_{2(m_0)}}|R_{1(m_0)} + R_{2(m_0)} < m)$
$E(\frac{m_1-S_{1(m_1)}-S_{2(m_1)}}{m-R_{1(m_0)}-R_{2(m_0)}}|R_{1(m_0)} + R_{2(m_0)} < m)$ is a monotone nondecreasing function of m_1. $\alpha^*_{1m} \approx \alpha_{1m}$ and $\alpha^*_{2m} \approx \alpha_{2m}$,

$$\begin{aligned}
Pr \quad & (R_{m_0} < m) - Pr(R_{m_0-1} < m) \\
= \quad & (1 - Pr(R_{m_0} = m)) - (1 - Pr(R_{m_0-1} = m)) \\
\leq \quad & exp(-m_0(\alpha_{1m} + \alpha_{2m}) - m_1(\alpha^*_{1m} + \alpha^*_{2m})) \cdot \frac{[1 - exp(-(\alpha^*_{1m} - \alpha_{1m}) - (\alpha^*_{2m} - \alpha_{2m}))]}{m!} \\
\times \quad & ((m_1 + 1)(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + (m_0 - 1)(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m}))^m \\
\approx \quad & 0
\end{aligned}$$

The dominating term $E(\frac{m_1-S_{1(m_1)}-S_{2(m_1)}}{m-R_{1(m_0)}-R_{2(m_0)}}|R_{1(m_0)} + R_{2(m_0)} < m)$ is a monotone nondecreasing function of m_1.Thus, $FNR^{(2)}$ is a monotone nondecreasing function of m_1

2.2.9 Control of FDR

FWER represents the upper bound of FDR, determining each stage's significance level. We want to determine each stage's significance level so that the overall significance level is equal to α.

$$\begin{aligned}
& FWER \\
= \quad & Pr(V_{1(m_0)} + V_{2(m_0)} \geq 1) \\
= \quad & 1 - Pr(V_{1(m_0)} + V_{2(m_0)} = 0) \\
= \quad & 1 - Pr(V_{1(m_0)} = 0) \cdot Pr(V_{2(m_0)} = 0 | V_{1(m_0)} = 0) \\
= \quad & 1 - exp(-m_0 \cdot \alpha_{1m}) \cdot exp(-m_0 \cdot \alpha_{2m}) \\
= \quad & 1 - exp(-m_0(\alpha_{1m} + \alpha_{2m})) = \alpha
\end{aligned}$$

Each stage's significance level is determined such that $\alpha_{1m} + \alpha_{2m} = -\frac{ln(1-\alpha)}{m_0}$. Since FWER is always smaller than FDR, we can ensure that FDR is controlled at preassigned level α by any combination of α_{1m} and α_{2m}.

2.2.10 Estimation procedure

Storey (2002) argues that estimation is often preferred over control because it is difficult to pre-specify an appropriate control level (Stan Pounds and Cheng Cheng, 2006). Instead of fixing the error rate and then estimating the rejection region, we attempt to use a new approach to fix the rejection region and estimate the error rate. The full details of the estimation and inference for proposed FDR are given in algorithm below.

Proposed FDR=$\frac{1}{1+\frac{m_1(\alpha^*_{1m}\cdot e^{\alpha^*_{2m}+\alpha^*_{2m}})}{m_0(\alpha_{1m}\cdot e^{\alpha_{2m}+\alpha_{2m}})}}$ (1)

1. For the m hypothesis tests, calculate their respective p-values $p_1, \ldots, p_m$.
2. Estimate α_{1m} by min($\frac{V_{1mo}}{m_0}, c_\alpha$)
3. Estimate α^*_{1m} by max($\frac{S_{1mo}}{m_1}, c_\alpha$)
4. For fixed α_{1m}, compute α_{2m} by max(-$\alpha_{1m} - \frac{log(1-\alpha)}{m_0}, 0$)
5. For fixed α^*_{1m} and α_{2m}, pick up α^*_{2m} for some range $(\alpha_{2m}, \alpha^*_{1m})$
6. For B number of values α^*_{2m}, average their respective FDRs.

In (1), there still remains some space of improvement for tighter control if we know π_0. To estimate π_0, Storey and Tibshirani (2001) use the fact that null p-values are distributed uniformly on [0,1] and then plug it in estimating FDR. It was shown that the violation of uniformity of p-values could bias the estimate of π_0 upward. Discrete p-values become encountered in practice as categorical genomic data discussed in next

chapter. These p-values may be stochastically larger than uniform, thus violating the assumption of uniformity. Gene expression levels and most of are typically continuous. Thus, the estimation of π_0 perform better in DNA microarray gene expression datasets.

CHAPTER 3

FALSE DISCOVERY RATE IN GENOMIC SEQUENCES

3.1 Introduction

High dimension, low sample size data may appear in various areas of science: the dimension tends to ∞ while the sample size is small. This data models are abound in genomic studies, in particular, where sample size n may be small and there are different epidemiologic strata $G(> 2)$, so that classical MANOVA(multivariate analysis of variance) may be pertinent. An important task of this study is to identify the most siginifcant genes(or positions) among a number of genes: Which positions are differentially expressed across the groups? The feature of this study is that the number of genes in a sequence(K) is much larger than the number of sequences(n). As we have seen before, control of the FWER is too conservative when there are many hypotheses such as microarray experiments. False discovery rate (FDR) procedure is better than the FWER procedure to handle multiple testing problem in large scale assciation study. However, most of the FDR procedures has not been used extensively in genomic studies compared to gene expression studies. On the other hand, in

multiple testing problems, the response variables are continuous, but may be count or discrete, or purely qualitative responses, that is, high-dimensional low sample size categorical data setups, complicating the multiplicity problems. In SARS epidemic models in chapter 5 for illustration, we have 900 genes (or positions) for each sequence and 14 samples downloaded from four locations, Guadong, Beijing, Hongkong and Taiwan. The response variables are a,c,g, and t, having even not ordered categories.

Suppose we have a general model comprise $G(> 2)$ groups of sequences. Each sequence has K positions, and in each position, there is a categorical response with C categories. n_{gkc} denotes the number of responses in category c at site k in the g-th group, $c = 1, \ldots, C; k = 1, \ldots, K$ and $g = 1, \ldots, G$. There is a set C of C^K joint labels $\mathbf{c} = (c_1, \ldots, c_K)$ in which each c_k takes on value $1, \ldots, C$. The number of observations in the g-th group with the label combination $\mathbf{c}$ is denoted by $n_g(\mathbf{c}), \mathbf{c} \in C$. We also have $\sum_{\mathbf{c}\in C} n_g(\mathbf{c}) = n_g$ and $\sum_{\mathbf{c}\in C} \pi_g(\mathbf{c}) = 1, \forall g = 1, \ldots, G$. The full multi-dimensional multinomial law is

$$\prod_{g=1}^{G} \{ \frac{n_g!}{\prod_{\mathbf{c}\in C}(n_g(\mathbf{c})!} \prod_{\mathbf{c}\in C} [\pi_g(\mathbf{c})]^{n_g(\mathbf{c})} \}.$$

The total number of unknown parameters is $q^0 = G(C^K - 1)$, but q^0 is too large compared to sample size n, where $n = \sum_{g=1}^{G} n_g$. Because of this problem, this law may not be reasonable. That's why the standard multivariate approach may be of limited utilities. In these sense, the (pseudo) marginal diversity measures may be combined into a composite measure providing a less stringent way of CATAMANOVA (categorical MANOVA). In the SNP model, the categories are not ordered, and hence, a stochastic ordering may not be feasible. However, the Hamming distance may have ordering. Even in that case, individual statistics, even coordinatewise ones, do not have a known null hypothesis distribution. That's why we have to use there jackknife

variance estimation and permutation distribution to construct some permutation tests. A pseudo-marginal approach based on Hamming distance provides some promising test statistic. Proposed FDR procedure along with the associated test statistic may be a useful tool for genomic studies. These perspectives are appraised in a nonstandard statistical analysis, using the 2002-03 SARS epidemic model.

3.2 A Pseudo Marginal Model

As stated before, the full multisample, multi-dimensional multinomial law may not be reasonable. For geographically separated sequences, the assumption of independence among G groups may be tenable but within group sequences may not be independent. For each sequence, the K positions may not have independent responses nor identically distributed. Under this assumption of inter-position stochastic dependence, we need to consider another measure of variation. The Gini Simpson biodiversity index has found useful applications in genetics and in bioinformatics. Mostly, categorical data models, without an ordering of the categories, appear, which preempts use of measures of quantitative diversity analysis. Without much stringent structural regularity assumptions, the Hamming distance exploits the idea of Gini-Simpson diversity index in a variety of multidimensional setups, We exploit the following Gini-Simpson index: $\mathbf{I}(\pi) = 1 - \pi^{\mathbf{t}}\pi = 1 - \sum_{c=1}^{C} \pi_c^2$, where $\pi = (\pi_1, \ldots, \pi_C)^t$ for a single multinomial population with C cells. Define $\mathbf{I}(\pi_{gk})$ for each k=1,...,K and every g=1,...,G. For each g(=1,...,G) and k(=1,...,K),
$\mathbf{I}(\pi)_{\mathbf{gk}} = 1 - (\pi_{\mathbf{gk}})^{\mathbf{t}}\pi_{\mathbf{gk}} = 1 - \sum_{c=1}^{C}(\pi_{gkc})^2$. Also, define $\mathbf{I}(\pi_k)$ in the pooled sample, for each k=1,...,K. Define $H(\prod_g) = 1/K \sum_{k=1}^{K} \mathbf{I}(\pi)_{gk}, g = 1, \ldots, G$ as the Hamming distance based measure. In genomic studies, the following multiple hypotheses are represented in terms of the Hamming distance.

$$H_0 : I(\pi)_{1k} \equiv I(\pi)_{2k} \equiv \cdots I(\pi)_{Gk}, \quad , k(= 1, \ldots, K)$$

vs

H_1 : There are at least one of k's that $I(\pi)_{gk} \neq I(\pi)_{g'k} \quad , 1 <= g < g' <= G.$

3.2.1 Proposed Test Statistics and P-values

Let us consider the following asymptotic distribution of the test statistic. For each k(=1,. . . ,K) and each g(=1,. . . ,G), the estimate of $\mathbf{I}(\pi_{\mathbf{gk}})$ is

$$\begin{aligned} U_{gk} &= \binom{n_g}{2}^{-1} \sum_{1<=i<j<=n_g} I(X_{gik} \neq X_{gjk}) \\ &= \sum_{c=1}^{C} \frac{n_{gkc}(n_g - n_{gkc})}{n_g(n_g - 1)} \end{aligned}$$

This is a U-statistic based on a kernel of degree 2, an unbiased estimator of Gini-Simpson index. In the pooled sample,

$$\begin{aligned} U_k &= \binom{n}{2}^{-1} \sum_{1<i<j<=n} I(X_{ik} = X_{jk}) \\ &= \sum_{c=1}^{C} \frac{n_{kc}(n - n_{kc})}{n(n-1)}. \end{aligned}$$

,where $n_{kc} = \sum_{g=1}^{G} n_{gkc}$. This statistic has the following asymptotic distribution only if the n_g's are large.

$$\sqrt{n_g}(U_{gk} - I_{gk}) \sim N(0, 4\zeta_{1gk}),$$

where $\zeta_{1gk} = E\{I(X_{gik} \neq X_{gjk})I(X_{gi'k} \neq X_{gj'k}) - E(I(X_{gik} \neq X_{gjk})I(X_{gi'k} \neq X_{gj'k}))\}$.

This ζ_{1gk} can be replaced by $\zeta_{1.k}$ from the pooled sample. $\zeta_{1.k}$ is estimated by the Jacknife variance estimator of U_k, which is $\hat{\zeta}_{1.k}(=\frac{1}{n-1}\sum_{i=1}^{n}(U_{nk,i}-U_k)^2)$, because the Jacknife variance estimator is more stable than other variance estimator.

For each each k(=1,...,K), the test statistic L_k is defined as $\sum_{g=1}^{G} n_g[U_{gk}-U_k]^2/(4\hat{\zeta}_{1.k})$. By virtue of Cochran's theorem, it has χ^2 distribution with degree G-1. But a conclusion based on this asymptotic distribution, whenever sample size is small, may give us misleading results. Moreover, for n not adequately large, the p-values have discrete distribution without assuming uniform distribution for the associated p-values under the null hypothesis. Hence, it might be better to simulate the permutation distribution of the marginal test statistic L_k. At least for small to moderate values of the sample sizes, $n_1,\ldots,n_G$, the permutation distribution can be generated by considering all possible n!(equally likely) permutations of the combined sample observations among the G groups of (sizes $n_1,\ldots,n_G$). Hence, conditionally distribution-free tests may be constructed for the test statistic L_k. The corresponding p-value is defined as below.

$$Pr(L_k > l_k|H_0)$$

, where L_k is a test statistic from the permuted distribution. Under the null hypothesis, the permutation distribution of L_k may be symmetric about 0, with mean $E_0(L_k)=0$. Under the alternative hypothesis, the distribution is tilted to the right. That's why we use a right-sided test. However, though the distribution freeness hold under the null hypothesis, such distributions are more complex to evaluate.

3.3 Discussion

3.3.0.1 cFDR

Tsai et al.(2003) discuss another measure of false discovery rate, the conditional FDR (cFDR), defined as

$$cFDR = E(\frac{V}{R}|R = r) = \frac{E(V|R = r)}{r}$$

Under Storey's(2002) mixture model, Tsai et al.(2003) show that

$$cFDR(c) = pFDR(c) = \frac{\pi \times c}{F(c)},$$

where F(c)=$Pr(p <= c)$. Let $V_{1(m_0)} + V_{2(m_0)}$be V_{m_0} , $S_{1(m_1)} + S_{2(m_1)}$ be S_{m_1}, and $R_{m_0} = V_{m_0} + S_{m_1}$. As we have seen in the earlier chapter,

$$\begin{aligned} Pr & (V_{m_0} = v|R_{m_0} = r) \\ = & Pr(V_{m_0} = v|V_{m_0} + S_{m_1} = r) \\ = & \binom{r}{v} p^v (1-p)^{r-v} \end{aligned}$$

where p=$\frac{m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}{m_1(\alpha^*_{1m} \cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m} \cdot e^{\alpha_{2m}} + \alpha_{2m})}$. Hence,

$$\begin{aligned} cFDR & = \frac{E(V_{m_0}|R_{m_0} = r)}{r} \\ & = r \times \frac{p}{r} \\ & = p \end{aligned}$$

This is exactly the same as the Proposed FDR procedure.

Theorem 3.3.1 *In the asymptotic sample setting, Proposed FDR is a conservative point estimate of the FDR over all siginificance regions simultaneously.*

Proof.

$lim_{m\to\infty}\alpha_{2m} = 0$ and $lim_{m\to\infty}\alpha^*_{2m} \in (0, c)$.

$$\begin{aligned} E(FDR^{(2)}(c)) &= E(\frac{m_0(\alpha_{1m}\cdot e^{\alpha_{2m}} + \alpha_{2m})}{m_1(\alpha^*_{1m}\cdot e^{\alpha^*_{2m}} + \alpha^*_{2m}) + m_0(\alpha_{1m}\cdot e^{\alpha_{2m}} + \alpha_{2m})}) \\ &\to E(\frac{V(c)}{V(c) + S(c)\alpha^*_{2m} + m_1 * \alpha^*_{2m}}) \\ &<= E(\frac{V(c)}{V(c) + S(c)}) \\ &<= E(\frac{V(c)}{R(c)\bigvee 1}) \end{aligned}$$

3.3.1 False discovery rate optimality and Average Power

Storey et al. (2005) introduced the optimal discovery procedure which is the procedure to maximize the expected number of true positives (ETP) for each fixed expected number of false positives (EFP). This proposed optimality criterion is related to optimality in terms of FDRs and misclassification rates. For FDRs, $FDR \approx \frac{EFP}{EFP+ETP}$. It has been suggested that this FDR optimality should be defined in terms of the proportion of true alternatives among the tests not called significant Genovese and Wasserman (2002). This quantity has been called the "false non-discovery rate"' Genovese and Wasserman (2002) and the "miss rate" (Taylor et al. 2005); We call it the "missed discovery rate(MDR)". A procedure is optimal if for each fixed FDR level, the MDR is minimized: $MDR \equiv E[\frac{FN}{FN+TN}] \approx \frac{EFN}{EFN+ETN}$, where EFN is expected number of false negatives and ETN is the expected number of true negatives. Now, applying this to our proposed FDR procedure, since α_{1m} and α^*_{1m} is

estimated by V/m_0 and S/m_1, respectively,

$$
\begin{aligned}
MDR \approx \frac{EFN}{EFN+ETN} & \\
&= \frac{m-m_0-ETP}{m-ETP-EFP} \\
&= \frac{m_1-m_1(\alpha^*_{1m}+\alpha^*_{2m}-\alpha^*_{1m}\times\alpha^*_{2m})}{m1-m_1(\alpha^*_{1m}+\alpha^*_{2m}-\alpha^*_{1m}\times\alpha^*_{2m})+m_0-m_0(\alpha_{1m}+\alpha_{2m}-\alpha^*_{1m}\times\alpha_{2m})} \\
&= \frac{(m_1-S)(1-\alpha^*_{2m})}{(m_1-S)(1-\alpha^*_{2m})+(m_0-V)(1-\alpha_{2m})} \\
&= \frac{1}{1+\frac{(m_0-V)(1-\alpha_{2m})}{(m_1-S)(1-\alpha^*_{2m})}}
\end{aligned}
$$

Our FDR procedure is considered to be optimal if for fixed V and S, there is a big difference between α_{2m} and α^*_{2m}.

On the other hand, comparing the first stage FDR(AV_1) with two stage FDR(Proposed FDR)(AV_2) in terms of average power,

$AV_1 : E(S_1)/m_1 = \alpha^*_{1m}$

AV_2: $E(S_1+S_2)/m_1 = \alpha^*_{1m}+\alpha^*_{2m}(1-\alpha^*_{1m})$. Obviously, $AV_1 < AV_2$. By taking the second stage testing procedure from the first stage FDR, the increase in power that we achieve is greater. For fixed α^*_{1m} and α_{1m}, as α increases, α_{2m} and α^*_{2m} increases. Thus, the average power of two stage FDR is a monontone nondecreasing function of α.

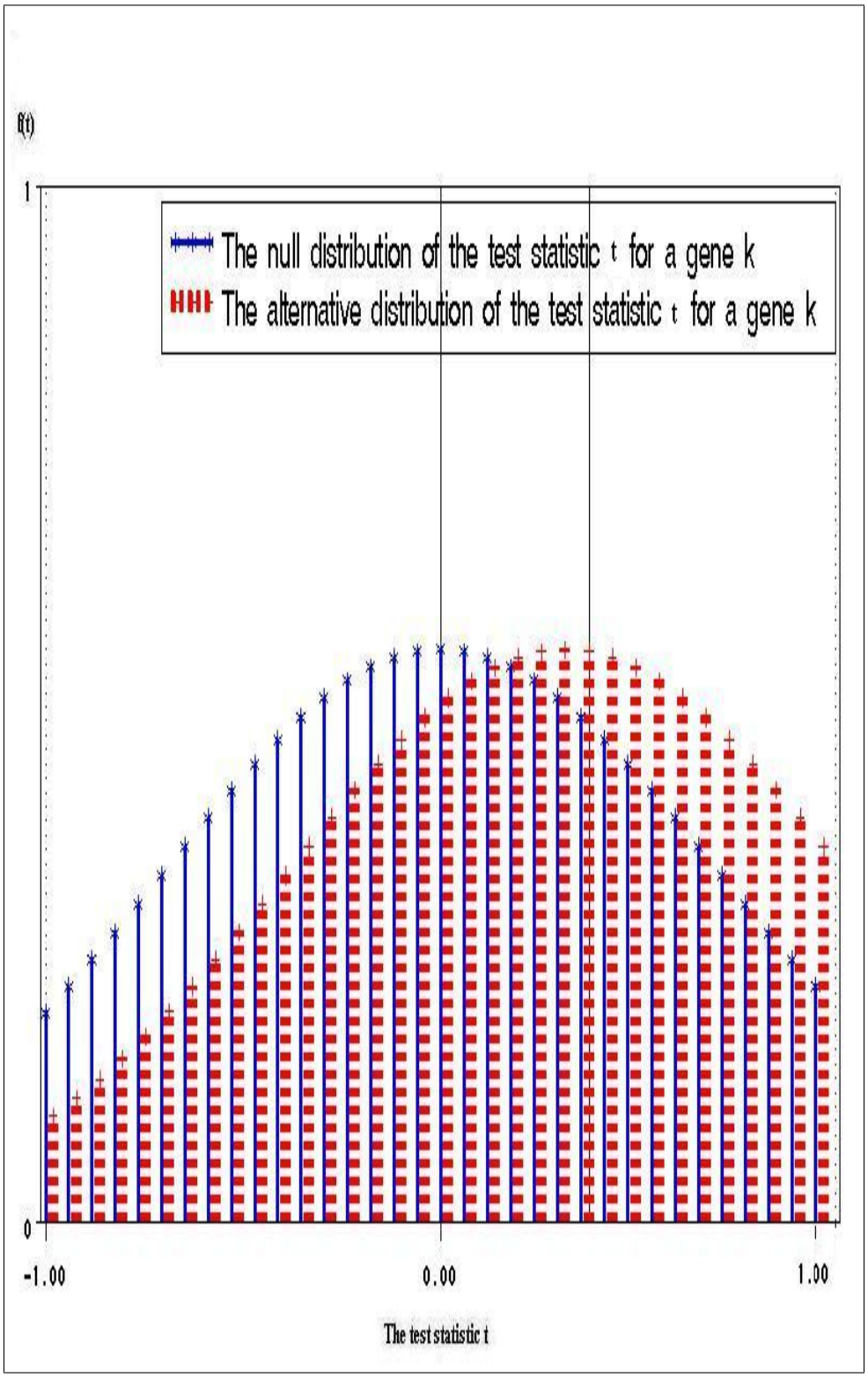

FIGURE I: Comparison of the null distribution with the alternative distribution

CHAPTER 4

CLASSIFICATION OF GENES

4.1 Introduction

In a genomic data, the goal of class discovery is to partition a set of subjects into groups relatively similar, in the sense that subjects in the same group are more alike than subjects in different groups. In a large number of correlated genes with heterogeneity amidst a smaller sample like DNA microarray data, order-restricted inference problems often appears in complex ways. To study gene expression pattern across various treatment groups with this order constraint weakens the effectiveness of standard statistical inference and as a result, calls for different perspectives. For this reason, nonstandard robust methods are proposed to classify genes reflecting the concept of order-restricted inference without any assumptions of specific forms. M-estimator based on Union-Intersection principle (UIP), proposes distribution-insensitive clustering of genes. It might be also possible to construct a locally most powerful rank test using a suitable rank scores along with UIP, though it is too difficult to construct an optimal test based on UMP. The Kendall-tau statistics may be utilized to construct a distribution-free test. Gene expression levels may be compared among more than 2 groups using exact tests of homogeneity; associations among the variables assessed using the Kendall's tau-b statistic.

By using exact permutation distrubution theory, conditionally distribution-free test based upon these three proposed test statistics is used to generate p-values and as a result is amenable in small sample size setup. It is also computationally tractable and statistically robust.

4.2 Proposed Test Statistics and P-values

4.2.1 Preliminary notation

Consider a DNA microarray experiment having expression data on K genes for n mRNA samples. The gene expression data are in a $K \times n$ matrix $X = (X_{ki})$, with rows corresponding to genes and columns corresponding to individual microarry experiments,where x_{ji} denotes the expression measure of gene k in sample i, i = 1, . . . ,n, k = 1,. . . ,K. The expression measures x_{ki} are assumed to be preprocessed data. For comparing several groups, a general model consists of G (> 2) groups of subjects, each subjects having K genes. For simplicity, we assume that there are no missing values resulting in $n_{gk} = n_g, \forall k$. Let $n = \sum_{g=1}^{G} n_g$ be the total number of subjects in the pooled sample. A row vector $\mathbf{X}_k = (X_{1,k}, X_{2,k}, \ldots, X_{n_1,k}, \ldots, X_{n,k})$ represents the pooled sample at gene k. In this pooled sample, define
$R_k = (R_{1,k}, \ldots, R_{n_1,k}, R_{n_1+1,k}, \ldots, R_{n,k})$, where $R_{i,k}$ is the rank of $X_{i,k}$ in the pooled sample among all the n observations in the kth gene.

4.2.2 Linear Rank Statistics

We want to find out a gene's true profile to one of a specified set of candidate profiles. Two common inequality profiles (nondecreasing/nonincreasing) in terms of mean gene expression levels are introduced here. Without loss of generality, we focus on

monotone increasing pattern among the groups.

Let $\mu_{ki} = E(X_{ki})$ denote the mean expression level of the kth gene in the ith observation. For the kth gene (or position), we can formulate H_{0k} vs H_{1k} as below.

$$H_{0k} : \mu_{1k} = \mu_{2k} = \cdots = \mu_{Gk} \qquad H_{1k} : \mu_{1k} \leq \mu_{2k} \leq \cdots \leq \mu_{Gk}$$

where

$$\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{Gk})'$$

The $(G-1) \times G$ matrix

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & \ldots \\ 0 & -1 & 1 & 0 & 0 & \ldots \\ 0 & 0 & -1 & 1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ldots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix}$$

These hypotheses can be restated as the following two hypotheses.

$$H_{0k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\mu}_k = \bigcap_{j=1}^{G-1} H_{0jk} = \mathbf{0}$$

vs

$$H_{1k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\mu}_k = \bigcup_{j=1}^{G-1} H_{1jk} \geq \mathbf{0}$$

where $H_{0jk} : \theta_{jk} = \mu_{j+1,k} - \mu_{j,k} = 0$ vs $\quad H_{1jk} : \theta_{jk} = \mu_{j+1,k} - \mu_{j,k} \geq 0$. These hypotheses are written in terms of Finite UI principle. But an infinite UIT will be formulated as well.

These hypotheses can be restated as the following two hypotheses. For a given $\mathbf{a}$,

$$H_{0k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\mu}_k = \bigcap_{\mathbf{a}\in\mathcal{R}^{+G}} H_{0\mathbf{a}k} = \mathbf{0}$$

vs

$$H_{1k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\mu}_k = \bigcup_{\mathbf{a}\in\mathcal{R}^{+G}} H_{1\mathbf{a}k} \geq \mathbf{0}$$

where $H_{0\mathbf{a}k} : \mathbf{a}'\boldsymbol{\theta}_k = 0, \quad H_{1\mathbf{a}k} : \mathbf{a}'\boldsymbol{\theta}_k \geq 0.$

This UI (Union-Intersection) principle assumes that for testing $H_{0\mathbf{a}k}$ vs $H_{1\mathbf{a}k}$, we have a optimal test. However, the underlying density of gene expression levels $X_{ik}, i = 1, \ldots, n, k = 1, \ldots, K$ are completely unknown with unknown variance. In this framework, it is hard to construct either an optimal test based on UMP or a similar test using UMPI. In these sense, nonparametrics might yield robust statistical inference procedures that are distribution free.

Fortunately, the null hypothesis H_{0k} is a hypothesis of invariance (under suitable groups of transformation that map the sample space onto itself). Then it might be possible to construct a test for $H_{0\mathbf{a}k}$vs $H_{1\mathbf{a}k}$, a locally most powerful rank test (LMPR)test for each $\mathbf{a}$. By definition, a test is LMPR if among the class of rank tests, it is uniformly most powerful (UMP)for H_0 against a class $H_1(\epsilon)$ of alternatives that are indexed by a parameter Δ, such that $0 < \Delta < \epsilon, \epsilon > 0$ Silvapulle and Sen (2004). LMPR properties may not be available for restricted alternatives. However, UIT-based LMPR test can handle such a problem.

Even though each sample size n_g differs by group, all the n$(=\sum_{g=1}^{G} n_g)$ observations $\mathbf{X}_k = (X_{1,k}, \ldots, X_{n_1,k}, X_{n_1+1,k} \ldots, X_{n,k})$ for each gene k in the pooled sample are i.i.d r.v's under the null hypothesis. Under the null hypothesis of homogeneity, the joint distribution of n observations for each gene k, remains invariant under any

permutation. This permutation distribution $\mathcal{P}_n$ can be obtained by considering every possible $n!$ permutations of the pooled sample observations among the G groups. Hence, conditionally distribution-free tests can be constructed by an appeal to this permutational invariance. We denote this condtional probability law by $\mathcal{P}_n$.

For each gene k, define a multivariate linear rank statistics
$T_{gk}, g = 1, \ldots, G, k = 1, \ldots, K$ as follow. For a suitable rank scores $a(k)$,

$$T_{gk} = \sum_{i=1}^{n} (c_{ig} - \bar{c}_n) a(R_{i,k}) = \sum_{i=1}^{n} c_{ig} a(R_{i,k})$$

where

$$c_{ig} = \begin{cases} \frac{1}{n_g} & \text{if } i = \sum_{l=1}^{g-1} n_l + 1, \ldots, \sum_{l=1}^{g} n_l \\ 0 & \text{otherwise} \end{cases}$$

and $\mathbf{T}_k = (T_{1k}, \ldots, T_{Gk})'$.

Without loss of generality, assume that $\sum_{i=1}^{n} a(R_{i,k}) = 0$. The mean of T_{gk} is

$$\begin{aligned} E_{\mathcal{P}_n}(T_{gk}) &= (E_{\mathcal{P}_n}(a(R_{i,k}))(\sum_{i=1}^{n}(c_{ig} - \bar{c}_n)) \\ &= (\frac{1}{n}\sum_{i=1}^{n} a(R_{i,k}))(\sum_{i=1}^{n} c_{ig}) \\ &= 0 \end{aligned}$$

The variance of T_{gk} is

$$\begin{aligned}
V_{\mathcal{P}_n}(T_{gk}) &= E_{\mathcal{P}_n}(T_{gk})^2 \\
&= V_{\mathcal{P}_n}(a(R_{i,k})\sum_{i=1}^{n}(c_{ig})^2 + \sum_{1\leq i\neq i'\leq n}(c_{ig})(c_{i'g})E_{\mathcal{P}_n}(a(R_{i,k}a(R_{i',k}) \\
&= (\frac{1}{n}\sum_{i=1}^{n}a^2(R_{i,k}))\sum_{i=1}^{n}(c_{ig})^2 + \sum_{1\leq i\neq i'\leq n}c_{ig}c_{i'g}\times(-\frac{1}{n(n-1)}\sum_{i=1}^{n}a^2(R_{i,k})) \\
&= (\frac{1}{(n-1)}\sum_{i=1}^{n}a^2(R_{i,k}))\times((\frac{(n-1)}{n}\sum_{i=1}^{n}(c_{ig})^2) - (\frac{1}{n}\sum_{1\leq i\neq i'\leq n}c_{ig}c_{i'g})) \\
&= (\mathbf{A}_n^2)(\frac{(n-n_g)}{n\cdot n_g})
\end{aligned}$$

where

$$\begin{aligned}
\frac{(n-1)}{n}\sum_{i=1}^{n}(c_{ig})^2 - \frac{1}{n}\sum_{1\leq i\neq i'\leq n}c_{ig}c_{i'g} &= \frac{(n-1)}{n}\sum_{i=1}^{n}(c_{ig})^2 - \frac{1}{n}(\sum_{i=1}^{n}(c_{ig})^2 - \sum_{i=1}^{n}(c_{ig})^2) \\
&= \sum_{i=1}^{n}(c_{ig})^2 - \frac{1}{n}\sum_{i=1}^{n}(c_{ig})^2 \\
&= \frac{1}{n_g} - \frac{1}{n} \\
&= \frac{(n-n_g)}{n\cdot n_g}
\end{aligned}$$

and $\mathbf{A}_n^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}a^2(R_{i,k})$.

For $1 \leq g \neq g' \leq G$, the covariance of T_{gk} and $T_{g'k}$ is

$$\begin{aligned} Cov_{\mathcal{P}_n}(T_{gk}, T_{g'k}) &= E_{\mathcal{P}_n}(T_{gk}, T_{g'k}) \\ &= E_{\mathcal{P}_n}(\sum_{i=1}^{n} c_{ig} a(R_{i,k}) \sum_{i=1}^{n} c_{ig'} a(R_{i,k}) \\ &= E_{\mathcal{P}_n} \sum_{i=\sum_{g=1}^{g-1} n_g - 1}^{\sum_{g=1}^{g} n_g} c_{ig} a(R_{i,k}) \sum_{i'=\sum_{g=1}^{g'-1} n_g - 1}^{\sum_{i'=1}^{g'} n_g} c_{i'g'} a(R_{i',k})) \\ &= (\sum_{i=\sum_{g=1}^{g-1} n_g - 1}^{\sum_{g=1}^{g} n_g} c_{ig})(\sum_{i'=\sum_{g=1}^{g'-1} n_g - 1}^{\sum_{i'=1}^{g'} n_g} c_{i'g'})(E_{\mathcal{P}_n}(a(R_{i,k})a(R_{i',k}))) \\ &= \mathbf{A}_n^2(-\frac{1}{n}) \end{aligned}$$

Hence, the permutation variance of $\mathbf{T}_k$ is

$$\begin{aligned} \mathbf{V}_k &= var(\mathbf{T}_k) \\ &= \mathbf{A}_n^2 \mathbf{C}_n \end{aligned}$$

where

$$\begin{aligned} \mathbf{C}_n &= \sum_{i=1}^{n} (\mathbf{c}_i - \bar{c}_n \mathbf{1}_n)(\mathbf{c}_i - \bar{c}_n \mathbf{1}_n)' \\ &= ((\frac{\delta_{g,g'} n - n_g}{n \cdot n_g})) \end{aligned}$$

,where

$$\delta_{g,g'} = \begin{cases} 1 & \text{if } 1 \leq g = g' \leq G \\ 0 & \text{otherwise} \end{cases}$$

$\mathbf{c}_i = (c_{i1}, \ldots, c_{iG})'$, a $n \times 1$ matrix $\mathbf{1}_n = (1, \ldots, 1)'$, $\mathbf{I}_G$ is a $G \times G$ Identity matrix,

and

$$\mathbf{C}_n = \begin{pmatrix} \frac{n-n_1}{n \cdot n_1} & -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots \\ -\frac{1}{n} & \frac{n-n_2}{n \cdot n_2} & -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots \\ -\frac{1}{n} & -\frac{1}{n} & \frac{n-n_3}{n \cdot n_3} & -\frac{1}{n} & -\frac{1}{n} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & \frac{n-n_G}{n \cdot n_G} \end{pmatrix}$$

If we define $\mathbf{T}_k$ in terms of the vector $\mathbf{c}_i$, $\mathbf{T}_k$ is$\sum_{i=1}^{n}(\mathbf{c}_i - \bar{c}_n \mathbf{1}_n)a(R_{i,k})$.

The mean of $\mathbf{T}_k$ is

$$\begin{aligned} E_{\mathcal{P}_n}(\mathbf{T}_k) &= (\mathbf{c}_i - \ c_n \mathbf{1}_n) E_{\mathcal{P}_n} a(R_{i,k}) \\ &= \mathbf{0} \end{aligned}$$

For $1 \leq k \leq k' \leq K$,the covariance matrix of $\mathbf{T}_k$ and $\mathbf{T}'_k$ is

$$Cov_{\mathcal{P}_n}(\mathbf{T}_k, \mathbf{T}'_k) = \mathbf{C}_n \times \sqsubseteq_{k,k'}$$

where $\sqsubseteq_{k,k'} = \frac{1}{(n-1)} \sum_{i=1}^{n}(a(R_{i,k}) - \bar{a}_n)(a(R_{i,k'}) - \bar{a}_n)$. The matrix $\mathbf{V}_n (= ((\sqsubseteq_{k,k'})))$ is a $\mathcal{P}_n$-invariant and completely known matrix.

$\mathbf{T}_n = (\mathbf{T}_1, \ldots, \mathbf{T}_K)'$. Define the $G \times K$ matrix $\mathbf{T}_n^0 = \sum_{i=1}^{n}(\mathbf{c}_i - \ c_n \mathbf{1}_n)\mathbf{a}_n(\mathbf{R}_i)$ as the transpose matrix of $\mathbf{T}_n$,where $\mathbf{a}_n(\mathbf{R}_i) = (a_{n_1} R_{i,1}, \ldots, a_{n_K}(R_{i,K}))'$. By using the concept of a multivariate linear rank statistics, the mean and the covariance matrix of $\mathbf{T}_n^0$ are defined as below.

$$E_{\mathcal{P}_n}(\mathbf{T}_n^0) = \mathbf{0}_{G \times K}$$

$$Cov_{\mathcal{P}_n}(\mathbf{T}_n^0) = \mathbf{C}_n \bigotimes \mathbf{V}_n.$$

In general, multivariate models, LMPR tests, by construction, might be conditionally distribution free (CDF). Given the invariance of $\mathbf{V}_n$ under $\mathcal{P}_n$, we adapt the UIP to

formulate a rank test for $H_{0k} : \boldsymbol{\theta}_k = \mathbf{0}$ vs $H_{1k} : \boldsymbol{\theta}_k \geq \mathbf{0}$. Let $\mathbf{Z}_k = \mathbf{A}\mathbf{T}_k$ and $\mathbf{S}_k = \mathbf{A}\mathbf{V}_k\mathbf{A}^t$. Let $\mathcal{G} = \{1, \ldots, G-1\}$, and for every a : $\emptyset \subseteq a \subseteq \mathcal{G}$, let a' be its complement and $|a|$ its cardinality. For each a : $\emptyset \subseteq a \subseteq \mathcal{G}$, partition $\mathbf{Z}_k$ and $\mathbf{S}_k$ as

$$\mathbf{Z}_k = \begin{pmatrix} \mathbf{Z}_{ka} \\ \mathbf{Z}_{ka'} \end{pmatrix} \qquad \mathbf{S}_k = \begin{pmatrix} \mathbf{S}_{kaa} & \mathbf{S}_{kaa'} \\ \mathbf{S}_{ka'a} & \mathbf{S}_{ka'a'} \end{pmatrix}$$

and write

$$\mathbf{Z}_{ka:a'} = \mathbf{Z}_{ka} - \mathbf{Z}_{kaa'}\mathbf{S}_{ka'a'}^{-1}\mathbf{Z}_{ka'},$$
$$\mathbf{S}_{kaa:a'} = \mathbf{S}_{kaa} - \mathbf{S}_{kaa'}\mathbf{S}_{ka'a'}^{-1}\mathbf{S}_{ka'a}$$

Then the test statistics for the kth gene is,

$$L_k = \Sigma_{\emptyset \subseteq a \subseteq \mathcal{G}} I(\mathbf{Z}_{ka:a'} > \mathbf{0}, \mathbf{S}_{ka'a'}^{-1}\mathbf{Z}_{ka'} \leq \mathbf{0})(n\mathbf{Z}'_{ka:a'}\mathbf{S}_{kaa:a'}^{-1}\mathbf{Z}'_{ka:a'})$$

and rejecting the null hypothesis for large positive values. By reference to the $\frac{n!}{n_1!\cdots n_G!}$ conditionally(permutationally) equally likely realizations of R_k for each k, we can enumerate $\mathbf{T}_k$ (and hence L_k); this generates the exact conditional (permutational) null distribution $\mathcal{P}_n$ of L_k, so that the test based on L_k is CDF (Conditionally Distribution Free). Now, p-value can be computed as below.

$$P_k = Pr(L_k \geq l_k)$$

where L_k is a test statistic from the permuted distribution and l_k is an observed test statistics. The behavior of L_k under alternatives depends on the stochastic ordering of $\boldsymbol{\mu}_k$ and these statistics may not be exact distribution-free nor have identical probabilitiy laws. However, for every $i < i'$, $X_{i'k} - X_{ik}$ has a distribution tilted to the

right so that

$$E\{L_k|H_{1k}\} \geq 0, k = 1, \ldots, K.$$

This motivates us to use tests based on L_k using the right hand side critical region. A proper multiple testing procedure may be applied to the set of dependent p-values. The proposed FDR procedure was shown to work well for this kinds of p-values. The procedure is used to determine which gene has monotone increasing pattern among the groups.

Choice of rank scores $a(k)$ determine if a test statistic is locally most powerful. For example, the Wilcoxon rank test is LMPR when the density is logistic and the normal score test is LMPR when the density is normal. In Chapter 5, we thoroughly investigate this aspect. For the test for linear trend, the Jonckheere test might be tenable. But without the linear ordering or the logistic density, the LMPR property might work for the Jonckheere test.

4.2.3 A Marginal Model Based On Kendall tau statistics

Following the same multisample (ordered alternative) model described in earlier section, define a design variate t_i in the ith array, for i=$1, \ldots, n$. We don't assume linear or specific parametric ordering of them. We can divide sample size n into G subsets of sizes $n_1, \ldots, n_G$. X_{ik} represents a gene expression level in the ith array and forms a K-vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})'$, for $i = 1, \ldots, n$. $F_i(\mathbf{x})$ is the joint distribution of $\mathbf{X}_i$. If a gene k is NDG, the F_{ik}, $i = 1, \ldots, n$ are assumed to be the same. For a DG k, for $i < i'$, $X_{ik} < X_{i'k}$, for $i < i'$, the F_{ik} has some monotone pattern:
$F_{1k} \geq F_{2k} \geq \cdots \geq F_{nk}$. We are interested in the following hypotheses.

$$H_0 = \bigcap_{k=1}^{K} H_{0k} \qquad vs \qquad H_1 = \bigcup_{k=1}^{K} H_{1k}$$

Considering possible dependence of the test statistics, Roy's Union Intersection principle may have an appeal. Now we define the Kendall tau statistics as

$$T_{nk} = \binom{n}{2}^{-1} \sum_{1 \le i < i' \le n} sign(X_{i'k} - X_{ik}) sign(t_{i'} - t_i).$$

In fact, this statistics is a generalized U-statistic of degree 2. Conveniently, we may define $S_n = \{(i, i') : t_i < t_{i'}; 1 \le i < i' \le n\}$, where N is the cardinality of the set S. Since the variation of T_{nk} ranges from -1 and 1, we can find the modified Kendall tau as

$$T^0_{nk} = N^{-1} \sum_{S} sign(X_{i'k} - X_{ik}).$$

. In fact, for any $k(= 1, \ldots, K)$, under H_{0k}, for every $i \neq i'$, $X_{i'k} - X_{ik}$ has symmetric distribution around 0 so that $E_0(T^0_{nk}) = 0, k = 1, \ldots, K$. For small values of n, by using S, we obtain the exact null distribution of T^0_{nk}. For $k(= 1, \ldots, K)$, under H_{0k}, for every $i \neq i'$, $X_{i'k} - X_{ik}$, under alternatives has tilted distribution to the right so that $E(T^0_{nk}) \ge 0, k = 1, \ldots, K$. For n small, the null distribution of T^0_{nk} is in fact discrete. In these sense, we simulate the permutation distribution of any marginal test statistics T^0_{nk}.

4.2.4 Robust M-test

For the kth gene(or position), consider the linear model $\mathbf{Y}_k = \mathbf{X}\boldsymbol{\beta}_k + \mathbf{E}_k$.
where $\mathbf{Y}_k = (Y_{1k}, \ldots, Y_{nk})'$ is the vector of gene expression levels across G groups in

the kth position and The *(known)Design matrix* of the $n \times G$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 1 & 1 & & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 1 & 0 & 1 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{pmatrix},$$

The *unknown parameter* of the $G \times 1$ matrix

$$\boldsymbol{\beta}_k = (\mu_{1k}, \delta_{2k}, \delta_{3k}, \dots, \delta_{Gk}),$$

where μ_{1k} denotes the average expression level at site k in the first group and δ_{jk} refers to the difference between μ_{1k} and the average expression level at site k in the jth group, $j = 2, \dots, G$. The *vector of independent and identically distributed (i.i.d.) errors with a distribution F* of the $n \times 1$ matrix

$$\mathbf{E}_k = (E_{1k}, E_{2k}, E_{3k}, \dots, E_{nk})$$

A paramter of interest $\boldsymbol{\beta}_k$ should be estimated for each k=1,...,K. However, gene expression data usually has many outliers, and is highly probable to be noisy. The small sample sizes used in typical microarray experiments result in unreliable estimation of variance. Because of the large number of genes and small number of arrays, and higher signal-noise ratio in microarray data, many traditional approaches seem improper. Robust statistics methods (Tukey 1977 ; Huber 1981) provide tools for this statistics problem in which an underlying distribution F are unknwon. A

robust procedure should be insensitive to departures from underlying assumptions caused by, for example, outliers. That is, it should have good performance under the underlying assumptions and the performance deteriorates as the situation departs from the assumptions. There are several types of robust estimators. Among them are M-estimator (maximum likelihood type estimator), L-estimator (linear combinations of order statistics) and R-estimator (estimator based on rank transformation) (Huber 1981); RM estimator (repeated median) (Siegel 1982) and LMS estimator (estimator using the least median of squares) (Rousseeuw 1984). We are concerned with the M-estimator, because even when a sample size is small, it still provides a good estimate.

Let $\rho : \mathcal{R}_p \times X \rightarrow R$ be a measurable function. We define an M-estimator $\mathbf{M}_n$ as a solution of the minimization with respect to $\mathbf{t} \in \mathcal{R}_p$.

$$\sum_{i=1}^{n} \rho((Y_i - \mathbf{x}_i'\mathbf{t})),$$

where $\mathbf{x}_i'(= (x_{i1}, \ldots, x_{iG}))$ is the ith row of $\mathbf{X}, i = 1, \ldots, n$. $\mathbf{M}_n$ should be not only regression equivalent: $\mathbf{M}_n(\mathbf{Y} + \mathbf{Xb}) = \mathbf{M}_n(\mathbf{Y}) + \mathbf{b})$ for $\mathbf{b}$ in $\mathcal{R}_p$, but also scale equivalent:$\mathbf{M}_n(c\mathbf{Y}) = c\mathbf{M}_n(\mathbf{Y})$ for $c > 0$. In general, the second condition is not met. Studentization leads to $\mathbf{M}_n$ scale as well as regression equivalent. Define an studentized M-estimator $\mathbf{M}_n$ of $\boldsymbol{\beta}_k$ as a solution of the minimization

$$\sum_{i=1}^{n} \rho((Y_i - \mathbf{x}_i'\mathbf{t})/S_n)$$

with respect to $\mathbf{t}(G\times$ 1 matrix) where $\mathbf{x}_i$' the ith row of $\mathbf{X}$, $i = 1, \ldots, n$ and $S_n = S_n(\mathbf{Y})$ is an appropriate scale statistic.

The linear model above is the classical one-way ANOVA model except that the

distribution $Y_{1k}, \ldots, Y_{nk}$ may not be normal but is of the form F and the G groups are stochastically ordered. Within this framework, we consider the null hypothesis H_0 that the G groups in the kth gene are statistically homogeneous and the alternative hypothesis H_1 refers to the fact that the G groups in the kth gene are ordered in increasing level of dominance. It is plausible to construct H_{0k} and H_{1k} as below.

$$H_{0k} : \delta_{2k} = \delta_{3k} = \cdots = \delta_{Gk} = 0$$

vs

$$H_{1k} : 0 \leq \delta_{2k} \leq \delta_{3k} \leq \cdots \leq \delta_{Gk}$$

These hypothesis can be rephrased as the following two hypotheses.

$$H_{0k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\beta}_k = \mathbf{0} \qquad H_{1k} : \boldsymbol{\theta}_k = \mathbf{A}\boldsymbol{\beta}_k \geq \mathbf{0}$$

where the $(G-1) \times G$ matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 1 & 0 & 0 & 0 & \ldots \\ 0 & -1 & 1 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ldots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix},$$

For testing the null hypothesis, it may be intended to consider alternatives that the vector $\boldsymbol{\theta}_k$ belongs to the nonnegative orthant space $\Re^{+(G-1)}$. In the univariate case, an optimal UMP test exists for such one-sided alternative. However, in such a multivariate case, UMP tests do not exist. For example, the Hotelling T^2 will result in a larger set of confidence interval and will entail some loss of efficiency. It's

therefore interesting to appraise statistical inference under such restricted setups. UIP (Union-Intersection Principle) formulation of Roy (1953) could be well tailored for Statistical inference under the one-sided multivariate alternative hypothesis

Let the first derivative of ρ function be ψ. $\mathbf{M}_n$ is a median-unbiased estimator of $\boldsymbol{\beta}_k$. Skewsymmetry of ψ and symmetry of F are necessary for this median-unbiasedness of $\mathbf{M}_n$. For this reason, Huber loss function may be a good candidate for ψ function. Hence, minimiazation leads to the estimator that is scale as well as regression equivalent.

Define the Huber function as

$$\rho(t) = \begin{cases} c|t| - (1/2) \times c^2 & \text{if } |t| > c \\ (1/2) \times t^2 & \text{if } |t| <= c \end{cases}$$

The derivative of the Huber function ψ is

$$\psi(t) = \begin{cases} c \times sign(t) & \text{if } |t| > c \\ t & \text{if } |t| \leq c \end{cases}$$

ψ function can be decomposed into the sum

$$\psi = \psi_a + \psi_b + \psi_c$$

where ψ_a is absolutely continous function having absolutely continous derivative, ψ_c is a continous, piecewise linear function which is constant in a neighborhood of $\pm\infty$, and ψ_s is a nondecreasing step function. In case of Huber loss function, $\psi_a = \psi_c = 0$. Now this function satisfy the following condtions. Jurečková and Sen (1996)

- **M1** : $S_n(\mathbf{Y})$ is regression invariant and scale invariant, $S_n > 0$ a.s. and $n^{\frac{1}{2}}(S_n - S) = \mathbf{O}_p(1)$

- **M2** : $H(t) = \int \rho((z-t)/S)dF(z)$ has the unique minimum at t=0.

- **M3** : For some $\delta > 0$ and $\eta > 1$,

$$\int_{\infty}^{\infty} [|z| sup_{|u|\leq\delta} sup_{|v|\leq\delta} |\psi_a''(e^{-v}(z+u)/S|]^{\eta} dF(z) < \infty$$

and

$$\int_{\infty}^{\infty} [|z|^2 sup_{|u|\leq\delta} |\psi_a''(e^{-v}(z+u)/S|]^{\eta} dF(z) < \infty$$

where $\psi_a'(z) = (d/dz)\psi_a(z)$ and $\psi_a''(z) = (d^2/dz^2)\psi_a(z)$.

- **M4** : ψ_c is a continous, piecewise linear function with knots at $\mu_1, \ldots, \mu_k$, which is constant in a neighborhood of $\pm\infty$. Hence the derivative ψ_c' is a step function

$$\psi_c'(z) = \alpha_\nu, \qquad \mu_\nu < z < \mu_{\nu+1}, \nu = 0, 1, \ldots, k,$$

where $\alpha_0, \alpha_1, \ldots, \alpha_k \in \Re_1, \alpha_0 = \alpha_k = 0$ and
$-\infty = \mu_0 < \mu_1 < \cdots < \mu_k < \mu_{k+1} < \infty$. Assume that f(z) is bounded in neighborhoods of $S\mu_1, \ldots, S\mu_k$.

- **M5** : $\psi_s(z) = \lambda_\nu$ for $q_\nu < z \leq q_{\nu+1}, \nu = 1, \ldots, m$ where
$-\infty = q_0 < q_1 < \cdots < q_{m+1} = \infty, -\infty < \lambda_0 < \lambda_1 < \cdots < \lambda_m < \infty$.

Assume that f(z) and f'(z) are bounded in neigborhood $Sq_1, \ldots, Sq_m$. The asymptotic representation of $\mathbf{M}_n$ is involved in the functionals

$$\begin{aligned} \gamma_1 &= S^{-1} \int_{-\infty}^{\infty} (\psi_a'(z/S) + \psi_c'(z/S)) dF(z) \\ \gamma_2 &= S^{-1} \int_{-\infty}^{\infty} z(\psi_a'(z/S) + \psi_c'(z/S)) dF(z) \end{aligned}$$

Moreover the following condtions are satisfied.

$$\begin{aligned} X1 \quad &. \quad x_{i1} = 1, i = 1, \ldots, n \\ X2 \quad &. \quad n^{-1} \sum_{i=1}^{n} ||\mathbf{x}_i||^4 = \mathbf{O}_p(1) \\ X3 \quad &. \quad lim_{n\to\infty} \mathbf{Q}_n = \mathbf{Q} \end{aligned}$$

where $\mathbf{Q}_n = n^{-1}\mathbf{X}'\mathbf{X}$ and $\mathbf{Q}$ is a positive definite $p \times p$ matrix. Then under these conditions, M_n is a solution of the system of equations

$$\sum_{i=1}^{n} \mathbf{x}_i \psi(\frac{Y_i - \mathbf{x}_i'\mathbf{t}}{S_n}) = \mathbf{0}.$$

To make $S_n(\mathbf{Y})$ regression invariant and scale invariant, $S_n(\mathbf{Y})$ is computed in the following manner. We use regression scores defined below. For $\alpha \in (0, 1)$, $\hat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n1}(\alpha), \ldots, \hat{a}_{nn}(\alpha))'$ is the optimal solution to maximize

$$\sum_{i=1}^{n} Y_i \hat{a}_{ni}(\alpha)$$

with the constraint:

$$\sum_{i=1}^{n} x_{ij} \hat{a}_{ni}(\alpha) = (1 - \alpha) \sum_{i=1}^{n} x_{ij}, j = 1, \ldots, G$$

.

Hajék(1965) proposed scores:

$$a_n^*(R_i, \alpha) = \begin{cases} 0 & \text{if } R_i/n < \alpha \\ R_i - n\alpha & \text{if } (R_i - 1)/n < \alpha < R_i/n \\ 1 & \text{if } \alpha < (R_i - 1)/n \end{cases}$$

Select a nondecreasing, square integrable function $\phi : (0,1) \to \mathcal{R}_1$ such that $\phi(\alpha) = -\phi(1-\alpha), 0 < \alpha < 1$. For a fixed number $\alpha_0 (0 < \alpha_0 < 1/2)$, assumed that ϕ is standardized

$$\int_{\alpha_0}^{1-\alpha_0} \phi^2(\alpha) d\alpha = 1.$$

Define the regression scores generated by ϕ.

$$\hat{b}_{ni} = -\int_{\alpha_0}^{1-\alpha_0} \phi(\alpha) d\hat{a}_{ni}(\alpha), i = 1, \ldots, n.$$

That is,

$$\hat{b}_{ni} = \begin{cases} n \int_{(R_i-1)/n}^{R_i/n} \phi(\alpha) a_n^{*'}(R_i, \alpha) d\alpha & \text{if } \alpha \le (R_i - 1)/n \le 1 - \alpha, R_i/n \le 1 - \alpha \\ n \int_{(R_i-1)/n}^{1-\alpha} \phi(\alpha) a_n^{*'}(R_i, \alpha) d\alpha & \text{if } \alpha \le (R_i - 1)/n \le 1 - \alpha, R_i/n > 1 - \alpha \\ n \int_{\alpha}^{R_i/n} \phi(\alpha) a_n^{*'}(R_i, \alpha) d\alpha & \text{if } \alpha > (R_i - 1)/n, R_i/n \le 1 - \alpha \\ 0 & \text{if } 1 - \alpha < (R_i - 1)/n \\ n \int_{\alpha}^{1-\alpha} \phi(\alpha) a_n^{*'}(R_i, \alpha) d\alpha & \text{else} \end{cases}$$

S_n is defined as

$$n^{-1} \sum_{i=1}^{n} Y_i \hat{b}_{ni} = n^{-1} \mathbf{X}' \hat{\mathbf{b}}_n.$$

Suppose that γ_1 is not equal to zero. The following theorem tells us about the asymptotic distribution of $\mathbf{M}_n$.

Theorem 4.2.1 *The sequence*

$$n^{\frac{1}{2}} \{\hat{\gamma}_1 (\mathbf{M}_n - \boldsymbol{\beta}) + \hat{\gamma}_2 (\frac{S_n}{S} - 1) \mathbf{e}_1\}$$

has the asymptotic G-dimensional normal distribution $\mathcal{N}_G(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1})$*, where* $\sigma^2 = \int_{-\infty}^{\infty} \psi^2(z/S) dF(z)$.

Jurečková and Sen (1996) The asymptotic variance of $n^{\frac{1}{2}}(\mathbf{M}_n - \boldsymbol{\beta} + c)$ is $\boldsymbol{\Delta}_k$, where $c = \hat{\gamma}_2(\frac{S_n}{S} - 1)\mathbf{e}_1$ and $\boldsymbol{\Delta}_k$ is $(\hat{\gamma}_1)^{-2}\hat{\sigma}^2\mathbf{Q}^{-1}$. The scale factor $(\hat{\gamma}_1)^{-2}\hat{\sigma}^2$ is the same for every possible permutation and does not affect the partitions of the following $\mathbf{Z}_k$ and $\mathbf{V}_k$. For simplicity, the term c could be disregarded for deriving a test statistics.
We are tempted to use the UIP to formulate a robust M-test for

$$H_{0k} : \boldsymbol{\theta}_k = \mathbf{0} \qquad H_{1k} : \boldsymbol{\theta}_k \geq \mathbf{0}$$

Let $\mathbf{Z}_k = \mathbf{A}\mathbf{M}_n$ and $\mathbf{V}_k = \mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^t$. Let $\mathcal{G} = \{1, \ldots, G-1\}$, and for every a : $\emptyset \subseteq a \subseteq \mathcal{G}$, let a' be its complement and $|a|$ its cardinality. For each a : $\emptyset \subseteq a \subseteq \mathcal{G}$, partition $\mathbf{Z}_k$ and $\mathbf{V}_k$ as

$$\mathbf{Z}_k = \begin{pmatrix} \mathbf{Z}_{ka} \\ \mathbf{Z}_{ka'} \end{pmatrix} \qquad \mathbf{V}_k = \begin{pmatrix} \mathbf{V}_{kaa} & \mathbf{V}_{kaa'} \\ \mathbf{V}_{ka'a} & \mathbf{V}_{ka'a'} \end{pmatrix}$$

and write

$$\mathbf{Z}_{ka:a'} = \mathbf{Z}_{ka} - \mathbf{Z}_{kaa'}\mathbf{V}^{-1}_{ka'a'}\mathbf{Z}_{ka'},$$
$$\mathbf{V}_{kaa:a'} = \mathbf{V}_{kaa} - \mathbf{V}_{kaa'}\mathbf{V}^{-1}_{ka'a'}\mathbf{V}_{ka'a}$$

By virtue of weak convergence of $n^{\frac{1}{2}}(\mathbf{M}_n - \boldsymbol{\beta})$ to a G-variate normal law, for n very large, we got $(n\mathbf{V}_k^{-1})^{(1/2)}(\mathbf{Z}_k - \boldsymbol{\theta}_k) \to^{\mathcal{D}} \mathcal{N}_{G-1}(\mathbf{0}, \mathbf{I})$
Let then

$$L_k = \Sigma_{\emptyset \subseteq a \subseteq \mathcal{G}} I(\mathbf{Z}_{ka:a'} > \mathbf{0}, \mathbf{V}^{-1}_{ka'a'}\mathbf{Z}_{ka'} \leq \mathbf{0})(n\mathbf{Z}'_{ka:a'}\mathbf{V}^{-1}_{kaa:a'}\mathbf{Z}'_{ka:a'})$$

and rejecting the null hypothesis for large positive values.
Typically, we are dealing with high dimension(K) low sample size dataset. Our case

pertains to the case when n is small and the asymptotic normality does not hold. On the other hand, the permutation distribution theory is valid for small sample size setup. Under the null hypothesis of homogeneity, the joint distribution of all n observations remains invariant under any permutation, leading to manageable testing procedures. There are all possible $\frac{n!}{n_1!n_2!\cdots n_G!}$ equally likely permutations, which is a large number to overcome this problem. Hence, conditionally distribution-free tests can be constructed by using the permutational invariance structure. Now, p-value can be computed as below.

$$P_k = Pr(L_k \geq l_k), k = 1, \ldots, K$$

where L_k is a test statistic from the permuted distribution and l_k is an observed test statistics. And then an appropriate multiple testing procedure may be applied to the K p-values. As a multiple testing procedure, proposed FDR procedure is used to determine which gene has monotone increasing pattern across the groups.

CHAPTER 5

NUMERICAL STUDY

5.1 Numerical Study of FDR in DNA microarray experiment

Simulation study can provide a concrete description of performance of FDR estimators. It is useful for numerical evaluation of performance in a large number of hypotheses, that is, many genes. We assess the performance of our FDR procedure with that of other procedures in terms of FDR control and power: Storey's FDR, the Benjamini-Hochberg procedure (BH), the Benjamini and Liu procedure (BL), and Lehmann and Romano's FDR (Lehmann). In this simulation study, these five different FDR procedures are computed for different values of α and the proportion of true null hypotheses, π_0. In particular, we examine the amount of improvement offered by the different procedures in terms of controlling FDR. We also present the performance of proposed pFDR compared to Storey's pFDR. First, numerical results are shown in both an independent p-value example and a dependent p-value example. In each of these examples, the average power is defined to be the average probability of rejecting the false null hypotheses, $E(S)/m_1$. At the final point, proposed FDR is applied to real data: Leukemia study of Golub.et al to illustrate the performance.

Suppose we collect data from n microarrys with the same m genes. We observe the vectors $\mathbf{X}(j) = (X_{1j}, \ldots, X_{mj})$, $j = 1, \ldots, n$. We assume that the X_{ij} are independent across the n arrays or n observations for each gene i, but they are not necessarily independent or identically distributed across m genes of the vector for each j. In other words, the data may be expressed as a $m \times n$ matrix $\mathbf{X}$ with dependent rows and indepedent columns. We construct a test statistic T_i , a function of $X_{i1}, \ldots, X_{in}$.

5.1.1 Independence example

We consider a multiple hypothesis testing situation where each independent random variable T_i has mean μ_i and the same variance 1, i=1, . . . , 1000. The problem is to test 1000 one-sided hypothesis of $\mu = 0$ against $\mu > 0$ with the null distribution N(0,1) and alternative distribution N(2,1). Each individual hypothesis is tested by a z-test. We let $m_0 = 100, 400, 700$ and generated 1000 indepedent sets of 1000 normal random variables for each m_0-value. Proposed FDR and other procedures are computed at the FDR level $\alpha = 0.1, 0.05$, and 0.01, respectively. The actual (true) FDR is estimated by averaging the Q values over 1000 iterations Storey (2002)Yekutieli and Benjamini (1999)Pawitan et al. (2006). We also present a numerical study to compare the average power of our proposed FDR controlling procedure with other procedures. Table I presents different FDR procedures such as Storey's FDR, the Benjamini and Hochberg procedure (BH), the Benjamini and Liu procedure (BL), Lehmann and Romano's FDR (Lehmann) with Proposed FDR. All FDR procedures seem to control the FDR at α under independence. The increase in the proportion of the true null hypotheses (π_0), the greater FDRs are. All FDR procedures increase as α increases. It can be seen that Proposed FDR is relatively close to the actual FDR. The another point is that we don't want to report a smaller false discovery rate than truly exists. Hence, proposed FDR is a consistently

conservative point estimate of the FDR at all levels simultaneously. Proposed FDR performs better than other procedures, because the average power of proposed FDR is always greater than other procedures in figure I. As α increases, average power also increases. Proposed FDR offers a more powerful alternative to the traditional Benjamini and Hochberg procedure. We lose no power regardless of the value of π_0 and α. Table II presents Proposed pFDR gets better than Storey's pFDR.

TABLE I: Comparison of different FDR procedures (Independence)

α	π_0	Storey's FDR	BH	BL	Lehmann	Actual FDR	Proposed FDR
0.1	0.1	0.00072	0.0099	0.0004	0.0004	0.00075	0.00078
	0.4	0.00408	0.04023	0.00124	0.00124	0.00424	0.00395
	0.7	0.01063	0.07101	0.00768	0.00768	0.01615	0.01199
0.05	0.1	0.00051	0.00494	0.00016	0.00016	0.00047	0.00045
	0.4	0.00297	0.01986	0.00121	0.00125	0.00215	0.00276
	0.7	0.01048	0.03522	0.00447	0.00456	0.01009	0.00997
0.01	0.1	0.00024	0.00108	0	0	0.00017	0.0003
	0.4	0.00143	0.00378	0.00088	0.00088	0.0013	0.0012
	0.7	0.00522	0.00645	0.00138	0.00138	0.00431	0.0037

TABLE II: Comparison of different pFDR procedures (Independence)

α	π_0	pFDR	Proposed pFDR
0.1	0.1	0.00129	0.00108
	0.4	0.00759	0.00355
	0.7	0.026561	0.01199
0.05	0.1	0.00154	0.00075
	0.4	0.00902	0.00276
	0.7	0.03178	0.00797
0.01	0.1	0.00309	0.00032
	0.4	0.01859	0.00119
	0.7	0.067889	0.00367

5.1.2 Dependence example

Our second numerical study illustrates the performance of our FDR procedure compared to the other procedures including Benjamini and Hochberg procedure under a certain form of dependence. The null statistics have N(0,1) marginal distributions and the alternative distributions have marginal distribution N(3,1) with different value of m_0. 1000 dependent random variable T_i with mean μ_i ,the same variance 1 and common correlation ρ, i=1, . . . , 1000 are generated. 1000 one-sided hypothesis tests of $\mu = 0$ against $\mu > 0$ are tested by a z-test. We let $m_0 = 100, 400, 700$ and generated 1000 sets of 1000 normal random variables for each m_0-value. Our proposed FDR and other procedures are applied at each $\rho = 0.3, 0.5$, and 0.7 and $\alpha = 0.1$. In this section, we will show that our FDR procedure performs better under all configurations of dependence structures.

Table III compares different FDR procedures such as Storey's FDR, the Benjamini and Hochberg procedure (BH), the Benjamini and Liu procedure (BL), Lehmann and Romano's FDR (Lehmann) with proposed FDR. All five FDR procedures seem to control the FDR when the proportion of true null hypotheses is less than 1. It is shown that Proposed FDR procedure is very close to the actual FDR. As ρ increases, all FDR procedures decreases. Proposed FDR procedure controls the FDR at all levels under all configurations of dependence structures. Figure II is the graphical summary to show that proposed FDR is more powerful than other procedures. Table IV compares the performance of Proposed pFDR with Storey's pFDR. Proposed pFDR is always smaller than Storey's pFDR.

TABLE III: Comparison of different FDR procedures (Dependence)

ρ	π_0	Storey's FDR	BH	BL	Lehmann	Actual FDR	Proposed FDR
0.3	0.1	0.00428	0.00959	0.00024	0.00025	0.00569	0.00467
	0.4	0.02410	0.03943	0.00054	0.00059	0.03382	0.02784
	0.7	0.08264	0.06069	0.00215	0.00228	0.09142	0.08446
0.5	0.1	0.00355	0.01079	0.00085	0.00076	0.00689	0.00583
	0.4	0.01887	0.04229	0.00096	0.00117	0.03527	0.03135
	0.7	0.06129	0.06920	0.01113	0.01099	0.09963	0.09013
0.7	0.1	0.00279	0.01155	0.00357	0.00332	0.00779	0.00736
	0.4	0.02499	0.03526	0.00769	0.00775	0.03042	0.03484
	0.7	0.11931	0.05009	0.01224	0.01226	0.07246	0.09323

TABLE IV: Comparison of different pFDR procedures (Dependence)

ρ	π_0	pFDR	Proposed pFDR
0.3	0.1	0.00467	0.00428
	0.4	0.02784	0.02410
	0.7	0.0846	0.08264
0.5	0.1	0.00583	0.00355
	0.4	0.03887	0.01887
	0.7	0.1000	0.06129
0.7	0.1	0.00736	0.00279
	0.4	0.02499	0.03484
	0.7	0.11931	0.10323

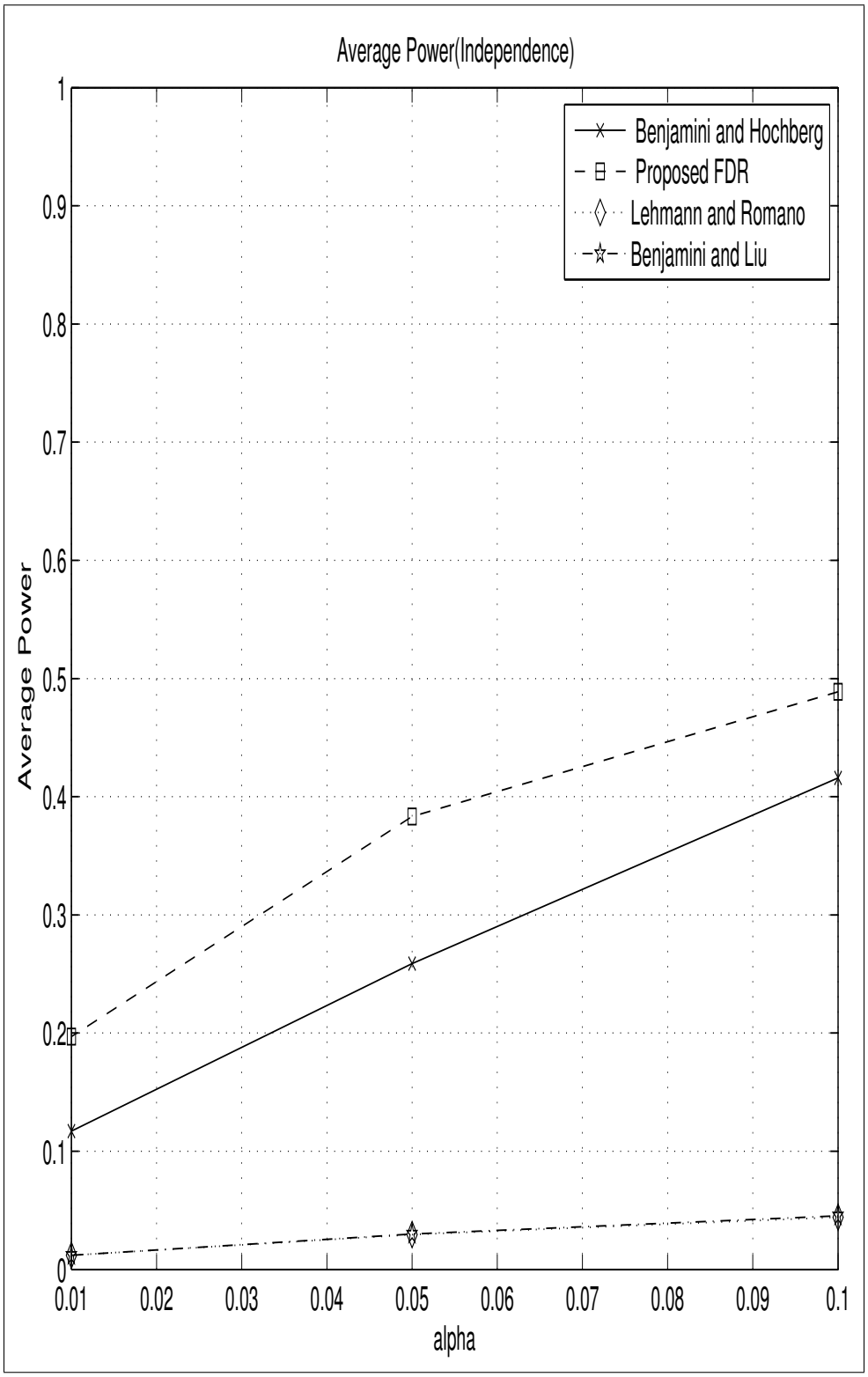

FIGURE I: Comparison of Average Power for different FDR procedures (Independence)

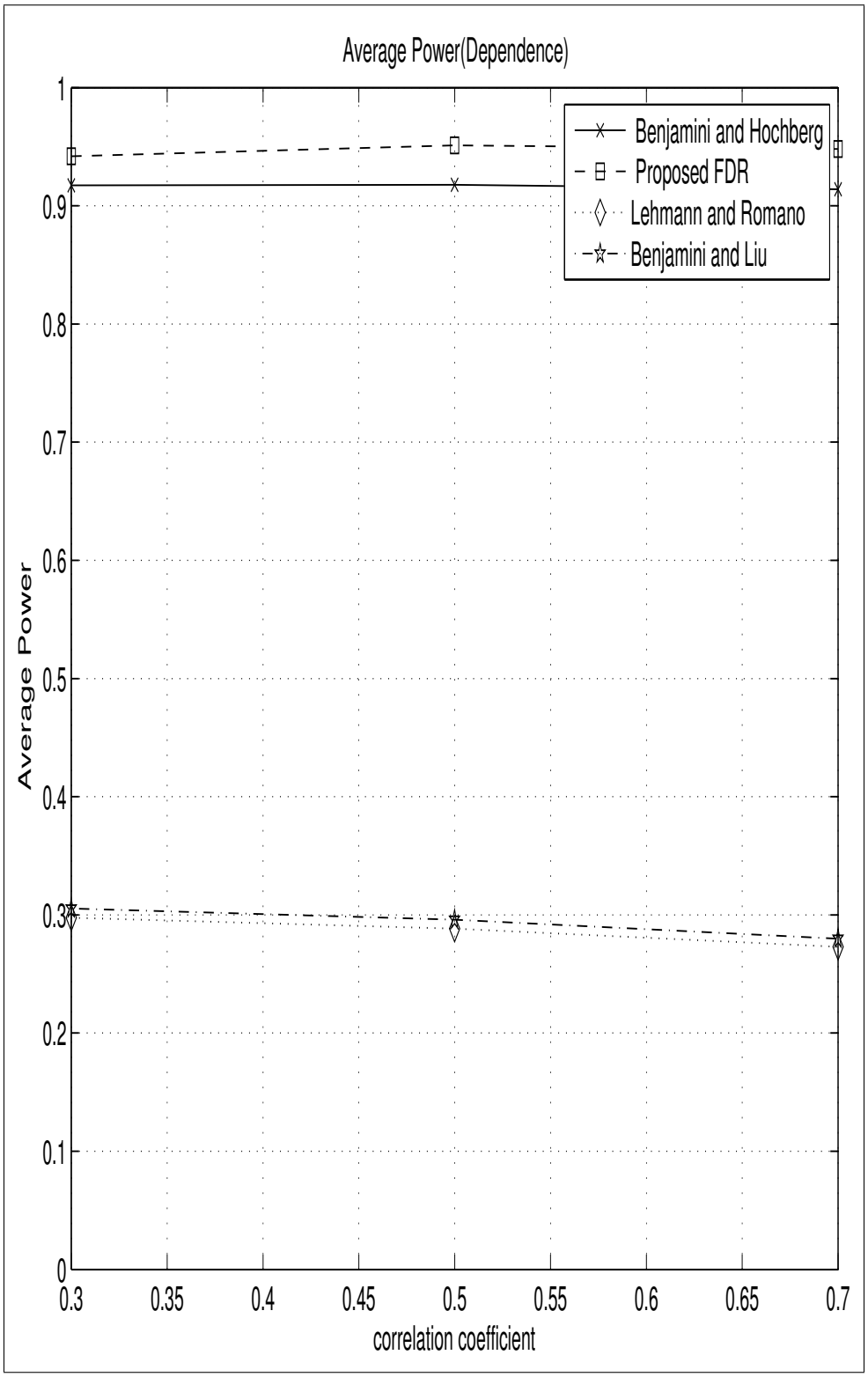

FIGURE II: Comparison of Average Power for different FDR procedures (Dependence)

5.1.3 Application to Real Data: Leukemia study

Correct diagnosis of neoplasia malignancies is necessary for proper treatment. Microarray technologies provided the means by which neoplasms can be more accurately classified, thus leading to effective treatment. Golub et al. (1999) studied two hematologic malignancies: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). They expected these two malignancies could be identified based on microarray gene expression measures. They measured gene expression levels using Affymetrix high-density oligonucleotide chips. The goal of this study is to identify differentially expressed genes between the two diseases. Gene expression data has 7129 genes and 38 tumor mRNA samples. Pre-processing was done as described in Dudoit et al. (2002). Differentially expressed genes in ALL and AML patients were identified.

Figure III shows the randomness assumption of array effects are reasonable. Figure IV compares the original expression level with signed square root transformation(Z score transformation) of the expression level. By transforming the original expression level, this Z score transformation, provides a way of standardizing data and allows the comparison of microarray data independent of the original hybridization intensities. Data normalized by Z score transformation can be used in the calculation of significant changes in gene expression between different samples and conditions Cheadle et al. (2003).

Each p-value is computed based on welch two sample t-statistics for each gene. We assumed that the $P_i's$ are uniformly distributed among the m_0 genes. This assumption is appropriate and realistic. Figure V displays that p-values, greater than 0.001, are uniformly disributed and the assumptions among the non-disease genes, m_0 genes are reasonable.

We will numerically compare the performance of proposed FDR procedure with others for different values of m_0. The study was performed for π_0=0.2,0.4,0.6. Table V compares different FDR procedures such as Storey's FDR (FDR), the Benjamini and Hochberg procedure (BH) and Proposed FDR. It can be shown that the increase in the proportion of the true null hypotheses (π_0), that is, non-disease genes is greater, the greater FDRs and pFDRs are. Proposed FDR controls the FDR at all levels α=0.1 (threshold=0.0004),0.05 (threshold=0.0002),0.01 (threshold=0.000175) ,whereas the Benjamini and Hochberg procedure having greater power fails to control the FDR. It turns out that Proposed FDR is more amenable in real microarray data structures. Storey's method of estimating π_0 works well in the continous gene expression levels data. Table VII displays Modified FDR using the estimate of π_0 (=0.4). Table VIII presents Storey's pFDR (pFDR) with Proposed pFDR. Proposed pFDR is smaller than Storey's pFDR. Table VI display the 30 most significant genes in the dataset at FDR level=0.1.

TABLE V: Different FDRs In Real Data

α	π_0	Storey's FDR	BH	Proposed FDR
0.1	0.20	0.00197	0.0687	0.01686
	0.40	0.0025	0.14964	0.0434
	0.60	0.0036	0.2658	0.0997
0.05	0.20	0.0013	0.0534	0.0115
	0.40	0.00169	0.1302	0.02879
	0.60	0.00256	0.2445	0.04877
0.01	0.20	0.0012	0.04279	0.0081
	0.40	0.00159	0.10617	0.00934
	0.60	0.00236	0.19248	0.0100

TABLE VI: Displaying the 30 most significant genes at FDR=0.1

pvalues	gene.names
1.381111e-10	C-myb gene extracted from Human (c-myb) gene, comp
2.138241e-10	FAH Fumarylacetoacetate
3.837362e-09	Zyxin
6.082366e-09	Leukotriene C4 synthase (LTC4S) gene
2.221575e-08	TCF3 Transcription factor 3 (E2A immunoglobulin en
2.517146e-08	RETINOBLASTOMA BINDING PROTEIN P48
3.740919e-08	CTPS CTP synthetase
5.867391e-08	CCND3 Cyclin D3
6.796881e-08	Clone 22 mRNA, alternative splice variant alpha-1
8.590343e-08	MB-1 gene
8.639399e-08	LEPR Leptin receptor
9.888047e-08	Thrombospondin-p50 gene extracted from Human throm
1.352416e-07	PROTEASOME IOTA CHAIN
1.820797e-07	RPA1 Replication protein A1 (70kD)
1.890900e-07	MYL1 Myosin light chain (alkali)
2.368127e-07	TOP2B Topoisomerase (DNA) II beta (180kD)
2.574041e-07	ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 s
2.796545e-07	Cytoplasmic dynein light chain 1 (hdlc1) mRNA
3.576030e-07	CST3 Cystatin C (amyloid angiopathy and cerebral h
4.776810e-07	GB DEF = Homeodomain protein HoxA9 mRNA
5.193354e-07	LYN V-yes-1 Yamaguchi sarcoma viral related oncoge
5.590720e-07	PRG1 Proteoglycan 1, secretory granule
6.749931e-07	Transcriptional activator hSNF2b
6.875291e-07	CYP2C18 Cytochrome P450, subfamily IIC (mephenytoi
7.288953e-07	Liver mRNA for interferon-gamma inducing factor(IG
7.733038e-07	Inducible protein mRNA
8.367065e-07	Catalase (EC 1.11.1.6) 5primeflank and exon 1 mapp
8.565317e-07	CD33 CD33 antigen (differentiation antigen)
9.520948e-07	CARCINOEMBRYONIC ANTIGEN PRECURSOR
9.716227e-07	MCM3 Minichromosome maintenance deficient (S. cere

TABLE VII: Modified FDR at $\pi_0 = 0.4$

α	Modified FDR
0.1	0.0422
0.05	0.02819
0.01	0.009224

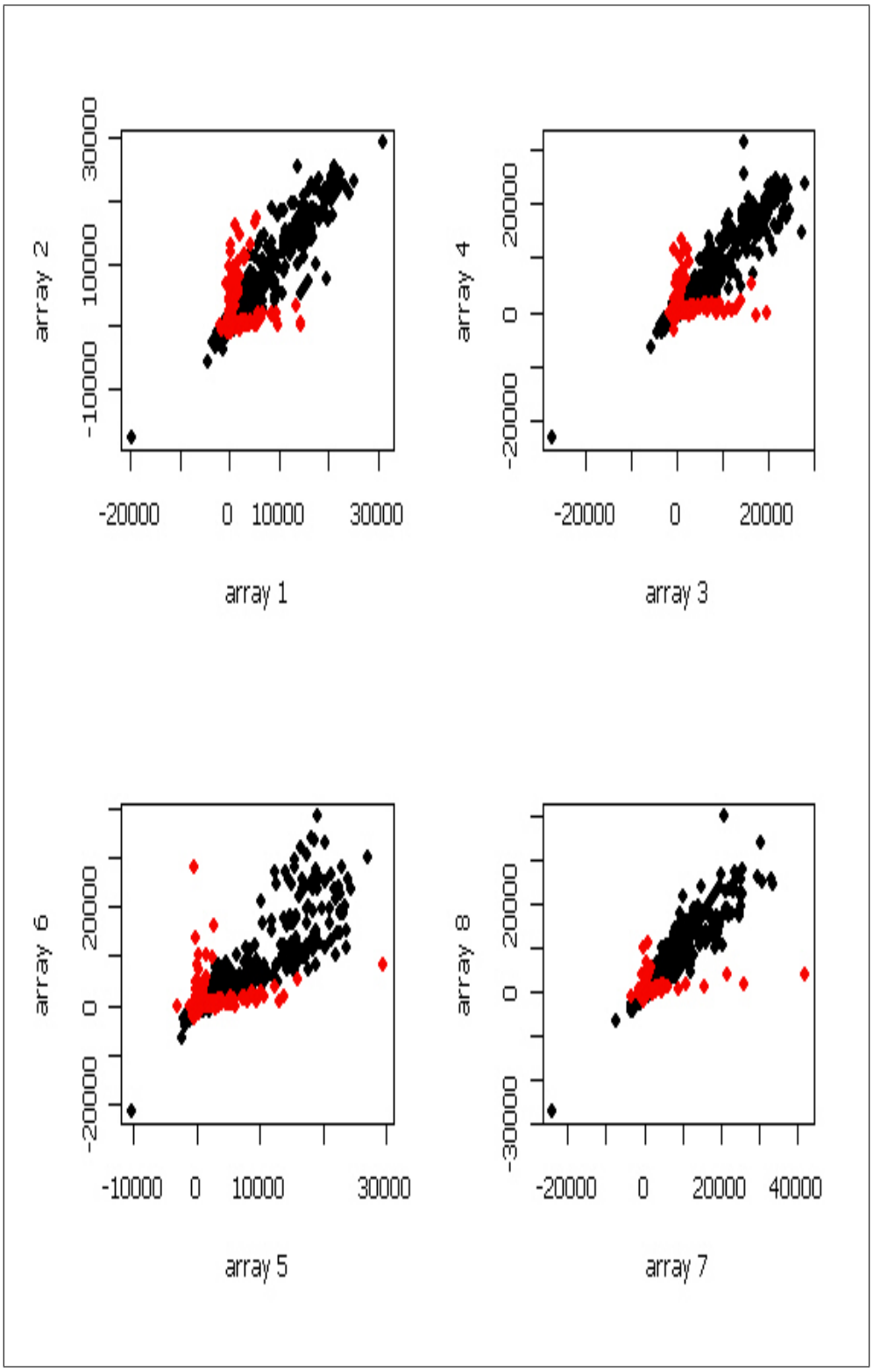

FIGURE III: Comparison of arrays

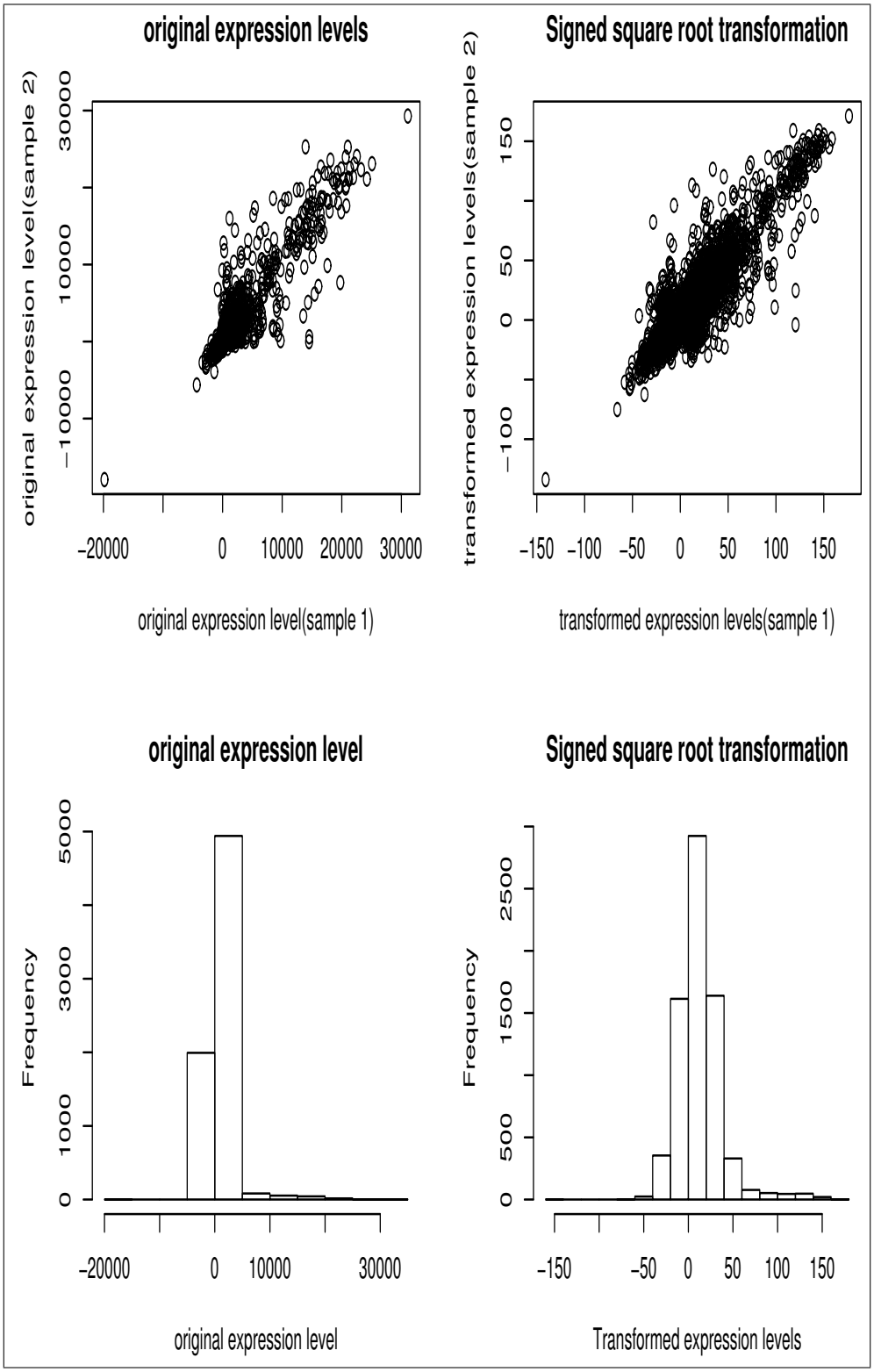

FIGURE IV: Comparions of expression level with signed square root transformation of expression level

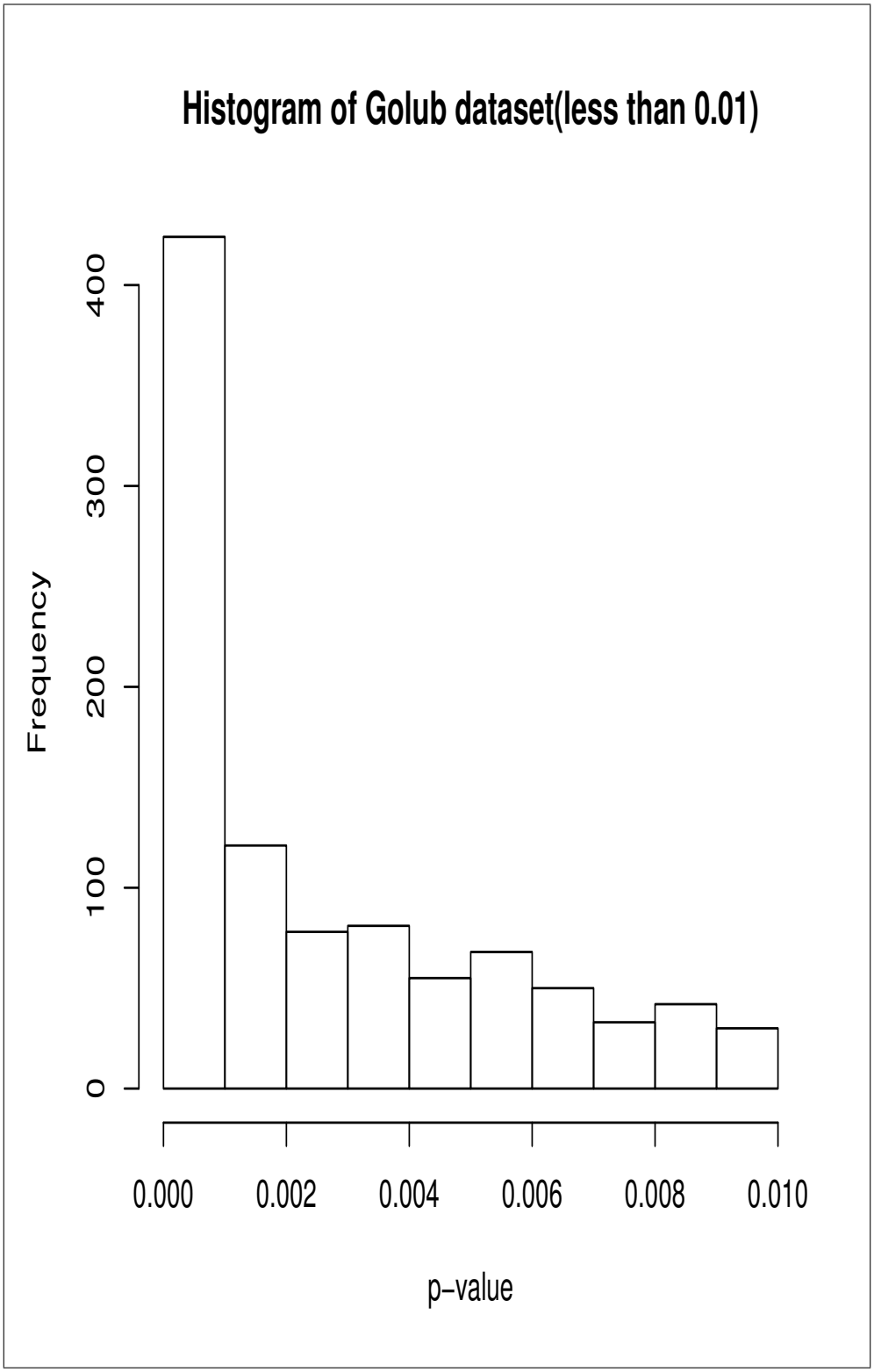

FIGURE V: Distribution of p-value (Real data)

TABLE VIII: Comparison of different pFDR procedures (Real data):Golub et al.

α	π_0	pFDR	Proposed pFDR
0.1	0.20	0.02558332	0.0169
	0.40	0.03257848	0.04339
	0.60	0.04633	0.09975
0.05	0.20	0.0328	0.0115
	0.40	0.04318	0.02879
	0.60	0.06519	0.0588
0.01	0.20	0.0349	0.0081
	0.40	0.0464	0.01934
	0.60	0.0687	0.04907

5.2 FDR in genomic sequences

5.2.1 Application to The SARSCoV RNA Genome

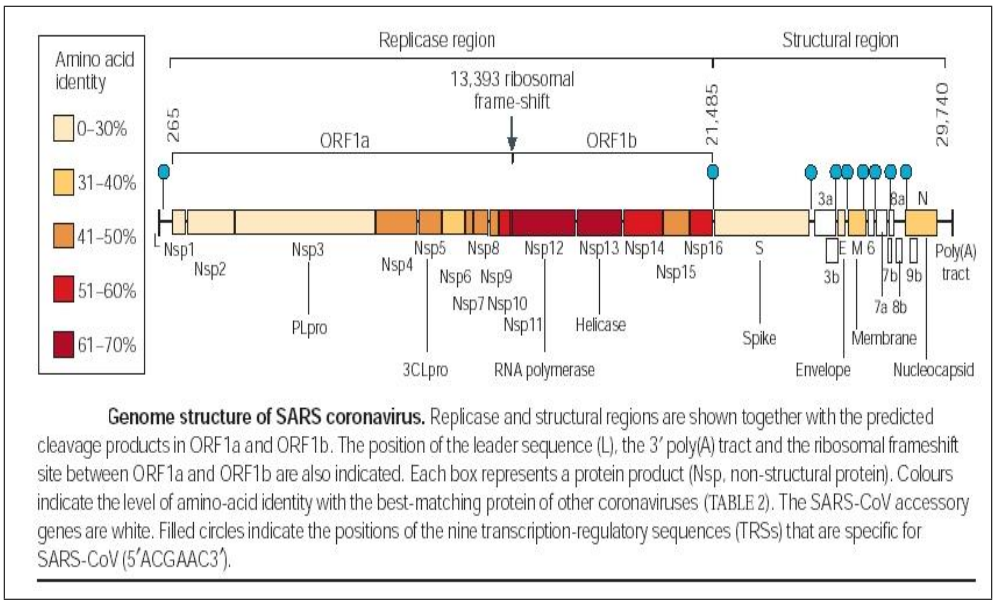

FIGURE VI: The SARSCoV RNA Genome

Following its origin in Southern China, the SARS epidemic resulted in 8422 infected people and 916 deaths. The SARS causative agent was identified as a coronavirus

(SARSCoV) with GenBank database. The SARS epidemic has an identified single-stranded and positive-sense RNA virus with large genome size and moderate mutation rate. For these sequences, an enormously high-dimensional purely qualitative categorical model is constructed. n (=14) SARS complete sequences are downloaded ,isolated from Guadong, Beijing, Hongkong and Taiwan: $n_1(= 5)$ from Guadong, $n_2(= 4)$ from Beijing, $n_3(= 3)$ from Hongkong and $n_4(= 2)$ from Taiwan. To simplify the measurement of variation, the sequences with no nucleotide changes are removed. The responses consist of not even ordered categories, a,c,g,and t and an ordering may not be feasible. The Hamming distance give us a stochastic ordering. But individual statistics using Hamming distance do not have a known null hypothesis distribution in general. For these reasons, we use jackknife variance estimation $\hat{\zeta}_{1.k}$ and permutation distribution to construct some permutation tests. There are K=900 genes (or positions) for each sequence and for each position, the test statistic described and corresponding p-value are computed. The test statistic L_k is defined as $\sum_{g=1}^{4} n_g[U_{gk} - U_k]^2/(4\hat{\zeta}_{1.k}), \quad k = 1, \ldots, 900$. The permutation distribution can be generated by considering 14!(equally likely) permutations of the combined sample observations among four groups of (sizes 5,4,3 and 2). Based on 900 p-values from the positons, each p-value are computed from the permuted distribution. Storey's FDR procedure, Benjamini and Hochberg procedure, Benjamini and Liu procedure, and Proposed FDR are computed and compared with $\alpha = 0.1$ (threshold=0.003),0.05 (threshold=0.0015) in Table IX. It turns out that the Benjamini and Hochberg (BH) and the Benjamin and Liu (BL) procedures don't work well in this genomic data. Table X presents proposed pFDR with Storey's pFDR. Proposed pFDR is always smaller than Storey's pFDR.

TABLE IX: Comparison of different FDR procedures-Hamming distance

α	π_0	Storey's FDR	BH	BL	Proposed FDR
0.1	0.40	0.037	0.038	0	0.0134
	0.60	0.054	0.049	0	0.0384
	0.80	0.103	0	0	0.0898
0.05	0.40	0.0209	0.034	0	0.007
	0.60	0.0313	0.0203	0	0.0157
	0.80	0.0616	0	0	0.0402

TABLE X: Comparison of different pFDR procedures-Hamming distance

α	π_0	pFDR	Proposed pFDR
0.1	0.40	0.037	0.0234
	0.60	0.054	0.0484
	0.80	0.1032	0.0898
0.05	0.40	0.022	0.007
	0.60	0.032	0.0257
	0.80	0.064	0.0402

5.3 Numerical Study in Classification Of Genes

Mitogenesis in hormone-responsive breast cancer cells may be stimulated by the steroid hormone estrogen. The cDNA microarray gene expression levels of a hormone-responsive breast cancer epithelial cell line with a mitogenic dose of estrogen without other confounding growth factors in serum ,were examined. Gene expression changes were measured at 6 time points 1, 4, 12, 24, 36, and 48 hours after estrogen stimulation. The expression levels of DNA replication fork genes stimulated by estrogen, without growth factors in serum, shows that the steroid hormone estrogen plays a important role of generating Mitogenesis. (Molecular Endocrinology 16, 2002). For the purpose of illustration, the data set in Lobenhofer et al. (2002) is analyzed. The data consists of 1900 genes measured at 6 time points with 8 observations (n=8) each time point. Gene expression levels are log-transformed. But the dataset to which we applied the analysis contains 1000 genes and 5 time points (1, 4, 12, 24, 36 hours

after estrogen stimulation), at which each group has 4,3,2,2,and 1 observations, respectively. Figure VII and VIII shows mean expression level changes at 5 time points. The patterns over time include monotone nondecreasing, monotone nonincreasing, up-down, and down-up profiles. The pattern of interest is the monotone nondecreasing profile over time. We then express these in term of inequalities between the expected expression levels at 5 time points.

5.3.1 Application To the Breast Cancer Study

We generate p-values in terms of 6 test statistics: Proposed FDRs (PLF) using 3 rank score statistics in linear rank statistics (uniform(Wilcoxon)(U), Normal(N),and logistic(L),respectively) Proposed FDRs(PMF) using robust M-estimator with the regression scores generated by ϕ which can be Normal (N) or uniform (U) and Kendall-tau statistic. Proposed pFDRs are defined similarly.

Table XI displays comparison of FDR procedures with application to Breast data. The study was performed for $\pi_0 = 0.3, 0.5$, and 0.70. Proposed FDR always controls the FDR at all levels α=0.1 (threshold=0.001), 0.05 (threshold=0.0005), regardless of test statistics used. The increase in the proportion of the true null hypotheses(π_0), the greater FDRs and pFDRs are, except for the Benjamini-Hochberg procedure (BH). The Benjamini-Hochberg procedure (BH) and Storey's FDR fail to control the FDR at some α. Moreover, they have higher variability of the standard estimate of the false discovery rate, so these FDR methodologies are far from the true FDP. Figure X shows that proposed FDR using p-values generated by linear rank statistics with normal scores not only attains greater powers regardless of α but also reports smaller FDR estimates. When the distribution of the test statistics is Gaussian, this will be better. Proposed FDR is more feasible in this microarray data. Table XII presents Proposed pFDR is always smaller than Storey's pFDR.

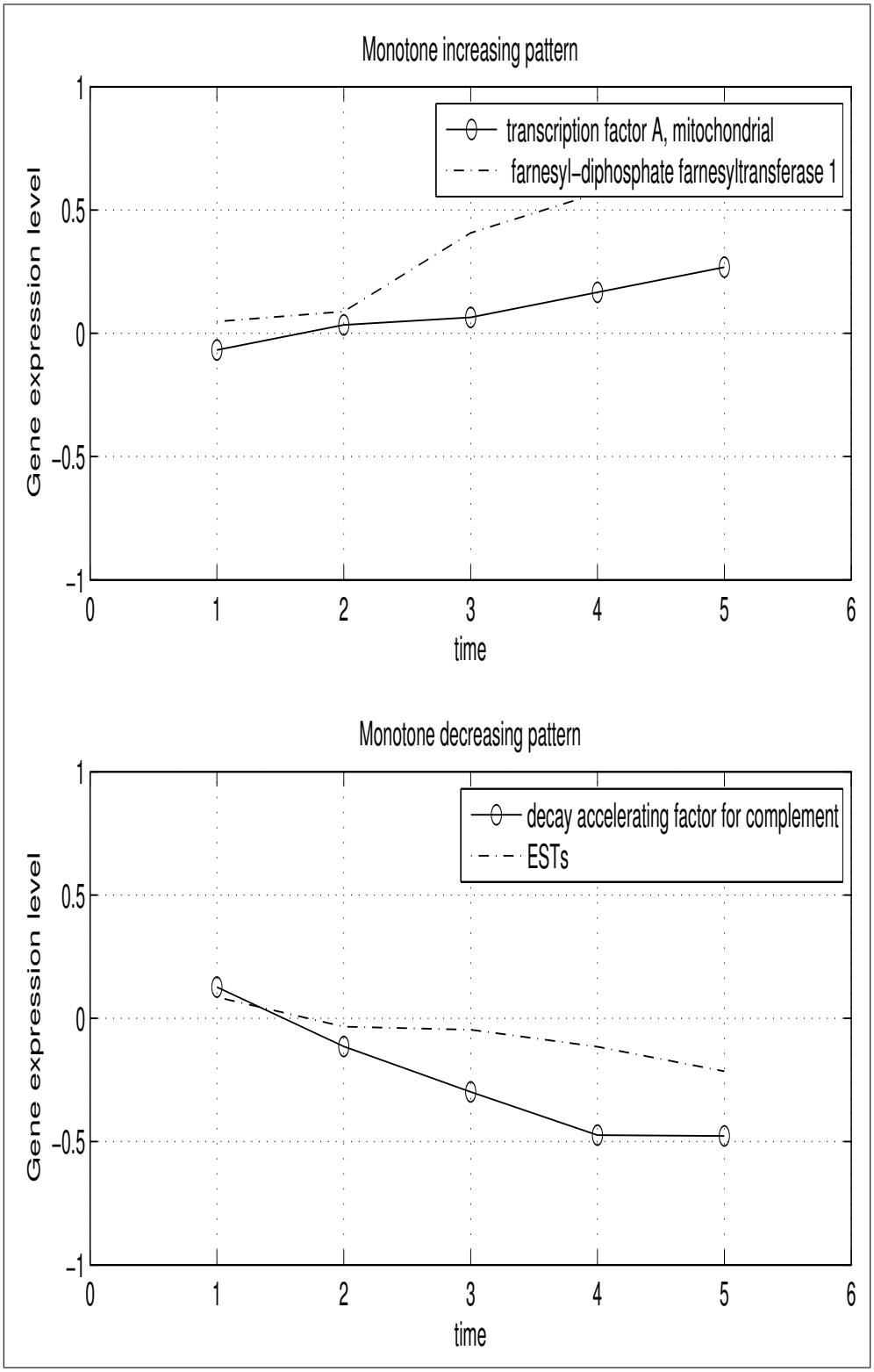

FIGURE VII: Mean expression levels for monotone profiles

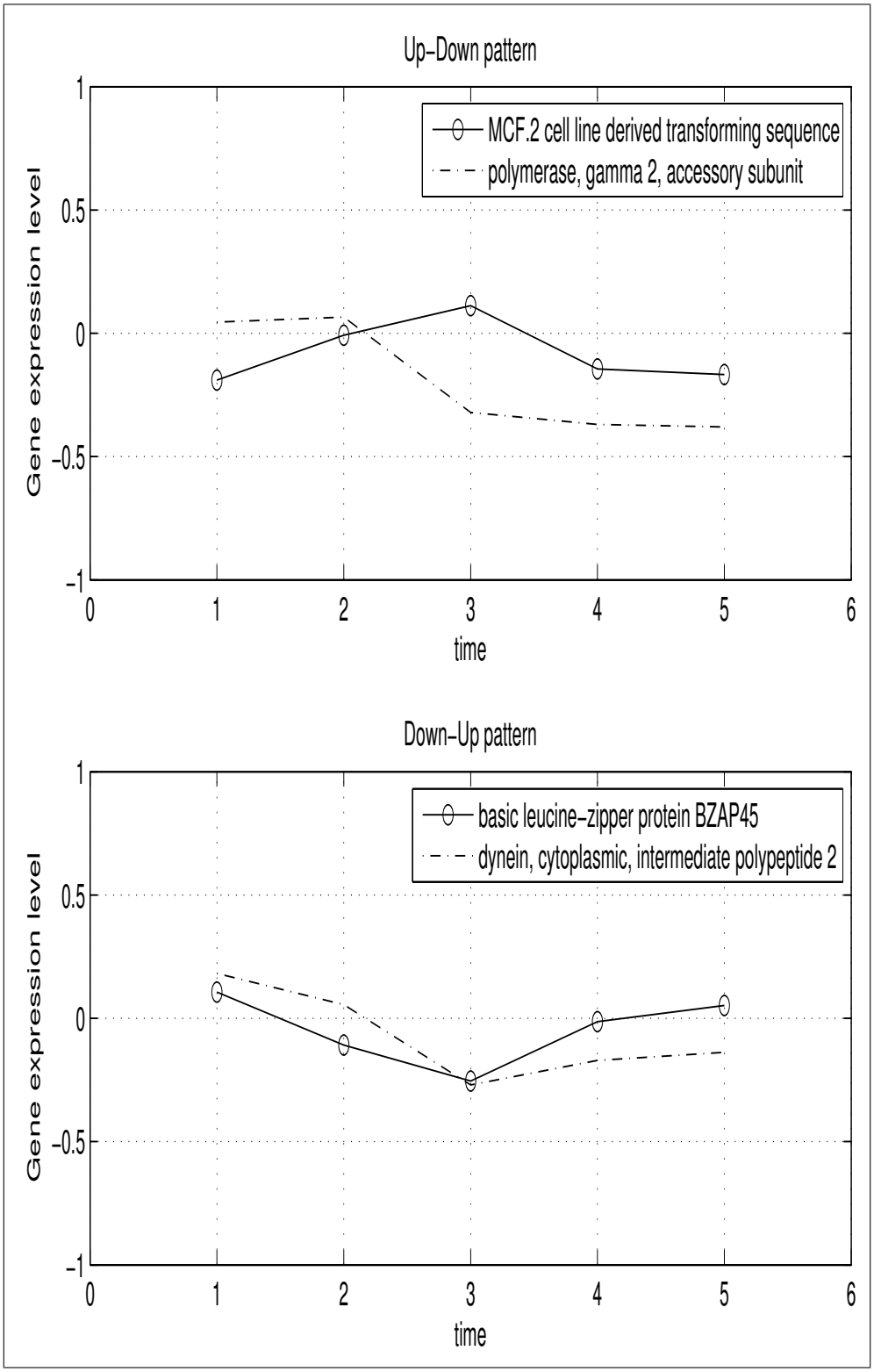

FIGURE VIII: Mean expression levels for other profiles

TABLE XI: Comparison of different FDR procedures (Breast data):

α	π_0	Test statistics	Storey's FDR	BH	Proposed FDR
0.1	0.3	linear rank statistics(Uniform)	0.049	0.089	0.021
		...(Normal)	0.049	0.085	0.020
		...(logistic)	0.084	0.003	0.030
		M-estimator(Normal)	0.038	0.168	0.027
		...(Uniform)	0.039	0.164	0.026
		Kendall-tau	0.039	0.163	0.025
	0.5	linear rank statistics(Uniform)	0.066	0.041	0.051
		...(Normal)	0.066	0.040	0.037
		...(logistic)	0.112	0.000	0.049
		M-estimator(Normal)	0.052	0.123	0.052
		...(Uniform)	0.052	0.125	0.051
		Kendall-tau	0.050	0.122	0.049
	0.7	linear rank statistics(Uniform)	0.098	0.00	0.078
		...(Normal)	0.100	0.00	0.078
		...(logistic)	0.170	0.000	0.100
		M-estimator(Normal)	0.074	0.007	0.095
		...(Uniform)	0.073	0.006	0.093
		Kendall-tau	0.070	0.004	0.089
0.05	0.3	linear rank statistics(Uniform)	0.043	0.00	0.019
		...(Normal)	0.044	0.00	0.022
		...(logistic)	0.082	0.000	0.030
		M-estimator(Normal)	0.038	0.000	0.023
		...(Uniform)	0.038	0.000	0.042
		Kendall-tau	0.035	0.001	0.040
	0.5	linear rank statistics(Uniform)	0.058	0.000	0.036
		...(Normal)	0.059	0.000	0.040
		...(logistic)	0.113	0.000	0.047
		M-estimator(Normal)	0.050	0.000	0.042
		...(Uniform)	0.050	0.000	0.042
		Kendall-tau	0.048	0.002	0.040
	0.7	linear rank statistics(Uniform)	0.088	0.00	0.050
		...(Normal)	0.088	0.00	0.047
		...(logistic)	0.174	0.000 0.050	
		M-estimator(Normal)	0.082	0.000	0.047
		...(Uniform)	0.082	0.006	0.047
		Kendall-tau	0.079	0.0045	0.045

TABLE XII: Comparison of different pFDR procedures (Breast data)

α	π_0	Test statistics	Storey's FDR	Proposed pFDR
0.1	0.3	linear rank statistics(Uniform)	0.052	0.021
		...(Normal)	0.050	0.020
		...(logistic)	0.084	0.003
		M-estimator(Normal)	0.040	0.027
		...(Uniform)	0.041	0.026
		Kendall-tau	0.040	0.023
	0.5	linear rank statistics(Uniform)	0.066	0.041
		...(Normal)	0.069	0.040
		...(logistic)	0.118	0.059
		M-estimator(Normal)	0.054	0.052
		...(Uniform)	0.054	0.052
		Kendall-tau	0.050	0.049
	0.7	linear rank statistics(Uniform)	0.103	0.078
		...(Normal)	0.105	0.079
		...(logistic)	0.178	0.100
		M-estimator(Normal)	0.078	0.095
		...(Uniform)	0.077	0.093
		Kendall-tau	0.074	0.090
0.05	0.3	linear rank statistics(Uniform)	0.049	0.020
		...(Normal)	0.050	0.022
		...(logistic)	0.094	0.030
		M-estimator(Normal)	0.043	0.023
		...(Uniform)	0.042	0.022
		.Kendall-tau	0.040	0.021
	0.5	linear rank statistics(Uniform)	0.066	0.036
		...(Normal)	0.067	0.040
		...(logistic)	0.128	0.050
		M-estimator(Normal)	0.058	0.042
		...(Uniform)	0.058	0.042
		Kendall-tau	0.057	0.041
	0.7	linear rank statistics(Uniform)	0.099	0.050
		...(Normal)	0.100	0.049
		...(logistic)	0.197	0.050
		M-estimator(Normal)	0.088	0.045
		...(Uniform)	0.088	0.043
		Kendall-tau	0.085	0.040

CHAPTER 6

SUMMARY AND FUTURE RESEARCH

6.1 Summary and Conclusion

My thesis consists of three topics: False discovery rate in microarray studies, False discovery rate in genomic sequences, and Classification of genes. For estimating false discovery rates in microarray setting, we wanted to consider more general dependence structures among tested genes. We utilized the Chen-Stein method to derive the Poisson distributions of V and R, respectively. This was derived from fairly mild regularity conditions regarding the dependence of the genes: the classification into two subsets of non-differentially expressed genes and differentially expressed genes crucial to sort plausible dependence patterns out. These estimation procedure has an advantage of not needing unfeasible conditions under which central limit theorems apply and it prevents the standard estimate of FDR from being increased due to ignoring high correlations. A primary goal of developing FDR procedure under this framework is to minimize FDR level and increase the associated power. Two-stage FDR procedure by adding one more rejection procedure has these desirable

properties. Besides, this proposed FDR procedure is always controlled at preassigned overall significance level α. We also developed proposed pFDR procedure as well. In the simulated data example and real data example, the proposed FDR procedure provides exact estimation to actual FDR and has greater power than other conventinal FDR procedures. Proposed pFDR procedure has smaller values than Storey's pFDR procedure

Secondly, We considered high dimension low sample size genomic sequences without ordering of response categories. When constructing an appropriate test statistics in this model, the classical MANOVA approach may not be tenable due to too large number of parameters and too small sample size. In thsese sense, a pseudo marginal model based on the Hamming distance were presented. The Hamming distance utilizes the idea of Gini-Simpson diversity index in a variety of multidimensional setups. For small sample size, the permutation distribution was generated by considering all possible n!(equally likely) permutations of the combined sample observations among the G groups of (sizes $n_1, \ldots, n_G$). We applied proposed FDR procedure developed earlier to SARS epidemic genomic sequences. This procedure along with the associated test statistics for each gene worked well in the set of p-values generated from the exact permutation theory and controls the FDR at any level α. Proposed pFDR procedure was smaller than Storey's pFDR procedure

Finally, these previous setups may fall into classification of genes. This classification may involve complex order-restricted inference. For this problem, Roy's (1953) union-intersection principle have some advantanges (Silvapulle and Sen 2004, Tsai and Sen 2005). We presented three appropriate test statistics: linear rank statistics, a M-estimator, and kendall-tau statistics. The test statistic based on linear rank statistics using a suitable rank scores has the property of achieving a locally most powerful test, instead of the most powerful test. The M-estimator accounting for

outlier arrays provides robust test statistic, that is, distribution-insensitive clustering of genes. The Kendall-tau statistic may be utilized to construct a distribution-free test, not depending on any nuisance parameters. By exact permutation distrubution theory, conditionally distribution-free test based upon each test statistic generated corresponding p-values in small sample size setup. We assessed the performance of proposed FDR associated with each test statistic to Lobenhofer et al's breast cancer study (2002). The linear rank statistic using a normal score has smaller FDR level compared to other FDR procedures.

6.2 Discussion and Future Research

The statistical properties proposed in FDR procedure may depend on choice of appropriate parameters, α_{1m} and α_{2m}, based on two-stage estimator. Simulation studies suggest that the proposed procedure along with parameters outperform the conventional FDR procedures. We considered two-stage FDR procedure only, but still one may ask about a FDR procedure accomodating more rejection stages. In this case, choice of multiple parameters may be complicated but may definitely help to minimize FDR level and increase the associate power compared to two-stage procedure.

Average power has been mostly used in assessing the performance of FDR procedure. We assessed the performance of proposed FDR procedure in terms of average power. However, balancing FDR procedure and FNR procedure may be of greater importance in statistical practice of high-throughput screeing data analysis like microarray experiment, that is, controlling the FNR level while maintaining fixed FDR level. We already developed two-stage FNR procedure but one may need to evaluate power in terms of this procedure in simulated data example and in real data example

A pseudo marginal approach based on the Hamming distance seeks to find a

distribution-free test. In fact, it still depends on unknown parameter, maybe preempting an appropriate method in generating p-values. And the distribution of the test statistic based on linear rank statistic with appropriate constants still depends on values of constants. In these sense, Kendall-tau statistics may be a promising alternative to them, because the distribution of this statistic is free of unknown nuisance parameters. In view of using exact distribution-free tests, more sophisticated methods must be taken into account. However, Kang and Sen (2007) presented more general version of Kendall's tau statistics. to utilize a hybrid of Kendall's tau and linear rank statistics. It incorporated the sign function which have invariance property under the monotone transformations of observations. It showed not only sign of the difference between two observations but also the magnitude of the difference. They evaluated the performance of this Kendall's tau- type linear rank statistics and Kendall's tau statistics considered in Sen (2007-2008) with the real data, Lobenhofer et al. (2002). This resulted in smaller FDR procedure associated with two-stage FDR procedure presented in my thesis. More simulation studies are expected under more various situations. For example, we used two means of gene expression level according to two hypotheses, but by varying this mean expression levels, we can evaluate the performance of proposed FDR along with associated test statistics very well. We also conduct simulation studies with more complex dependence structures, such as complicated Markov chain structures or positive regression dependence structures (PRDS).

We considered two-stage FDR procedure but this result can be extended to multi-stage FDR procedure incorporating the Chen-Stein method. We expect that this procedure will provide less stringent FDR estimation as well as more power in the multiple testing context, even though it is mathematically and computationally complicated to implement.

Appendix

[proof of Theorem 2.1.1] $M_0 = \{1, 2, \ldots, m_0\}$ and $M = \{1, 2, \ldots, m\}$. Assume that $Cov(P_i, P_j) \approx 0$ when $i, j \in M_0$ or $(i \in M_0, j \in M - M_0)$ or $(j \in M_0, i \in M - M_0)$. Let q_i denote $Pr(P_i < c_\alpha)$. $B_2\alpha$ is defined as any subset of inactive genes out of m_0 genes. $B_1\alpha$ is defined to be any subset of active genes out of $m - m_0$ genes.

$V_{m_0} = \sum_{i \in M_0} I(P_i < c_\alpha)$ is approximately Poisson variable with $EV_{m_0} = \mu_{m_0} = m_0\alpha_m$.

By the Chen-stein method,

$$b_1 = \sum_{\alpha \in M_0} \sum_{j \in B_{2\alpha}} q_i q_j = m_0^2(\alpha_m)^2 = o(1),$$

$$b_2 = \sum_{i \in M_0} \sum_{i \neq j \in B_{2\alpha}} q_{ij} = m_0(m_0 - 1)\alpha_m^2 \approx m_0^2(\alpha_m)^2 = o(1),$$

if and only if $\alpha_m = o(\frac{1}{m_0})$ and $b_3 = 0$.

Let Z_1 be a Poisson random variable with $EZ = EV_{m_0} = m_0\alpha_m$.

$||\mathcal{L}(V_{m_0}) - \mathcal{L}(Z_1)||$=$2sup_A|P(V_{m_0} \in A) - P(Z_1 \in A)|$ =$\leq 2(b_1 + b_2 + b_3) = o(1)$.

$S_{m_1} = \sum_{i \in M - M_0} I(P_i < c_\alpha)$ is approximately Poisson variable with

$ES_{m_1} = (m - m_0)\alpha_m^*$.

By the Chen-stein method,

$$b_1 = \sum_{\alpha \in M - M_0} \sum_{j \in B_{1\alpha}} q_i q_j = (m - m_0)^2(\alpha_m^*)^2 = o(1)$$

$$b_2 = \sum_{i \in M - M_0} \sum_{i \neq j \in B_{1\alpha}} q_{ij} = (m - m_0)(m - m_0 - 1)\alpha_m^2 \approx (m - m_0)^2(\alpha_m)^2 = o(1)$$

if and only if $\alpha_m = o(\frac{1}{m_1})$ and b_3=0.

Let Z_2 be a Poisson random variable with $EZ = ES_{m_1} = (m - m_0)\alpha_m^*$.

$||\mathcal{L}(S_{m_1}) - \mathcal{L}(Z_2)||$=$2sup_A|P(S_{m_1} \in A) - P(Z_2 \in A)|$ =$\leq 2(b_1 + b_2 + b_3) = o(1)$. Thus,

$R_{m_0} = \sum_{i=1}^{m} I(P_i < c_\alpha) = V_{m_0} + S_{m_1}$ is approximately Poisson variable with

$ER_{m_0} = \mu^*_{m_0} = m_0\alpha_m + (m - m_0)\alpha^*_m$. [proof of Theorem 2.2.1]

Let M_0 be $\{1, 2, \ldots, m_0\}$ and M be $\{1, 2, \ldots, m\}$. Assume that $Cov(P_i, P_j) \approx 0$ when $i, j \in M_0$ or $(i \in M_0, j \in M - M_0)$ or $(j \in M_0, i \in M - M_0)$. Let q_i denote $Pr(P_i < c_\alpha)$. $B_2\alpha$ is defined as any subset of inactive genes out of m_0 genes. $B_1\alpha$ is defined to be any subset of active genes out of $m - m_0$ genes.

$V_{1(m_0)} = \sum_{i \in M_0} I(P_i < c_\alpha)$ is approximately Poisson variable with

$EV_{1(m_0)} = \mu_{1(m_0)} = m_0\alpha_{1m}$.

$$\begin{aligned} b_1 &= \sum_{\alpha \in M_0} \sum_{j \in B_{2\alpha}} q_i q_j = m_0^2(\alpha_{1m})^2 = o(1), \\ b_2 &= \sum_{i \in M_0} \sum_{i \neq j \in B_{2\alpha}} q_{ij} = m_0(m_0 - 1)\alpha_{1m}^2 \approx m_0^2(\alpha_{1m})^2 = o(1) \end{aligned}$$

if and only if $\alpha_{1m} = o(\frac{1}{m_0})$ and $b_3 = 0$.

Let Z_1 be a Poisson random variable with $EZ = EV_{1(m_0)} = m_0\alpha_1 m$.

$||\mathcal{L}(V_{1(m_0)}) - \mathcal{L}(Z_1)||$=$2sup_A|P(V_{1(m_0)} \in A) - P(Z_1 \in A)|$ =$\leq 2(b_1 + b_2 + b_3) = o(1)$.

$S_{1(m_1)} = \sum_{i \in M - M_0} I(P_i < c_\alpha)$ is approximately Poisson variable with

$ES_{1(m_1)} = (m - m_0)\alpha^*_{1m}$.

By the Chen-stein method,

$$\begin{aligned} b_1 &= \sum_{\alpha \in M - M_0} \sum_{j \in B_{1\alpha}} q_i q_j = (m - m_0)^2(\alpha^*_{1m})^2 = o(1), \\ b_2 &= \sum_{i \in M - M_0} \sum_{i \neq j \in B_{1\alpha}} q_{ij} = (m - m_0)(m - m_0 - 1)\alpha_{1m}^2 \approx (m - m_0)^2(\alpha_{1m})^2 = o(1) \end{aligned}$$

if and only if $\alpha_{1m} = o(\frac{1}{m_1})$ and b_3=0.

Let Z_2 be a Poisson random variable with $EZ = ES_{1(m_1)} = (m - m_0)\alpha^*_{1m}$.

$||\mathcal{L}(S_{1(m_1)}) - \mathcal{L}(Z_2)||$=$2sup_A|P(S_{1(m_1)} \in A) - P(Z_2 \in A)|$ =$\leq 2(b_1 + b_2 + b_3) = o(1)$.

Thus, $R_{1(m_0)} = \sum_{i=1}^{m} I(P_i < c_\alpha) = V_{1(m_0)} + S_{1(m_1)}$ is approximately Poisson variable

with $ER_{1(m_0)} = \mu^*_{m_0} = m_0\alpha_{1m} + (m - m_0)\alpha^*_{1m}$.

REFERENCES

Arratia, R., Goldstein, L. and Gordon, L. (1990). [poisson approximation and the chen-stein method]: Rejoinder. *Statistical Science* **5**, 432–434.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate:a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57**, 289–300.

Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure. *Journal of Statistical Planning and Inference* **82**, 163–170.

Benjamini, Y. and Yekutieli1, D. (2005). Quantitative trait loci analysis using the false discovery rate. *Genetics* **171**, 783–790.

Benjamini1, Y. and Yekutiel, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188.

Black, M. (2004). A note on the adaptive control of false discovery rates. *Journal of the Royal Statistical Society: Series B* **66**, 297–304.

Cheadle, C., Vawter, M. P., Freed, W. J. and Becker, K. G. (2003). Analysis of microarray data using z score transformation. *Journal of Molecular Diagnostics* **5**.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *The Annals of Probability* **3**, 534–545.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.

Fujisawa, H., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y., Muto, T. and Matsuura, M. (2004). Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics* **20**, 718–726.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 499–517.

Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from microarray experiments. *Bioinformatics* **59**, 992–1000.

Ghosh, D., Barette, T. R., Rhodes, D. and Chinnaiyan, A. M. (2003). *Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer*. Springer, Berlin / Heidelberg.

Goldstein, D., Ghosh, D. and Conlon, E. M. (2002). Statistical issues in the clustering of gene expression data. *Statistica Sinica* **12**, 219–240.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, R. C., Gaasenbeek, M., Mesirov, J. P., Coller, G. H., Loh, M. L., Downing, J. R., Caligiuri, M. A. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley Series in Probability and Statistics, New York, USA.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York, USA.

Jain, N., Cho, H., O'Connell, M. and Lee, J. K. (2005). Rank-invariant resampling based estimation of false dis- covery rate for analysis of small sample microarray data. *BMC Bioinformatics* **6**.

Jurečková, J. and Sen, P. K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations*. Wiley Series, New York, USA.

Kang, M. and Sen, P. K. (2007-08). Kendall's tau-type rank statistics in genome data. *(Submitted)* .

Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. ii. multivariate reverse rule distributions. *Journal of Multivariate Analysis* **286**, 499–516.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* **30**, 81–93.

Korn, E. L., Troendle, J. F., McShane, L. M. and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* **124**, 379–398.

Krishnaiah, P. and Sen, P. K. (1985). *Handbook OF Statistics 4: nonparametric methods*. North-Holland, Netherlands.

Lee, M.-L. T. (2004). *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers Group, Netherlands.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Annals of Statistics* **33**, 1138–1154.

Lehmann, E. L., Romano, J. P. and Shaffer, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *Annals of Statistics* **33**, 1084–1108.

Lobenhofer, E. K., Bennett, L., Cable, P. L., Li, L., Bushel, P. R. and Afshari, C. A. (2002). Regulation of dna replication fork genes by 17-estradiol. *Molecular Endocrinology* **16**, 1215–1229.

Marchini1, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413–417.

McLachlan, G. J., Bean, R. W. and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.

McLachlan, G. J., Do, K.-A. and Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Wiley Series, New York.

Moore, J. H. and Ritchie, M. D. (2004). The challenges of whole-genome approaches to common diseases. *The journal of the American Medical Association* **291**, 1642.

Odeh, R. E. (1972). On the power of jonckheere's k-sample test against ordered alternatives. *Biometrika* **59**, 467–471.

Pawitan, Y., Calza, S. and Ploner, A. (2006). Estimation of false discovery proportion under general dependence. *Bioinformatics* **22**, 3025–3031.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A. and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* **21**, 3017–3024.

Peddada, S., Harris, S. E., Zajd, J. and Harvey, E. (2005). Oriogen: Order restricted inference for ordered gene expression data. bioinformatics. *Bioinformatics* **21**, 3933–3934.

Peddada, S., Lobenhofer, E. K., Li, L., Afshari, C. A., Weinberg, C. and Umbach, D. M. (2003). Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* **19**, 834–841.

Pinheiro, A., Sen, P. K. and Pinheiro, H. P. (2002). Decomposability of high-dimensional diversity measures: Quasi u-statistics, martingales and nonstandard asymptotics. . *Submitted for Publication* .

Pinheiro, H. P., Pinheiro, A. and Sen, P. K. (2005). Comparison of genomic sequences by hamming distance. *Journal of Statistical Planning and Inference* **130**, 325–349.

Pounds, S. and Cheng., C. (2004). Improving false discovery rate estimation. *Bioinformatics* **20**, 1737–1745.

Pounds, S. and Morris, S. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236–1242.

Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375.

Sarkar, S. K. (1998). Probability inequalities for ordered mtp2 random variables: A proof of the simes conjecture. *The Annals of Statistics* **26**, 494–504.

Sarkar, S. K. (2000). A note on the monotonicity of the critical values of a step-up test. *Journal of Statistical Planning and Inference* **87**, 241–249.

Sarkar, S. K. (2004). Fdr-controlling stepwise procedures and their false negative rates. *Journal of Statistical Planning and Inference* **125**, 119–137.

Sarkar, S. K. (2006). False discovery and false nondiscovery rates in single-step multipe testing procedures. *The Annals of Statistics* **34**, 394–415.

Sarkar, S. K. and Chang, C.-K. (1997). The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* **92**, 1601–1608.

Schaid, D. J., McDonnel, S. K., Hebbring, S. J., Cunningham, J. M. and Thibodeau, S. N. (2005). Nonparametric tests of association of multiple genes with human disease. *The American Society of Human Genetics* **76**, 780–793.

Sebastiani, P., Gussoni, E., Kohane, I. S. and F.Ramoni, M. (2003). Statistical challenges in functional genomics. *Statistical science* **18**, 1–131.

Sen, P. K. (2005). Gini diversity index, hamming distance, and curse of dimensionality. *Metron - International Journal of Statistics* **LXIII**, 329–349.

Sen, P. K. (2006). Robust statistical inference for high-dimensional data models with application to genomics. *Austrian journal of statistics* **35**, 197–214.

Sen, P. K. (2007-08). Kendall's tau in high-dimension genomics parsimony (in press). *IMS Collection (Edited by B.S.Clarke and S.Ghosal)* **2**.

Sen, P. K. and Tsai, M.-T. (2005). Asymptotically optimal tests for parametric functions against ordered functional alternatives. *Journal of Multivariate Analysis* **95**, 37–49.

Sen, P. K., Tsai, M.-T. and Jou, Y.-S. (2007). High-dimension low sample size perspectives in constrained statistical inference:the sarscov rna genome in illustration. *Journal of American Statistical Association* **102**, 686–694.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology* **46**, 561–584.

Sidak, Z., Sen, P. K. and Hajek, J. (1999). *Theory of Rank Tests(Probability and Mathematical Statistics)*. Academic Press, San Diego, CA.

Silvapulle, M. J. and Sen, P. K. (2004). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. Wiley Series, New York, USA.

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

Storey, J. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics* **3**, 2013–2035.

Storey, J. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society, Series B* **69**, 1–22.

Storey, J., Dai, J. and Leek, J. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* **8**, 414–432.

Storey, J., E.Taylor, J. and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B* **66**, 187–205.

Storey, J. and Tibshirani, R. (2003). Estimating false discovery rates under dependence, with applications to dna microarrays. *PNAS* **100**, 9440–9445.

Tukey (1977). *Explortary Data Analysis*. Addison-Wesley, MA, USA.

Westfall, P. H. and Wright, P. (1993). On adjusting p-values for multiplicity.

Biometric **49**, 941–945.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82**, 172–199.