# Part I

# Large-Scale Multiple Testing

# Chapter 1

# A Compound Decision-Theoretic Approach to Large-Scale Multiple Testing

T. Tony Cai * and Wenguang Sun †

**Abstract**

In this article, we discuss topics in large-scale multiple testing and present a compound decision theoretical framework for false discovery rate (FDR) analysis. It is shown that conventional multiple testing procedures that threshold $p$-values can be much improved by a class of powerful data-driven procedures that exploit relevant information of the sample, including the proportion of non-nulls, the null and alternative distributions, the correlation structures as well as possible external information. Our discussion reveals the special features of large-scale inference problems and provides additional insights into the classic statistical decision theory. Both simulated and real data examples are presented for illustration of ideas and comparison of different procedures. Some important open problems for future research are also discussed.

**Keywords:** Compound decision theory; dependence; false discovery rate; grouped hypotheses; hidden Markov models; large-scale multiple testing.

## 1 Introduction

Large-scale multiple testing is an important area in modern statistics with a wide range of applications including genome-wide association studies, DNA microarray analysis, brain imaging studies and astronomical surveys. In these applications, one often tests thousands or even millions of hypotheses simultaneously. The analysis of these large-scale problems poses many statistical challenges not present in smaller scale studies. We discuss in this article the challenging statistical issues and new methodological developments in this field.

*Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA. Research supported in part by NSF Grant DMS-0604954 and NSF FRG Grant DMS-0854973.

†Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.

## 1.1 Setting and Notation

We shall use DNA microarray studies to illustrate the typical setting and notation in large-scale multiple testing problems. In DNA microarray experiments, a standard technique for comparison of genes across two conditions is *differential analysis* (Dudoit et al. 2002; Sebastiani et al. 2003), where gene expression data are first collected on the same $m$ genes for the two groups of subjects $\{X_1, \cdots, X_{n1}\}$ and $\{Y_1, \cdots, Y_{n_2}\}$, and then a two sample $t$-statistic $t_i$ is calculated for each gene. The $p$-value and $z$-value of each test can be obtained using appropriate transformations: $p_i = 2F(-|t_i|)$ and $z_i = \Phi^{-1}\{F(t_i)\}$, where $F$ and $\Phi$ are respectively the cdf's of the $t$-variable and standard normal variable. The data notation is summarized in Table 1. For example, Hedenfalk et al. (2001) analyzed a breast cancer

Table 1: Summary of the data notation

| | Cond. I | | | Cond. II | | | $t$ | $z$ | $p$ |
|---|---|---|---|---|---|---|---|---|---|
| | $X_1$ | $\cdots$ | $X_{n_1}$ | $Y_1$ | $\cdots$ | $Y_{n_2}$ | $T$ | $Z$ | $P$ |
| Gene 1 | $x_{11}$ | $\cdots$ | $x_{n_1 1}$ | $y_{11}$ | $\cdots$ | $y_{n_2 1}$ | $t_1$ | $z_1$ | $p_1$ |
| Gene 2 | $x_{12}$ | $\cdots$ | $x_{n_1 2}$ | $y_{12}$ | $\cdots$ | $y_{n_2 2}$ | $t_2$ | $z_2$ | $p_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| Gene $m$ | $x_{1m}$ | $\cdots$ | $x_{n_1 m}$ | $y_{1m}$ | $\cdots$ | $y_{n_2 m}$ | $t_m$ | $z_m$ | $p_m$ |

(BRC) study where gene expression data were measured on the same $m = 3226$ genes for 15 breast cancer patients, 7 with BRCA1 mutation and 8 with BRCA2 mutation. Van't Wout et al. (2003) carried out a human immunodeficiency virus (HIV) study where gene expression levels are measured on the same $m = 7680$ genes for 4 HIV positive cases and 4 HIV negative controls. In both the BRC and HIV analyses, the $m$ genes can be assumed to come from a mixture of two populations (null and non-null): in the null population, the gene expression levels are not associated with the disease condition, while in the non-null population, genes are differentially expressed between the two conditions. The goal is to test for differential gene expression and hence to identify genes associated with the disease status.

## 1.2 Type I error rates

When performing a single test, two types of errors may be committed: rejecting a hypothesis when it is a true null (type I error or false positive) or accepting it when it is a non-null (type II error or false negative). The outcomes of $m$ simultaneous tests can be summarized in Table 2.

The consequences of the two types of errors are often different. In single hypothesis testing, it is desirable to control the Type I error rate at a prespecified level $\alpha$ and minimize the type II error rate. When multiple tests are considered

Table 2: Classification of tested hypotheses

| | Claimed non-significant | Claimed significant | Total |
|---|---|---|---|
| Null | $N_{00}$ | $N_{10}$ | $m_0$ |
| Non-null | $N_{01}$ | $N_{11}$ | $m_1$ |
| Total | $S$ | $R$ | $m$ |

jointly, a procedure that tests each hypothesis at level $\alpha$, referred to as the per-comparison error rate (PCER) procedure, in general leads to the inflation of type I errors. The family wise error rate (FWER), defined as FWER $= P(N_{10} \geqslant 1)$, has been widely used to avoid misleading inferences caused by the multiplicity in simultaneous testing. Here "family" refers to the collection of all tests being conducted. Instead of controlling the PCER at level $\alpha$, an $\alpha$-level FWER controlling procedure guarantees that the probability of making one or more type I errors in the family will not exceed the prespecified level $\alpha$. A well known FWER procedure is the Bonferroni method that tests each hypothesis at level $\alpha/m$. We refer to Hochberg and Tamhane (1987) and Shaffer (1995) for a review of FWER procedures. A natural extension of the FWER is the $k$-FWER, defined as the probability of making $k$ or more false rejections in the family. Recently, some $k$-FWER procedures have been discussed in Lehmann and Romano (2005), Romano and Shaikh (2006) and Sarkar (2007).

However, the power to reject a non-null hypothesis while controlling for the FWER is greatly reduced as the number of tests increases. In situations where both the number of hypotheses and the number of true non-nulls are large, it is cost-effective to tolerate some type I errors, provided that the number is small compared to the total number of rejections. These considerations lead to a more powerful approach which calls for controlling the false discovery rate (FDR, Benjamini and Hochberg 1995). The FDR, defined as the expected proportion of false rejections among all rejections, provides a novel way to combine the errors in multiple comparisons. Using the notation in Table 2, the FDR can be defined as

$$\mathrm{FDR} = E\left(\frac{N_{10}}{R}\middle| R>0\right) P(R>0) = E\left(\frac{N_{10}}{R\vee 1}\right). \tag{1.1}$$

According to the definition, the FDR is zero when no hypotheses are rejected. Other similar measures include the positive false discovery rate (Storey 2002) and the marginal false discovery rate (Genovese and Wasserman 2002; Sun and Cai 2007), respectively defined as

$$\mathrm{pFDR} = \mathrm{E}(\frac{N_{10}}{R}|R>0) \ \text{ and } \ \mathrm{mFDR} = \frac{\mathrm{E}(N_{10})}{\mathrm{E}(R)}.$$

The pFDR and mFDR are equivalent when test statistics come from a random mixture of the null and non-null distributions (Storey 2003). Genovese and Wasserman (2004) showed that, under mild conditions, mFDR = FDR + $O(m^{-1/2})$, where $m$ is the number of hypotheses.

The FDR controlling procedures are more appropriate in large-scale multiple comparison problems and have been successfully applied in different scientific areas such as multi-stage clinical trials, microarray experiments, genome-wide association studies, brain imaging studies and astronomical surveys, among others (Weller et al. 1998; Efron et al. 2001; Miller et al. 2001; Tusher et al. 2001; Storey and Tibshirani 2003; Dudoit et al. 2002; Sabatti et al. 2003; Menshausen and Rice 2006; Schwartzman et al. 2008). This article discusses the recent theoretical and methodological developments in the FDR field. The main goal is to develop a compound decision theoretical framework for large-scale multiple testing and to introduce a new class of powerful data-driven FDR procedures that are particularly suitable for testing a large number of hypotheses.

## 1.3 Optimality

In single hypothesis testing, the power is defined as the probability of correctly rejecting a non-null hypothesis. In the Neyman-Pearson testing framework, the *most powerful test* maximizes the power subject to a constraint on the type I error rate. The power can be generalized in different ways as we move from single hypothesis testing to multiple hypothesis testing. For FDR control, the most widely used measure is the false negative (or non-discovery) rate (FNR; Genovese and Wasserman 2002), the expected proportion of non-nulls among all non-rejections. Using the notation in Table 2, we have

$$\mathrm{FNR} = E\left(\frac{N_{01}}{S}\middle| S>0\right)P(S>0) = E\left(\frac{N_{01}}{S\vee 1}\right). \tag{1.2}$$

The FNR is a natural dual quantity to the FDR and will be used in this paper. Other quantities, including the missed discovery rate (MDR), the expected true positives (ETP) and the average power (AP), have also been considered in the literature (Spjøtvoll 1972; Storey 2007; Efron 2007). An FDR procedure is said to be *valid* if it controls the FDR at a prespecified level $\alpha$ and *optimal* if it has the smallest FNR among all valid FDR procedures at level $\alpha$.

## 1.4 $P$-values and adjusted $p$-values

The $p$-value is a measure of how strongly the observed data contradict the null hypothesis. In single hypothesis testing, the $p$-value of a test can be interpreted as the probability under the null of observing a test statistic that is as extreme as or more extreme than the observed value in the direction of rejection. Therefore a statistical decision can be made by simply comparing the $p$-value with a given test level $\alpha$. A similar concept, the adjusted $p$-value, was developed in the multiple testing context (Rosenthal and Rubin 1983; Wright 1992; Westfall and Young 1993). Given a testing procedure, the adjusted $p$-value for hypothesis $i$ is the level of the entire testing procedure at which $H_i$ would *just* be rejected, given the values of all test statistics. For example, if the interest is in controlling the FWER, the Bonferroni adjusted $p$-value for hypothesis $i$ is $\tilde{p}_i = mp_i$, where $p_i$ is the unadjusted $p$-value and $m$ is the number of tests. Then $H_i$ is rejected at nominal FWER $\alpha$ if $\tilde{p}_i \leqslant \alpha$.

## 1.5 Stepwise FWER procedures

The Bonferroni method is a *single-step* procedure, which evaluates each hypothesis using a common critical value that is essentially independent of other test statistics. Other single-step procedures include the Šidák procedure and the minP procedure (cf. Westfall and Young 1993; Dudoit et al. 2003). The single-step procedures is conservative and can be improved by procedures that have a more complicated structure. Two class of stepwise procedures, namely step-down and step-up procedures, have been developed in the literature. Due to their data-dependent nature, the stepwise procedures are capable of improving the power of a single-step procedure at the same FWER level without making additional assumptions. The most well known stepwise procedures for FWER control include the Holm procedure, the Simes procedure and the Hochberg procedure.

**Holm procedure.** Let $p_{(1)} \leqslant \cdots \leqslant p_{(m)}$ be the ordered $p$-values and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses. The Holm procedure (Holm 1979) operates as follows: if $p_{(1)} \geqslant \alpha/m$, then accept all hypotheses and stop. Otherwise reject $H_{(1)}$ and test the remaining $m-1$ hypotheses at level $\alpha/(m-1)$. If $p_{(2)} \geqslant \alpha/(m-1)$, accept the remaining hypotheses and stop. Otherwise reject $H_{(2)}$ and test the remaining $m-2$ hypotheses at level $\alpha/(s-2)$. And so on. The adjusted $p$-value for the Holm procedure is

$$\tilde{p}_{(i)} = \max_{k=1,\cdots,i}[\min\{(m-k+1)p_{(k)}, 1\}],$$

for $i = 1, \cdots, m$. It can be shown that the Holm procedure controls the FWER at level $\alpha$ for all possible constellations of true and false hypotheses. This is referred to as the *strong control* of FWER. In addition, the Holm procedure starts with the most significant $p$-value and continues rejecting hypotheses as long as their $p$-values are small. This is called a *step-down* procedure.

**Simes-Hochberg procedure.** In contrast to step-down procedures, step-up procedures begin by looking at the least significant $p$-value and then move to the more significant ones. A well known step-up procedure is Simes-Hochberg procedure (Hochberg 1988), which operates as follows. Let $k = \max\{i : p_{(i)} \leqslant \alpha/(m-i+1)\}$, then reject hypotheses $H_1, \cdots, H_{(k)}$. The adjusted $p$-value for the Simes-Hochberg procedure is

$$\tilde{p}_{(i)} = \min_{k=i,\cdots,m}[\min\{(m-k+1)p_{(k)}, 1\}], \tag{1.3}$$

for $i = 1, \cdots, m$. We refer to Hochberg and Tamhane (1987) and Shaffer (1995) and for a comprehensive review of other issues in FWER control.

The rest of this chapter is organized as follows. Traditional FDR procedures are introduced in Section 2. In Section 3 we formulate the multiple testing problem in a compound decision theoretic framework and discuss an oracle and an adaptive procedure for FDR control. The simultaneous testing of grouped hypotheses and multiple testing under dependence are considered in Sections 4 and 5, respectively. We conclude the article with a discussion of some open problems and future directions.

# 2 FDR controlling procedures based on $p$-values

In contrast to the restrictive FWER criterion, an FDR procedure allows making more than one Type I errors as long as the total number of false rejections is small relative to the total number of rejections. Since the seminar work of Benjamini and Hochberg (BH 1995), the FDR procedures have been widely used in large-scale multiple comparison problems.

In this section, we first introduce the well known BH step-up procedure and then present several improvements over the BH procedure, including a $q$-value procedure (Storey 2002), an adaptive $p$-value procedure (Benjamini and Hochberg 2000), an oracle $p$-value procedure and the corresponding plug-in $p$-value procedure (Genovese and Wasserman 2004).

## 2.1 BH step-up procedure

Let $\alpha$ be an FDR level. Denote by $p_{(1)}, \cdots, p_{(m)}$ the ordered individual $p$-values and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses. Benjamini and Hochberg (1995), designated by BH hereinafter, proposed the following step-up procedure:

$$\text{Let } k = \max\{i : P_{(i)} \leqslant i\alpha/m\}, \text{ then rejects all } H_{(i)}, i \leqslant k. \tag{2.1}$$

BH showed that the step-up procedure (2.1) controls the FDR at the nominal level $\alpha$ when the tests statistics are independent. The BH procedure is more powerful than FWER controlling procedures at the same level.

The BH step-up procedure is distribution-free, which guarantees that the FDR is controlled at level $\alpha$ regardless of the $p$-value distribution. Thus it provides *strong control* of the FDR. However, with the gain of robustness, we pay a price for ignoring the distributional information in the sample. This issue was first raised in BH (2000), where it is argued that the BH step-up procedure is conservative when some of the hypotheses are in fact non-nulls. Specifically, the BH step-up procedure controls the FDR at level $(1-p)\alpha$, instead of the nominal level $\alpha$, where $p$ is the proportion of non-nulls. Next we shall discuss several approaches that exploit the information of $p$ to improve the classical BH procedure .

## 2.2 Adaptive $p$-value procedure

BH (2000) proposed a modified procedure for FDR control that is *adaptive* to the unknown non-null proportion. Let $\hat{p}$ be a conservative estimate of $p$. The adaptive $p$-value procedure operates as follows.

$$\text{Let } k = \max\{i : P_{(i)} \leqslant i\alpha/[(1-\hat{p})m]\}, \text{ then rejects all } H_{(i)}, i \leqslant k, \tag{2.2}$$

A graphical implementation of the BH adaptive procedure was presented in BH (2000) and it was shown that the adaptive procedure is more powerful than the BH step-up procedure.

## 2.3 Storey's $q$-value procedure

The conservativeness of the BH step-up procedure is also noted by Storey (2002), who suggested a direct approach to the FDR control. Instead of fixing the FDR level $\alpha$ and estimate the cutoff, Storey (2002) proposed to estimate the FDR conservatively for a given cutoff. Storey's approach also indicates that the efficiency of an FDR procedure can be improved by exploiting the information of $p$.

Let $p$ denote the proportion of non-nulls and $G$ the marginal distribution of the $p$-value. It can be shown that for the random mixture model, the pFDR for a given $p$-value cutoff $\lambda$ is

$$\mathrm{pFDR}(\lambda) = E\left(\frac{N_{10}}{R}\middle| R>0\right) = \frac{(1-p)\lambda}{G(\lambda)}. \tag{2.3}$$

Consider a set of tests conducted with independent $p$-values. For an observed $p$-value, its $q$-value is defined as

$$q(p_i) = \inf_{\gamma \geqslant p_i}\{\mathrm{pFDR}(\gamma)\} = \inf_{\gamma \geqslant p_i}\left\{\frac{(1-p)\gamma}{G(\gamma)}\right\}. \tag{2.4}$$

The $q$-value can be explained as the minimum pFDR level such that a hypothesis with the $p$-value of $p_i$ is *just* rejected. Thus the $q$-value in multiple testing can be viewed as the pFDR analogue of the $p$-value in single hypothesis testing.

In practice, the non-null proportion $p$ and $G$ are unknown. Given a cutoff $\lambda$, an underestimate for $p$ is proposed by Storey (2002): $\hat{p}(\lambda) = 1 - \frac{\#\{p_i>\lambda\}}{(1-\lambda)m}$. The estimate $\hat{p}(\lambda)$ is conservative because the largest $p$-values are most likely to come from the null. Next, an empirical estimate of $G$ is $\hat{G}(\lambda) = \frac{\#\{p_i<\lambda\}}{m}$. Therefore the $q$-value can be estimated as

$$\hat{q}\left(p_{(i)}\right) = \widehat{\mathrm{pFDR}}\left(p_{(i)}\right) = \frac{(1-\hat{p})p_{(i)}}{\hat{G}\left(p_{(i)}\right)}. \tag{2.5}$$

In practice, we can calculate the $q$-values for all individual $p$-values and determine the rejection region. The $q$-value procedure is equivalent to the BH procedure when $\hat{p}$ is estimated as zero, and is equivalent to the adaptive $p$-value procedure when the same $\hat{p}$ is used.

## 2.4 Oracle and plug-in $p$-value procedures

Let $G_1(t)$ denote the non-null distribution of $p$-value and $\pi$ the proportion of non-nulls. We assume that $G_1$ is concave, then according to Genovese and Wasserman (2002), the *oracle p-value procedure* rejects all hypotheses whose $p$-value is less than $u^*$, the solution to the equation $\{(1-p)u\}/\{(1-p)u + pG_1(u)\} = \alpha$ or equivalently,

$$G_1(u)/u = (1/p-1)\,(1/\alpha-1)\,. \tag{2.6}$$

The cutoff $u^*$ is *optimal* in the sense that it has the smallest FNR among all $p$-value based procedures at FDR level $\alpha$.

The idea that a testing problem is connected to an estimation problem is further developed in Genovese and Wasserman (2004), designated by GW hereinafter. Let $\hat{G}$ and $\hat{p}$ be estimates of the $p$-value cdf and the proportion of non-nulls, respectively. The FDR for a given cutoff $t$ can be estimated as $\hat{Q}(t) = (1-\hat{p})t/\hat{G}(t)$. A class of plug-in FDR procedures were constructed (GW 2004) based on the $p$-value cutoff

$$t(\hat{p}, \hat{G}) = \sup\{t : \hat{Q}(t) \leqslant \alpha\}. \tag{2.7}$$

The BH step-up procedure, BH adaptive procedure and Storey's FDR approach can be identified as special cases when different estimates for $p$ and $G$ are chosen.

Although numerical results are promising, no theoretical supports for the uses of BH adaptive $p$-value procedure or Storey's $q$-value procedure were provided in the original works of BH (2000) and Storey (2002). GW (2004) developed a stochastic process framework for multiple testing and showed that, when consistent estimates of $G$ and $p$ are chosen, the class of plug-in procedures (2.7) controls the FDR at level $\alpha + o(1)$. Therefore the validity of BH adaptive procedure, GW plug-in procedure and Storey's FDR procedure, which can be viewed as special cases of (2.7), are established in an asymptotic sense.

### 2.5 Other issues

Resampling approaches (e.g., bootstrap, permutation) can be used to estimate the $p$-values and adjusted $p$-values without making any parametric assumptions on the joint distribution of the test statistics, and the correlation structure and distributional characteristics of the gene expression can be preserved. Algorithms for computing adjusted $p$-values are introduced, for example, in Westfall and Young (1993) and Dudoit et al. (2003). Permutation based methods have been applied to the significance analysis of microarrays (SAM). A widely used SAM procedure was proposed in Tusher et al. (2001).

The theoretical properties and operating characteristics of $p$-value based FDR procedures have been studied extensively in the literature. Here we mention a few references, Finner and Roters (2002), Sarkar (2002, 2006), Lehmann and Romano (2005), Finner et al. (2009), for interested readers.

## 3 Oracle and adaptive compound decision rules for FDR control

Developing the optimal FDR procedure is important from both theoretical and practical perspectives. The construction of an optimal multiple testing procedure involves two important steps: deriving an optimal test statistic $T$ and setting a cutoff for $T$ for a given FDR level. However, the focus of the FDR literature has been exclusively on the second step on how to threshold the $p$-values so that the FDR can be controlled, while the more fundamental problem on how to choose $T$ is ignored.

In single hypothesis testing, $p$-value is a fundamental statistic for deciding whether a hypothesis should be rejected or accepted. This $p$-value testing

framework is almost universally used in the FDR literature. For example, the FDR procedures reviewed in the previous section essentially involve first ranking the $p$-values from individual tests and then choosing a cutoff along the rankings. However, testing procedures built on ranked $p$-values fail to exploit all important distributional information in the sample (e.g., the symmetry of distribution and correlation in the sample), and hence are inefficient. Sun and Cai (2007) showed that $p$-value is not a fundamental building block in large-scale multiple testing, and proposed an adaptive data-driven procedure based on $z$-values that uniformly improves all $p$-value based FDR procedures.

In this section, we study the optimality issue in a compound decision theoretic framework. Our strategy for deriving an *optimal* testing procedure essentially involves three steps.

1. The first step is to derive an *oracle* test statistic $\boldsymbol{T}_{OR}$ that gives the optimal significance rankings of all tests. "Oracle" is used here to reflect the fact that we have assumed in this step that all distributional information of the sample is known. A useful technique in the derivation is to make connections between the multiple testing and *weighted classification* problems, then solve the former problem via finding the optimal classification rule.
2. The second step is to choose an optimal cutoff $c_{OR}$ along the rankings produced by $\boldsymbol{T}_{OR}$ so that the FDR is controlled at the nominal level $\alpha$. The essential idea is to first evaluate the distributions of $\boldsymbol{T}_{OR}$, then calculate the FDR level for a given cutoff $c$, and finally choose the largest cutoff $c_{OR}$ that controls the FDR. The resulting testing procedure $\boldsymbol{\delta}(\boldsymbol{T}_{OR}, c_{OR}\mathbf{1}) = I(\boldsymbol{T}_{OR} < c_{OR}\mathbf{1})$ is referred to as the *oracle procedure*, which has the smallest FNR or the largest power to detect non-null cases among valid FDR procedures.
3. The third step is to develop a *data-driven* procedure that mimics the oracle procedure by plugging-in estimates of the unknown parameters. To achieve this goal, we need to (i) construct good estimates of the unknown parameters from the sample, and (ii) establish the *asymptotic validity and optimality* of the data-driven procedure by showing that it achieves the performance of the oracle procedure asymptotically.

## 3.1 Two-group random mixture model

A two-component random mixture model provides a convenient and efficient framework for large-scale multiple testing and has been widely used in the FDR literature (Efron et al. 2001; Storey 2002; Newton et al. 2004; Sun and Cai 2007). Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$ be independent Bernoulli($p$) variables, where $\theta_i = 1$ indicates that hypothesis $i$ is a non-null and $\theta_i = 0$ otherwise. It is assumed that $\boldsymbol{x} = (x_1, \cdots, x_m)$ are observations generated conditional on $\boldsymbol{\theta}$:

$$X_i|\theta_i \sim (1-\theta_i)F_0 + \theta_i F_1, \tag{3.1}$$

where $F_0$ and $F_1$ are the conditional cumulative distribution functions (cdf) of $X_i$ under the null and alternative, respectively. Let $F(x) = (1-p)F_0(x) + pF_1(x)$ be the mixture cdf and $f(x) = (1-p)f_0(x) + pf_1(x)$ the mixture density, with $f_0$ and $f_1$ the corresponding conditional probability distribution functions (pdf).

## 3.2 Compound decision problem

Consider the random mixture model (3.1). In a multiple testing problem, we are interested in separating the non-null cases ($\theta_i = 1$) from the null cases ($\theta_i = 0$). A solution to this problem can be represented by a general decision rule $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_m) \in \{0,1\}^m$, where $\delta_i = 1$ indicates that we claim case $i$ is a non-null and $\delta_i = 0$ otherwise. In an FDR analysis, the $m$ decisions are combined and evaluated integrally; this is referred to as a *compound decision problem* (Robbins 1951). The decision rule $\boldsymbol{\delta}$ is *simple* if $\delta_i$ is only a function of $x_i$, i.e., $\delta_i(\mathbf{x}) = \delta_i(x_i)$. The simple rules correspond to solving the $m$ component problems separately. In contrast, $\boldsymbol{\delta}$ is *compound* if $\delta_i$ depends on other $x_j$'s, $j \neq i$. A decision rule $\boldsymbol{\delta}$ is *symmetric* if $\boldsymbol{\delta}(\pi(\mathbf{x})) = \pi(\boldsymbol{\delta}(\mathbf{x}))$ for all permutation operators $\pi$.

Robbins (1951) considered the following compound decision problem. Let $X_i$, $i = 1, \cdots, n$, be independent normal random variables with unit variance and mean $\theta_i$, where each $\theta_i$ is 1 or $-1$. It is desired to classify each $\theta_i$ according to its sign, and the risk of a classification procedure is taken to be the expected number of errors, i.e., the risk for a classification rule $\boldsymbol{\delta} \in \{-1,1\}^m$ is

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = E\left\{\sum_{i=1}^{m} I(\theta_i = 1)I(\delta_i = -1) + I(\theta_i = -1)I(\delta_i = 1)\right\}, \tag{3.2}$$

Robbins showed that the *simple rule*

$$\boldsymbol{\delta}^S = [\mathrm{sgn}(x_i) : i = 1, \cdots, n] \tag{3.3}$$

is the unique minimax decision rule with constant risk, say $r_n$. At the same time, he argued that minimax did not equal *best* by exhibiting another decision rule

$$\boldsymbol{\delta}^C = \left[\mathrm{sgn}\left(x_i - \frac{1}{2}\log\frac{1-\bar{x}}{1+\bar{x}}\right) : i = 1, \cdots, n\right], \tag{3.4}$$

which has a much lower risk $r_n^*$ when $p_0$ approaches 0 or 1, and only exceeds $r_n$ slightly near 0.5. As $n \to \infty$, $r_n^* - r_n \to 0$ at point 0.5, so it is *subminimax*. We shall give a graphical illustration of this interesting result shortly. An important feature of the decision rule $R^*$ is that it classifies $\theta_i$ depending on the whole vector $x = (x_1, \cdots, x_n)$, not on $x_i$ alone. Thus $R^*$ is a compound rule.

The subminimax rule can be motivated as follows. Assume that an oracle knows the proportion of $\theta = 1$, say $p$, then it can be shown that the optimal classification rule (Bayes oracle rule) is

$$\boldsymbol{\delta}^* = \left[\mathrm{sgn}\left\{x_i - \log\left(\frac{1-p}{p}\right)\right\} : i = 1, \cdots, n\right]. \tag{3.5}$$

The subminimax rule can be obtained by replacing the unknown $p$ using its unbiased estimate

$$\hat{p} = \frac{1}{n}(\text{no. of } i \text{ for which } x_i > 0) = \frac{1+\bar{x}}{2} \tag{3.6}$$

The subminimax rule is a compound decision rule, which improves the efficiency of simple rules by utilizing the information of $\hat{p}$ that is estimated from the entire

sample $\boldsymbol{x}$. Some calculations show that the classification risks for the minimax rule, the Bayes oracle rule, and the subminimax rule are

$$\begin{aligned} R_{\text{Minimax}} &= \Phi(-1), \\ R_{\text{Oracle}} &= p\Phi\left(-1+\frac{1}{2}\log\frac{1-p}{p}\right)+(1-p)\Phi\left(-1-\frac{1}{2}\log\frac{1-p}{p}\right), \text{ and} \\ R_{\text{Subminimax}} &= r(p)+\frac{1}{n}p\left[\left\{1+\frac{1}{4p(1-p)}\right\}^{2}-1\right]\Phi\left(-1+\frac{1}{2}\log\frac{1-p}{p}\right), \end{aligned}$$

respectively. We plot the classification risks as functions of the proportion $p$. The results are shown in Figure 3.1. It is easy to see that the risk of the subminimax rule approaches the risk of the Bayes oracle rule as $n \to \infty$. For large $n$, the subminimax rule is better than the minimax rule for most values of $p$ except the values near 0.5. The efficiency gain is the largest as $p$ approaches 0 and 1.

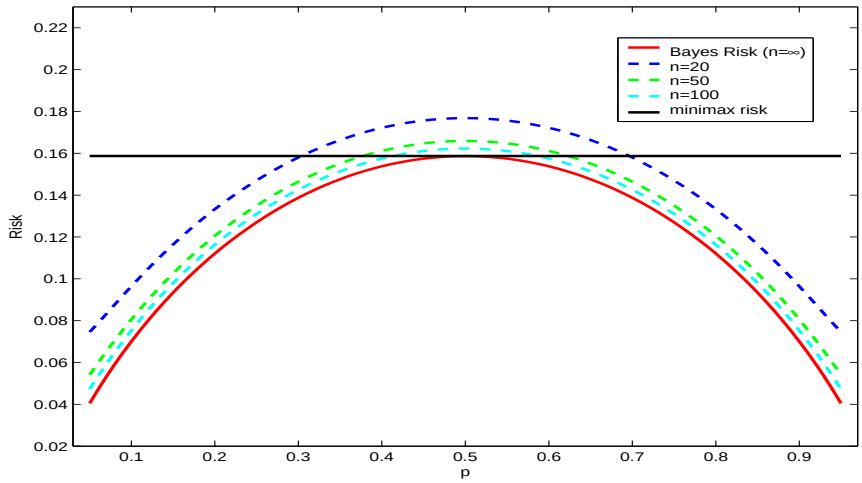


Figure 3.1: The classification risks of the minimax, subminimax and Bayesian oracle rules.

An important implication of Robbin's results is that the precision of individual decisions can be much increased by pooling information from different samples. The distributional information among all hypotheses, such as the proportion of non-nulls, is important for construction of efficient testing procedures. Therefore we anticipate that, in large-scale multiple testing, conventional testing procedures may be much improved by an estimation-testing combined procedure.

## 3.3 Derivation of the optimal test statistic

In the loss function (3.2) that was considered by Robbins (1951), it is assumed that a false positive and a false negative have the same cost. However, in practice, a false positive is often considered to be more serious than a false negative. Therefore

it is desirable to treat the two types of errors differently. Let $\lambda$ be the known relative cost of a false positive to a false negative, then a weighted classification rule $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_m) \in \{0,1\}^m$, where $\delta_i = 1$ indicates that we classify $\theta_i$ as a non-null and $\delta_i = 0$ otherwise, can be used to separate the non-nulls from the nulls. The false positives and false negatives of the $m$ simultaneous decisions can be combined using the following loss function

$$L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{m} \sum_i \left[\lambda(1-\theta_i)\delta_i + \theta_i(1-\delta_i)\right]. \tag{3.7}$$

The goal of a weighted classification problem is to find a decision rule $\boldsymbol{\delta}^\lambda$ that minimizes the weighted classification risk $R_\lambda = E[L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta})]$.

However, it may be difficult to prespecify the relative cost of a false positive to a false negative, especially in many large-scale studies where the non-nulls are very sparse. We can use the FDR and FNR to combine the respective false positives and false negatives instead, and apply a multiple testing procedure to select non-null cases from the nulls. In a multiple testing problem, the two types of errors are also treated differently, where a higher penalty on false positives is achieved by prespecifying a smaller FDR level $\alpha$. The goal of multiple testing is to find a decision rule $\boldsymbol{\delta}^\alpha \in \{0,1\}^m$ that has the smallest FNR among all FDR procedures at level $\alpha$. We define a multiple testing procedure in terms of a function $T$ and a constant $c$ such that $\boldsymbol{\delta}(T, c) = [I(T(x_i) < c) : i = 1, \cdots, m]$. In conventional $p$-value based FDR procedures, $T(x_i)$ is taken as $F_0(-|x_i|)$ or $2F_0(-|x_i|)$, where $F_0$ is the null cdf of $X_i$. We consider a wider class of testing procedures in which $T$ is allowed to also depend on other quantities. We are interested in finding the optimal choices of $T$ and its corresponding cutoff $c$.

**A monotone ratio condition.** The goal of both the multiple testing and weighted classification problems is to separate the non-null cases from the null cases, and the solution to both problems can be represented by a decision rule of the form

$$\boldsymbol{\delta}(\boldsymbol{T}, c\mathbf{1}) = \{I(T_i < c) : i = 1, \cdots, m\}, \tag{3.8}$$

where $\boldsymbol{T}$ is a classifier or a test statistic and $c$ is a cutoff. In the multiple testing literature, the following assumption has been used (e.g., Genovese and Wasserman 2004; Storey 2004):

$$\text{the FDR level yielded by } \boldsymbol{\delta}(\boldsymbol{T}, c\mathbf{1}) \text{ is increasing in } c; \tag{3.9}$$

$$\text{the FNR level yielded by } \boldsymbol{\delta}(\boldsymbol{T}, c\mathbf{1}) \text{ is decreasing in } c. \tag{3.10}$$

Assumptions (3.9) and (3.10) are desirable for developing multiple testing procedures since it implies that in order to minimize the FNR, we should choose the largest cutoff $c$ that satisfies FDR$\leqslant \alpha$. Let $G_0$ and $G_1$ be the null and non-null cdfs of $T$, respectively. Denote by $G = (1-\pi)G_0 + \pi G_1$ the marginal cdf of $T$. A general condition that guarantees (3.9) and (3.10) is the monotone ratio condition (MRC), which assumes that

$$g_1(t)/g_0(t) \text{ is monotonically decreasing in } t. \tag{3.11}$$

See Sun and Cai (2007) for a proof. Denote by $\mathcal{T}$ the collection of all test statistics that satisfy (3.11). The MRC class $\mathcal{T}$ is fairly general. Let $T_i = p_i$, the $p$-value of an individual test based on the observation $x_i$. Assume that $p(x_i) \sim G = (1-p)G_0 + pG_1$, where $G_0$ and $G_1$ are the $p$-value distributions under the null and the alternative, respectively. Next assume that $G_1(t)$ is twice differentiable. Note $g_0(t) = G_0'(t) = 1$, the assumption $p(.) \in \mathcal{T}$ implies that $G_1''(t) = g_1'(t) < 0$, i.e., $G_1(t)$ is concave. Therefore the MRC condition (3.11) can be viewed as a generalized version of the concavity assumption on $G_1$ for $p$-value (Storey 2002, Genovese and Wasserman 2004). In addition, other test statistics, including the local false discovery rate (Lfdr, Efron et al. 2001) and the local index of significance (Sun and Cai 2009) also belong to $\mathcal{T}$. See Sun and Cai (2007) for more discussions of the MRC.

**Connection between multiple testing and weighted classification.** An important step for our derivation is to show that the multiple testing and weighted classification problems are "equivalent" when the MRC holds (Sun and Cai 2007). Specifically, let $\mathcal{D}_\alpha$ be the collection of all $\alpha$-level FDR procedures of the form $\boldsymbol{\delta} = I(\boldsymbol{T} < c\mathbf{1})$. Suppose that the classification risk with the loss function defined in (3.7) is minimized by $\delta^\lambda\{\boldsymbol{T}, c(\lambda)\}$, so that $\boldsymbol{T}$ is optimal in the weighted classification problem. If $\boldsymbol{T} \in \mathcal{T}$, then $\boldsymbol{T}$ is also optimal in the multiple testing problem, in the sense that for each FDR level $\alpha$, there exists a unique $\lambda(\alpha)$, and hence $c\{\lambda(\alpha)\} = c(\alpha)$, such that $\delta^{\lambda(\alpha)}\{\boldsymbol{T}, c(\alpha)\}$ controls the FDR at level $\alpha$ with the smallest FNR level among all testing rules in $\mathcal{D}_\alpha$. This implies that the more complicated multiple testing problem can be solved by studying an equivalent weighted classification problem.

The next step is to derive the optimal classification rule. We first study an ideal setup where there is an oracle that knows $p, f_0$ and $f_1$. Then the oracle rule in this weighted classification problem gives the optimal choice of $\delta$.

**Theorem 3.1.** (***The oracle rule for weighted classification.***) *Consider the random mixture model (3.1). Suppose $p, f_0, f_1$ are known. Then the classification risk $E[L_\lambda(\theta, \delta)]$ with loss function (3.7) is minimized by $\delta^\lambda(\Lambda, 1/\lambda) = \{\delta_1, \cdots, \delta_m\}$, where*

$$\delta_i = I\left\{\Lambda(x_i) = \frac{(1-p)f_0(x_i)}{pf_1(x_i)} < \frac{1}{\lambda}\right\}, i = 1, \cdots, m. \tag{3.12}$$

## 3.4 Oracle testing procedure

We have shown that $\delta^\lambda(\Lambda, 1/\lambda) = [I\{\Lambda(x_1) < 1/\lambda\}, \cdots, I\{\Lambda(x_m) < 1/\lambda\}]$ is the oracle rule in the weighted classification problem. The equivalence between multiple testing and weighted classification implies the optimal testing rule is also of the form $\delta^{\lambda(\alpha)}[\Lambda, 1/\lambda(\alpha)]$ if $\Lambda \in \mathcal{T}$, although the cutoff $1/\lambda(\alpha)$ is not obvious. Note that $\Lambda(x) = \text{Lfdr}(x)/[1 - \text{Lfdr}(x)]$ is monotonically increasing in $\text{Lfdr}(x)$, where $\text{Lfdr}(\cdot) = (1-p)f_0(\cdot)/f(\cdot)$ is the local false discovery rate (Lfdr) introduced by Efron et al. (2001) and Efron (2004), so the optimal rule for mFDR control is of the form $\delta(\text{Lfdr}(\cdot), c) = \{I[\text{Lfdr}(x_i) < c] : i = 1, \cdots, m\}$. The Lfdr has been widely used in the FDR literature to provide a Bayesian version of the frequentist FDR measure and interpret results for individual cases (Efron 2004). We rediscover it

here as the optimal (oracle) statistic in the multiple testing problem in the sense that the thresholding rule based on $\text{Lfdr}(X)$ controls the mFDR at the nominal level with the smallest mFNR.

The MRC implies that in order to minimize the mFNR level, we should choose the largest threshold for the Lfdr statistic. Therefore the *oracle testing procedure* is

$$\boldsymbol{\delta}(\text{Lfdr}, c_{OR}) = \{I[\text{Lfdr}(x_i) < c_{OR}] : i = 1, \cdots, m\}, \tag{3.13}$$

where the oracle threshold $c_{OR} = \sup\{c \in (0,1) : \text{mFDR}(c) \leqslant \alpha\}$. The oracle procedure (3.13) provides an ideal target for evaluating different multiple testing procedures. In particular, it is more efficient than the $p$-value oracle procedure proposed in Genovese and Wasserman (2002). Hence the $z$-value oracle procedure is more efficient than all $p$-value based FDR procedures.

The Lfdr statistic is defined in terms of the *z-values*, which can be converted from other test statistics including the $t$-statistic and $\chi^2$-statistic using appropriate transformations. Note that the non-null proportion $p$ is a global parameter. The expression $\text{Lfdr(z)} = (1-p)f_0(z)/f(z)$ therefore implies that we actually rank the relative importance of the observations according to their likelihood ratios, and that the rankings are generally different from the rankings of $p$-values. An interesting consequence of using the Lfdr statistic in multiple testing is that an observation located farther from the null may have a lower significance level. It is therefore possible that the test accepts a more extreme observation while rejecting a less extreme observation, which implies that the rejection region is asymmetric. This is not possible for a testing procedure based on the individual $p$-values, whose rejection region is always symmetric about the null.

## 3.5 A data-driven procedure

The oracle procedure is not applicable in practice because the distributional information is usually unknown. This section first discusses the estimation and the non-null proportion in large-scale multiple comparisons. Then we introduce a data-driven procedure that mimics the oracle procedure.

Efron (2004) raised an important issue that in many large-scale studies the usual assumption that the null distribution is known is incorrect, and seemingly negligible differences in the null may result in large differences in subsequent studies. It was demonstrated that the null distribution should be estimated from data instead of being assumed known. Besides the null distribution, the proportion of non-null effects $p$ is also an important quantity. The implementation of many FDR procedures requires the knowledge of $p$ (BH 2000; Storey 2002; GW 2004). Developing good estimators for the proportion of non-nulls is a challenging task. Recent work includes that of Genovese and Wasserman (2004), Langaas, Lindqvist and Ferkingstad (2005), Meinshausen and Rice (2006), Cai, Jin and Low (2007), and Jin and Cai (2007).

Jin and Cai (2007) developed an approach based on the empirical characteristic function and Fourier analysis for simultaneous estimation of both the null distribution $f_0$ and proportion of non-null effects $p$. The estimators are shown

to be uniformly consistent over a wide class of parameters. Numerical results also showed that the estimators perform favorably in comparison to other existing methods. This method will be used in our data-driven procedure.

Next we outline the steps for an intuitive derivation of the adaptive $z$-value based procedure. The derivation essentially involves mimicking the operation of the $z$-value oracle procedure and evaluating the distribution of $T_{OR}(z)$ empirically. Let $z_1, \cdots, z_m$ be a random sample from the mixture model (3.1) with the CDF $F = (1-p)F_0 + pF_1$ and PDF $f = (1-p)f_0 + pf_1$. Let $\hat{p}$, $\hat{f}_0$ and $\hat{f}$ be consistent estimates of $p, f_0$ and $f$. Such estimates are provided, for example, in Jin and Cai (2006). Define $\hat{T}_{OR}(z_i) = [(1-\hat{p})\hat{f}_0(z_i)/\hat{f}(z_i)] \wedge 1$. The mFDR of decision rule $\boldsymbol{\delta}(T_{OR}, \lambda) = \{I[T_{OR}(z_i) < \lambda] : i = 1, \cdots, m\}$ is given by $Q_{OR}(\lambda) = (1-p)G^0_{OR}(\lambda)/G_{OR}(\lambda)$, where $G_{OR}(t)$ and $G^0_{OR}(t)$ are the marginal cdf and null cdf of $T_{OR}$, respectively. Let $S_\lambda = \{z : T_{OR}(z) < \lambda\}$ be the rejection region. Then $G_{OR}(\lambda) = \int_{S_\lambda} f(z)dz = \int 1\{T_{OR}(z) < \lambda\} f(z)dz$. We estimate $G_{OR}(\lambda)$ by $\hat{G}_{OR}(\lambda) = \frac{1}{m}\sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) < \lambda\}$. The numerator of $Q_{OR}(\lambda)$ can be written as $(1-p)G^0_{OR}(\lambda) = (1-p)\int_{S_\lambda} f_0(z)dz = \int 1\{T_{OR}(z) < \lambda\} T_{OR}(z) f(z)dz$ and we estimate this quantity by $\frac{1}{m}\sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) < \lambda\}\hat{T}_{OR}(z_i)$. Then $Q_{OR}(\lambda)$ can be estimated as

$$\hat{Q}_{OR}(\lambda) = [\sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) < \lambda\}\hat{T}_{OR}(z_i)]/[\sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) < \lambda\}].$$

Set the estimated threshold as $\hat{\lambda}_{OR} = \sup\{t \in (0,1) : \hat{Q}_{OR}(t) \leqslant \alpha\}$ and let $R$ be the set of the ranked $\hat{T}_{OR}(z_i)$: $R = \{\hat{\text{Lfdr}}_{(1)}, \cdots, \hat{\text{Lfdr}}_{(m)}\}$. We only consider the discrete cutoffs in set $R$, where the estimated mFDR is reduced to $\hat{Q}_{OR}(\hat{\text{Lfdr}}_{(k)}) = \frac{1}{k}\sum_{i=1}^k \hat{\text{Lfdr}}_{(i)}$. We propose the following adaptive step-up procedure:

$$\text{Let } k = \max\{i : \frac{1}{i}\sum_{j=1}^i \hat{\text{Lfdr}}_{(j)} \leqslant \alpha\}, \text{ then reject all } H^{(i)},\ i = 1, \cdots, k. \qquad (3.14)$$

In the FDR literature, $z$-value based methods such as the Lfdr procedure (Efron, 2004) are only used to calculate individual significance levels whereas the $p$-value based procedures are used for global FDR control to identify non-null cases. It is also notable that the goals of global error control and individual case interpretation are naturally unified in the adaptive procedure. The procedure (3.14) is more *adaptive* than the BH adaptive procedure in the sense that it adapts both to the global feature ($p$) and local feature ($f_0/f$). In contrast, the BH method only adapts to the global feature $p$. Suppose we use the theoretical null $N(0,1)$ in the expression of $\hat{\text{Lfdr}} = (1-\hat{p})f_0/\hat{f}$. The $p$-value approaches treat points $-z$ and $z$ equally, whereas the $z$-value approaches evaluate the relative importance of $-z$ and $z$ according to their estimated densities. For example, if there is evidence in the data that there are more non-nulls around $-z$ (i.e., $\hat{f}(-z)$ is larger), then observation $-z$ will be correspondingly ranked higher than observation $z$.

The following theorem shows that the adaptive procedure (3.14) asymptotically attains the performance of the oracle procedure based on the $z$-values in the

sense that both the mFDR and mFNR levels achieved by the oracle procedure are also asymptotically achieved by the adaptive $z$-value procedure.

**Theorem 3.2** (Asymptotic validity and optimality of the adaptive procedure)**.** *Consider the random mixture model (3.1). Suppose $f$ is continuous and positive on the real line. Assume $T_{OR}(z_i) = (1-p)f_0(z_i)/f(z_i)$ is distributed with the marginal PDF $g = (1-p)g_0 + pg_1$ and $T_{OR} \in \mathcal{T}$ satisfies the SMLR assumption. Let $\hat{p}$, $\hat{f}_0$, $\hat{f}$ be estimates of $p$, $f_0$ and $f$ such that $\hat{p} \xrightarrow{qm} p$, $E\|\hat{f} - f\|^2 \to 0$ and $E\|\hat{f}_0 - f_0\|^2 \to 0$. Define test statistic $\hat{T}_{OR}(z_i) = (1-\hat{p})\hat{f}_0(z_i)/\hat{f}(z_i)$. Let $L\hat{f}dr_{(1)}, \cdots, L\hat{f}dr_{(m)}$ be the ranked values of $\hat{T}_{OR}(z_i)$, then the FDR level of the adaptive procedure (3.14) is $\alpha + o(1)$, and the mFNR level of the adaptive procedure (3.14) is $\tilde{Q}_{OR}(\lambda_{OR}) + o(1)$, where $\tilde{Q}_{OR}(\lambda_{OR})$ is the mFNR level achieved by the oracle oracle procedure.*

## 3.6 Numerical Results

We now turn to the numerical performance of our adaptive $z$-value procedure. When the Lfdr statistic is needed to be estimated, $f_0$ is chosen to to be the theoretical null density $N(0,1)$, $p$ is estimated consistently using the approach of Jin and Cai (2007), and $f$ is estimated using the kernel density estimator. The new data-driven procedure is compared with the BH step-up procedure and the adaptive $p$-value procedure (BH 2000; GW 2004). These three procedures are designated respectively by SC, BH and AP hereinafter.

*Example* 3.3. We generate $m = 3000$ observations from the normal mixture model $0.8N(0,1) + p_1 N(\theta_{1i}, 1) + (0.2 - p_1)N(\theta_{2i}, 1)$, where $\theta_{1i}$ and $\theta_{2i}$ are randomly generated from uniform distributions $U(\mu_1 - \epsilon_1, \mu_1 + \epsilon_1)$ and $U(\mu_2 - \epsilon_2, \mu_2 + \epsilon_2)$. We apply the BH, AP and SC with FDR $= 0.10$, $\mu_1 = -3$, $\epsilon_1 = 0.4$ and $\epsilon_2 = 0.2$. The comparison results are displayed in Figure 3.2. In panel (a), we set $\mu_2 = 3$ and plot the FNR's by BH, GW and SC as functions of $p_1$. In panel (b), we set $p_1 = 0.18$ and plot the FNR's by BH, AP and SC as functions of $\mu_2$. We can see that the BH is dominated by AP, which is again dominated by SC. The efficiency gain of SC becomes more prominent when the alternative distribution is more asymmetric. □

Next we illustrate our method in the analysis of the microarray data from an HIV study. The goal of the HIV study (van't Wout et al., 2003) is to discover differentially expressed genes between HIV positive patients and HIV negative controls. Gene expression levels were measured for four HIV positive patients and four HIV negative controls on the same $m = 7680$ genes. A total of $m$ two sample $t$-tests were performed and the corresponding two-sided $p$-values were obtained. The $z$-values were then converted from the $t$-statistics using the transformation $z_i = \Phi^{-1}[G_0(t_i)]$, where $\Phi$ and $G_0$ are the CDFs of a standard normal and a $t$ variable with six degrees of freedom, respectively. The histograms of the $z$-values and $p$-values were presented in Figure 3.3. An important feature in this data set is that the $z$-value distribution is asymmetric about the null. The distribution is skewed to the right.

When the null hypothesis is true, the $p$-values and $z$-values should follow their *theoretical null distributions*, which are uniform and standard normal, respectively.

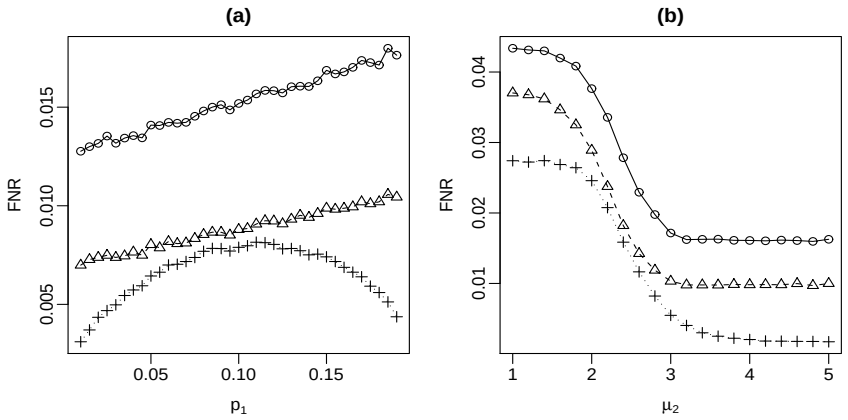


Figure 3.2: The comparison of BH, AP and SC when the alternative is concentrated ('$\circ$': BH; '$\triangle$': AP; '$+$': SC). The FDR level is set at 0.1. (a). The FNR versus $p_1$; (b) The FNR versus $\mu_2$. Both the BH and AP are dominated by SC.

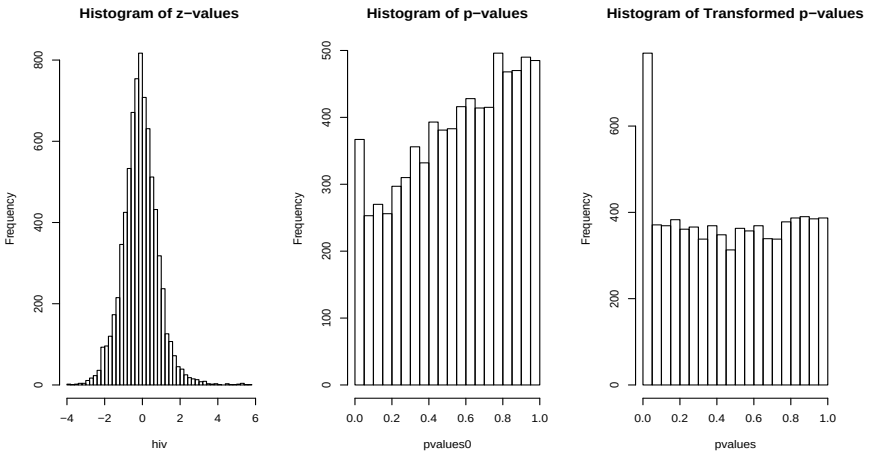


Figure 3.3: The histograms of the HIV data: $p$-values and $z$-values. The transformed $p$-values are approximately distributed as uniform (0, 1) for the null cases.

However, the theoretical nulls are usually quite different from the empirical nulls for the data arising from microarray experiments. We take the approach in Jin and Cai (2006) to estimate the null distribution as $N(\hat{\mu}_0, \hat{\sigma}_0^2)$. The estimates $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ are consistent. We then proceed to estimate the proportion of the non-nulls $\hat{p}$ based on $\hat{\mu}_0$ and $\hat{\sigma}_0^2$. The marginal density $f$ is estimated by a kernel density estimate $\hat{f}$ with the bandwidth chosen by cross validation. The Lfdr statistics are then calculated as $\widehat{\text{Lfdr}}(z_i) = (1-\hat{p})\hat{f}_0(z_i)/\hat{f}(z_i)$. The transformed $p$-values are obtained as $\hat{F}_0(z_i)$, where $\hat{F}_0$ is the estimated null CDF $\Phi(\frac{x-\hat{\mu}_0}{\hat{\sigma}_0})$. As we can see from the right panel of Figure 3.3, after transformation, the distribution of the transformed $p$-values is approximately uniform when the null is true.

We compare the BH, AP and SC using both the theoretical nulls and estimated nulls. We calculate the number of rejections for each mFDR level and the results are shown in Figure 3.4. For the left panel, $f_0$ is chosen to be the theoretical null $N(0,1)$ and the estimate for the proportion of nulls is 1. The BH and AP procedures therefore yield the same number of rejections. For the right panel, the estimated null distribution is $N(-0.08, 0.77^2)$ with estimated proportion of nulls $\hat{p}_0 = 0.94$. Transformed $p$-values as well as the Lfdr statistics are calculated according to the estimated null. The following observations can be made from the results displayed. (i) The number of rejections is increasing as a function of the mFDR. (ii) For both the $p$-value and $z$-value based approaches, more hypotheses are rejected by using the estimated null. (iii) Both comparisons show that SC is more powerful than the BH and AP that are based on $p$-values.

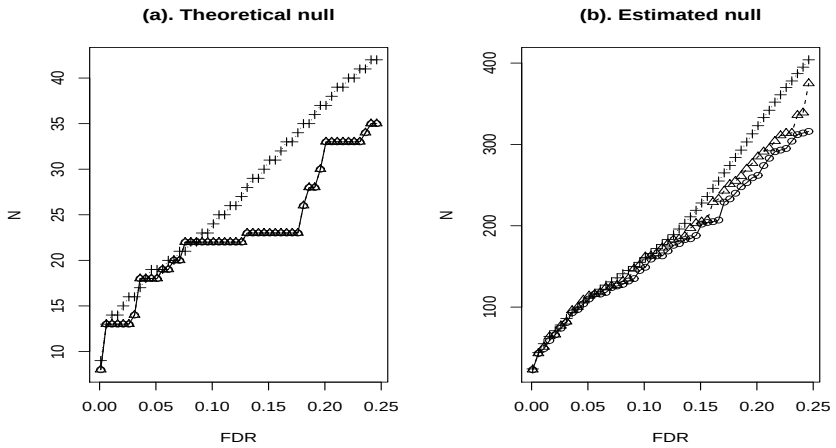


Figure 3.4: Analysis of the HIV data: Number of rejections versus FDR levels: '$\circ$': BH; '$\triangle$': AP; '$+$': SC.

# 4 Simultaneous Testing of Grouped Hypotheses

So far we have made two important assumptions: all hypotheses are independent and all observations come from a homogeneous distribution. Next we shall deal with more complicated situations where these assumptions do not hold. In this section, we consider the multiple testing problem when hypotheses come from heterogeneous groups. In Section 5, we discuss the multiple testing problem under dependence. For both problems, we will focus on the motivation and the main ideas of our solution. More technical details are given in Cai and Sun (2009) and Sun and Cai (2009).

## 4.1 Motivating examples

Conventional multiple testing procedures, such as the false discovery rate analyses (Benjamini and Hochberg 1995; Efron et al. 2001; Storey 2002; Genovese and Wasserman 2002; van der laan et al. 2004), implicitly assume that data are collected from repeated or identical experimental conditions, and hence the hypotheses are exchangeable. However, in many applications, data are known to be collected from heterogeneous sources and hypotheses intrinsically form into different groups.

Consider the following two examples. The adequate yearly progress (AYP) study compares the academic performances of social-economically advantaged (SEA) versus social-economically disadvantaged (SED) students of California high schools (Rogosa 2003). Standard tests in mathematics were administered to 7867 schools and a $z$-value for comparing SEA and SED students was obtained for each school. The estimated null densities of the $z$-values for small, medium and large schools are plotted on the left panel of Figure 4.1. It is interesting to see that the null density of the large group is much wider than those of the other two densities. The differences in the null distributions have significant effects on the outcomes of a multiple testing procedure. Another example is the brain imaging study analyzed in Schwartzman et al. (2005). In this study, 6 dyslexic children and 6 normal children received diffusion tensor imaging brain scans on the same 15443 brain locations (voxels). A $z$-value (converted from a two-sample $t$-statistic) for comparing dyslexic versus normal children was obtained for each voxel. The right panel in Figure 4.1 plots the estimated null densities of the $z$-values for the front and back halves of the brain. We can see that the null cases from two groups centered on different means, and the density of the back half is narrower. There are many other examples where the hypotheses are naturally grouped. For instance, in analysis of geographical survey data, individual locations are aggregated into several large clusters; and in meta-analysis of large biomedical studies, the data are collected from different clinical centers. An important common feature of these examples is that data are collected from heterogeneous sources and the hypotheses being considered are grouped and no longer exchangeable. We shall see that incorporating the grouping information is important for optimal simultaneous inference with samples collected from different groups.

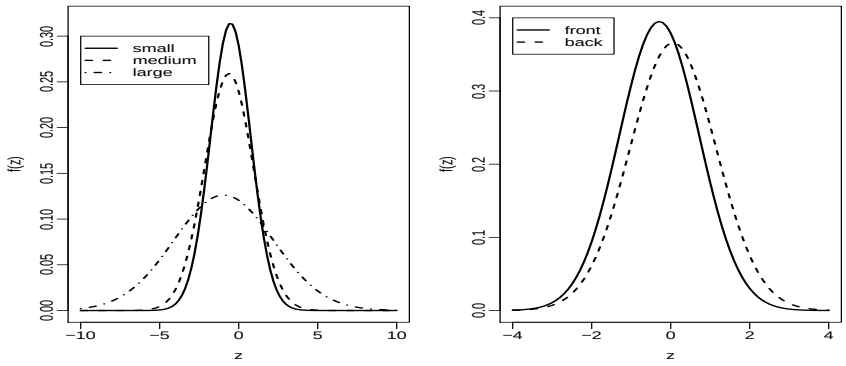


Figure 4.1: Empirical null densities of the AYP study and the brain imaging study. The null density of the large group is much wider than those of the other two densities. In the right panel, the null densities of the front and back halves of the brain are $N(0.06, 1.09^2)$ and $N(-0.29, 1.01^2)$, respectively, which are centered at different means.

## 4.2 The multiple-group model

The multiple-group random mixture model (Efron 2008a; see Figure 4.2) extends the previous random mixture model (3.1) (for a single group) to cover the situation where the $m$ cases can be divided into $K$ groups. It is assumed that within each group, the random mixture model (3.1) holds separately.

Let $\boldsymbol{g} = (g_1, \cdots, g_K)$ be a multinomial variable with probabilities $\{\pi_1, \cdots, \pi_K\}$, where $g_i = k$ indicates that case $i$ belongs to group $k$. We assume that prior to analysis, the group labels $\boldsymbol{g}$ have been determined by external information derived from other data or *a priori* knowledge. Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$ be Bernoulli variables, where $\theta_i = 1$ indicates that case $i$ is a non-null and $\theta_i = 0$ otherwise. Given $\boldsymbol{g}$, $\boldsymbol{\theta}$ can be grouped as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K) = \{(\theta_{k1}, \cdots, \theta_{km_k}) : k = 1, \cdots, K\}$, where $m_k$ is the number of hypotheses in group $k$. Different from $\boldsymbol{g}$, $\boldsymbol{\theta}$ are unknown and need to be inferred from observations $\boldsymbol{x}$. Let $\theta_{ki}, i = 1, \cdots, m_k$, be independent Bernoulli $(p_k)$ variables and $\boldsymbol{X} = (X_{ki})$ be generated conditional on $\boldsymbol{\theta}$:

$$X_{ki}|\theta_{ki} \sim (1 - \theta_{ki})F_{k0} + \theta_{ki}F_{k1},\ i = 1, \cdots, m_k,\ k = 1, \cdots, K. \tag{4.1}$$

Hence within group $k$, the $X_{ki}$'s, $i = 1, \cdots, m_k$, are i.i.d. observations with mixture distribution $F_k = (1 - p_k)F_{k0} + p_kF_{k1}$. Denote by $f_k$ the mixture density of group $k$, the null and non-null densities by $f_{k0}$ and $f_{k1}$, respectively. Then $f_k = (1 - p_k)f_{k0} + p_kf_{k1}$.

## 4.3 Conventional FDR procedures

We first consider the problem in an ideal setting where all distributional information is assumed to be known. This section considers two conventional FDR approaches: pooled and separate analyses.

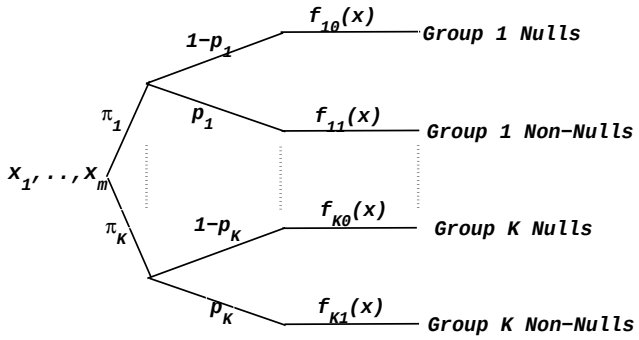


Figure 4.2: The multiple group model: the $m$ hypotheses are divided into $K$ groups with prior probability $\pi_k$; the random mixture model (3.1) holds separately within each group, with possibly different $p_k$, $f_{k0}$ and $f_{k1}$.

**Pooled FDR analysis.** A naive approach to testing grouped hypotheses is to simply ignore the group labels and combine all cases into a pooled sample. Denote by $f$ the mixture density,

$$f = \sum_k \pi_k[(1-p_k)f_{k0} + p_k f_{k1}] = (1-p)f_0^* + pf_1^*,$$

where $p = \sum_k \pi_k p_k$ is the non-null proportion, $f_0^* = \sum_k[(\pi_k - \pi_k p_k)/(1-p)]f_{k0}$ and $f_1^* = \sum_k(\pi_k p_k/p)f_{k1}$ are the pooled or global null and non-null densities, respectively. Denote the pooled null distribution by $F_0^* = \sum_k[(\pi_k - \pi_k p_k)/(1-p)]F_{k0}$. In a pooled analysis, the group labels are ignored and one tests against the common pooled null distribution $F_0^*$ in all individual tests. Define the pooled Lfdr statistic (PLfdr) by

$$\text{PLfdr}(x_i) = \frac{(1-p)f_0^*(x_i)}{f(x_i)},\ i = 1, \cdots, m. \tag{4.2}$$

The results in Section 3 imply that among all testing procedures that adopt the pooled-analysis strategy, the optimal one is

$$\boldsymbol{\delta}(\mathbf{PLfdr}, c_{OR}(\alpha)\mathbf{1}) = [I\{\text{PLfdr}(x_i) < c_{OR}(\alpha)\} : i = 1, \cdots, m], \tag{4.3}$$

where $c_{OR}(\alpha)$ is the largest cutoff for the PLfdr statistic that controls the overall FDR at level $\alpha$. Let $\text{PLfdr}_{(1)}, \cdots, \text{PLfdr}_{(m)}$ be the ranked PLfdr values and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses. An asymptotically equivalent version of (4.3) is the *PLfdr procedure*:

$$\text{Reject all } H_{(i)},\ i = 1, \cdots, l, \quad \text{where } l = \max\left\{ i : (1/i)\sum_{j=1}^{i} \text{PLfdr}_{(j)} \leqslant \alpha \right\}. \tag{4.4}$$

The following result shows that the PLfdr procedure is valid for FDR control when testing against the pooled null distribution $F_0^*$.

**Theorem 4.1.** *Consider the mixture model (4.1). Let* $PLfdr_{(i)}, i = 1, \cdots, m$ *be the ranked PLfdr values defined in (4.2). Then the PLfdr procedure (4.4) controls the FDR at level* $\alpha$ *when testing against the pooled null distribution* $F_0^*$.

We should emphasize here that a pooled analysis makes sense only when the null distributions $F_{k0}$ are the same for all groups, in which case $F_0^*$ coincides with the common group null. When $F_{k0}$ are different across groups, in general the pooled null distribution $F_0^*$ differs from any of the group null $F_{k0}$. In this case a pooled analysis is not appropriate at all because for each individual case a rejection against $F_0^*$ does not imply rejection against a null distribution $F_{k0}$ for a given group. To further illustrate this important point, let us take the most extreme case. Consider two groups where the null distribution of the first group is the alternative distribution of the second, and vice versa. It is then impossible to decide whether a case is a null or non-null without knowing the grouping information. In this case $F_0^*$ is not the right null distribution to test against for any individual tests and therefore it is entirely inappropriate to perform a pooled analysis.

**Separate FDR analysis.** Another natural approach to testing grouped hypotheses is the *separate analysis* where each group is analyzed separately at the same FDR level $\alpha$. Define the conditional Lfdr for group $k$ as

$$\text{CLfdr}^k(x_{ki}) = \frac{(1-p_k)f_{k0}(x_{ki})}{f_k(x_{ki})}, i = 1, \cdots, m_k;\ k = 1, \cdots, K. \tag{4.5}$$

Again implied by the results in Sun and Cai (2007), the optimal procedure for testing hypotheses from group $k$ is of the form

$$\boldsymbol{\delta}^k(\mathbf{CLfdr}^k, c_{OR}^k(\alpha)\mathbf{1}) = [I\{\text{CLfdr}^k(x_{ki}) < c_{OR}^k(\alpha)\} : i = 1, \cdots, m_k], k = 1, \cdots, K, \tag{4.6}$$

where $c_{OR}^k(\alpha)$ is the largest cutoff for CLfdr statistic that controls the FDR of group $k$ at level $\alpha$. By combining testing results from separate groups together, we have $\boldsymbol{\delta} = (\boldsymbol{\delta}^1, \cdots, \boldsymbol{\delta}^K)$.

Similarly we can propose the separated Lfdr *(SLfdr) procedure* that is asymptotically equivalent to (4.6). Denote by $\text{CLfdr}_{(1)}^k, \cdots, \text{CLfdr}_{(m_k)}^k$ the ranked CLfdr values in group $k$ and $H_{(1)}^k, \cdots, H_{(m_k)}^k$ the corresponding hypotheses. The testing procedure for group $k$ is:

$$\text{Reject all } H_{(i)}^k,\ i = 1, \cdots, l_k, \text{ where } l_k = \max\left\{ i : (1/i)\sum_{j=1}^{i} \text{CLfdr}_{(j)}^k \leqslant \alpha \right\}. \tag{4.7}$$

The final rejection set of the SLfdr procedure is obtained by combining the $K$ rejection sets from all separate analyses: $\mathcal{R}_{\text{SLfdr}} = \cup_{k=1}^K \{H_{(i)}^k : i = 1, \cdots, l_k\}$. The next theorem shows that the SLfdr procedure is also valid for global FDR control.

**Theorem 4.2.** *Consider the random mixture model (4.1). Let* $CLfdr^k_{(i)}, i = 1, \cdots, m_k$, $k = 1, \cdots, K$, *be the ranked CLfdr values defined by (4.5) for group* $k$. *Then the SLfdr procedure (4.7) controls the global FDR at level* $\alpha$.

## 4.4 Optimal FDR procedures for grouped tests

The pooled and separate analyses are inefficient in reducing the overall FNR. In this section, we begin by considering an ideal setting where all distributional information is known and propose an optimal (oracle) FDR procedure that uniformly outperforms both the pooled and separate procedures. We then turn to the situation where the distributions are unknown and introduce a data-driven procedure that is asymptotically valid and optimal.

Consider a weighted classification problem with loss function

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = (1/m) \sum_{k=1}^{K} \sum_{i=1}^{m_k} \lambda(1 - \theta_{ki})\delta_{ki} + \theta_{ki}(1 - \delta_{ki}). \tag{4.8}$$

The goal in a weighted classification problem is to find $\boldsymbol{\delta} \in \{0, 1\}^m$ that minimizes the classification risk $E[L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta})]$. Cai and Sun (2009) showed that the multiple testing and weighted classification problems are "equivalent" under mild conditions for model (4.1). Consider an ideal setting where an oracle knows $p_k$, $f_{k0}$ and $f_{k1}$, $k = 1, \cdots, K$. The optimal classification rule is given by the next theorem.

**Theorem 4.3.** *Consider the random mixture model (4.1). Suppose* $p_k$, $f_{k0}$, $f_{k1}$ *are known. Then the classification risk with loss function (4.8) is minimized by* $\boldsymbol{\delta}^\lambda = (\delta_{ki})$, *where*

$$\delta_{ki} = I\left\{\Lambda^k(x_{ki}) = \frac{(1 - p_k)f_{k0}(x_{ki})}{p_k f_{k1}(x_{ki})} < \frac{1}{\lambda}\right\}. \tag{4.9}$$

Note that $\Lambda^k(x) = \mathrm{CLfdr}^k(x)/[1-\mathrm{CLfdr}^k(x)]$ is strictly increasing in $\mathrm{CLfdr}^k(x)$, where $\mathrm{CLfdr}^k(x)$ is the conditional local false discovery rate defined in (4.5), an equivalent optimal test statistic is $\mathbf{CLfdr} = [\mathrm{CLfdr}^k(x_{ki}) : i = 1, \cdots, m_k, k = 1, \cdots, K]$. Therefore the optimal testing procedure is of the form $\boldsymbol{\delta}[\mathbf{CLfdr} < c(\alpha)]$. The MRC implies the cutoff should be chosen as $c_{OR}(\alpha) = \sup\{c \in (0, 1) : \mathrm{mFDR}(c) \leqslant \alpha\}$. Therefore the optimal (oracle) procedure for multiple group hypothesis testing is the following *CLfdr oracle procedure:*

$$\boldsymbol{\delta}[\mathbf{CLfdr}, c_{OR}(\alpha)\mathbf{1}] = [I\{\mathrm{CLfdr}^k(x_{ki}) < c_{OR}(\alpha)\} : i = 1, \cdots, m_k, k = 1, \cdots, K], \tag{4.10}$$

Note that different from (4.6), the oracle procedure (4.10) suggests using a universal cutoff for all CLfdr statistics regardless of their group identities.

For a given FDR level, it is difficult to calculate the optimal cutoff $c_{OR}(\alpha)$ directly. Also, the CLfdr oracle procedure requires the distributional information of all individual groups, which is usually unknown in practice. Cai and Sun (2009) derived the following CLfdr procedure that is asymptotically equivalent to the oracle procedure (4.10). Let $\hat{p}_k$, $\hat{f}_{k0}$ and $\hat{f}_k$ be estimates obtained for separate groups. The CLfdr procedure involves the following three steps:

1. Calculate the plug-in CLfdr statistic $\widehat{\text{CLfdr}}^k(x_{ki}) = (1-\hat{p}_k)\hat{f}_{k0}(x_{ki})/\hat{f}_k(x_{ki})$.
2. Combine and rank the plug-in CLfdr values from all groups. Denote by $\widehat{\text{CLfdr}}_{(1)}, \cdots, \widehat{\text{CLfdr}}_{(m)}$ the ranked values and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses.
3. Reject all $H_{(i)}$, $i = 1, \cdots, l$, where $l = \max\left\{i : (1/i)\sum_{j=1}^{i} \widehat{\text{CLfdr}}_{(j)} \leqslant \alpha\right\}$.

The next theorem shows that the data-driven procedure is *asymptotically valid and optimal* in the sense that both the FDR and FNR levels of the oracle procedure are asymptotically achieved by the data-driven procedure.

**Theorem 4.4.** *(Cai and Sun 2009). Consider the multiple group model (4.1). Let $\hat{p}_k$, $\hat{f}_{k0}$ and $\hat{f}_k$ be consistent estimates of $p_k$, $f_{k0}$ and $f_k$ such that $\hat{p_k} \xrightarrow{p} p_k$, $E\|\hat{f}_{k0} - f_{k0}\|^2 \to 0$, $E\|\hat{f}_k - f_k\|^2 \to 0$, $k = 1, \cdots, K$. Let*

$$\widehat{CLfdr}^k(x_{ki}) = (1-\hat{p})\hat{f}_0(x_{ki})/\hat{f}(x_{ki}),$$

*for $i = 1, \cdots, m_k, k = 1, \cdots, K$. Combine all test statistics from separate groups and let $\widehat{CLfdr}_{(1)}, \cdots, \widehat{CLfdr}_{(m)}$ be the ranked values. Then*

*(i). The FDR and FNR levels of the data-driven procedure are respectively $\alpha + o(1)$ and $FNR_{OR} + o(1)$, where $FNR_{OR}$ is the FNR level of the oracle procedure (4.10).*

*(ii). The FDR level of the data driven procedure in group $k$ can be consistently estimated as $\widehat{FDR}^k = (1/R_k)\sum_{i=1}^{R_k} \widehat{CLfdr}^k_{(i)}$. In addition, $\widehat{FDR}^k = FDR^k_{OR} + o(1)$, where $FDR^k_{OR} + o(1)$ is the FDR level of the oracle procedure (4.10) in group $k$.*

It is important to note that in Step 1, the external information of group labels is utilized to calculate the CLfdr statistic; this is the feature from a separate analysis. However, in Steps 2 and 3, the group labels are dropped and the rankings of all hypotheses are determined globally; this is the feature from a pooled analysis. Therefore the CLfdr procedure is a hybrid strategy that enjoys features from both pooled and separate analyses.

Unlike for the separate analysis, the group-wise FDR levels of the CLfdr procedure are in general different from $\alpha$. In addition to its validity, one may be interested in knowing the actual group-wise FDR levels $\text{FDR}^k$ yielded by the CLfdr procedure; this can be conveniently obtained based on the quantities that we have already calculated. Specifically, let $R_k$ be the number of rejections in group $k$. The actual $\text{FDR}^k$'s can be consistently estimated by $\widehat{\text{FDR}}^k = \frac{1}{R_k}\sum_{i=1}^{R_k} \text{CLfdr}^k_{(i)}$.

## 4.5 Simulation studies

Consider the following two-group normal mixture model:

$$X_{ki} \sim (1-p_k)N(\mu_{k0}, \sigma^2_{k0}) + p_k N(\mu_k, \sigma^2_k), \quad k = 1, 2. \tag{4.1}$$

The numerical performances of the PLfdr, SLfdr and CLfdr procedures are investigated in the next simulation study. The nominal global FDR level is 0.10.

*Example* 4.5. The null distributions of both groups are fixed as $N(0,1)$. Three simulation settings are considered: (i) The group sizes are $m_1 = 3000$ and $m_2 = 1500$; the group mixture pdf's are $f_1 = (1-p_1)N(0,1) + p_1 N(-2,1)$ and $f_2 = 0.9N(0,1) + 0.1N(4,1)$. We vary $p_1$, the proportion of non-nulls in group 1, and plot the FDR and FNR levels as functions of $p_1$. (ii) The groups sizes are also $m_1 = 3000$ and $m_2 = 1500$; the group mixture pdf's are $f_1 = 0.8N(0,1) + 0.2N(\mu_1, 1)$ and $f_2 = 0.9N(0,1) + 0.1N(2, 0.5^2)$. The FDR and FNR levels are plotted as functions of $\mu_1$. (iii) The marginal pdf's are $f_1 = 0.8N(0,1) + 0.2N(-2, 0.5^2)$ and $f_2 = 0.9N(0,1) + 0.1N(4,1)$. The sample size of group 2 is fixed at $m_2 = 1500$, the FDR and FNR levels are plotted as functions of $m_1$. The simulation results with 500 replications are given in Figure 4.3. The top row compares the actual FDR levels of the three procedures; the results for setting (i), (ii) and (iii) are shown in Panels (a), (b) and (c), respectively. The group-wise FDR levels of the CLfdr procedure are also provided (the dashed line for group 1 and dotted line for group 2). The bottom row compares the FNR levels of the three procedures; the results for setting (i), (ii) and (iii) are shown in Panels (d), (e) and (f), respectively.

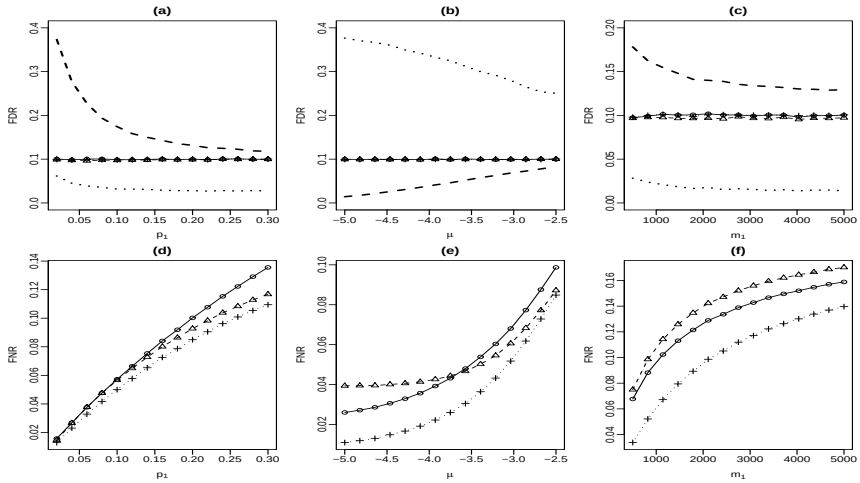


Figure 4.3: Results for Simulation Study 4.5: the top row compares the FDR levels and the bottom row compares the FNR levels ($\circ$ , PLfdr; $\triangle$, SLfdr; $+$, CLfdr). The optimal group-wise FDR levels suggested by the CLfdr procedure are provided together with the global FDR levels (dashed line, Group1; dotted line, Group2).

We can see that all three procedures control the global FDR level at the nominal level 0.10, indicating that all three procedures are valid. It is important to note that the CLfdr procedure chooses group-wise FDR levels automatically (dashed and dotted lines in Panels (a)-(c)), and the levels are in general different from the nominal level 0.10. The relative efficiency of PLfdr versus SLfdr is inconclusive (depends on simulation settings). For example, the SLfdr procedure yields lower FNR levels in Panel (d), but higher FNR levels in Panel (f). However, all simulations show that both the PLfdr and SLfdr procedures are uniformly

dominated by the CLfdr procedure.

## 4.6 A case study

We now return to the adequate yearly progress (AYP) study mentioned in the introduction. In this section, we analyze the data collected from $m = 7867$ of California high schools (Rogosa 2003) by using the PLfdr, SLfdr and CLfdr procedures.

One goal of the AYP study is to compare the success rates in Math exams of social-economically advantaged (SEA) versus social-economically disadvantaged (SED) students. Since the average success rates of the SEA students are in general (7370 out of 7867 schools) higher that the SED students, it is of interest to identify a subset of schools in which the advantaged-disadvantaged performance differences are unusually small or large. Denote by $X_i$ and $Y_i$ the success rates, and $n_i$ and $n_i'$ the numbers of scores reported for SEA and SED students in school $i$, $i = 1, \cdots, m$. Define the centering constant $\Delta = \text{median}(X_i) - \text{median}(Y_i)$. A $z$-value for comparing the SEA students versus the SED students can be computed for each school:

$$z_i = \frac{X_i - Y_i - \Delta}{\sqrt{X_i(1-X_i)/n_i + Y_i(1-Y_i)/n_i'}}, \tag{4.1}$$

for $i = 1, \cdots, m$. We claim school $i$ is "interesting" if the observed $|z_i|$ is large.

The AYP data has been analyzed by Efron (2007 and 2008b), where he first estimated the global null density $\hat{f}_0$, then searched for interesting cases in the tail areas of $\hat{f}_0$. This pooled-analysis strategy ignores the fact that the hypotheses formed for different schools are not exchangeable. In particular, the number of scores reported by each school varies from less than a hundred to more than ten thousands. A pooled analysis tends to over-select too many large schools, which often express themselves as "very significant" in the tail areas due to small denominators in (4.1). In contrast, small schools are likely to be hidden in the central area of $\hat{f}_0$ and appear "uninteresting". This is not desirable because, in practice, investigators are interested in identifying significant differences from all schools, not only from large schools. As we shall see, an important feature of the AYP data is that the empirical null distributions of the $z$-values are substantially different for small and large schools, therefore a pooled analysis is inappropriate and one should perform a separate analysis to take into account the effect of school size. Based on a preliminary cluster analysis, we divide all schools into three groups according to the number of scores reported ($n_i + n_i'$): small schools ($n_i + n_i' \leqslant 120$), medium schools ($120 < n_i + n_i' \leqslant 900$) and large schools ($n_i + n_i' > 900$). The group characteristics are summarized in Table 3, where the empirical null distributions are estimated using Jin and Cai (2007)'s method. Note that the variance of the empirical null distribution for the scores from the large schools is more than four times than those for the scores from the other two groups. See also Figure 4.1 in the introduction.

We then apply the PLfdr, SLfdr and CLfdr procedures to the AYP data at different FDR levels. The PLfdr procedure claims the most discoveries, followed by the CLfdr and then SLfdr procedure. It is important to emphasize that the PLfdr

Table 3: Group characteristics in the AYP data: 7867 schools in total. The global null density is $\hat{f}_0 = N(-0.59, 1.59^2)$

| Group | Group Definition | Group Size | Proportion | Empirical Null |
|---|---|---|---|---|
| Small | $n_i + n_i' \leqslant 120$ | 516 | 6.6% | $\hat{f}_{10} = N(-0.51, 1.27^2)$ |
| Medium | $120 < n_i + n_i' \leqslant 900$ | 6514 | 80.6% | $\hat{f}_{20} = N(-0.61, 1.54^2)$ |
| Large | $n_i + n_i' > 900$ | 837 | 12.8% | $\hat{f}_{30} = N(-0.95, 3.16^2)$ |

procedure is inappropriate here because the pooled null distribution is not the correct null to test against. The PLfdr procedure is too liberal for the large group yet too conservative for the small group: around 50%-70% significant schools come from the large group, although its population proportion is only 13%; in contrast, only around 1% interesting cases come from the small group, although its population proportion is more than 6%. The SLfdr procedure considers the groups separately; large schools are no longer over-selected and more small schools are identified. The CLfdr procedure further improves the SLfdr procedure by efficiently exploiting the important grouping information and weighting the numbers of discoveries among groups. The optimal group-wise FDR levels estimated by the CLfdr procedure at different nominal FDR levels are plotted in Figure 4.4, suggesting that we should choose higher FDR levels for the medium group and lower FDR level for the large group. Note that the SLfdr procedure uses the same FDR level for all groups, the CLfdr procedure usually identifies more cases from the medium group, but fewer cases from the large group.

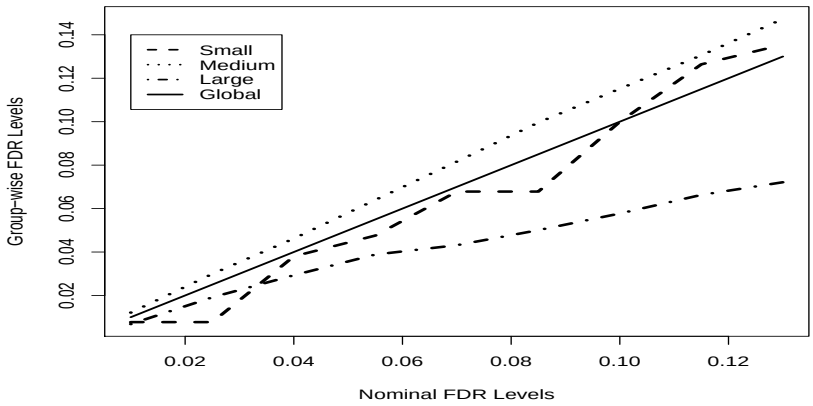


Figure 4.4: AYP study. Optimal group-wise FDR levels estimated by the CLfdr procedure.

# 5 Large-Scale Multiple Testing under Dependence

Observations arising from large scale multiple comparison problems are often dependent. For example, in microarray experiments, different genes may cluster into groups along biological pathways and exhibit high correlation. In public health surveillance studies, the observed data from different time periods and locations are often serially or spatially correlated. Correlation has big effects on a multiple testing procedure. Finner and Roters (2002) and Owen (2005) showed that both the expectation and variance of the number of Type I errors are greatly affected by the correlation among the hypotheses. Qiu et al. (2005) noted that the correlation effects can substantially deteriorate the performance of many FDR procedures. The correlation effects on the $z$-value null distribution is studied by Efron (2007), who suggested that an adjusted FDR estimate should be combined with the use of an Lfdr procedure to remove the bias caused by the correlation. Nevertheless, the works by Benjamini and Yekutieli (2001), Farcomeni (2006) and Wu (2009) show that the FDR is controlled at the nominal level by the BH step-up and adaptive $p$-value procedure under different dependence assumptions, supporting the "do nothing" approach.

Among the suggestions with respect to the correlation effects on an FDR procedure, the validity issue is overemphasized, and the efficiency issue is ignored. The FDR procedures developed under the independence assumption, even valid, may suffer from substantial efficiency loss when the dependence structure is highly informative. These situations include the geographical disease mapping studies, multiple-stage clinical trials, functional Magnetic Resonance Imaging analyses and comparative microarray experiments, where the non-null cases are often structured in some way, e.g., correlated temporally, spatially or functionally. Benjamini and Heller (2007) and Genovese et al. (2005) suggested incorporating scientific or spatial information into a multiple testing procedure to improve the efficiency. However, their approaches essentially rely on prior information, such as well defined clusters or prespecified weights, and the correlation structure among the hypotheses is not modeled.

We study multiple testing under dependency in a compound decision-theoretic framework. An important dependency structure, the hidden Markov model (HMM), is considered. The HMM is an effective tool for modeling the dependency structure and has been widely used in areas such as speech recognition, signal processing (Rabiner 1989; Ephraim and Merhav 2002). Also see Churchill (1992), Krogh et al. (1994) for its applications in analyzing biological sequences and processes.

In this section, we first propose an oracle testing procedure in an ideal setting where the HMM parameters are assumed to be known. Under mild conditions, the oracle procedure is shown to be optimal in the sense that it minimizes the FNR subject to a constraint on the FDR. This approach is distinguished from the conventional methods in that the proposed procedure is built on a new test statistic (local index of significance, LIS) instead of the $p$-values. Unlike $p$-values, the LIS takes into account the observations in adjacent locations by exploiting the local dependency structure in the HMM. The precision of individual tests is hence improved by utilizing the dependency information.

We then introduce a data-driven procedure that mimics the oracle procedure by plugging in consistent estimates of the unknown HMM parameters. The data-driven procedure is shown to be *asymptotically optimal* in the sense that it attains both the FDR and FNR levels of the oracle procedure asymptotically. Simulation studies conducted in Section 5.4 indicate the favorable performance of the LIS procedure. Our findings show that the correlation among hypotheses is highly informative in simultaneous inference and can be exploited to construct more efficient testing procedures. The LIS procedure is illustrated in analyzing the SNP data from a genome-wide association study of T1D disease in Section 5.5.

## 5.1 The hidden Markov model

Let $\boldsymbol{\theta} = (\theta)_1^m = (\theta_1, \cdots, \theta_m)$ be a sequence of Bernoulli random variables and distributed as a stationary Markov chain, where $\theta_i = 1$ indicates that case $i$ is a non-null and $\theta_i = 0$ otherwise. Assume that observations $\boldsymbol{x} = (x_1, \cdots, x_m)$ are generated according to the following conditional probability model:

$$P(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{F}) = \prod_{i=1}^{m} P(x_i|\theta_i, \mathcal{F}), \tag{5.1}$$

where $P(x_i < x|\theta_i = j) = F_j(x)$, $j = 0, 1$ and $\mathcal{F} = (F_0, F_1)$. Denote by $f_0$ and $f_1$ the corresponding pdf's. The HMM can be illustrated in Figure 5.1.

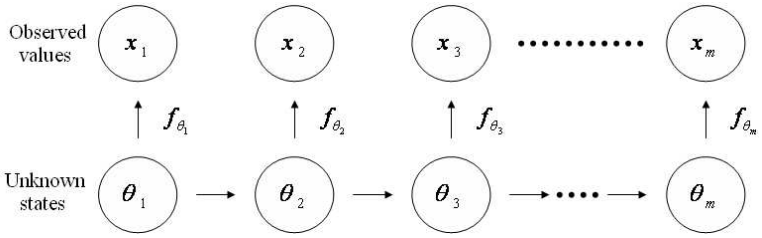


Figure 5.1: Graphical representation of an HMM

We assume that the Markov chain $(\theta_i)_1^m = (\theta_1, \cdots, \theta_m)$ is stationary, irreducible and aperiodic. Specifically, the transition probabilities are homogeneous and bounded away from 0 and 1. That is, $a_{jk} = P(\theta_i = k|\theta_{i-1} = j), 0 \leqslant j, k \leqslant 1$, do not depend on $i$, with the standard stochastic constraints $0 < a_{jk} < 1, a_{j0} + a_{j1} = 1$. The convergence theorem of a Markov chain implies that $(1/m)\sum_{i=1}^{m} I(\theta_i = j) \to \pi_j$ almost surely as $m \to \infty$. The Bernoulli variables $\theta_1, \cdots, \theta_m$ are identically distributed (but correlated) with $P(\theta_i = j) = \pi_j$. Denote by $\mathcal{A} = \{a_{jk}\}$ the transition matrix, $\boldsymbol{\pi} = (\pi_0, \pi_1)$ the stationary distribution, $\mathcal{F} = \{F_0, F_1\}$ the observation distribution, and $\vartheta = (\mathcal{A}, \boldsymbol{\pi}, \mathcal{F})$ the collection of all HMM parameters.

## 5.2 The oracle procedure

Let $\boldsymbol{\delta} \in \{0,1\}^m$ be a general decision rule defined as before. Sun and Cai (2009) showed that under the HMM dependency and a monotone ratio condition, the multiple testing problem is equivalent to a weighted classification problem with loss function

$$L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{m} \sum_i [\lambda(1-\theta_i)\delta_i + \theta_i(1-\delta_i)]. \tag{5.2}$$

It can be shown that the optimal solution to the weighted classification problem is $\boldsymbol{\delta}(\boldsymbol{\Lambda}, 1/\lambda) = (\delta_1, \cdots, \delta_m)$, where

$$\Lambda_i(\boldsymbol{x}) = \frac{P_\vartheta(\theta_i = 0|\boldsymbol{x})}{P_\vartheta(\theta_i = 1|\boldsymbol{x})} \tag{5.3}$$

and $\delta_i = I\{\Lambda_i(\boldsymbol{x}) < 1/\lambda\}$ for $i = 1, \cdots, m$.

*Remark* 5.1. Given $\vartheta$, the oracle classification statistic $\Lambda_i(\boldsymbol{x})$ can be expressed in terms of the forward and backward density variables, which are defined as $\alpha_i(j) = f_\vartheta[(x_t)_1^i, \theta_i = j]$ and $\beta_i(j) = f_\vartheta[(x_t)_{i+1}^m | \theta_i = j]$, respectively (note that the dependence of $\alpha_i(j)$ on $(x_t)_1^i$ has been suppressed, similarly for $\beta_i(j)$). It can be shown that $P_\vartheta(\boldsymbol{x}, \theta_i = j) = \alpha_i(j)\beta_i(j)$ and hence $\Lambda_i(\boldsymbol{x}) = [\alpha_i(0)\beta_i(0)]/[\alpha_i(1)\beta_i(1)]$. The forward variable $\alpha_i(j)$ and backward variable $\beta_i(j)$ can be calculated recursively using the *forward-backward procedure* (Baum et al. 1970 and Rabiner 1989). Specifically, we initialize $\alpha_1(j) = \pi_j f_j(x_1)$, $\beta_m(j) = 1$, then by induction we have $\alpha_{i+1}(j) = \left[\sum_{k=0}^1 \alpha_i(k) a_{kj}\right] f_j(x_{i+1})$ and $\beta_i(j) = \sum_{k=0}^1 a_{jk} f_k(x_{i+1}) \beta_{i+1}(k)$.

Since $\Lambda_i(\boldsymbol{x})$ is increasing in $P_\vartheta(\theta_i = 0|\boldsymbol{x})$, an optimal multiple-testing rule in an HMM can be written in the form of $\boldsymbol{\delta} = [I\{P_\vartheta(\theta_i = 0|\boldsymbol{x}) < t\} : i = 1, \cdots, m]$. Define the *local index of significance* (LIS) for hypothesis $i$ by

$$\mathrm{LIS}_i = P_\vartheta(\theta_i = 0|\boldsymbol{x}). \tag{5.4}$$

The LIS depends only on $x_i$ and reduces to Efron's local false discovery rate (Lfdr) in the independent case, i.e., $\mathrm{LIS}_i(\boldsymbol{x})$ simplifies to $\mathrm{Lfdr}(x_i) = (1-\pi)f_0(x_i)/f(x_i)$, where $\pi$ is the proportion of non-nulls and $f$ is the marginal pdf. The oracle testing procedure is

$$\boldsymbol{\delta} = [I\{\mathbf{LIS} < c_{OR}\mathbf{1}\} : i = 1, \cdots, m],$$

where $c_{OR}$ is the largest cutoff for LIS that controls the FDR at level $\alpha$.

*Remark* 5.2. It is important to note that the conventional testing procedures essentially involve ranking and thresholding $p$-values, whereas under our framework the optimal statistic is the LIS. Now we compare $p$-value and LIS from a compound decision theoretic view. Let $\boldsymbol{\delta}$ be a general decision rule, then $\boldsymbol{\delta}$ is *symmetric* if $\boldsymbol{\delta}(T(\boldsymbol{x})) = T(\boldsymbol{\delta}(\boldsymbol{x}))$ for all permutation operators $T$. In situations where one expects the non-null hypotheses appear in clusters, it is natural to treat differently a hypothesis surrounded by non-nulls from one surrounded by nulls. However, these two hypotheses are exchangeable when a symmetric rule is applied. The

FDR procedures that threshold the $p$-value or Lfdr are symmetric rules; so are not desirable when hypotheses are correlated. In contrast, we consider decision rule $\boldsymbol{\delta}(\mathrm{LIS}, \lambda) = \{I(\mathrm{LIS}_i(\boldsymbol{x}) < \lambda) : i = 1, \cdots, m\}$. It is easy to see that $\boldsymbol{\delta}(\mathrm{LIS}, \lambda)$ is asymmetric, and the order of the sequence $(x_i)_1^m$ is accounted for in deciding the significance level of hypothesis $i$. In particular, the local dependency structure is captured by the HMM, and the operation of the forward-backward procedure implies that a large (small) observation will increase (decrease) the significance level of its neighbors. The performance of the testing procedure is hence improved by pooling information from adjacent locations. In addition, the signal to noise ratio is increased since the information from the whole sequence is integrated to calculate the LIS value of a single hypothesis. Therefore, the LIS is more robust against local disturbance, which further increases the efficiency of our testing procedure.

## 5.3 A data-driven procedure for dependent tests in an HMM

The oracle procedure is difficult to implement since $c_{OR}$ is difficult to calculate. In addition, the HMM parameters $\vartheta$ are usually unknown. Sun and Cai (2009) derived a data-driven procedure that mimics the oralce procedure. We first estimate the unknown quantities by $\hat{\vartheta}$, then plug-in $\hat{\vartheta}$ to obtain $\widehat{\mathrm{LIS}}_i$. The maximum likelihood estimate (MLE) is commonly used and is strongly consistent and asymptotically normal under certain regularity conditions (Baum and Petrie, 1966; Leroux, 1992; Bickel et al., 1998). The MLE can be computed using the EM algorithm or other standard numerical optimization schemes, such as the gradient search, or downhill simplex algorithm. These methods are reviewed by Ephraim and Merhav (2002). In many practical applications, the number of components in the non-null mixture $L$ is unknown, yet the information is needed by the algorithms used to maximize the likelihood function. Consistent estimates of $L$ can be obtained using the method proposed by Kiefer (1993) and Liu and Narayan (1994), among others. Alternately, one can use likelihood based criteria, such as Akaike or Bayesian information criterion (BIC) to select the number of components in the normal mixture.

Let $\hat{\vartheta}$ be an estimate of the HMM parameter $\vartheta$. Define the plug-in test statistic $\widehat{\mathrm{LIS}}_i(\boldsymbol{x}) = P_{\hat{\vartheta}}(\theta_i = 0|\boldsymbol{x})$. For given $\hat{\vartheta}$, $\widehat{\mathrm{LIS}}_i$ can be computed via the forward-backward procedure. Denote by $\widehat{\mathrm{LIS}}_{(1)}(\boldsymbol{x}), \cdots, \widehat{\mathrm{LIS}}_{(m)}(\boldsymbol{x})$ the ranked plug-in test statistics and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses. In light of the oracle procedure, we propose the following data-driven procedure:

$$\text{Let } k = \max\left\{ i : \frac{1}{i}\sum_{j=1}^{i} \widehat{\mathrm{LIS}}_{(j)}(\boldsymbol{x}) \leqslant \alpha \right\}, \text{ then reject all } H_{(i)},\ i = 1, \cdots, k. \quad (5.5)$$

The testing procedure given in (5.5) is referred to as the *LIS procedure*. We shall show that the performance of OR is asymptotically attained by LIS under some standard assumptions on the HMM. The asymptotic properties of the LIS procedure are studied by the following theorems. Theorem 5.3 shows that the rejection sets yielded by OR and LIS are asymptotically equivalent in the sense

that the ratio of the number of rejections and the ratio of the number of true positives yielded by the two procedures approach 1 as $m \to \infty$.

**Theorem 5.3.** *Consider an HMM defined as in (5.1). Let $R$ and $\hat{R}$, $V$ and $\hat{V}$ be the number of rejections and number of false positives yielded by OR and LIS procedures, respectively. Under some regularity conditions (see Sun and Cai 2009), we have $\hat{R}/R \xrightarrow{p} 1$, $\hat{V}/V \xrightarrow{p} 1$.*

Theorem 5.4 below, together with the validity of the oracle procedure, implies that the FDR is controlled at level $\alpha + o(1)$ by LIS, so the LIS procedure is asymptotically valid. Theorem 5.4 also shows that the performance of OR is attained by the LIS procedure asymptotically in the sense that the FNR level yielded by LIS approaches that of OR as $m \to \infty$, therefore the LIS procedure is asymptotically efficient.

**Theorem 5.4.** *Consider an HMM defined as in (5.1). Let $FDR_{OR}$ and $FDR_{LIS}$, $FNR_{OR}$ and $FNR_{LIS}$ be the FDR levels and FNR levels yielded by OR and LIS, respectively. Under some regularity conditions (see Sun and Cai 2009) $FDR_{OR} - FDR_{LIS} \to 0$. In addition, if at least a fixed proportion of hypotheses are not rejected, then $FNR_{OR} - FNR_{LIS} \to 0$ as $m \to \infty$.*

## 5.4 Simulation studies

We first assume that $L$, the number of components in non-null mixture, is known or estimated correctly from the data. The situation where $L$ is misspecified is considered in Sun and Cai (2009). In all simulations, we choose the number of hypotheses $m = 3000$ and the number of replications $N = 500$.

*Example* 5.5. The Markov chain $(\theta_i)_1^m$ is generated with the initial state distribution $\boldsymbol{\pi}^0 = (\pi_0, \pi_1) = (1, 0)$ and transition matrix $\mathcal{A} = [0.95, 0.05; 1 - a_{11}, a_{11}]$. The observations $(x_i)_1^m$ are generated conditional on $(\theta_i)_1^m$: $x_i|\theta_i = 0 \sim N(0, 1)$, $x_i|\theta_i = 1 \sim N(\mu, 1)$. Figure 5.2 compares the performance of BH, AP, OR and LIS. In the top row we choose $\mu = 2$ and plot the FDR, FNR and average number of true positives (ATP) yielded by BH, AP, OR and LIS as functions of $a_{11}$. In the bottom row, we choose $a_{11} = 0.8$ and plot the FDR, FNR and ATP as functions of $\mu$. The nominal FDR in all simulations is set at level 0.10.

From Panel (a), we can see that the FDR levels of all four procedures are controlled at 0.10 asymptotically, and the BH procedure is conservative. From Panels (b) and (c), we can see that the two lines of the oracle procedure and LIS procedure are almost overlapped, indicating that the performance of the oracle procedure is attained by the LIS procedure asymptotically. In addition, the two $p$-value based procedures are dominated by the LIS procedure and the difference in FNR and ATP levels becomes larger as $a_{11}$ increases. Note that $a_{11}$ is the transition probability from a non-null case to a non-null case, therefore it controls how likely the non-null cases cluster together. It is interesting to observe that the $p$-value procedures have higher FNR levels as the non-nulls cluster in larger groups. In contrast, the FNR levels of the LIS procedure decreases as $a_{11}$ increases. This observation shows that if modeled appropriately, the positive dependency

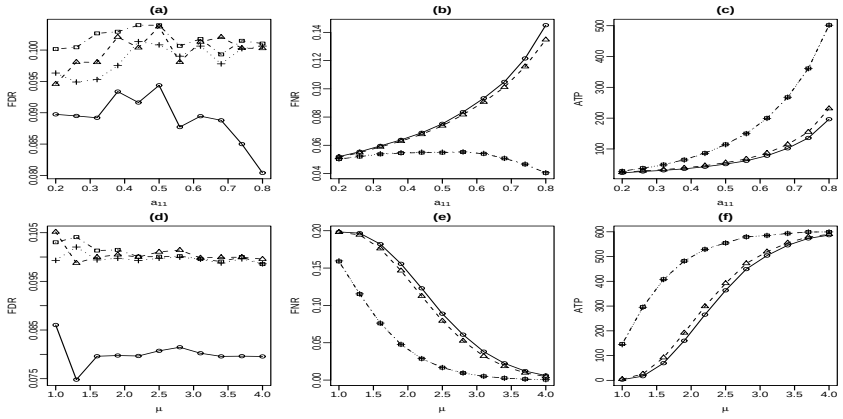


Figure 5.2: A comparison of the BH, AP, OR and LIS in an HMM with simple alternative (○ , BH; △, AP; +, OR; □, LIS). (a)-(c) The FDR, FNR and ATP versus $a_{11}$. (d)-(f) The FDR, FNR and ATP versus $\mu$. The FDR level is set at 0.10.

is a blessing (the FNR level decreases in $a_{11}$); but if it is ignored, the positive dependency may become a disadvantage. In situations where the non-null cases are prevented from forming into clusters ($a_{11} < 0.5$), the LIS procedure is still more efficient than BH and AP, although the gain in efficiency is not as much as the situation where $a_{11} > 0.5$.

Panel (d) similarly shows that all procedures are valid and BH is conservative. On Panels (e) and (f), we plot the FNR and ATP levels as functions of the non-null mean $\mu$. We can see that BH and AP are dominated by LIS, and the difference is large when $\mu$ is small to moderate. This is due to the fact that the LIS procedure can integrate information from adjacent locations, so is still very efficient even when the signals are weak.

The superiority of LIS is achieved by incorporating the informative dependency structure; hence more efficient rankings of the SNPs are produced. The LIS rankings are fundamentally different from the rankings by BH. Table 4 compares the outcomes of the LIS and BH procedures for testing two clusters of non-null SNPs, where ○ denotes a null hypothesis or an acceptance and ● denotes a non-null hypothesis or a rejection. It is interesting to note that BH suggests site 2720 be less significant than site 2723. In contrast, LIS suggests that site 2720, being surrounded by significant neighbors, be more significant than site 2723. LIS tends to identify disease-associated SNPs in clusters, while the BH procedure only identifies sporadic suspicious SNPs. Using HMM, LIS efficiently increases the signal to noise ratio by integrating information from adjacent locations. The precision is greatly improved in the sense that (1) the number of false positives is greatly reduced and (2) the statistical power to reject a non-null is substantially increased. This indicates that dependence can make the testing problem "easier" and is a *blessing* if incorporated properly in a testing procedure.

Table 4: Significance levels suggested by BH & LIS

| Sequence | States | $p$-values | LIS Values | BH *Procedure* | LIS *Procedure* |
|---|---|---|---|---|---|
| 2694 | ● | 3.62e-01 | 1.59e-03 | ○ | ● |
| 2695 | ● | 1.52e-03 | 6.85e-05 | ● | ● |
| 2696 | ● | 6.62e-04 | 9.29e-04 | ● | ● |
| 2697 | ● | 7.19e-01 | 1.00e-01 | ○ | ● |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2718 | ● | 1.23e-02 | 1.65e-04 | ○ | ● |
| 2719 | ● | 3.37e-03 | 8.19e-05 | ● | ● |
| 2720 | ● | 5.59e-01 | 2.19e-03 | ○ | ● |
| 2721 | ● | 2.07e-04 | 2.67e-05 | ● | ● |
| 2722 | ● | 3.42e-02 | 7.42e-04 | ○ | ● |
| 2723 | ● | 2.88e-01 | 6.19e-03 | ○ | ● |

## 5.5 A case study

Because of the recent advancements in comprehensive genomic information and cost-effective genotyping technologies, genome-wide association studies (GWAS) have become a popular tool to detect genetic variants that contribute to complex diseases. However, the established genetic associations with T1D only explain around 50% of the genetic risk for T1D. Many more genes with small to moderate effects remain to be discovered (Todd et al., 2007 and Hakonarson et al., 2007). Finding these unknown "weak" genes is important for improving the understanding of the pathology of T1D disease.

It has been appreciated that genomic dependency information can significantly improve the efficiency in analysis of large-scale genomic data. We expect that the information of the SNP dependency can be exploited to construct more efficient tests. From a biological point of view, the SNP dependency is informative in constructing more efficient association tests because, when a SNP is associated with a disease, it is likely that the neighboring SNPs are also disease-associated (due to the co-segregation). Therefore, when deciding the significance level of a SNP, the neighboring SNPs should be taken into account. The dependency of adjacent SNPs is captured using an HMM.

To search systematically for these unknown loci, Hakonarson et al. (2007) performed a genome-wide association study, where a discovery cohort, including 563 cases and 1146 controls, was collected. All participants were of European ancestry and recruited through paediatric diabetes clinics in Philadelphia, Montreal, Toronto, Ottawa and Winnipeg. The study subjects were genotyped using the Illumina HumanHap550 BeadChip at the Children's Hospital of Philadelphia (CHOP). A replication cohort, consisting of 483 parents-offspring trios with affected children, was also collected and genotyped (Hakonarson et al., 2007; Grant et al., 2008). A series of standard quality control procedures was performed to

eliminate markers with minor allele frequency less than 1%, with hardy-Weinberg Equilibrium $p$-values lower than 1e-6, or with genotype nocall rate higher than 5%. After the quality control, 534213 markers on 23 chromosomes in the discovery cohort are eligible for further analysis. The same quality control procedure was applied to the replication cohort. Additional markers showing excessive (>families) Mendelian inconsistencies were eliminated. After the screening, 532662 markers over 23 chromosomes in the replication cohort are eligible for further analysis.

We first conducted a $\chi^2$-test for each SNP to assess the association between the allele frequencies and the disease status, then obtain $p$-values and $z$-values using appropriate transformations. Each chromosome is modeled separately to obtain chromosome-specific HMM parameters $\Psi_k$ and LIS values $\widehat{\text{LIS}}_{ki}$. We assume that the null distribution is standard normal $N(0,1)$ and the non-null distribution is a normal mixture $\sum_{l=1}^{L} c_l N(\mu_l, \sigma_l^2)$. The number of components $L$ in the non-null distribution is determined by the BIC criterion, BIC $= \log\{P(\hat{\Psi}_L|\boldsymbol{z})\} - \frac{|\Psi_L|}{2}\log(m)$, where $P(\Psi_L|\boldsymbol{z})$ is the likelihood function, $\hat{\Psi}_L$ is the MLE of HMM parameters, and $|\Psi_L|$ is the number of HMM parameters. We vary $L$ and evaluate different choices of $L$ for each chromosome. The high transition probabilities ($a_{00}$ and $a_{11}$) indicate that the genomic dependency is strong.

A meta-analysis based on recent GWAS has confirmed 15 T1D susceptibility loci, among which four are identified by BH and seven are identified by the LIS procedure. Detailed results are provided in Table 5. In contrast, LIS does not claim two loci, namely the gene for collagen type 1 a2 (COL1A2; rs10255021) and rs672797, in the vicinity of latrophilin 2 (LPHN2), are disease-associated, while the p-value based approaches even claimed their significant association under the Bonferroni correction. However, these two loci failed to replicate in follow-up studies. A closer look at the nearby regions shows that these two SNPs are both surrounded by insignificant SNPs; hence the "significance" is more likely to be "noise". By borrowing information from nearby locations, LIS successfully classifies them as non-disease associated SNPs.

Table 5: The 7 known T1D susceptibility loci identified by PLIS.

| Chr. | T1D Loci | SNP | Dist. | $D'$ | LIS Stat. | p Value |
|---|---|---|---|---|---|---|
| 1 | rs2476601 | rs2476601 | 0 | | 3.19e-09 | 1.22e-12 |
| 6 | rs3129871 | rs3129871 | 0 | | 2.42e-89 | 2.91e-90 |
| 16 | rs725613 | rs725613 | 0 | | 2.73e-08 | 4.92e-09 |
| 11 | rs4244808 | rs4320932 | 8 | 1 | 8.40e-05 | 5.07e-04 |
| 12 | rs1701704 | rs10876864 | 11 | 0.955 | 3.26e-04 | 4.80e-07 |
| 10 | rs12251307 | rs4147359 | 15 | 0.610 | 8.48e-05 | 1.55e-04 |
| 6 | rs3757247 | rs10498965 | 42 | 1 | 1.96e-04 | 8.25e-04 |

Three loci are identified directly; four loci are identified by nearby SNPs (within 50Kb) in LD with them. Chr., chromosome; Dist., distance of the significant SNPs to the known T1D loci, in Kb; D', disequilibrium scores of the significant SNPs to the known T1D loci, derived from HapMap CEPH Utah (CEU) data.

# 6 Open Problems

We conclude with discussions on some open problems and possible directions for future research in large-scale multiple testing.

It is of great importance from both theoretical and practical perspectives to develop efficient multiple testing procedures under general dependency. The problem is challenging. The asymptotic optimality of a data-driven procedure requires the estimates of the unknown model parameters to be consistent. However, to the best of our knowledge, such theoretical results for other dependency structures, such as for a higher dimensional random field, have not been developed in the literature. Hence the optimality of the LIS procedure may be lost in the estimation step. In addition, the implementation of the data-driven procedure for other correlation structures may be very complicated. The forward-backward procedure and EM algorithm for an HMM is known to be efficient and relatively easy to program. However, such efficient algorithms may not exist for other dependency structures.

Another important direction is to extend the testing procedure in Section 4 to deal with the situation when the grouping information is unknown. One needs to take into account the interplay among grouping, estimation and testing in developing an efficient procedure. On the one hand, if we use many small groups instead of several large groups, the hypotheses within each small group will be more homogeneous, which will increase the precision of a testing procedure. On the other hand, the estimated test statistic will become less precise if a group consists of too few cases; hence the efficiency may be lost in the estimation step as the number of groups increases. A problem of particular interest in this direction is multiple testing with covariates.

The false discovery exceedance (FDX) control is an alternative approach to the FDR control. The false discovery proportion (FDP) is the proportion of false positives among all rejections and the FDX is the tail probability that the FDP exceeds a specified bound (Genovese and Wasserman 2006). In many applications, the FDP in each realization has very large deviations from the nominal FDR level. This unfavorable situation is allowed by FDR controlling procedures. Instead of controlling the average error rate, an FDX procedure aims to control the tail probability of FDP $> \alpha$ below a pre-specified level $\gamma$. The FDX procedures are argued to be more appropriate since the variability of FDP is taken into account. The procedures related to the FDX/FDP control are discussed in Pacifico et al.(2004), Lehmann and Romano (2005) and Genovese and Wasserman (2006). However, these procedures are essentially based on thresholding $p$-values. We anticipate that more efficient testing procedures can be developed under the compound decision theoretic framework.

Finally, we mention a few other interesting directions that are worth pursuing for future research with related references: multiple testing with weights (Genovese et al 2006; Roeder et al. 2007); conjunction and partial conjunction analysis of sets of hypotheses ( Friston et al. 2005; Pacifico et al. 2007; Benjamini and Heller 2007); multiple testing with a hierarchical structure (Meinshausen 2008; Yekutieli 2008; Bickel et al. 2009).

# References

[1] Baum, L., and Petrie, T. (1966), Statistical Inference for Probalistic Functions of Finite Markov Chains, *Ann. Math. Statis.,* 37, 1554-1563.

[2] Benjamini, Y., and Heller R. (2007), "False Discovery Rates for Spatial Signals", *Journal of the American Statistical Association*, **102**: 1272-1281.

[3] Benjamini Y., and Heller R. (2008), "Screening for Partial Conjunction Hypotheses", *Biometrics*, **64**: 1215-1222.

[4] Benjamini, Y., Hochberg, Y. (1995), "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society* , Series B, **57**: 289-300.

[5] Benjamini, Y., and Hochberg, Y. (2000), "On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics," *Journal of Educational and behavioral statistics,* **25**: 60-83.

[6] Benjamini, Y., and Yekutieli, D. (2001), "The Control of False Discovery Rate in Multiple Testing under Dependency," *The Annals of Statistics*, **29**:1165-1188.

[7] Bickel, P., Ritov, Y. and Rydèn, T. (1998), "Asymptotic Normality of the Maximum Likelihood Estimator for General Hidden Markov Models", *The Annals of Statistics,* 26, 1614-1635.

[8] Cai, T. Jin, J., and Low, M. (2007). "Estimation and confidence sets for sparse normal mixtures", *The Annals of Statistics*, **35**, 2421-2449.

[9] Cai, T., and Sun, W. (2009), "Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks", *Journal of the American Statistical Association,* to appear.

[10] Churchill, G. (1992), Hidden Markov Chains and the Analysis of Genome Structure. *Computers in Chemistry,* 16, 107-115.

[11] Copas, J. (1974), "On Symmetric Compound Decision Rules for Dichotomies", *The Annals of Statistics,* 2, No. 1, 199-204.

[12] Dudoit, S., Shaffer, J., and Boldrick, J. (2003). "Multiple hypothesis testing in microarray experiments", *Statistical Science*, 18, 71-103.

[13] Dudoit, S., Yang, Y., Callow, M., and Speed, T. (2002), "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments," *statistica Sinica*, **12**: 111-139.

[14] Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association,* **96**: 1151-1160.

[15] Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: the Choice of a Null Hypothesis," *Journal of the American Statistical Association,* **99**: 96-104.

[16] Efron, B. (2007), "Correlation and Large-Scale Simultaneous Testing", *Journal of the American Statistical Association,* **102**: 93-103.

[17] Efron, B. (2008a), "Simultaneous inference: When Should Hypothesis Testing Problems Be Combined?", *The Annals of Applied Statistics*, **1**: 197-223.

[18] Efron, B. (2008b), "Microarrays, empirical Bayes and the two-groups model", *Statistical Science*, 23, 1-22.

[19] Ephraim, Y. and Merhav, N. (2002), "Hidden Markov Processes", *invited paper, IEEE transactions on Information Theory*, 48, 1518-1569.

[20] Farcomeni, A. (2007), "Some Results on the Control of the False Discovery Rate under Dependence", *Scandinavian Journal of Statistics*, **34**: 275-297.

[21] Ferkinstad, E., Frigessi, A., Thorleifsson, G. and Kong, A. (2008), "Unsupervised Empirical Bayesian Multiple Testing with External Covariates", *The Annals of Applied Statistics*, **2**: 714-735.

[22] Finner, H., and Roters, M. (2002), "Multiple hypotheses testing and expected number of type I errors", *The Annals of Statistics*, **30**, 220-238.

[23] Finner, H., Dickhaus, T., and Roters, M. (2009), "On the false discovery rate and an asymptotically optimal rejection curve", **37**, 596-618.

[24] Friston, K., Penny, W., and Glaser, D. (2005), "Conjunction revisited", *NeuroImage* **25**: 661667.

[25] Genovese, C., Roeder, K. and Wasserman, L. (2006), "False Discovery Control with $p$-value Weighting, " *Biometrika*, **93**: 509-524.

[26] Genovese, C. and Wasserman, L. (2002), "Operating Characteristic and Extensions of the False Discovery Rate Procedure, " *Journal of the Royal Statistical Society ,* Sries B, **64**: 499-517.

[27] Genovese, C. and Wasserman, L. (2004a), "A Stochastic Process Approach to False Discovery Control," *Annals of Statistics,* **32**: 1035-61.

[28] Genovese, C. and Wasserman, L. (2006), "Exceedance control of the false discovery proportion," *Journal of the American Statistical Association*, **101**, 1408-1417.

[29] Genovese, C., Roeder, K., Wasserman, L. (2006), "False Discovery Control with $p$ value Weighting. " *Biometrika*, 93, 509-524.

[30] Hakonarson, H., Grant, S., Bradfield, J.; et al. (2007) "A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene", *Nature*, **448**:591-594.

[31] Hedenfalk, I., Duggen, D., Chen, Y., et al. (2001), "Gene Expression Profiles in Hereditary Breast Cancer," *New England Journal of Medicine*, **344**, 539-548.

[32] Holm, S. (1979). "A simple sequentially rejective multiple test procedure", *Scandinavian Journal of Statistics*, 6, 65-70.

[33] Hochberg, Y. (1988)." A sharper Bonferroni procedure for multiple tests of significance", *Biometrika*, 75,800-3.

[34] Hochberg, Y., and Tamhane, A. (1987), "Multiple Comparison Procedures," Wiley, New York.

[35] Jin, J. and Cai, T. (2007), "Estimating the Null and the Proportion of Non-Null Effects in Large-scale Multiple Comparisons," *Journal of the American Statistical Association*, 102, 495-506.

[36] Kieffer, J. (1993), Strongly Consistent Code-Based Identification and Order Estimation for Constrained Finite-State Model Classes, *IEEE Transactions on Information Theory,* 39, 893-902.

[37] Krogh, A., Brown, M., Mian, I., Sjölander, K., and Haussler, D. (1994), Hidden Markov Models in Computational Bilogy Applications to Protein Modeling, *Journal of Molecular Biology,* 235, 1501-1531.

[38] Langaas M., Lindqvist, B., and Ferkingstad, E. (2005), "Estimating the proportion of true null hypotheses, with application to DNA microarray data", *Journal of the Royal Statistical Society ,* Series B, 67, 555-572.

[39] Lehmann, E., and Romano, J. (2005), "Generalizations of the familywise error rate," *The Annals of Statistics*, **33**, 1138-1154.

[40] Leroux, B. (1992), "Maximum-likelihood Estimation for Hidden Markov Models", *Stochastic Processes and their applications*, 40, 127-143.

[41] Liu, C., and Narayan, P. (1994), Order Estimation and Sequential Universal Data Compression of a Hidden Markov Source by the Method of Mixtures, *IEEE Transactions on Information Theory,* 40, 1167-1180.

[42] Meinshausen, N., and Rice, J. (2006), "Estimating the Proportion of False Null Hypotheses among a Large Number of Independently Tested Hypotheses", *The Annals of Statistics*, **34**: 373-393.

[43] Miller, C., Genovese, C., Nichol, R., Wasserman, L., Connolly, A., Reichart, D., Hopkins, D., Schneider, J., and Moore, A. (2001), "Controlling the False-Discovery Rate in Astrophysical Data Analysis", *The Astronomical Journal*, **122**: 3492-3505.

[44] Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting differential gene expression with a semiparametric hierachical mixture method," *Biostatstics*, 5, 155-176.

[45] Owen, A. (2005), "Variance of the number of false discoveries," *Journal of the Royal Statistical Society Series B,* **67**: 411-426.

[46] Pacifico M., Genovese C., Verdinelli, I., and Wasserman, L. (2004), "False Discovery Control for Random Fields", *Journal of the American Statistical Association*, **99**: 10021014.

[47] Qiu, X., Klebanov, L., and Yakovlev, A. (2005), "Correlation Between Gene Expression Levels and Limitations of the Empirical Bayes Methodology for Finding Differentially Expressed Genes", *Statistical Applications in Genetics and Molecular Biology*, 4, Article 34. Available at: "http://www.bepress.com/sagmb/vol4/iss1/art34".

[48] Rabiner, L. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Procedings of the IEEE*, **77**: 257-286.

[49] Robbins, H. (1951), "Asymptotically Subminimax solutions of compound statistical decision problems," *Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press, Berkeley.

[50] Roeder, K., Devlin, B., and Wasserman, L. (2007), "Improving Power in Genome-Wide Association Studies: Weights Tip the Scale", *Genetic Epidemiology*, **31**: 741-747

[51] Romano, J., and Shaikh, A. (2006), "Step-up procedures for control of generalizations of the familywise error rate," *The Annals of Statistics*, **34**, 1850-1873.

[52] Rogasa, D. (2003), "Accuracy of API index and school base report elements: 2003 Academic Performance Index, California Department of Education", Technical Report, Department of Statistics and School of Education, Stanford University, available at "http://www.cde.ca.gov/ta/ac/ap/researchreports.asp".

[53] Rosenthal R., and Rubin D. (1983). "Ensemble-adjusted p-values", *Psychological Bulletin*, 94, 540-41.

[54] Sabatti, C., Service, S., and Freimer, N. (2003), "False Discovery Rate in Linkage and Association Genome Screens for Complex Disorders", *Genetics*, **164**: 829-833

[55] Sarkar, S. (2002), "Some results on false discovery rate in stepwise multiple testing procedures", *Annals of statistics,* **30**, 239-257.

[56] Sarkar, S. (2006), "False Discovery and False Nondiscovery Rates in Single-Step Multiple Testing Procedures", *Annals of statistics,* **34**: 394-415.

[57] Sarkar, S. K. (2007), "Step-up procedures controlling generalized FWER and generalized FDR," *The Annals of Statistics*, **35**, 2405-2420.

[58] Sebastiani, P., Gussoni, E., Kohane, I., and Ramoni, M. (2003), "Statistical Challenges in Functional Genomics," *Statistical Science*, 18, 33-70.

[59] Shaffer (1995), "Multiple hypothesis testing", *Annual Review of Psychology*, **46**, 561-584.

[60] Spjøtvoll, E. (1972), "On the Optimality of Some Multiple Comparison Procedures," *The Annals of Mathematical Statistics,* 43, 398-411.

[61] Storey, J. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society ,* Sries B, **64**: 479-498.

[62] Storey, J. (2003), "The Positive False Discovery Rate: a Bayesian Interpretation and the Q-value," *The Annals of Statistics,* **31**:2012-35.

[63] Storey, J., and Tibshirani, R. (2003), "Statistical significance for genome-wide studies", *Proceedings of the National Academy of Sciences*, **100**: 9440-9445

[64] Storey, J. (2007), "The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing," *Journal of Royal Statistical Society*, Series B, 69, 347-368.

[65] Schwartzman, A., Dougherty, R., Taylor, J. (2008) "False Discovery Rate Analysis of Brain Diffusion Direction Maps," *Annals of Applied Statistics*, **2**: 153-175.

[66] Sun, W., and Cai, T. (2007), "Oracle and adaptive compound decision rules for false discovery rate control", *Journal of the American Statistical Association*, **102**: 901-912.

[67] Sun, W., and Cai, T. (2009), "Large-scale multiple testing under dependence", *Journal of the Royal Statistical Society*, Series B, **71**:393-424.

[68] Todd JA, Walker NM, Cooper JD; et al. (2007) "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes", *Nature Genetics*, **39**:857-864.

[69] Tusher, V., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response", *Proceedings of National Academy of Science,* USA **98**: 5116-5121.

[70] van der Laan, M., Dudoit, S., and Pollard, K. (2004), "Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives," *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 15.

[71] van't Wout, A., Lehrman, G., Mikheeva, S., O'Keeffe, G., Katze, M., Bumgarner, R., Geiss, G., and Mullins, J. (2003), "Cellular Gene Expression upon Human Immunodeficiency Virus Type 1 Infection of $CD4^+$-T-Cell Lines", *Journal of Virology,* 77, 1392-1402.

[72] Weller, J., Song, J., Heyen, D., Lewin, H., and Ron, M. (1998). "A New Approach to the Problem of Multiple Comparisons in the Genetic Dissection of Complex Traits", *Genetics* **150**: 1699-1706.

[73] Westfall P., and Young S. (1993), "Resamplingbased Multiple Testing." New York, Wiley.

[74] Wright S. (1992), "Adjusted p-values for simultaneous inference", *Biometrics*, 48, 1005-13.

[75] Wu, W. (2009), "On False Discovery Control under Dependence", *The Annals of statistics,* **36**: 364-380.