# Optimal Multiple Testing Procedure under a General Loss Funtion

September 14, 2011

## 1 Introduction

In a single hypothesis testing problem, one tries to find the most powerful test under the restriction that the probability of making a Type I error is controlled at level $\alpha$. This can be viewed as a special case of the general decision theoretical problem. Suppose we want to conduct a test on some parameter $\mu$: $H_0$: $\mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, based on the observations $\mathbf{X}$. Let $\theta$ be the indicator that $H_0$ is false. Let $\delta(\mathbf{X})$ be the decision rule: $H_0$ is rejected if $\delta(\mathbf{X}) = 1$ and not rejected if $\delta(\mathbf{X}) = 0$ . Define the following two functions $L_1$ and $L_2$:

$$L_1(\delta,\theta) = \begin{cases} \lambda_0 & \text{if } \delta = 1 \text{ and } \theta = 0 \\ 0 & \text{Otherwise} \end{cases} \quad \text{and} \quad L_2(\delta,\theta) = \begin{cases} \lambda_1 & \text{if } \delta = 0 \text{ and } \theta = 1 \\ 0 & \text{Otherwise} \end{cases}$$

Here $\lambda_0$ is the cost for a Type I error and $\lambda_1$ is the cost for a Type II error. Let $\lambda = \lambda_0/\lambda_1$. Then $\lambda$ is the relative cost of a Type I to Type II error. The weighted 0-1 loss is defined as below:

$$L_\lambda(\theta,\delta) = \lambda(1-\theta)\delta + (1-\delta)\theta \tag{1}$$

Note that the weighted 0-1 loss function is equivalent to $L_1(\delta, \theta) + L_2(\delta, \theta)$. The rule that minimizes the expected loss function $L_\lambda(\theta, \delta)$ in (1) is the one that minimizes $E(L_2)$ subject to controlling $E(L_1)$ at a certain level $\alpha(\lambda)$.

The weighted 0-1 loss function (1) is very simple in the sense that it only judges whether a decision is right or wrong and it gives the same penalty for all type II errors under different values of $\mu$ and does not reflect how serious a wrong decision is. In reality, it is often the case that some wrong decisions are more (less) serious than others. For example, in clinical trials, the researcher may want to find out whether a possible side effect of a new drug increases the blood pressure of the patients. Due to the many combinations of high blood pressure, the incorrect acceptance of a false null hypothesis is a much more serious mistake if the patients' blood pressure is raised much higher than the standard values. In such cases, it may not be sensible to give the same penalty for accepting the null hypothesis when the true parameter $\mu$ is equal to $\mu_0 + 1$ and when it is equal to $\mu_0 + 5$. Hence the more the alternative value differs from the null value, the more serious is the corresponding Type II error. On the other hand, if true value for the parameter is very close to the null, then the mistake of accepting the alternative hypothesis may not be considered very serious. For example, a new drug has been developed and its efficacy is compared to that of some conventional drug. If the improvement in efficacy of the new drug over the conventional one is very slight, then the new drug may not be worthwhile considering the cost it will induce to produce it and the possible side effects associated with it. Hence accepting the null hypothesis may not a serious mistake, if the true value $\mu$ is equal to $\mu + 0.5$. These observations motivate our following consideration, in determining the penalty for a Type II error, the distance between the alternative and null

value for the tested parameter should play a role. Accordingly, we define the following:

$$L_2^*(\theta,\delta) = \begin{cases} \lambda_1 s(|\mu-\mu_0|) & \text{if } \delta = 0 \text{ and } \theta = 1 \\ 0 & \text{Otherwise} \end{cases}$$

Note $s(\cdot)$, called the severity function for a Type II error, is a nondecreasing function, and $\lambda_1 s(|\mu-\mu_0|)$ is the cost for a Type II error. The larger the value of $|\mu-\mu_0|$, the more severe is the Type II error and hence the more the cost is associated with it. The choice of the $s(\cdot)$ depends on how fast we want the cost of the Type II error to increase as $\mu$ moves away from $\mu_0$. Again write $\lambda = \lambda_0/\lambda_1$. Then the following loss function can be considered.

$$L_{\lambda,s}(\theta,\delta) = \lambda(1-\theta)\delta + s(|\mu-\mu_0|)(1-\delta)\theta \tag{2}$$

Duncan [5] considered a special case of (2) with $s(\cdot) = |\cdot|$. The relative cost of a Type I to a Type II error in (2) is then $\dfrac{\lambda}{s(|\mu-\mu_0|)}$, which is a function of $|\mu-\mu_0|$. For different choices of $s(\cdot)$ function, this relative cost is shown in Figure 1.

Simultaneous inferences can also be conducted in the form of a decision theoretical analysis. In particular, by doing so, optimal multiple testing procedures can often be derived. Assume testing $m$ null hypotheses $H_1, H_2, \ldots, H_m$ simultaneously based on the observations $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_m})$, where $H_i$: $\mu_i = \mu_{i0}$. Consider $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$, with $\theta_i \in \{0,1\}$, where $\theta_i = 1$ indicates $H_i$ is false and $\theta_i = 0$ indicates otherwise. Also let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ be the decision rule, with $\delta_i \in \{0,1\}$, where $\delta_i = 1$ corresponds to rejecting $H_i$ and $\delta_i = 0$ otherwise.

The following average weighted 0-1 loss function has often been considered.

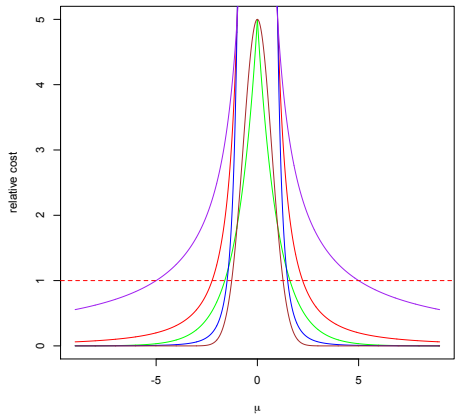


Figure 1: Relitive cost for different choices of $\lambda_1$: $s(|x|) = |x|$ (purple), $s(|x|) = x^2$ (red), $s(|x|) = x^4$ (blue), $s(|x|) = e^{|x|}$ (green), $s(|x|) = e^{|x|^2}$ (brown). $\lambda$ is chosen to be 5.

See (Spjotvoll [13], Storey [14], Sun and Cai [15]).

$$L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{m} \sum_i \{\lambda(1-\theta_i)\delta_i + \theta_i(1-\delta_i)\} \tag{3}$$

Following the argument in the single hypothesis testing case, we propose to consider the following loss function

$$L_{\lambda,s}(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{m} \sum_i \{\lambda(1-\theta_i)\delta_i + s(|\mu_i - \mu_{i0}|)\theta_i(1-\delta_i)\} \tag{4}$$

Note that the weighted 0-1 loss function (3) is a special case of our proposed loss function (4) with $s(\cdot) = 1$. We think the loss function (4) is often more reasonable and more sensitive to signals, resulting in more powerful tests. As a motivating example, we performed the following numerical study. We generated m(=1000) i.i.d. Bernoulli random variables $\theta_1, \ldots, \theta_m$ with success probability $1 - \pi_0$. We then generate $\mu_1, \ldots, \mu_m$ according to the model $(1-\theta_i)N(0,1) + \theta_i N(2,1)$. Then the observations $x_i$'s are independently

generated from the $N(\mu_i, 1)$ distribution. We considered testing $H_i : \mu_i = 0, i = 1, \ldots, m$. We made our decision, respectively, based on the loss funtion (3) and (4) when choosing $s(|\mu|) = \mu^2$. The results are demonstrated in Figure 2. We first fixed $\pi_0$ to be 0.8 and let $\lambda$ vary between 5 and 12. For each simulation, we calculated proportions of correctly rejected hypotheses among all the rejections and the correctly accepted hypotheses among all acceptances, based on the two classification rules that minimize the two loss functions (3) and (4) respectively. We then repeated the experiment 1000 times and calculated the average true rejection proportion (true rejection rate) and average true acception proportion (true acception rate) and plotted them in the top panels. The top panels show that both the true rejection rate and true acception rate resulting from the loss function (4) (red) is larger than those resulting from the loss function (3) (blue). Similar phenomenon occurs when fixing $\lambda$ to be 10 and letting $\pi_0$ vary from 0.2 to 0.8 with an increment of 0.2, as shown in the bottom two panels. Consequently, the loss function (4) with the penalty of Type II error relying on the severity function seems preferable to the weighted 0-1 loss (3).

The remaining of this paper is organized as follows: Section 2 gives a brief review of the basics of multiple testing and some relevant literature. In section 3, we specify our model framework and derive the decision rule that minimizes the expectation of the loss function (4). In section 4, we deriven an optimal multiple testing procedure under the loss function (4). In section 5, we derive a data driven procedure that performs as well as the oracle procedure asymptotically. In Section 6, we present the results of simulation studies and real data analyses. Proofs of the results are given in Appendix.

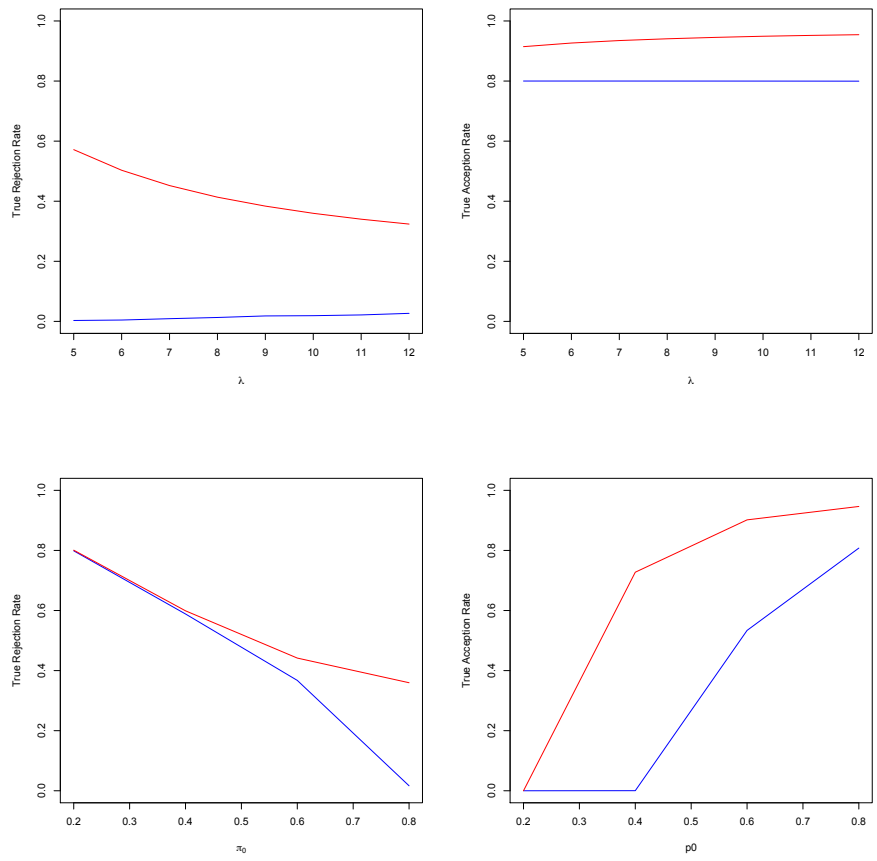


Figure 2: Comparison of the Bayes rules based on the two loss functions $L_\lambda$ (3) and $L_{\lambda,s}$ (blue) (4) (red)

## 2 Preliminaries

Again, consider testing $m$ null hypotheses $H_1, H_2, \ldots, H_m$ simultaneously based on the observations $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_m})$. Let $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ be defined as in Section 1. Define $V_i = I(\theta_i = 0, \delta_i = 1)$, $S_i = I(\theta_i = 1, \delta_i = 1)$, $U_i = I(\theta_i = 0, \delta_i = 0)$, $K_i = I(\theta_i = 1, \delta_i = 0)$, and define $R_i = I(\delta_i = 1) = 1 - A_i$. Table 1 illustrates the categorization of the tested hypotheses of any procedure, where $R = \sum R_i$ is the total number of rejections, $A = \sum A_i$ is the total number of acceptions, $V = \sum V_i$ is the number of Type I errors, and $K = \sum K_i$ is the number of Type II errors. Popular

Table 1: **Classifications of Tested Hypotheses**

| Hypotheses | Claimed non-significant | Claimed significant | Total |
|---|---|---|---|
| Null | $U$ | $V$ | $m_0$ |
| Nonnull | $K$ | $S$ | $m_1$ |
| Total | $A$ | $R$ | $m$ |

joint measures of Type I errors include the false discovery rate(FDR), defined as $E\left(V/R \vee 1\right)$, and the marginal false discovery rate ($m$FDR), defined as $E(V)/E(R \vee 1)$. When the number of hypotheses is very large, which is often the case in many scientific investigations, the $m$FDR is equivalent to the FDR, in the sense that $m$FDR=FDR+$O(1/\sqrt{m})$, (see Genovese and Wasserman [10] and Xie et al. [17]). Accordingly, joint measures of Type II errors include the false nondiscovery rate (FNR), defined as $E\left(K/A \vee 1\right)$, and the marginal false nondiscovery rate (mFNR), $E(K)/E(A \vee 1)$. Also, as shown in Genovese and Wasserman [10] and Xie et al. [17], $m$FNR=FNR+$O(1/\sqrt{m})$. That is, $m$FNR is asymptotically equivalent to the FNR, for large $m$.

The following mixture model has been widely used in multiple testing literature, expecially those involving controlling the FDR, (see Efron [8] and

Efron [9]).

**Definition 2.1** *Suppose that $\theta_i$, $i = 1, 2, \ldots, m$, are independent Bernoulli random variables with $P(\theta_i = 1) = \pi_1 = 1 - \pi_0$. Conditioned on $\theta_i$, $X_i$ is distributed as*

$$X_i \mid \theta_i \sim (1 - \theta_i)F_0 + \theta_i F_1 \tag{5}$$

*Then the marginal cumulative distribution function of $X_i$ is the mixture distribution $F(x) = \pi_0 F_0(x) + \pi_1 F_1(x)$, and the probability density function is $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$.*

Usually, when deriving a multiple testing procedure, one aims to controll a chosen error rate at a prespecified level by determining a rejection region for the test statistics. For example, Benjamini and Hochberg [1] proposed a procedure (known as the BH procedure) that was shown to control the FDR under a certain positive dependence assumption on the test statistics, as shown in Benjamini and Yekutieli [2]. On the other hand, decision theoretical analysis can provide one way to achieve optimality in multiple testing. Specifically, multiple testing procedures that minimize a measure of Type II errors while controlling a corresponding measure of Type I errors can be developed. Spjotvoll [13] and Storey [14], respectively, derived an optimal procedure that maximizes the number of expected true positives $E(S)$, for each fixed number of expected false positives $E(V)$.

Genovese and Wasserman [10] constructed an optimal procedure which minimizes the FNR subject to a bound on the FDR, under the mixture model (5). Their procedure is based on the $p$-values for individual tests. Specifically,

the threshold $\mu^*$ of their oracle procedure is defined as below:

$$\mu^* = \sup\left\{t : Pr(P \leq t) = \frac{\pi_0 t}{\alpha}\right\}.$$

Under the mixture model (5), Sun and Cai [15] developed a compound decision theoretical framework for multiple testing, and derived an optimal procedure that minimizes the $m$FNR subject to controlling the $m$FDR. They considered the weighted 0-1 loss function (3). The corresponding decision rule is essentially based on the local fdr statistic, defined as the posterior probability $P(\theta_i = 0 \mid X_i = x_i) = \pi_0 f_0(x_i)/f(x_i)$ which was introduced by Efron et al. [7] and has been widely used to interpret results for individual cases. They have further shown that, if the test statistic satisfies a monotone likelihood ratio condition, the optimal multiple testing procedure that minimizes the $m$FNR while controlling the $m$FDR is based on the same test statistics on which the compound decision rule is based. Sun and Cai [15] derived an oracle testing procedure and a data-driven adaptive procedure that asymptotically attains the performance of the oracle procedure. Their procedure was shown to dominate many others, such as those of Benjamini and Hochberg [1] and Genovese and Wasserman [10]. Further, Xie et al. [17] defined short range dependence and showed that the results in Sun and Cai [15] still holds under the short range dependence.

# 3 The Model and The Decision Rule

We consider the following hierarchical Bayes model, for $i = 1, \ldots, m$,

$$\begin{aligned} \theta_i &\sim Bernoulli(\pi_1) \\ \mu_i \mid \theta_i &\sim (1-\theta_i)h_0 + \theta_i h(\mu_i) \\ \mathbf{X} \mid \boldsymbol{\mu} &\sim f(\mathbf{x}; \boldsymbol{\mu}) \end{aligned} \tag{6}$$

Here $h_0$ is a point mass at 0. Note that, in this model, we do not impose any dependence restriction on $\mathbf{X}$, $\boldsymbol{\mu}$ or $\boldsymbol{\theta}$.

Let $w_i(\mathbf{x})$ be defined as

$$w_i(\mathbf{x}) = E(s(|\mu_i|) \mid \theta_i = 1, \mathbf{x}) \tag{7}$$

which is viewed as the average severity conditional on the observations $\mathbf{X}$.

The following theorem provides the Bayes decision rule $\boldsymbol{\delta}^*$ which minimizes the conditional expectation of the loss function (4) given observations $\mathbf{X}$.

**Theorem 3.1** *Assume model (6) and the loss function (4). Let* $\boldsymbol{\delta}^* = (\delta_i^*)$ *be the Bayes decision rule, then*

$$\delta_i^*(\mathbf{x}) = \begin{cases} 1 & \text{if } P(\theta_i = 0 \mid \mathbf{x}) \leq \dfrac{w_i(\mathbf{x})}{\lambda} P(\theta_i = 1 \mid \mathbf{x}) \\ 0 & \text{Otherwise} \end{cases} \tag{8}$$

We prove the theorem below.

**Proof**

$$
\begin{aligned}
& E(L_{\lambda,s}(\boldsymbol{\theta},\boldsymbol{\delta}) \mid \mathbf{x}) \\
= \; & \frac{1}{m}\sum_{i=1}^{m}\{\lambda\delta_i P(\theta_i=0 \mid \mathbf{x}) + (1-\delta_i)E(s(|\mu_i|) \mid \theta_i=1,\mathbf{x})P(\theta_i=1 \mid \mathbf{x})\} \\
= \; & \frac{1}{m}\sum_{i=1}^{m}\{E(s(|\mu_i|) \mid \theta_i=1,\mathbf{x})P(\theta_i=1 \mid \mathbf{x}) \\
& +\delta_i\left[\lambda P(\theta_i=0 \mid \mathbf{x}) - E(s(|\mu_i|) \mid \theta_i=1,\mathbf{x})P(\theta_i=1 \mid \mathbf{x})\right]\}
\end{aligned}
$$

Note that the first term is a constant with respect to $\boldsymbol{\delta}$. Consequently, $\boldsymbol{\delta}^*$, given in (8), minimizes the expected loss function. ∎

# 4 The Oracle Procedure

## 4.1 Optimal Testing Procedure

In the previous section, we derive the Bayes decision rule which minimizes the expected loss function $L_{\lambda,s}$ in (4). In this section, we will show that an appropriate choice of $\lambda$ leads to a decision rule which is also optimal in multiple hypothesis testing. Before stating our result, we define two error rates, $m\text{FDR}^*(\boldsymbol{\delta})$ and $m\text{FNR}^*(\boldsymbol{\delta})$, as follows:

$$
m\text{FDR}^*(\boldsymbol{\delta}) = \frac{\sum_i E\left[\delta_i(\mathbf{X})P(\theta_i=0 \mid \mathbf{X})\right]}{\sum_i E\left[\delta_i(\mathbf{X})P(\theta=0 \mid \mathbf{X})\right] + \sum_i E\left[\delta_i(\mathbf{X})w_i(\mathbf{X})P(\theta_i=1 \mid \mathbf{X})\right]}
$$

$$
m\text{FNR}^*(\boldsymbol{\delta}) = \frac{\sum_i E\left[(1-\delta_i(\mathbf{X}))w_i(\mathbf{X})P(\theta_i=1 \mid \mathbf{X})\right]}{\sum_i E\left[(1-\delta_i(\mathbf{X}))w_i(\mathbf{X})P(\theta_i=1 \mid \mathbf{X})\right] + \sum_i E\left[(1-\delta_i(\mathbf{X}))P(\theta_i=0 \mid \mathbf{X})\right]}
$$

In both definitions, we weigh a type II error by the average severity $w_i(\mathbf{X})$.

**Theorem 4.1** *Assume the model given in (6). Let* $\boldsymbol{\delta}$ *be a testing procedure with* $\delta_i$ *defined in* (8) *and* $mFDR^*(\boldsymbol{\delta}) = \alpha$*, and let* $\boldsymbol{\delta}'$ *be any other rule such that* $mFDR^*(\boldsymbol{\delta}') \leq \alpha$*. Then* $mFNR^*(\boldsymbol{\delta}) \leq mFNR^*(\boldsymbol{\delta}')$.

This theorem indicates that the Bayes decision rule is also optimal in the multiple testing in the sense that it minimizes the mFNR* while controlling the mFDR*. When setting the severity function $s(\cdot) = 1$, the theorem implies that the procedure $\delta_i = \{\frac{P(\theta_i=0|\mathbf{X})}{P(\theta_i=1|\mathbf{X})} < \lambda\}$ is the optimal testing procedure in the sense that it minimizes the mFNR among all the procedure which controls the mFDR at a certain level, which is the procedure given in Sun and Cai [15]. We thus generalized the result of Sun and Cai [15] in two ways. First, Sun and Cai [15] has shown that their oracle procedure is optimal among the collection of decision rules based on test statistics satisfying monotone likelihood ratio property. Our result shows that such a restriction can be removed. Secondly, our result holds for arbitrary dependence structure.

According to Theorem 4.1, one knows that (8) is the optimal rule which minimizes the mFNR* at some level $\alpha(\lambda)$. The question now is how to choose the cutoff $\lambda$ such that (8) actually controls the $m$FDR* at the specified level, say 5%. We will resolve such an issue in what follows.

Define $\mathbf{T} = \{T_i(\mathbf{X}),\ i = 1, \ldots, m\}$ where

$$T_i(\mathbf{X}) = \frac{P(\theta_i = 0 \mid \mathbf{X})}{P(\theta_i = 0 \mid \mathbf{X}) + w_i(\mathbf{X})P(\theta_i = 1 \mid \mathbf{X})}.$$

Note that $T_i(\mathbf{X})$ is increasing with respect to $\dfrac{P(\theta_i = 0|\mathbf{X})}{w_i(\mathbf{X})P(\theta_i = 1|\mathbf{X})}$. Denote the *cdf* of $T_i(\mathbf{X})$ as $G_i(t) = \pi_0 G_{i,0}(t) + \pi_1 G_{i,1}(t)$, where $G_{i,0}$ is the distribution

of $T_i(\mathbf{X})$ conditioned on $\theta_i = 0$, and $G_{i,1}$ is its distribution conditioned on $\theta_i = 1$. Let $G_{i,\mu}(t)$ be the distribution of $T_i(\mathbf{X})$ conditioned on $\theta_i = 1$ and $\mu_i = \mu$. Then $G_{i,1}(t) = \int G_{i,\mu}(t)h(\mu)d\mu$. Denote the *pdf* of $T_i(\mathbf{X})$ as $g_i(t) = \pi_0 g_{i,0}(t) + \pi_1 g_{i,1}(t)$, where $g_{i,1}(t) = \int g_{i,\mu}(t)h(\mu)d\mu$. We first have the following claim which helps us define our oracle procedure.

**Claim 4.1** *Consider the decision rule* $\boldsymbol{\delta}(\mathbf{T}, c) = I(\mathbf{T} < c\mathbf{1})$. *Then the* $mFDR^*(\boldsymbol{\delta}(\mathbf{T}, c))$ *is monotonically increasing in* $c$.

**Theorem 4.2 (The Oracle Procedure)** *The oracle multiple testing procedure that controls the* $mFDR^*$ *at* $\alpha$ *and that minimizes the* $mFNR^*$ *is* $\boldsymbol{\delta}(\mathbf{T}, \mathbf{c}^*\mathbf{1}) = \{I(T_i < c^*), i = 1, \ldots, m\}$, *where*

$$c^* = \sup\left\{t : \frac{\sum \pi_0 G_{i,0}(t)}{\sum \pi_0 G_{i,0}(t) + \sum \pi_1 \int s(|\mu|) G_{i,\mu}(t) h(\mu) d\mu} \leq \alpha\right\} \tag{9}$$

A speical case of model (6) is the following:

$$\begin{aligned} \theta_i &\overset{\text{iid}}{\sim} Bernoulli(\pi_1) \\ \mu_i \mid \theta_i &\overset{\text{ind}}{\sim} (1-\theta_i)h_0 + \theta_i h(\mu_i) \\ X_i \mid \mu_i &\overset{\text{ind}}{\sim} I(\mu_i = 0) f_0(x_i) + I(\mu_i \neq 0) f(x_i; \mu_i). \end{aligned} \tag{10}$$

In this case, $T_i(\mathbf{X}) = T(X_i) = \dfrac{P(\theta_i = 0 \mid X_i)}{P(\theta_i = 0 \mid X_i) + w(X_i)P(\theta_i = 1 \mid X_i)}$. Since $(X_i, \theta_i, \mu_i)$, $i = 1, \ldots, m$, are i.i.d., the $T_i$'s have a common distribution, denoted as $G(t) = \pi_0 G_0(t) + \pi_1 G_1(t)$, where $G_1(t) = \int G_\mu(t)h(\mu)d\mu$. Let $g_0(t) = \dfrac{d}{dt}G_0(t)$, $g_1(t) = \dfrac{d}{dt}G_1(t)$ and $g_\mu(t) = \dfrac{d}{dt}G_\mu(t)$. The threshold $c^*$ of

the oracle procedure reduces to the following:

$$c^* = \sup\left\{t : \frac{\pi_0 G_0(t)}{\pi_0 G_0(t) + \pi_1 \int s(|\mu|) G_\mu(t) h(\mu) d\mu} \leq \alpha\right\} \tag{11}$$

In summary, we have derived the oracle multiple procedure (9) which is optimal in the sense that it minimizes the $m$FNR$^*$ among all procedures which controls the $m$FDR$^*$ at some pre-specified level, under arbitrary dependence structure.

## 4.2 Comparison of the Oracle Procedure with its Competitors

Choosing $s(|\mu|) = \mu^2$, we compare our oracle procedure defined in (11) with the oracle procedure in Sun and Cai [15] and the oracle $p$-value procedure in Genovese and Wasserman [10]. We calculate the acceptance regions, the mFDR*, and the mFNR* for all three approaches under the following model, which was also used in Example 1 in Section 3.2 of Sun and Cai [15].

$$Z \overset{\text{ind}}{\sim} \pi_0 N(0,1) + \pi_1 F_1 \quad \text{with} \quad F_1 = \pi_{11} N(\mu_1, 1) + \pi_{12} N(\mu_2, 1), \tag{12}$$

where $\pi_0 + \pi_1 = 1$ and $\pi_{11} + \pi_{12} = 1$. This is a special form of our model (10) with $h(\mu_1) = \pi_{11}$ and $h(\mu_2) = \pi_{12}$. As in Sun and Cai [15], we choose $\pi_0 = 0.8$, $\mu_1 = -3$, $\mu_2 = 4$ and let $\pi_{11}$ vary in $(0, 1)$.

Under this model, our oracle testing procedure $T \leq c^*$ (in (11)) corresponds to $\{z : Z \leq c_l \quad \text{or} \quad Z \geq c_u\}$. For a given $t$, solve the following equation for $z$ to obtain $c_l$ and $c_u$:

$$t\pi_1[\pi_{11}\mu_1^2 exp(\mu_1 z - \frac{1}{2}\mu_1^2) + \pi_{12}\mu_2^2 exp(\mu_2 z - \frac{1}{2}\mu_2^2)] - \pi_0(1-t) = 0$$

Then we calculate the $m$FDR* and $m$FNR* as

$$
\begin{aligned}
& mFDR^* \\
= \quad & \frac{\pi_0[\Phi(c_l) + \overline{\Phi}(c_u)]}{\pi_0[\Phi(c_l) + \overline{\Phi}(c_u)] + \pi_1\{\pi_{11}\mu_1^2[\Phi(c_l - \mu_1) + \overline{\Phi}(c_u - \mu_1)] + \pi_{12}\mu_2^2[\Phi(c_l - \mu_2) + \overline{\Phi}(c_u - \mu_2)]\}}
\end{aligned}
$$

$$
\begin{aligned}
& mFNR^* \\
= \quad & \frac{\pi_1\{\pi_{11}\mu_1^2[\Phi(c_u - \mu_1) - \Phi(c_l - \mu_1)] + \pi_{12}\mu_2^2[\Phi(c_u - \mu_2) - \Phi(c_l - \mu_2)]\}}{\pi_0[\Phi(c_u) - \Phi(c_l)] + \pi_1\{\pi_{11}\mu_1^2[\Phi(c_u - \mu_1) - \Phi(c_l - \mu_1)] + \pi_{12}\mu_2^2[\Phi(c_u - \mu_2) - \Phi(c_l - \mu_2)]\}}
\end{aligned}
$$

Repeat the process until we find $t$ such that $m$FDR* converges to $\alpha$.

For the oracle adaptive BH procedure, we calculate the $p$-values as follows:

$$P = 2\overline{\Phi}(|Z|), \tag{13}$$

where $\Phi$ is the distribution of $N(0, 1)$. Then the distribution of the $p$-values is then

$$Pr(P \leq t) = P\{|Z| \geq \Phi^{-1}(1 - t/2)\}$$

We then find the threshold $\mu^*$ according to the following:

$$\mu^* = \sup\left\{t : Pr(P \leq t) = \frac{\pi_0 t}{\alpha}\right\},$$

where $\alpha = 0.05$ and $\pi_0 = 0.8$ in our study. We then find the threshold on the $z$-value scale according to the transformation (13).

The result of this numerical study is shown in Figure 3 and Figure 4. In Figure 3, for different $\pi_{11}$ values, we plotted, at level 0.05, the acceptance

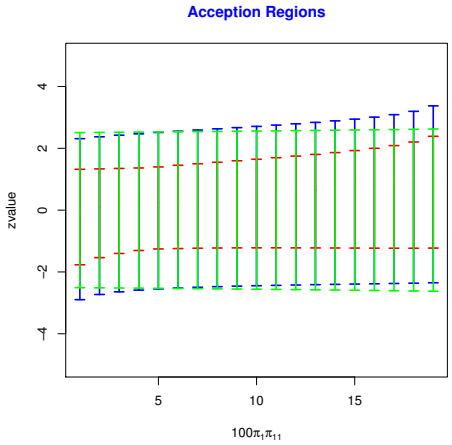


Figure 3: The acceptance regions: the oracle procedure controlling the $m$FDR at level 0.05 (blue) (Sun and Cai [15]), the oracle procedure controlling $m$FDR* at level 0.05 (red), and the oracle adaptive BH procedure at level 0.05 (green). The data are generated according to (12) with $\pi_0 = 0.8$, and $\pi_{11}$ varying from 0 to 1. The parameters $\mu_1$ and $\mu_2$ are set to be $-3$ and 4 respectively.

regions in terms of the $z$ values for our procedure controlling the $m$FDR*, for the procedure of Sun and Cai [15] controlling the $m$FDR and the oracle $p$-value procedure in Genovese and Wasserman [10]. It is clearly seen that our proposed approach has much wider rejection regions than both of the other two approaches. In Figure 4, we have further reported the values of the mFNR and mFNR* of the three approaches under the same setting. It can been seen that, for almost all values of $\pi_{11}$, Sun and Cai [15] has smaller $m$FNR and $m$FNR* value than the oracle BH procedure. Further, the proposed approach has the smallest mFNR and mFNR*, for each value of $\pi_{11}$. For instance, the ratio of the mFNR* of the proposed approach to that of the oracle procedure in Sun and Cai [15] can be as small as 0.15. It is thus demonstrated that our proposed approach is more powerful than the other two approaches.

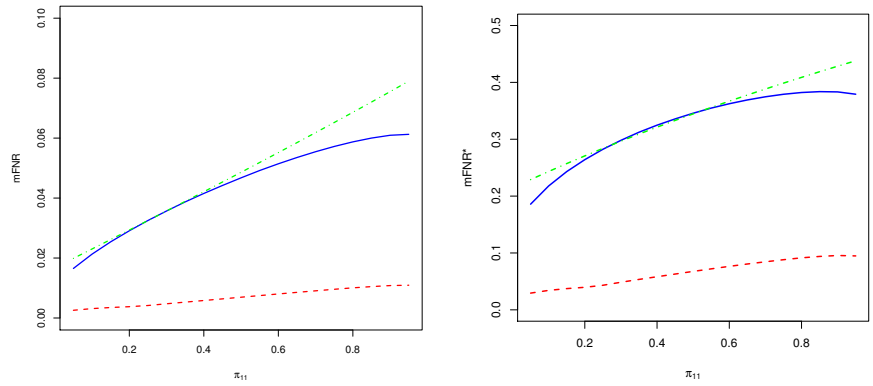


Figure 4: Comparisons of the three procedures in terms of their $m$FNR and $m$FNR*: the oracle procedure controlling the $m$FDR at level 0.05 (blue solid) (Sun and Cai [15]), the oracle procedure controlling $m$FDR* at level 0.05 (red dasehd), and the oracle adaptive BH procedure at level 0.05 (green dot-dashed). The setting is the same as described in the caption of Figure 3

# 5 The Data Driven Procedure

Note that the results in this section are derived under indepedence model (10). The oracle procedure defined in (11) depends on unknown quantities, such as $\int s(|\mu|)f(x;\mu)h(\mu)d\mu$. We need to estimate these quantities in order to make it applicable. In Section 5.3, we provide consistent estimates for such unknown quantities and derive a data driven procedure as described in Section 5.1. In Section 5.2, we show that the $m$FDR* and the $m$FNR* of our data driven procedure converge to those of the oracle procedure asymptotically.

## 5.1 The Procedure

First, the $m$FDR* of the oracle procedure $\boldsymbol{\delta}(\mathbf{T}, c^*)$, defined in (11), is the following:

$$m\text{FDR}^*(c^*) = Q^*(c^*) = \frac{\int I(T(x) < c^*)T(x)(\pi_0 f_0(x) + \pi_1 \int s(|\mu|)f(x;\mu)h(\mu)d\mu)dx}{\int I(T(x) < c^*)(\pi_0 f_0(x) + \pi_1 \int s(\mu|)f(x;\mu)h(\mu)d\mu)dx}$$

Let $q(x) = \pi_0 f_0(x) + \pi_1 \int f(x-\mu)h(\mu)d\mu$ be the marginal distribution of $x$, and let $q^*(x) = \pi_0 f_0(x) + \pi_1 \int s(|\mu|)f(x-\mu)h(\mu)d\mu$. Suppose $\hat{q}$, $\hat{q}^*$ and $\hat{T}$ are estimators of $q$, $q^*$ and $T$. Then an estimator of the $m$FDR* follows:

$$\hat{Q}^*(c) = \frac{\frac{1}{m}\sum I(\hat{T}(x_i) < c^*)\hat{T}(x_i)\frac{\hat{q}^*(x_i)}{\hat{q}(x_i)}}{\frac{1}{m}\sum I(\hat{T}(x_i) < c^*)\frac{\hat{q}^*(x_i)}{\hat{q}(x_i)}}$$

Let $\hat{c}^* = \sup\{t : \hat{Q}^*(t) \leq \alpha\}$. It is equivalent to consider only the set of the discrete thresholds in the set of estimated estimators $\{\hat{T}(x_i), 1 = 1, \ldots, m\}$. Then a data driven procedure can be given as follows.

**Definition 5.1 (The Data Driven Procedure)** *Define* $c_i = \dfrac{\hat{q}^*(x_i)}{\hat{q}(x_i)}$. *Let* $\hat{T}_{(1)}, \ldots, \hat{T}_{(m)}$ *be the ordered test statistics and* $c_{(1)}, \ldots, c_{(m)}$ *be the corresponding* $c_i$*'s. Let*

$$k = \max\left\{ j : \frac{\sum_{i=1}^{j} \hat{T}_{(i)} c_{(i)}}{\sum_{i=1}^{j} c_{(i)}} \leq \alpha \right\} \tag{14}$$

*Then reject all* $H_{(i)}, i = 1, \ldots, k$.

If we had chosen $s(\cdot) = 1$, then $q^*$ would reduce to $q$ and our oracle and adaptive procedures would reduce to the oracle and adaptive procedures derived

in Sun and Cai [15]. By choosing $s(\cdot)$ to be a nonconstant nondecreasing function of $|\mu|$, we take into account the severity of a type II error, resulting in a difference rejection set.

## 5.2 Asymptotic Properties

**Theorem 5.1 (Asymptotic Validity)** *Suppose $X_1, \ldots, X_m$ are independently distributed according to model (10). Assume both $q(x)$ and $q^*(x)$ are continuous and positive on $\mathcal{R}$. If $\hat{\pi}_0 \to_p \pi_0$, $E \parallel \hat{f}_0 - f_0 \parallel^2 \to 0$, and $E \parallel \hat{q}^* - q^* \parallel^2 \to 0$, then the mFDR* of the data driven procedure defined in (14) converges to $Q^*(c^*)$, where the $Q^*(c^*)$ is the mFDR* of the oracle procedure defined in* (11)*, as $m \to \infty$.*

**Theorem 5.2 (Asymptotic Optimality)** *Assume $X_1, \ldots, X_m$, $q$, $q^*$, $\hat{\pi}_0$, $\hat{f}_0$ and $\hat{q}^*$ satisfy the same conditions as in Theorem 5.1. Let the mFNR* of the oracle procedure defined in (11) be $\tilde{Q}^*(c^*)$. Then the mFNR* of the data driven procedure defined in (14) converges to $\tilde{Q}^*(c^*)$ as $m \to \infty$.*

## 5.3 Empirical Estimation of the Unknown Quantities

Previously, we have shown that the $m$FDR* and the $m$FNR* of the data driven procedure defined in Definition 5.1 converge respectively to those of the oracle procedure if there exist consistent estimators for the unknown quantities $q(x)$ and $q^*(x)$. Note that $q(x)$ is the marginal density and can be estimated by using the kernel density estimation. The estimation of $q^*(x)$ is more challenging because it involves $\int s(|\mu|) f(x;\mu) h(\mu) d\mu$.

Assume $X_i \sim N(0,1)$ under the null hypothesis. In applications where the test statistics follow another distribution $\Psi_0$ under null hypothesis, one

can apply the transformation $Y_i = \Phi^{-1}(\Psi_0(X_i))$ and $Y_i$ then follows a normal distribution under $H_0$. Let the severity function $s(x) = x^2$. Then $\int s(\mu)\phi(x-\mu)h(\mu)d\mu$ can be written as

$$q_1''(x) + 2xq_1'(x) + (x^2+1)q_1(x), \tag{15}$$

where $q_1(x) = \int \phi(x-\mu)h(\mu)d\mu$. It is thus sufficient to estimate $q_1(x)$ and its first and second derivatives. Let $\hat{\pi}_0$ is an estimate of $\pi_0$, and let $\hat{q}(x)$, be the kernel density estimate of $q(x)$. Similar to Brown and Greenshtein [3], we can easily get $\hat{q}'(x)$ and $\hat{q}''(x)$, estimators of the derivatives of the density. Since

$$q(x) = \pi_0 f_0(x) + \pi_1 q_1(x),$$

one can thus estimate $q_1(x)$, $q_1'(x)$, and $q_1''(x)$ by $\hat{q}_1(x) = (\hat{q}(x) - \hat{\pi}_0\phi(x))/(1-\hat{\pi}_0)$, $\hat{q}_1'(x) = (\hat{q}'(x) - \hat{\pi}_0\phi'(x))/(1-\hat{\pi}_0)$ and $\hat{q}_1''(x) = (\hat{q}''(x) - \hat{\pi}_0\phi''(x))/(1-\hat{\pi}_0)$, respectively. We further write $q_1^*(x) = \int \mu^2\phi(x-\mu)h(\mu)d\mu$. According to (15), one can further estimate $q_1^*$ and $q^*$. Specifically, we estimate $q_1^*$ as $\hat{q}_1^*(x) = \hat{q}_1''(x) + 2x\hat{q}_1'(x) + (x^2+1)\hat{q}_1(x)$ and $\hat{q}^*(x) = \hat{\pi}_0\phi(x) + (1-\hat{\pi}_0)\hat{q}_1^*(x)$. Eventually, we derive an estimate of $T(x)$ as $\hat{T} = \dfrac{\hat{\pi}_0\phi(x)}{\hat{q}^*(x)}$.

In a series of influential papers including Efron et al. [7] and Efron [8], the necessity of estimating the empirical null was emphasized, i.e., estimating $\mu_0$, and $\sigma_0^2$, the parameters of the null distribution. We can also utilize such information by replacing $\phi(x)$ by $\phi(\dfrac{x-\mu_0}{\sigma_0})$ when reliable estimators are available (See Jin and Cai [12], Cai and Jin [4], and Jin [11]).

When estimating the unknown quantities, we don't restrict ourselves to any specific forms for $h(\mu)$, the prior distribution of $\mu$ when $\mu$ is nonzero. Such a procedure can be viewed as a non-parametric empirical Bayes approach.

We derive the data-driven approach under the independence assumption. We will also show the performance of this approach under some dependence structure in our simulation study in Section 6.

# 6 Simulation Study and Real Data Analysis

## 6.1 Simulation Studies

In this subsection, we perform numerical studies to compare the performance of our data driven procedure with the data driven procedure in Sun and Cai [15] and the oracle adaptive BH procedure. When applied to a set of $p$-values, the oracle adaptive BH procedure is the step up procedure with a set of critical values $\frac{i\alpha}{m\pi_0}, i = 1, \ldots, m$. The oracle adaptive BH procedure controls the FDR at level $\alpha$, exactly under independence, and conservatively, under the positive dependence of the test statistics (See, for example, Benjamini and Yekutieli [2]).

Let the observations be $X_i \sim N(\mu_i, 1)$ and let $\theta_i$ be the indicator for $\mu_i = 0$, $i = 1, \ldots, m$. Further, conditioned on $\theta_i = 0$, we let $\mu_i = 0$; and when conditioned on $\theta_i = 1$, we let $\mu_i \sim h(\mu)$. In the first study, we choose $h(\mu) = \phi(\mu - 1)$, and we consider testing $H_i : \mu_i = 0$ against $\mu_i > 0$, $i = 1, \ldots, m$. In the second and third study, we choose $h(\mu) = \pi_{11}\phi(\mu - 2) + (1 - \pi_{11})\phi(\mu + 3)$, and we consider testing $H_i : \mu_i = 0$ against $\mu_i \neq 0$, $i = 1, \ldots, m$. Here $\phi$ is the density function of the standard normal random variable, and $\pi_{11} \in [0, 1]$. We fix the number of hypotheses $m$ to be 2000 and the number of repetitions $n$ to be 1000 in all the studies.

**Simulation Study I**

We let the proportion of true null hypotheses $\pi_0$ vary among $\{0.1, 0.2, \ldots, 0.9\}$. For each $\pi_0$, we generate $m$ i.i.d. $\theta_i$'s from the Bernoulli$(1-\pi_0)$ distribution. We then generate the mean vector $\boldsymbol{\mu}$ conditioned on the realized $\theta_i$'s; that is, if $\theta_i = 0$, then $\mu_i = 0$; if $\theta_i = 1, \mu_i$ is generated from the $N(1,1)$ distribution. The observed $X_i$'s are generated from the $N_m(\boldsymbol{\mu}, I)$ distribution, which is the multivariate normal distribution with the mean $\boldsymbol{\mu}$ and the covariance matrix the identity matrix. At level 0.05, we apply our data driven procedure, the data driven procedure in Sun and Cai [15], and the oracle adaptive BH procedure. We repeat the above process $n(= 1000)$ times, and calculate the average power, $m$FDR, $m$FNR, $m$FDR*, and $m$FNR* of the three procedures. Figure 5 shows the average power of the three procedures for the different $\pi_0$ values. Our proposed method, corresponding to the red line, is more powerful for all values of $\pi_0$. Figure 6 shows the $m$FDR, $m$FDR*, $m$FNR, and $m$FNR* of the three procedures for the different $\pi_0$ values. In the top right panel, we see that the proposed approach controls the $m$FDR$^*$ very well. In the bottom two panels, our proposed method has uniformly smaller $m$FNR and $m$FNR$^*$ than the other two methods.

**Simulation Study II**

We fix $\pi_0 = 0.5$ and let $\pi_{11}$ move in $\{0.1, 0.2, \ldots, 0.9\}$, where $\pi_{11}$ is the proportion of the alternative means which follow the $N(2,1)$ distribution. For each $\pi_{11}$ value, we generate $m$ i.i.d. $\theta_i$'s from the Bernoulli(0.5) distribution. If $\theta_i = 0$, then $\mu_i = 0$, and if $\theta_i = 1$, we generate $\gamma_i$ from the Bernoulli$(\pi_{11})$ distribution. Conditioned on $\gamma_i = 1$, we generate $\mu_i$ from the $N(2,1)$ distribution, and conditioned on $\gamma_i = 0$, $\mu_i$ is generated according to the $N(-3,1)$ distribution. The observed $X_i$'s are again generated from the $N_m(\boldsymbol{\mu}, I)$ distribution, conditioned on the realized mean vector $\boldsymbol{\mu}$. Figures 7

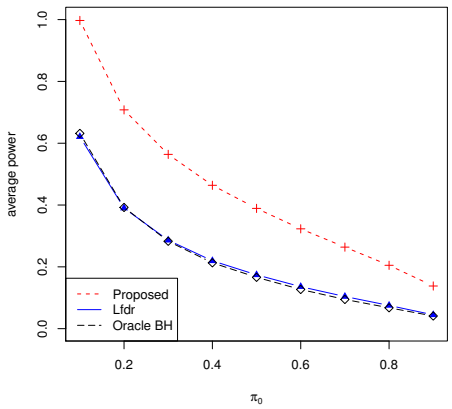


Figure 5: Comparison of the Data Driven Procedures in terms of Average Power: data driven procedure controlling $m$FDR at level 0.05 (blue solid) (Sun and Cai [15]), the proposed data driven procedure controlling $m$FDR* at level 0.05 (red dashed), and the oracle adaptive BH procedure (long-dashed black).

and 8 illustrate the results of the this simulation study. Figure 7 shows the average power of the three procedures for the different $\pi_{11}$ values. We see that our procedure is more powerful for all values of $\pi_{11}$, and the procedure in Sun and Cai [15] dominates the oracle adaptive BH procedure. Figure 8 shows the $m$FDR, $m$FDR*, $m$FNR, and $m$FNR* of the three procedures for the different $\pi_{11}$ values. We see that our procedure has the $m$FDR* controlled at level 0.05 and has the lowest $m$FNR and $m$FNR*. Also the procedure in [15] has better performance than the oracle BH procedure.

**Simulation Study III**

To see how our procedure behave under dependence compared with the other two procedures, we perfom the third simulation study. The setting is the same as in Study II except we now fix $\pi_{11} = 0.1$ and, conditioned on the realized mean vector $\boldsymbol{\mu}$, the observed $X_i$'s are generated from the $N_m(\boldsymbol{\mu}, \Sigma)$

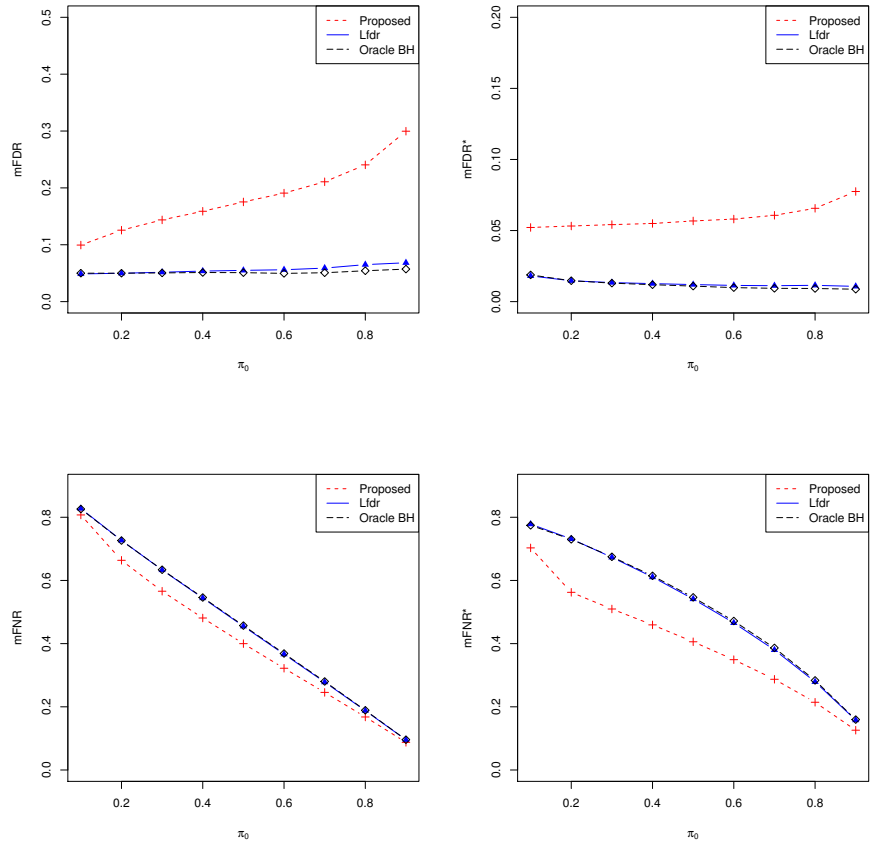


Figure 6: Comparison of the Data Driven Procedures: data driven procedure controlling $m$FDR at level 0.05 (blue solid) (Sun and Cai, 2007), the proposed data driven procedure controlling $m$FDR* at level 0.05 (red dashed), and the oracle adaptive BH procedure (long-dashed black).

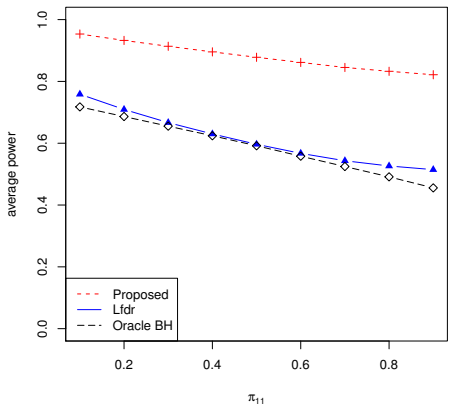


Figure 7: Comparison of the Data Driven Procedures in terms of Average Power: data driven procedure controlling the $m$FDR at level 0.05 (blue solid) (Sun and Cai [15]), the proposed data driven procedure controlling the $m$FDR* at level 0.05 (red dashed), and the oracle adaptive BH procedure (long-dashed black).

distribution. Here $\Sigma = \rho JJ' + (1-\rho)I$, and $J$ is the vector of 1's. Hence the $X_i$'s are multivariate normal with common correlation $\rho$. We let $\rho$ move in $(0.1, \ldots, 0.9)$. Figures 9 and 10 illustrate the results of the this simulation study. Figure 9 shows the average power of the three procedures for the different $\rho$ values. We see that our procedure is more powerful for all values of $\rho$, and the procedure in Sun and Cai [15] dominates the oracle adaptive BH procedure. Figure 10 shows the $m$FDR, $m$FDR*, $m$FNR, and $m$FNR* of the three procedures for the different $\rho$ values. We see that our procedure has the $m$FDR* controlled at level 0.05 and has the lowest $m$FNR and $m$FNR*. Also the procedure in Sun and Cai [15] has better performance than the oracle BH procedure.

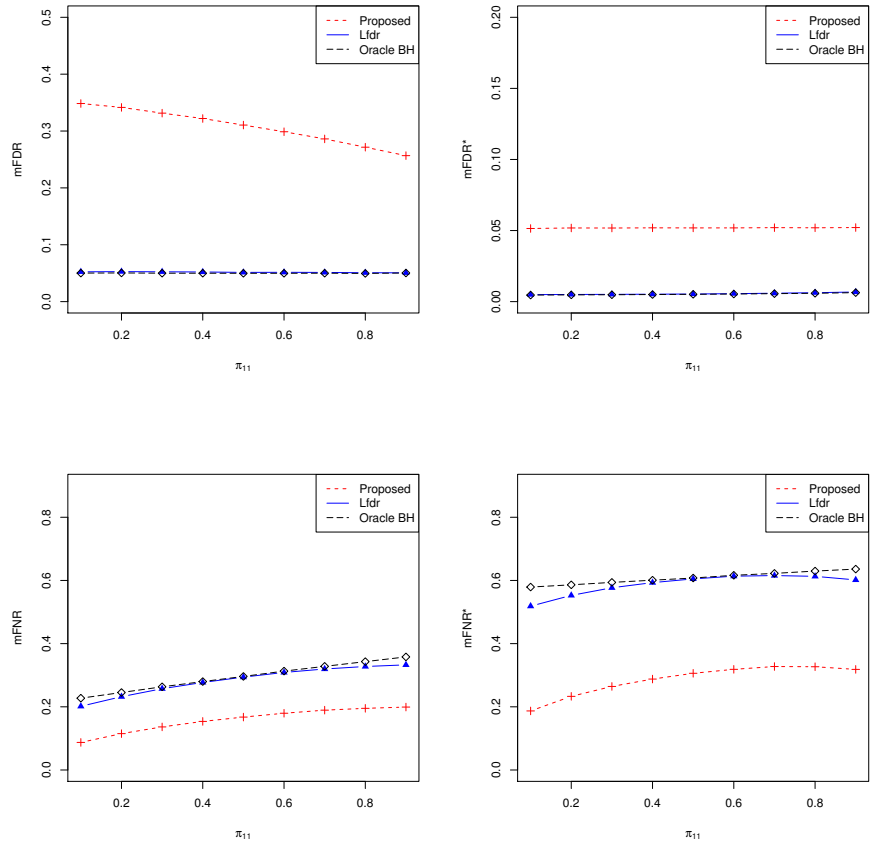


Figure 8: Comparison of the Data Driven Procedures: data driven procedure controlling the $m$FDR at level 0.05 (blue solid) (Sun and Cai [15]), the proposed data driven procedure controlling the $m$FDR* at level 0.05 (red dashed), and the oracle adaptive BH procedure (long-dashed black).

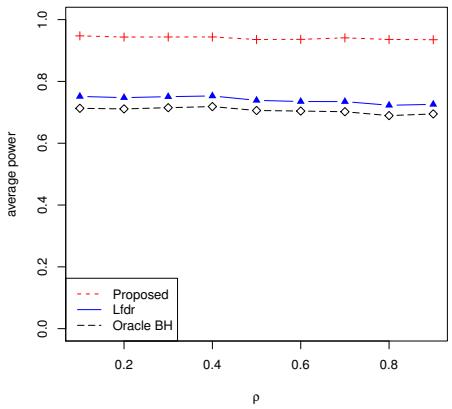


Figure 9: Comparison of the Data Driven Procedures in terms of Average Power: data driven procedure controlling the $m$FDR at level 0.05 (blue solid) (Sun and Cai [15]), the proposed data driven procedure controlling the $m$FDR* at level 0.05 (red dashed), and the oracle adaptive BH procedure (long-dashed black).

## 6.2 Applications to Real Data

### HIV Data

We apply our proposed data driven procedure to the HIV data considered in Sun and Cai [15] and Efron [6]. The HIV data compares 4 HIV positive patients versus 4 negative controls which provide the microarray of expression levels on $m = 7680$ genes (Van't Wout et al. [16]). The two-sample $t$-statistic for each gene, which compares the expression levels of positive and negative HIV patients, is transformed to the $z$-value using the $G_0$ with 6 degrees of freedom. At level 0.05, we applied our adaptive procedure controlling the $m$FDR*, the adaptive procedure in Sun and Cai [15] controlling the mFDR and the adaptive BH procedure. When applied to a set of $p$-values, the adaptive BH procedure is the step up procedure with a set of critical values

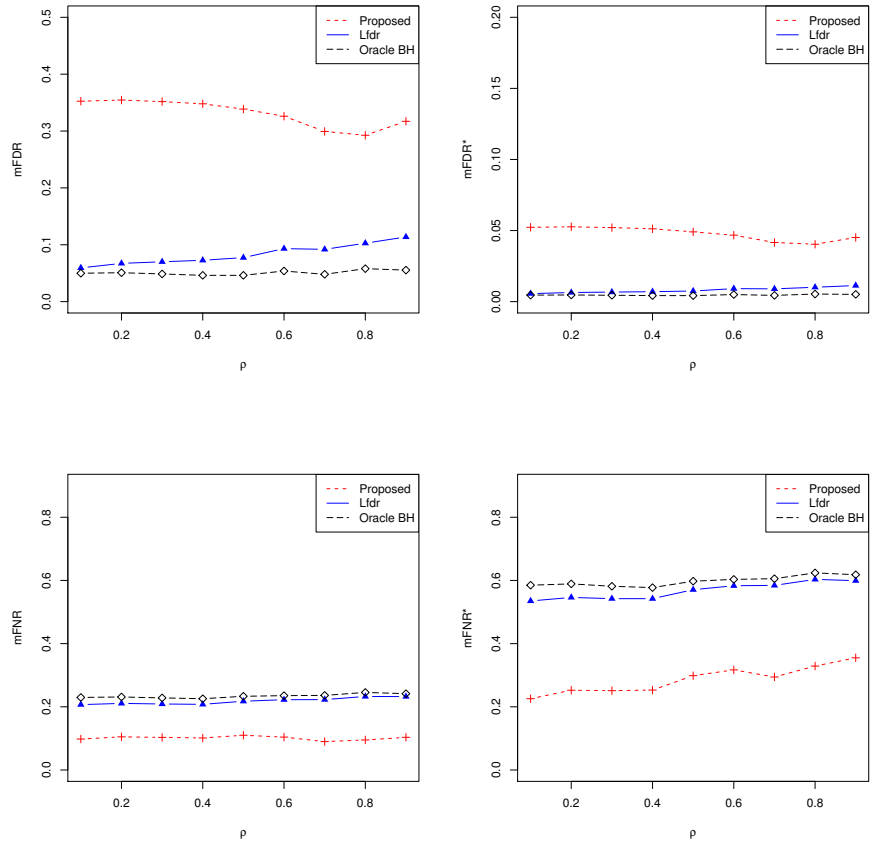


Figure 10: Comparison of the Data Driven Procedures: data driven procedure controlling the $m$FDR at level 0.05 (blue solid) (Sun and Cai [15]), the proposed data driven procedure controlling the $m$FDR* at level 0.05 (red dashed), and the oracle adaptive BH procedure (long-dashed black).

$\frac{i\alpha}{m\hat{\pi}_0}, i = 1, \ldots, m$, where $\hat{\pi}_0$ is an estimator of the proportion of true nulls $\pi_0$. We use the estimator $\pi_0$ given in Sun and Cai [15]. Using our data driven procedure to control the $m$FDR*, we found 73 significant genes, whereas 21 significant genes were found by using the method of Sun and Cai [15] to control the $m$FDR and 18 significant genes were found by the adaptive BH procedure. We also calculated local fdr statatistics based on the "empirical" null distribution for the $z$-values using the "locfdr" package in R (See Efron [6]) and found that there are 49 genes whose local fdr values are less than or equal to 0.02.

**Astromonical Data**

We also consider an astronomical data set. Our data set is obtained from the Palomar Transient Factory (PTF), which is a fully-automated and wide-field survey aimed at a systematic exploration of the optical transient sky (http://www.astro.caltech.edu/ptf/). Images of the sky were taken, where each image is $2048 \times 4096$ pixels, with 1.1 arcseconds per pixel giving an image about $1 \times 0.5$ degrees. The images were then processed to obtain the data set through some suitable transformation, where the high level of brightness of the image typically corresponds to a large data value. Our aim is to detect transient flashes of light in the sky, which may help explain various phenomina in the universe. Since we are not interested in detecting the galaxies such as bright small stars, we look at small subsets containing only 'noise' in the data matrix. Our data set is one of these subsets. We first standardized our data set so that the values in the data set have mean 0 and standard deviation 1. Altogether we have 16900 values, each of which corresponds to a z-value for a one sided test "$\mu = 0$" vs "$\mu > 0$". We then applied our data driven procedure, the data driven procedure of Sun and Cai

[15], and the adaptive BH method to the standardized data. At level 0.1, we found that our method, the method of Sun and Cai [15] and adaptive BH method detect 6, 3 and 2 significant hypotheses respectively. We also calculated local fdr statatistics based on the "empirical" null distribution for the $z$-values using the "locfdr" package in R (See Efron [6]) and found that there is 1 hypothesis whose local fdr value is less than or equal to 0.02.

# 7 Concluding Remarks

In this paper, we propose a loss function which is more elaborate and general, and possibly more reasonble, than the usual average weighted 0-1 loss function in multiple testing literature. The proposed loss function reflects the fact that the mistake of accepting a false null hypothesis is often getting more serious as the alternative moves away from the null hypothesis and not serious if the alternative value is close to the null value of the tested parameter. We reflect this fact by incorporating a severity function $s(\cdot)$ for the type II errors in the loss function. When conducting statistical inference in the form of decision theoretical analysis, the choice of the loss function is closedly related to the power of the tests. The multiple testing procedures derived based on the proposed loss function incorporate the distribution of the signal strength and hence has appealing features. Both our simulation studies and the real data analysis show that this new approach tends to be more powerful than all the alternatives considered, including the procedure of Sun and Cai [15] and the adaptive BH procedure.

# 8 Appendix

**Proof of Theorem 4.1:** According to (8), we have

$$\sum E[(\delta_i(\mathbf{X}) - \delta_i'(\mathbf{X}))(P(\theta_i = 0 \mid \mathbf{X}) - \frac{w_i(\mathbf{X})}{\lambda} P(\theta_i = 1 \mid \mathbf{X}))] \leq 0 \quad (16)$$

Also, by assumptions,

$$\sum E[(\delta_i(\mathbf{X}) - \delta_i'(\mathbf{X}))(P(\theta_i = 0 \mid \mathbf{X}) - \frac{\alpha}{1-\alpha} w_i(\mathbf{X}) P(\theta_i = 1 \mid \mathbf{X}))] \geq 0 \quad (17)$$

From (16) and (17), we have the following:

$$\sum E\left[(\delta_i(\mathbf{X}) - \delta_i'(\mathbf{X})) w_i(\mathbf{X}) P(\theta_i = 1 \mid \mathbf{X}) \left(\frac{1}{\lambda} - \frac{\alpha}{1-\alpha}\right)\right] \geq 0 \quad (18)$$

Since we know that $\dfrac{\sum E[\delta_i(\mathbf{X}) P(\theta_i = 0 \mid \mathbf{X})]}{\sum E[\delta_i(\mathbf{X}) w_i(\mathbf{X}) P(\theta_i = 1 \mid \mathbf{X})]} = \dfrac{\alpha}{1-\alpha}$ which gives $\dfrac{\alpha}{1-\alpha} \leq \dfrac{1}{\lambda}$, we have

$$\sum E[\delta_i(\mathbf{X}) w_i(\mathbf{X}) P(\theta_i = 1 \mid \mathbf{X})] \geq \sum E[\delta'(\mathbf{X}) w_i(\mathbf{X}) P(\theta_i = 1 \mid \mathbf{X})]$$

Then we have

$$E\left[\sum \left(\frac{1 - \delta_i'(\mathbf{X})}{E[(1 - \delta'(\mathbf{X})) w_i(\mathbf{X}) P(\theta_i = 1 \mid \mathbf{X})]} - \frac{1 - \delta_i(\mathbf{X})}{E[(1 - \delta(\mathbf{X})) w_i(\mathbf{X}) P(\theta_i = 1 \mid \mathbf{X})]}\right) (P(\theta_i = 0 \mid \mathbf{X}) - \frac{w_i(\mathbf{X})}{\lambda} P(\theta_i = 1 \mid \mathbf{X}))\right] \geq 0$$

This implies that $m\text{FNR}^*(\boldsymbol{\delta}) \leq m\text{FNR}^*(\boldsymbol{\delta}')$, as claimed. ∎

**Proof of Claim 4.1:** First we note the risk of $\boldsymbol{\delta}(\mathbf{T}, c)$ follows:

$$\frac{1}{m}\sum\{\lambda\pi_0 G_{i,0}(c) + \pi_1 \int s(|\mu|)(1 - G_{i,\mu}(c))h(\mu)d\mu\}.$$

The optimal cutoff $c_0$ that minimizes this risk must satisfy

$$\frac{\sum\int s(|\mu|)g_{i,\mu}(c_0)h(\mu)d\mu}{\sum g_{i,0}(c_0)} = \frac{\pi_0}{\pi_1}\lambda$$

By Theorem 3.1, we know that $c_0 = \dfrac{1}{1+\lambda}$. Then we have,

$$\frac{\sum\int s(|\mu|)g_{i,\mu}(c_0)h(\mu)d\mu}{\sum g_{i,0}(c_0)} = \frac{\pi_0}{\pi_1}\left(\frac{1}{c_0} - 1\right) \tag{19}$$

Hence the ratio $\dfrac{\sum\int s(|\mu|)g_{i,\mu}(c_0)h(\mu)d\mu}{\sum g_{i,0}(c_0)}$ is decreasing in $c_0$ for the test statistics $\mathbf{T}$.

Next we show that the following is true

$$\frac{\sum\int_{-\infty}^{c}\int s(|\mu|)g_{i,\mu}(t)h(\mu)d\mu dt}{\sum G_{i,0}(c)} > \frac{\sum\int s(|\mu|)g_{i,\mu}(c)h(\mu)d\mu}{\sum g_{i,0}(c)} \tag{20}$$

To see that (20) is true, we note the following:

$$\begin{aligned}
\frac{\sum\int_{-\infty}^{c}\int s(|\mu|)g_{i,\mu}(t)h(\mu)d\mu dt}{\sum G_{i,0}(c)} &= \frac{\sum\int_{-\infty}^{c}\sum g_{i,0}(t)\dfrac{\int s(|\mu|)g_{i,\mu}(t)h(\mu)d\mu}{\sum g_{i,0}(t)}dt}{\sum G_{i,0}(c)} \\
&> \frac{\sum\int s(|\mu|)g_{i,\mu}(c)h(\mu)d\mu}{\sum g_{i,0}(c)}\frac{\sum\int_{-\infty}^{c} g_{i,0}(t)}{\sum G_{i,0}(c)} \\
&= \frac{\sum\int s(|\mu|)g_{i,\mu}(c)h(\mu)d\mu}{\sum g_{i,0}(c)}
\end{aligned}$$

Also, we note that $m\text{FDR}^*(c) = \dfrac{H(c)}{1+H(c)}$, where $H(c) = \dfrac{\sum G_{i,0}(c)}{\sum\int_{-\infty}^{c}\int s(|\mu|)g_{i,\mu}(t)h(\mu)d\mu dt}$.

Then we have

$$H'(c) = \frac{\sum g_{i,0}(c) \sum \int_{-\infty}^{c} \int s(|\mu|) g_{i,\mu}(t) h(\mu) d\mu dt - \sum G_{i,0}(c) \sum \int s(|\mu|) g_{i,\mu}(c) h(\mu) d\mu}{(\sum \int_{-\infty}^{c} \int s(|\mu|) g^{\mu}(t) h(\mu) d\mu dt)^2}$$

By (20), $H'(c) > 0$. Hence $m$FDR*(c) is monotone increasing in $c$.

**Lemma 8.1** *Assume* $q(x) = \pi_0 f_0(x) + \pi_1 \int f(x-\mu) h(\mu) d\mu$ *is positive and continuous on* $\mathcal{R}$ *and* $q^*(x) = \pi_0 f_0(x) + \pi_1 \int \mu^2 f(x-\mu) h(\mu) d\mu$ *is continuous on* $\mathcal{R}$. *Let* $T = \frac{\pi_0 f_0}{q^*}$ *and* $\hat{T} = \frac{\hat{\pi}_0 \hat{f}_0}{\hat{q}^*}$. *If* $\hat{\pi}_0 \to_p \pi_0$, $E \parallel \hat{f} - f \parallel^2 \to 0$ *and* $E \parallel \hat{q}^* - q^* \parallel^2 \to 0$, *then* $E \parallel \hat{T} - T \parallel^2 \to 0$.

**Proof**

(1). First it can be shown that $f_0, \hat{f}_0, q^*, \hat{q}^*$ are all bounded except for an event with small probability.

Here $q$ is the marginal density function of $X$ and $q$ is continuous and positive. Hence $P(|X| \geq K) \to 0$, as $K \to \infty$. Choose $K_1$ large and let $I = [-K_1, K_1]$. Let $l = \inf_{x \in I} q^*(x)$ and $u = \sup_{x \in I} q^*(x)$. Let $A_\epsilon^1 = \{x : |q^* - \hat{q}^*| \geq l/2\}$. Then $(l/2)^2 P(A_\epsilon^1) \leq E \parallel q^* - \hat{q}^* \parallel^2$. By the assumption that $E \parallel \hat{q}^* - q^* \parallel^2 \to 0$, we have $P(A_\epsilon^1) \to 0$, as the number of hypotheses $m \to \infty$. Note that $l/2 \leq \hat{q}^* \leq l/2 + u$ for $x \notin A_\epsilon^1$. Similarly, $f, \hat{f}$ are bounded except in a set with small probability. Let $A_\epsilon$ be the set such that $q^*, q, f_0, \hat{f}_0$ are all bounded for $x \in A_\epsilon^c$. Since we also have $E \parallel \hat{f}_0 - f_0 \parallel^2 \to 0$, then $P(A_\epsilon) \to 0$, as $m \to \infty$.

(2). Note $T - \hat{T} = \frac{\hat{f}_0 q^*(\pi_0 - \hat{\pi}_0) + (1-\pi_0) q^*(\hat{f}_0 - f_0) + (1-p) f_0 (q^* - \hat{q}^*)}{\hat{q}^* q^*}$.
We also know that $f_0, \hat{f}_0, q^*, \hat{q}^*$ are all bounded in $A_\epsilon^c$. This implies that

$$(T - \hat{T})^2 \leq c_1(\pi_0 - \hat{\pi}_0)^2 + c_2(\hat{f}_0 - f_0)^2 + c_3(q^* - \hat{q}^*)^2$$

in $A_\epsilon^c$. Also, it can be seen that $E \parallel T - \hat{T} \parallel^2$ is bounded above, say, by $L$. Then

$$E \parallel T - \hat{T} \parallel^2 \leq LP(A_\epsilon) + (c_1 E(\pi_0 - \hat{\pi}_0)^2 + c_2 E \parallel \hat{f}_0 - f_0 \parallel^2 + c_3 E \parallel q^* - \hat{q}^* \parallel^2) P(A_\epsilon^c)$$

Since $\hat{\pi}_0 \rightarrow_p \pi_0$ implies $E(\hat{\pi}_0 - \pi_0)^2 \rightarrow 0$, then by the assumption that $E \parallel \hat{f}_0 - f_0 \parallel^2 \rightarrow 0$ and $E \parallel \hat{q}^* - q^* \parallel^2 \rightarrow 0$ , we have $E \parallel T - \hat{T} \parallel^2 \longrightarrow 0$.

**We assume the assumptions in Lemma 8.1 hold in Lemma 8.2 to Lemma 8.6**

**Lemma 8.2** *$E \parallel \hat{T} - T \parallel^2 \rightarrow 0$ implies that $\hat{T}(X) \rightarrow_p T(X)$. Similarly, $E \parallel \hat{q} - q \parallel^2 \rightarrow 0$ implies that $\hat{q}(X) \rightarrow_p q(X)$, and $E \parallel \hat{q}^* - q^* \parallel^2 \rightarrow 0$ implies that $\hat{q}^*(X) \rightarrow_p q^*(X)$.*

**Proof**

$\epsilon^2 P(|\hat{T} - T| \geq \epsilon) \leq E \parallel T - \hat{T} \parallel^2$. Since $\parallel T - \hat{T} \parallel^2 \rightarrow 0$, we have $P(|\hat{T} - T| \geq \epsilon) \rightarrow 0$. The rest of the proof follows.

**Lemma 8.3** *Suppose $\hat{T}(X) \rightarrow_p T(X)$, $\hat{q}(X) \rightarrow_p q(X)$, and $\hat{q}^*(X) \rightarrow_p q^*(X)$. Then we have the following:*

(1) $E\left(I(\hat{T}(X) < c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)}\right) \rightarrow E\left(I(T(X) < c)T(X)\frac{q^*(X)}{q(X)}\right)$; *and*

(2) $E\left(I(\hat{T}(X) < c)\frac{\hat{q}^*(X)}{\hat{q}(X) \vee d}\right) \rightarrow E\left(I(T(X) < c)\frac{q^*(X)}{q(X) \vee d}\right), \forall d > 0.$

**Proof**

(1). First we note the following: $\hat{T}(X) \rightarrow_p T(X)$ implies that $I(\hat{T}(X) < c) \rightarrow_p I(T(X) < c)$; $\hat{q}(X) \rightarrow_p q(X)$ implies that $\frac{1}{\hat{q}(X)} \rightarrow_p \frac{1}{q(X)}$, since

$q(\cdot) > 0$; Also, $\hat{q}^*(X) \to_p q^*(X)$. Hence we have

$$I(\hat{T}(X) < c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)} \to_p I(T(X) < c)T(X)\frac{q^*(X)}{q(X)};$$

which implies that

$$I(\hat{T}(X) < c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)} \quad \to_d \quad I(T(X) < c)T(X)\frac{q^*(X)}{q(X)}.$$

Also note that $I(\hat{T}(X) < c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)} = I(\hat{T}(X) < c)\frac{\hat{\pi}_0\hat{f}(X)}{\hat{q}(X)}$, which is bounded by 1. Then by the Portmanteau theorem,

$$E\left(I(\hat{T}(X) < c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)}\right) \quad \to \quad E\left(I(T(X) < c)T(X)\frac{q^*(X)}{q(X)}\right) \quad (21)$$

(2). $\hat{q}(X) \to_p q(X)$ implies that $\hat{q}(X) \vee d \to_p q(X) \vee d$, for any $d > 0$. Then we have

$$I(\hat{T}(X) < c)\frac{\hat{q}^*(X)}{\hat{q}(X) \vee d} \quad \to_p \quad I(T(X) < c)\frac{q^*(X)}{q(X) \vee d}, \forall d.$$

Since $E||\hat{q}^* - q^*||^2 \to 0$, we have $E||\hat{q}^*||^2 \leq E||q^*||^2 + C$ where $C$ is some constant for sufficiently large $m$. Further, one knows that $P(|\hat{q}^* - q^*| > \epsilon) \leq \frac{1}{\epsilon^2}E||\hat{q}^* - q^*||^2$. Let $A_{n,\epsilon} = \{|\hat{q}^* - q^*| > \epsilon\}$, then

$$EI(\hat{T}(x) < c)\frac{\hat{q}^*(x)}{\hat{q}(x) \vee d} \leq E1_{A_{n,\epsilon}}I(\hat{T}(x) < c)\frac{q^*(x) + \epsilon}{\hat{q}(x) \vee d} + E1_{A_{n,\epsilon}}\frac{1}{d}\hat{q}^*(x),$$

with

$$E1_{A_{n,\epsilon}}\frac{1}{d}\hat{q}^*(x) \leq \frac{1}{d}E1_{A_{n,\epsilon}}E|\hat{q}^*(x)|^2 \leq \frac{1}{d}(E|q^*(x)|^2 + C)E1_{A_{n,\epsilon}},$$

which goes to zero as $m \to \infty$.

Further, $E1_{A^c_{n,\epsilon}} I(\hat{T}(x) < c)\frac{q^*(x)+\epsilon}{\hat{q}(x)\vee d} \to EI(T(x) < c)\frac{q^*(x)+\epsilon}{q(x)\vee d}$ because the function $1_{A^c_{n,\epsilon}}\frac{q^*(x)+\epsilon}{\hat{q}(x)\vee d}$ is bounded. Let $\epsilon \to 0$, then we know that

$$\limsup EI(\hat{T}(x) < c)\frac{\hat{q}^*(x)}{\hat{q}(x)\vee d} \le EI(T(x) < c)\frac{q^*(x)}{q(x)\vee d}.$$

One can similarly show that

$$\liminf EI(\hat{T}(x) < c)\frac{\hat{q}^*(x)}{\hat{q}(x)\vee d} \ge EI(T(x) < c)\frac{q^*(x)}{q(x)\vee d}.$$

Consequently,

$$E\left(I(\hat{T}(X) < c)\frac{\hat{q}^*(X)}{\hat{q}(X)\vee d}\right) \;\to\; E\left(I(T(X) < c)\frac{q^*(X)}{q(X)\vee d}\right), \forall d. \quad (22)$$

**Lemma 8.4** *Write* $G^{0*}(c) = E\left(I(T(X) < c)T(X)\frac{q^*(X)}{q(X)}\right)$, $G^*_d(c) = E\left(I(T(X) < c)\frac{q^*(X)}{q(X)\vee d}\right)$, *and* $Q^*_d(c) = \frac{G^{0*}(c)}{G^*_d(c)}$. *Let* $\hat{G}^{0*}(c) = \frac{1}{m}\sum I(\hat{T}(X_i) < c)\hat{T}(X_i)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)}$, $\hat{G}^*_d(c) = \frac{1}{m}\sum I(\hat{T}(X_i) < c)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)\vee d}$, *and* $\hat{Q}^*_d(c) = \frac{\hat{G}^{0*}(c)}{\hat{G}^*_d(c)}$. *Then* $\hat{Q}^*_d(c) \to_p Q^*_d(c)$.

**Proof**

(1). Let $s_m = \sum_{i=1}^{m} I(\hat{T}(X_i) < c)\hat{T}(X_i)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)}$. First note that the following is true.

$$\begin{aligned} & I(\hat{T}(X_i) < c)\hat{T}(X_i)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)} I(\hat{T}(X_j) < c)\hat{T}(X_j)\frac{\hat{q}^*(X_j)}{\hat{q}(X_j)} \\ \to_d \;& I(T(X_i) < c)T(X_i)\frac{q^*(X_i)}{q(X_i)} I(T(X_j) < c)T(X_j)\frac{q^*(X_j)}{q(X_j)} \end{aligned}$$

Also, we know that $I(\hat{T}(X_i) < c)\hat{T}(X_i)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)} I(\hat{T}(X_j) < c)\hat{T}(X_j)\frac{\hat{q}^*(X_j)}{\hat{q}(X_j)}$ is

bounded. Hence, we have

$$\begin{aligned} & E\left(I(\hat{T}(X_i)<c)T(X_i)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)}I(\hat{T}(X_j)<c)T(X_j)\frac{\hat{q}^*(X_j)}{\hat{q}(X_j)}\right) \\ \to \quad & E\left(I(T(X_i)<c)T(X_i)\frac{q^*(X_i)}{q(X_i)}I(T(X_j)<c)T(X_j)\frac{q^*(X_j)}{q(X_j)}\right) \quad \text{and} \end{aligned}$$

$$\begin{aligned} \rho_m &= cov\left(I(\hat{T}(X_i)<c)\hat{T}(X_i)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)}, I(\hat{T}(X_j)<c)\hat{T}(X_j)\frac{\hat{q}^*(X_j)}{\hat{q}(X_j)}\right) \\ &\to cov\left(I(T(X_i)<c)T(X_i)\frac{q^*(X_i)}{q(X_i)}, I(T(X_j)<c)T(X_i)\frac{q^*(X_j)}{q(X_j)}\right)=0 \end{aligned}$$

(2). Also, we have the following:

$$\begin{aligned} Var\left(I(\hat{T}(X)<c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)}\right) &\le E\left(I(\hat{T}(X)<c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)}\right)^2, \quad \text{and} \\ E\left(I(\hat{T}(X)<c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)}\right)^2 &\to E\left(I(T(X)<c)T(X)\frac{q^*(X)}{q(X)}\right)^2<\infty. \end{aligned}$$

Hence $var(s_m/m)\to 0$.

(3). By the weak law of large numbers, $\frac{1}{m}\sum I(\hat{T}(X)<c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)} \to_p E\left(I(\hat{T}(X)<c)\hat{T}(X)\frac{\hat{q}^*(X)}{\hat{q}(X)}\right)$.

(4). By (21),

$$\frac{1}{m}\sum I(\hat{T}(X_i)<c)\hat{T}(X_i)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)} \to_p G^{0*}(c).$$

Similarly,

$$\frac{1}{m}\sum I(\hat{T}(X_i)<c)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)\vee d} \to_p G_d^*(c).$$

Hence $\hat{Q}_d^*(c)=\dfrac{\hat{G}^{0*}(c)}{\hat{G}_d^*(c)} \to_p Q_d^*(c)$.

**Lemma 8.5** *Let $G^{0*}(c)$, $\hat{G}^{0*}(c)$, $\hat{Q}_d^*(c)$, and $Q_d^*(c)$ be defined as in Lemma 8.4. Define $G^*(c) = E\left(I(T(X) < c)\frac{q^*(X)}{q(X)}\right)$, $Q^*(c) = \frac{G^{0*}(c)}{G_d^*(c)}$, $\hat{G}^*(c) = \frac{1}{m}\sum I(\hat{T}(X_i) < c)\frac{\hat{q}^*(X_i)}{\hat{q}(X_i)}$, and $\hat{Q}^*(c) = \frac{\hat{G}^{0*}(c)}{\hat{G}^*(c)}$. Then $\hat{Q}_d^*(c) \to_p Q_d^*(c)$ implies $\hat{Q}^*(c) \to_p Q^*(c)$.*

**Proof**

First note that we have the following:

$$\begin{aligned} & |\hat{Q}^*(c) - Q^*(c)| \leq |\hat{Q}^*(c) - \hat{Q}_d^*(c)| + |\hat{Q}_d^*(c) - Q_d^*(c)| + |Q_d^*(c) - Q^*(c)| \\ = \;& \hat{Q}_d^*(c) - \hat{Q}^*(c) + |\hat{Q}_d^*(c) - Q_d^*(c)| + |Q_d^*(c) - Q^*(c)|. \end{aligned}$$

Since $Q_d^*(c)$ is an increasing function with respect to $d$, then $Q_d^*(c) \to Q^*(c)$ as $d \to 0$. Also we assume that $\hat{Q}_d^*(c) \to_p Q_d^*(c)$. Then for any $\epsilon_1, \epsilon_2 > 0$, there exists $d_0 > 0$ such that $\forall d < d_0$,

$$0 < Q_d^*(c) - Q^*(c) < \epsilon_1.$$

and

$$P(|\hat{Q}_d^*(c) - Q_d^*(c)| < \epsilon_1) > 1 - \epsilon_2.$$

Further, for any $\omega$ from the underlying probability space, $\hat{Q}_d^*(c)$ is increasing with respect to $d$. Consequently,

$$\lim_{d \to 0} \hat{Q}_d^*(c) = \hat{Q}^*(c).$$

In other words, $\hat{Q}_d^*(c) \to \hat{Q}^*(c)$ pointwise, which implies that $\hat{Q}_d^*(c) \to_p$

$\hat{Q}^*(c)$. Consequently, for any $\epsilon_1$, and $\epsilon_2$, there exists $d_1$, such that $\forall d < d_1$,

$$P(|\hat{Q}_d^*(c) - \hat{Q}^*(c)| < \epsilon_1) > 1 - \epsilon_2.$$

All together, we know that for any $\epsilon_1, \epsilon_2 > 0$, there exists a $D = \min(d_0, d_1)$, such that $\forall d < D$,

$$\begin{aligned} & P(|\hat{Q}^*(c) - Q^*(c)| < 3\epsilon_1) \\ > \; & P(|\hat{Q}_d^*(c) - \hat{Q}^*(c)| < \epsilon_1, |\hat{Q}_d^*(c) - Q_d^*(c)| < \epsilon_1) \\ \geq \; & 1 - 2\epsilon_2. \end{aligned}$$

We therefore conclude that $\hat{Q}^*(c) \rightarrow_p Q^*(c)$.

**Lemma 8.6** *Assume that $\hat{Q}^*(c)$ and $Q^*(c)$ are as defined in Lemma 8.5. Let $\hat{\lambda} = \sup\{c \in (0,1) : \hat{Q}^*(c) \leq \alpha\}$ and $\lambda = \sup\{c \in (0,1) : Q^*(c) \leq \alpha\}$. If $\hat{Q}^*(c) \rightarrow_p Q^*(c)$, then $\hat{\lambda} \rightarrow_p \lambda$.*

**Proof**

(1). Let $T_{(i)}, i = 1, \ldots, m$ be the ordered values of $T(X_i), i = 1, \ldots, m$. Note that $\hat{Q}^*(c)$ is not continuous. Construct two functions as below. For $T_{(k)} < c < T_{(k+1)}$, define

$$\begin{aligned} \hat{Q}^{*-}(c) &= \hat{Q}^*(T_{(k)}) + (\hat{Q}^*(T_{(k)}) - \hat{Q}^*(T_{(k-1)}))\frac{T_{(k)} - c}{T_{(k+1)} - T_{(k)}} \quad \text{and} \\ \hat{Q}^{*+}(c) &= \hat{Q}^*(T_{(k)}) + (\hat{Q}^*(T_{(k+1)}) - \hat{Q}^*(T_{(k)}))\frac{T_{(k)} - c}{T_{(k+1)} - T_{(k)}} \end{aligned}$$

(2). Let $c_i = \dfrac{\hat{q}^*(x_i)}{\hat{q}(x_i)}$ and let $c_{(i)}$ be the value corresponding to $T_{(i)}$. Then

$$\begin{aligned}
\hat{Q}^*(T_{(k+1)}) - \hat{Q}^*(T_{(k)}) &= \frac{\sum_{i=1}^{k+1} c_{(i)} T_{(i)}}{\sum_{i=1}^{k+1} c_{(i)}} - \frac{\sum_{i=1}^{k} c_{(i)} T_{(i)}}{\sum_{i=1}^{k} c_{(i)}} \\
&= \frac{c_{(k+1)} T_{(k+1)}}{\sum_{i=1}^{k+1} c_{(i)}} + \sum_{i=1}^{k} \left( \frac{1}{\sum_{i=1}^{k+1} c_{(i)}} - \frac{1}{\sum_{i=1}^{k} c_{(i)}} \right) c_{(i)} T_{(i)} \\
&= \frac{c_{(k+1)} T_{(k+1)}}{\sum_{i=1}^{k+1} c_{(i)}} - \sum_{i=1}^{k} \frac{c_{(k+1)}}{\sum_{i=1}^{k+1} c_{(i)} \sum_{i=1}^{k} c_{(i)}} c_{(i)} T_{(i)} \\
&\geq 0
\end{aligned}$$

Hence $\hat{Q}^{*-}(c) \leq \hat{Q}^*(c) \leq \hat{Q}^{*+}(c)$. Let $\hat{\lambda}^- = \sup\{c \in (0,1) : \hat{Q}^{*-}(c) \leq \alpha\}$ and $\hat{\lambda}^+ = \sup\{c \in (0,1) : \hat{Q}^{*+}(c) \leq \alpha\}$. Then $\hat{\lambda}^- \geq \hat{\lambda} \geq \hat{\lambda}^+$.

(3). It can be shown, as done is Lemma A.4 in Sun and Cai (2007), that $\hat{Q}^{*-}(c) \to_{a.s.} \hat{Q}^{*+}(c)$, $\hat{\lambda}^- \to_p \lambda$, and $\hat{\lambda}^+ \to_p \lambda$. Hence $\hat{\lambda} \to_p \lambda$.

**Proof for Theorem 5.1**

By the Lemma 8.1, we have $\hat{T}_{OR} \to_p T_{OR}$ and by the Lemma 8.2 to Lemma 8.6, $\hat{\lambda}_{OR} \to_p \lambda_{OR}$. Then we have

$$\begin{aligned}
\int I(\hat{T}_{OR}(x) < \hat{\lambda}_{OR}) f(x) dx &\to \int I(T_{OR}(x) < \lambda_{OR}) f_0(x) dx, \\
\int I(\hat{T}_{OR}(x) < \hat{\lambda}_{OR}) q^*(x) dx &\to \int I(T_{OR}(x) < \lambda_{OR}) q^*(x) dx.
\end{aligned}$$

The $m$FDR of the data driven procedure is $\frac{\pi_0 \int I(\hat{T}_{OR}(x) < \hat{\lambda}_{OR}) f_0(x) dx}{\int I(\hat{T}_{OR}(x) < \hat{\lambda}_{OR}) q^*(x) dx}$ which converges to $Q^*(\lambda_{OR})$.

**Proof for Theorem 5.2**

Recall $q_1^*(x) = \int \mu^2 f(x-\mu) h(\mu) d\mu$. By Lemma 8.1 to Lemma 8.6, we have the following:

$$\begin{aligned} \int I(\hat{T}_{OR}(x) > \hat{\lambda}_{OR}) q_1^*(x) dx &\to \int I(T_{OR}(x) > \lambda_{OR}) q_1^*(x) dx; \\ \int I(\hat{T}_{OR}(x) > \hat{\lambda}_{OR}) q^*(x) dx &\to \int I(T_{OR}(x) > \lambda_{OR}) q^*(x) dx. \end{aligned}$$

The $m$FNR of the data driven procedure is $\frac{(1-\pi_0) \int I(\hat{T}_{OR}(x) > \hat{\lambda}_{OR}) q_1^*(x) dx}{\int I(\hat{T}_{OR}(x) > \hat{\lambda}_{OR}) q^*(x) dx}$, which converges to $\tilde{Q}^*(\lambda_{OR})$, where $\tilde{Q}^*(OR)$ is the the $m$FNR of the oracle procedure.

# References


[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL `http://www.jstor.org/stable/2346101`.

[2] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4): 1165–1188, 2001. ISSN 0090-5364.

[3] L.D. Brown and E. Greenshtein. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Arxiv preprint arXiv:0908.1712*, 2009.

[4] T.T. Cai and J. Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38(1):100–145, 2010. ISSN 0090-5364.

[5] D.B. Duncan. A bayesian approach to multiple comparisons. *Technometrics*, pages 171–222, 1965.

[6] B. Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, 2007. ISSN 0162-1459.

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

[8] Bradley Efron. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22, 2008. ISSN 0883-4237. URL `http://dx.doi.org/10.1214/07-STS236`.

[9] Bradley Efron. *Large-scale inference, empirical Bayes methods for estimation, testing, and prediction.* Cambridge University Press, 2010.

[10] C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002. ISSN 1467-9868.

[11] Jiashun Jin. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):461–493, 2008. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2007.00645.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2007.00645.x`.

[12] Jiashun Jin and T. Tony Cai. Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.*, 102(478):495–506, 2007. ISSN 0162-1459.

[13] E. Spjotvoll. On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics*, 43(2):398–411, 1972. ISSN 0003-4851.

[14] J.D. Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):347–368, 2007. ISSN 1467-9868.

[15] Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.*, 102 (479):901–912, 2007. ISSN 0162-1459.

[16] A.B. Van't Wout, G.K. Lehrman, S.A. Mikheeva, G.C. O'Keeffe, M.G. Katze, R.E. Bumgarner, G.K. Geiss, and J.I. Mullins. Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4+-T-cell lines. *Journal of Virology*, 77(2):1392, 2003. ISSN 0022-538X.

[17] Jichun Xie, T.T. Cai, John Maris, and Hongzhe Li. Optimal false discovery rate control for dependent data. 2011. submitted.