Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control

Wenguang SUN and T. Tony CAI

We develop a compound decision theory framework for multiple-testing problems and derive an oracle rule based on the z values that minimizes the false nondiscovery rate (FNR) subject to a constraint on the false discovery rate (FDR). We show that many commonly used multiple-testing procedures, which are p value–based, are inefficient, and propose an adaptive procedure based on the z values. The z value–based adaptive procedure asymptotically attains the performance of the z value oracle procedure and is more efficient than the conventional p value–based methods. We investigate the numerical performance of the adaptive procedure using both simulated and real data. In particular, we demonstrate our method in an analysis of the microarray data from a human immunodeficiency virus study that involves testing a large number of hypotheses simultaneously.

KEY WORDS: Adaptive procedure; Compound decision rule; False discovery rate; Local false discovery rate; Monotone likelihood ratio; Weighted classification.

1. INTRODUCTION

In large-scale multiple comparisons where hundreds or thousands of hypotheses are tested simultaneously, the goal is to separate the nonnull cases from the null cases. Commonly used multiple-testing procedures are typically based on the p values of the individual tests. For example, the well-known step-up procedure of Benjamini and Hochberg (1995), which aims to maximize the number of true positives while controlling the proportion of false positives among all rejections, thresholds the p values of the individual tests. Other examples include the adaptive procedure (Benjamini and Hochberg 2000), the plug-in procedure (Genovese and Wasserman 2004), and the augmentation procedure (van der Laan, Dudoit, and Pollard 2004). The operation of these procedures essentially involves first ranking the hypotheses based on the individual p values and then choosing a cutoff along the rankings.

The outcomes of a multiple testing procedure can be categorized as done in Table 1. The false discovery rate (FDR) is defined as $E(N_{10}/R|R > 0)\Pr(R > 0)$, with an FDR level of 0 when no hypotheses are rejected. Other similar measures include the positive false discovery rate (pFDR), $E(N_{10}/R|R > 0)$, and the marginal FDR (mFDR), $E(N_{10})/E(R)$. The pFDR and mFDR are equivalent when test statistics come from a random mixture of the null and nonnull distributions (Storey 2003). Genovese and Wasserman (2002) showed that under weak conditions, $\text{mFDR} = \text{FDR} + O(m^{-1/2})$, where m is the number of hypotheses.

Table 2 summarizes the data notation used in this article, which is standard for many microarray studies (see, e.g., Efron 2004a,b). Suppose that two groups of subjects, $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, each have measured expression levels for the same m genes. For $i = 1, \ldots, m$, a two-sample t statistic T_i is obtained for comparing the two groups on gene i, and is then transformed to a z value through $Z_i = \Phi^{-1}(F(T_i))$, where F and Φ are the cdf's of the t variable T_i and the standard normal variable. The corresponding p-values of all tests are recorded and denoted by $P_1, \ldots, P_m$.

In this article we develop a compound decision theory framework for multiple testing and derive an oracle rule based on the z values that, subject to a constraint on the FDR, minimizes the false nondiscovery rate (FNR), $E(N_{01}/S|S > 0)\Pr(S > 0)$. A major goal of this article is to show that the p value–based approaches generally do not lead to efficient multiple-testing procedures. The reason for the inefficiency of the p value methods can be traced back to work of Robbins (1951) that showed in a compound decision problem that simple rules (defined in Sec. 2) are usually inferior compared to compound decision rules. In addition, a result of Copas (1974) implies that the p value oracle procedure, defined in Section 3, is inadmissible in a compound decision problem. Robbins' argument that compound rules are superior to simple rules is especially important in large-scale multiple testing in which the precision of the tests can be increased by pooling information from different samples. Our approach is to use the z values to learn the distribution of the test statistics and use the information to construct a more efficient test.

In this article we first develop a z value–based oracle procedure that minimizes the mFNR subject to $\text{mFDR} \le \alpha$. We then compare this procedure with the p value oracle procedure proposed by Genovese and Wasserman (2002). A comparison of these two oracle procedures in Figure 1 shows that the p value oracle procedure is dominated by the z value oracle procedure, and that the gain in efficiency can be substantial when the alternative is asymmetric. More numerical examples are given in Section 5. It can be seen that the z value oracle procedure outperforms the p value oracle procedure in all cases except when the alternative is perfectly symmetric about the null, which implies that it is possible that p value–based procedures can be uniformly improved by using the z values.

We then develop a data-driven adaptive procedure based on the z values. We show that the z value–based adaptive procedure asymptotically attains the performance of the z value oracle procedure and is more efficient than the conventional p value–based methods, including the step-up procedure of Benjamini and Hochberg (1995) and the plug-in procedure of Genovese and Wasserman (2004). By treating multiple testing

Wenguang Sun is Ph.D. Candidate, Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *wgsun@mail.med.upenn.edu*). T. Tony Cai is Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *tcai@wharton.upenn.edu*). Cai's research was supported in part by National Science Foundation grants DMS-03-06576 and DMS-06-04954.

Journal of the American Statistical Association
September 2007, Vol. 102, No. 479, Theory and Methods
DOI 10.1198/016214507000000545

Table 1. Classification of tested hypothesis

	Claimed nonsignificant	Claimed significant	Total
Null	N_{00}	N_{10}	m_0
Nonnull	N_{01}	N_{11}	m_1
Total	S	R	m

as a compound decision problem, our results show that individual p values, although appropriate for testing a single hypothesis, fail to serve as the fundamental building block in large-scale multiple testing.

In addition to the theoretical properties, we study the numerical performance of the z value–based adaptive procedure using both simulated and real data. Simulations reported in Section 5 demonstrate that our z value procedure yields a lower mFNR than the p value–based methods at the same mFDR level. The gain in efficiency is substantial in many situations when the alternative is asymmetric about the null. We then apply our procedure to the analysis of microarray data from a human immunodeficiency virus (HIV) study in Section 6 and find that at the same FDR level, it rejects more hypotheses than the p value–based procedures. The type of asymmetry in the HIV data is commonly observed in microarray study, suggesting that our new z value–based adaptive procedure can aid in the discovery of additional new meaningful findings in many scientific applications.

The article is organized as follows. In Section 2 we develop a compound decision-theoretic framework and study various aspects of weighted classification and multiple-testing problems in this setting. In Sections 3 and 4 we propose oracle and adaptive testing procedures based on the z values for FDR control. In Section 5 we introduce numerical examples to show that our new z value–based procedure is more efficient than the traditional p value–based procedures. In Section 6 we illustrate our adaptive procedure with the analysis of microarray data in an HIV study. We conclude the article with a discussion of the results and some open problems. We provide proofs in the Appendix.

2. COMPOUND DECISION PROBLEM

Let Ω be the sample space and let Θ be the parameter space. Suppose that $\mathbf{x} = (x_1, \ldots, x_m) \in \Omega$ are observed and that we are interested in inference about the unknown $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \Theta$ based on $\mathbf{x}$. This involves solving m decision problems simultaneously and is called a *compound decision problem*. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ be a general decision rule. Then $\boldsymbol{\delta}$ is *simple* if δ_i is a function only of x_i, that is, $\delta_i(\mathbf{x}) = \delta_i(x_i)$. The simple rules correspond to solving the m component problems separately. In contrast, $\boldsymbol{\delta}$ is *compound* if δ_i depends on other x_j's,

Table 2. Summary of the data notation

Original data						T statistic	z value	p value
X_1	$\cdots$	X_{n_1}	Y_1	$\cdots$	Y_{n_2}	T	Z	P
X_{11}	$\cdots$	$X_{n_1 1}$	Y_{11}	$\cdots$	$Y_{n_2 1}$	T_1	Z_1	P_1
X_{12}	$\cdots$	$X_{n_1 2}$	Y_{12}	$\cdots$	$Y_{n_2 2}$	T_2	Z_2	P_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1m}	$\cdots$	$X_{n_1 m}$	Y_{1m}	$\cdots$	$Y_{n_2 m}$	T_m	Z_m	P_m

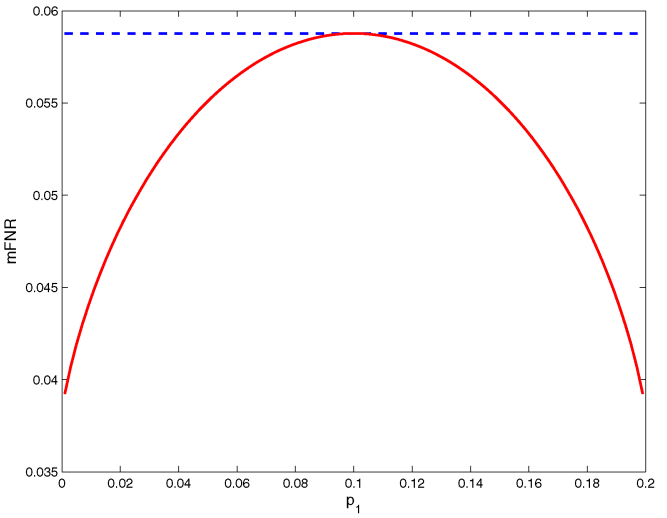

Figure 1. A comparison of the p value (- - -) and z value (——) oracle procedures with the mFDR level at .10. The test statistics have the normal mixture distribution $.8\mathrm{N}(0,1) + p_1\mathrm{N}(-3,1) + (.2 - p_1)\mathrm{N}(3,1)$. The mFNRs of the two procedures are plotted as a function of p_1.

$j \neq i$. A decision rule $\boldsymbol{\delta}$ is *symmetric* if $\boldsymbol{\delta}(\tau(\mathbf{x})) = \tau(\boldsymbol{\delta}(\mathbf{x}))$ for all permutation operators τ.

Robbins (1951) considered a compound decision problem where $X_i \sim \mathrm{N}(\theta_i, 1)$, $i = 1, \ldots, n$, are independent normal variables with mean $\theta_i = 1$ or -1. The goal is to classify each θ_i under the usual misclassification error. He showed that the unique minimax rule R: $\delta_i = \mathrm{sgn}(x_i)$ does not perform well in this compound decision problem by exhibiting a compound rule R^*: $\delta_i = \mathrm{sgn}(x_i - \frac{1}{2}\log\frac{1-\bar{x}}{1+\bar{x}})$ that substantially outperforms R when p, the proportion of $\theta_i = 1$, approaches 0 or 1, with only slightly higher risk near $p = .5$.

Let $\theta_1, \ldots, \theta_m$ be independent Bernoulli(p) variables and let X_i be generated as

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

The variables X_i are observed, and the variables θ_i are unobserved. The marginal cumulative distribution function (cdf) of X is the mixture distribution $F(x) = (1 - p)F_0(x) + pF_1(x)$, and the probability distribution function (pdf) is $f(x) = (1 - p)f_0(x) + pf_1(x)$, where f is assumed to be continuous and positive on the real line. In statistical and scientific applications, the goal is to separate the nonnull cases ($\theta_i = 1$) from the null ($\theta_i = 0$), which can be formulated either as a weighted classification problem or a multiple-testing problem. The solution to both problems can be represented by a decision rule, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m) \in \mathcal{I} = \{0, 1\}^m$.

In Section 6 we consider a problem in which we are interested in identifying a set of genes that are differentially expressed between HIV-positive patients and HIV-negative controls. This naturally gives rise to a multiple-testing problem in which the goal is to find as many true positives as possible while controlling the proportion of false positives among all rejections within level α. As in the work of Genovese and Wasserman (2002), we define the marginal FNR as $\mathrm{mFNR} = E(N_{01})/E(S)$, the proportion of the expected number of nonnulls among the expected number of nonrejections. The multiple testing problem is then to find δ that minimizes the mFNR

while controlling the mFDR at level α. On the other hand, in applications it is possible to rate the relative cost of a false positive (type I error) to a false negative (type II error). This naturally gives rise to a weighted classification problem with loss function

$$L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{m}\left\{\sum_i [\lambda I(\theta_i = 0)\delta_i + I(\theta_i = 1)(1 - \delta_i)]\right\}, \quad (2)$$

where $\lambda > 0$ is the relative weight for a false positive. The weighted classification problem is then to find $\boldsymbol{\delta}$ that minimizes the classification risk $E[L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta})]$. We develop a compound decision theory framework for inference about the mixture model (1) and make connections between multiple testing and weighted classification.

We consider $\boldsymbol{\delta}(\mathbf{x}) \in \{0, 1\}^m$, which is defined in terms of statistic $\mathbf{T}(\mathbf{x}) = [T_i(\mathbf{x}) : i = 1, \ldots, m]$ and threshold $\mathbf{c} = (c_1, \ldots, c_m)$ such that

$$\boldsymbol{\delta}(\mathbf{x}) = I(\mathbf{T} < \mathbf{c}) = \big[I(T_i(\mathbf{x}) < c_i) : i = 1, \ldots, m\big]. \quad (3)$$

Then $\boldsymbol{\delta}$ can be used in both weighted classification and multiple-testing problems. It is easy to verify that $\boldsymbol{\delta}$ given in (3) is symmetric if $T_i(\mathbf{x}) = T(x_i)$ and $\mathbf{c} = c\mathbf{1}$, where T is a function. In this section we allow T to depend on unknown quantities, such as the proportion of nonnulls and/or the distribution of X_i. We assume that $T(X_i) \sim G = (1 - p)G_0 + pG_1$, where G_0 and G_1 are the cdf's of $T(X_i)$ under the null and the alternative, respectively. The pdf of $T(X_i)$ is $g = (1 - p)g_0 + pg_1$. Let $\mathcal{T}$ be the collection of functions such that for any $T \in \mathcal{T}$, $T(X_i)$ has monotone likelihood ratio (MLR), that is,

$$g_1(t)/g_0(t) \text{ is decreasing in } t. \quad (4)$$

We call $\mathcal{T}$ the SMLR class. Let $\mathbf{x} = \{x_1, \ldots, x_m\}$ be a random sample from the mixture distribution F and let $\boldsymbol{\delta}(T, c) = \{I[T(x_i) < c] : i = 1, \ldots, m\}$. We call $\boldsymbol{\delta}(T, c)$ a SMLR decision rule if $T \in \mathcal{T}$, and let $\mathcal{D}_s$ denote the collection of all SMLR decision rules.

Suppose that $T(X_i) = p(X_i)$, 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. Assume that $G_1(t)$ is twice differentiable. Note that $g_0(t) = G_0'(t) = 1$, the assumption $p(\cdot) \in \mathcal{T}$ implies that $G_1''(t) = g_1'(t) < 0$, that is, $G_1(t)$ is concave. Therefore, the SMLR assumption can be viewed as a generalized version of the assumption of Storey (2002) and Genovese and Wasserman (2002, 2004) that the p value distribution under the alternative is concave. The following proposition shows that the SMLR assumption is a desirable condition in both multiple-testing and weighted classification problems.

Proposition 1. Let $\theta_i, i = 1, \ldots, m$, be iid Bernoulli($p$) variables and let $x_i|\theta_i$, $i = 1, \ldots, m$, be independent observations from the model (1). Suppose that $T \in \mathcal{T}$; then the implementation of $\boldsymbol{\delta}(T, c) = \{I[T(x_i) < c] : i = 1, \ldots, m\}$ in both the weighted classification and the multiple-testing problems implies that (a) $P(\theta_i = 1|T(X_i) \le c)$ is monotonically decreasing in threshold c, (b) the mFDR is monotonically increasing in c and the expected number of rejections r, (c) the mFNR is monotonically decreasing in c and r, (d) the mFNR is monotonically decreasing in the mFDR, and (e) in the classification problem, c (and thus r) is monotonically decreasing in the classification weight λ.

The following theorem makes connection between a multiple-testing problem and a weighted classification problem. In particular, the former can be solved by solving the latter, with an appropriately chosen λ.

Theorem 1. Let θ_i, $i = 1, \ldots, m$, be iid Bernoulli(p) variables and let $x_i|\theta_i$, $i = 1, \ldots, m$, be independent observations from the mixture model (1). Let $\Lambda \in \mathcal{T}$ and suppose that the classification risk with the loss function (2) is minimized by $\boldsymbol{\delta}^\lambda[\Lambda, c(\lambda)] = \{\delta_1^\lambda, \ldots, \delta_m^\lambda\}$, where $\delta_i^\lambda = I\{\Lambda(x_i) < c(\lambda)\}$. Then for any given mFDR level α in a multiple testing problem, there exists a unique $\lambda(\alpha)$ that defines a weighted classification problem. The optimal solution to the classification problem $\boldsymbol{\delta}^{\lambda(\alpha)}[\Lambda, c\{\lambda(\alpha)\}]$ is also optimal in the multiple testing problem in the sense that it controls the mFDR at level α with the smallest mFNR among all decision rules in $\mathcal{D}_s$.

The next step is to develop such an optimal rule $\boldsymbol{\delta}^\lambda(\Lambda, c(\lambda))$ as stated in Theorem 1. We first study an ideal setup in which 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 δ.

Theorem 2 (The oracle rule for weighted classification). Let θ_i, $i = 1, \ldots, m$, be iid Bernoulli(p) variables and let $x_i|\theta_i$, $i = 1, \ldots, m$, be independent observations from the mixture model (1). Suppose that p, f_0, and f_1 are known. Then the classification risk $E[L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta})]$ with L given in (2) is minimized by $\boldsymbol{\delta}^\lambda(\Lambda, 1/\lambda) = \{\delta_1, \ldots, \delta_m\}$, where

$$\delta_i = I\left\{\Lambda(x_i) = \frac{(1 - p)f_0(x_i)}{pf_1(x_i)} < \frac{1}{\lambda}\right\}, \qquad i = 1, \ldots, m. \quad (5)$$

The minimum classification risk is $R_\lambda^* \doteq \inf_{\boldsymbol{\delta}} E[L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta})] = p + \int_K [\lambda(1 - p)f_0(x) - pf_1(x)]\,dx$, where $K = \{x \in \Omega : \lambda(1 - p)f_0(x) < pf_1(x)\}$.

Remark 1. Let the likelihood ratio (LR) be defined as $L_i = f_0(x_i)/f_1(x_i)$. A compound decision rule is said to be *ordered* if for almost all $\mathbf{x}$, $L_i > L_j$ and $\delta_i(\mathbf{x}) > 0$ imply that $\delta_j(\mathbf{x}) = 1$. Copas (1974) showed that if a symmetric compound decision rule for dichotomies is admissible, then it is ordered. Note that (5) is a Bayes rule and so is symmetric, ordered, and admissible. But because the p value–based procedure $\boldsymbol{\delta}(p(\cdot), c) = \{I[p(X_i) < c] : i = 1, \ldots, m\}$ is symmetric but not ordered by the LR, it is inadmissible in the compound decision problem.

Remark 2. In practice, some of the p, f_0, and f_1 are unknown but estimable. Then we can estimate the unknown quantities and use the rule $\boldsymbol{\delta}^\lambda(\hat{\Lambda}, 1/\lambda)$. The *subminimax* rule, $\delta_i = \operatorname{sgn}(x_i - \frac{1}{2}\log\frac{1-\bar{x}}{1+\bar{x}})$, given by Robbins (1951), is recovered by letting $\lambda = 1$, $f_0 = \phi(\cdot + 1)$, $f_1 = \phi(\cdot - 1)$, and $\hat{p}_1 = (1 + \bar{x})/2$ in (5), where ϕ is the pdf of a standard normal.

Remark 3. Suppose that two weights $\lambda_1 < \lambda_2$ are chosen in the loss (2). Let $\Omega_1 = \{x : \lambda_1(1 - p)f_0(x) < pf_1(x)\}$ and $\Omega_2 = \{x : \lambda_2(1 - p)f_0(x) < pf_1(x)\}$. Then the classification risk R_λ is increasing in λ because $R_{\lambda_1}^* - R_{\lambda_2}^* = \int_{\Omega_1 \setminus \Omega_2} [\lambda_1(1 - p)f_0(x) - pf_1(x)]\,dx + \int_{\Omega_2} \{[(\lambda_1 - \lambda_2)(1 - p)f_0(x)]\}\,dx < 0$. Also, it is easy to see that $\Omega_1 \supset \Omega_2$. Thus the expected number of subjects classified to the nonnull population is decreasing in λ, demonstrating that (d) in Proposition 1 is satisfied by the classification rule $\boldsymbol{\delta}^\lambda(\Lambda, 1/\lambda)$.

3. THE ORACLE PROCEDURES FOR mFDR CONTROL

We have shown that $\boldsymbol{\delta}^{\lambda}(\Lambda, 1/\lambda) = [I\{\Lambda(x_1) < 1/\lambda\}, \ldots, I\{\Lambda(x_m) < 1/\lambda\}]$ is the oracle rule in the weighted classification problem. Theorem 1 implies that the optimal rule for the multiple-testing problem is also of the form $\boldsymbol{\delta}^{\lambda(\alpha)}[\Lambda, 1/\lambda(\alpha)]$ 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 FDR (Lfdr) introduced by Efron, Tibshirani, Storey, and Tusher (2001) and Efron (2004a), so the optimal rule for mFDR control is of the form $\boldsymbol{\delta}(\text{Lfdr}(\cdot), c) = \{I[\text{Lfdr}(x_i) < c] : i = 1, \ldots, m\}$. 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 2004a). 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 with the smallest mFNR.

The Lfdr statistic is defined in terms of the *z values*, which can be converted from other test statistics, including the t and chi-squared statistics, using inverse-probability transformations. Note that p is a global parameter. Therefore, the expression $\text{Lfdr(z)} = (1 - p) f_0(z)/f(z)$ implies that we actually rank the relative importance of the observations according to their LRs, and that the rankings are generally different from the rankings of p values unless the alternative distribution is symmetric about the null. 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. Therefore, it is possible that the test accepts a more extreme observation while rejecting a less extreme observation, implying that the rejection region is asymmetric. This is not possible for a testing procedure based on the individual p values, which has a rejection region always symmetric about the null. This phenomenon is illustrated in Figure 2 at the end of this section.

Setting the threshold for the test statistics has been the focus of the FDR literature (see, e.g., Benjamini and Hochberg 1995; Genovese and Wasserman 2004). Consider an ideal situation in which we assume that an oracle knows the true underlying distribution of the test statistics. Note that the SMLR assumption implies that the mFNR is decreasing in the mFDR; therefore, the oracle's response to such a thresholding problem is to "spend" all of the mFDR to minimize the mFNR. Besides providing a target for evaluating different multiple-testing procedures, the oracle procedure also sheds light on the development of the adaptive procedure that we propose in Section 4.

3.1 The Oracle Procedure Based on the p Values

Let $G_1(t)$ be the distribution of the p values under the alternative and let p be the proportion of nonnulls. We assume that G_1 is concave; then, according to Genovese and Wasserman (2002), the oracle procedure thresholds the p value at u^*, the solution to the equation $G_1(u)/u = (1/p - 1)(1/\alpha - 1)$. Here u^* is the optimal cutoff for a concave $G_1(t)$ in the sense that the rule $\boldsymbol{\delta}[p(\cdot), u^*] = \{I[P_i < u^*] : i = 1, \ldots, m\}$ yields the smallest mFNR among all p value–based procedures that control the mFDR at level α. Let $\tilde{G}(t) = 1 - G(t)$. The resulting mFDR of $(1 - p)u^*/[(1 - p)u^* + pG_1(u^*)]$ is just α, and the mFNR is given by $p\tilde{G}(u^*)/[p\tilde{G}_1(u^*) + (1 - p)(1 - u^*)]$.

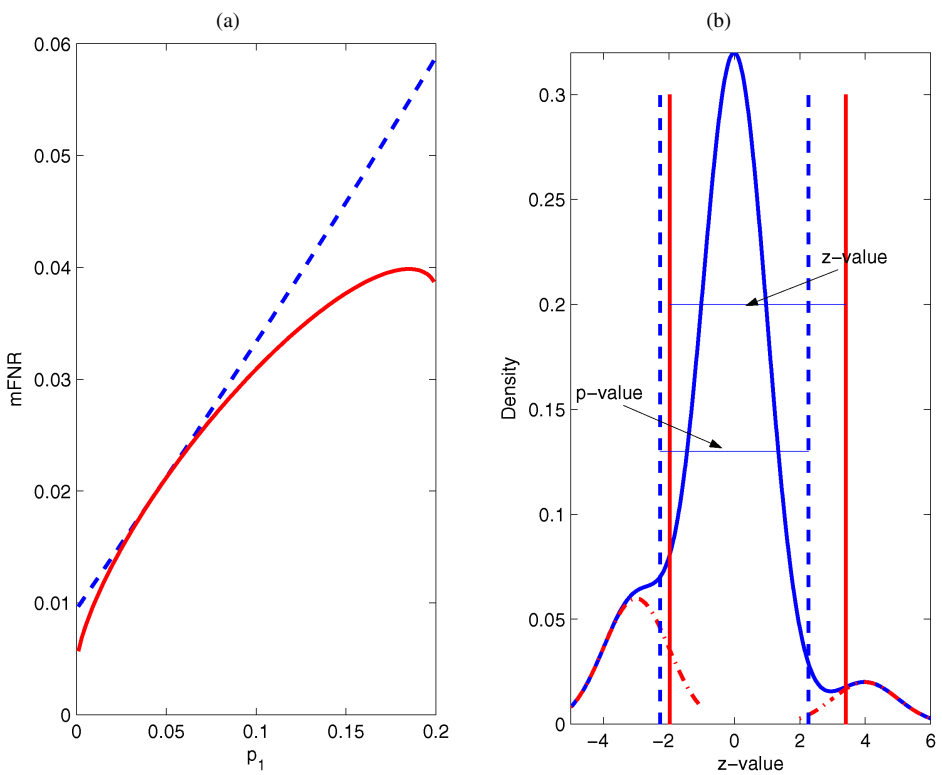

Figure 2. Symmetric rejection region versus asymmetric rejection region. (a) Comparison of oracle rules (- - -, p value; ——, z value). (b) Rejection regions. The normal mixture model is $.8\text{N}(0, 1) + p_1\text{N}(-3, 1) + (.2 - p_1)\text{N}(4, 1)$. Both procedures control the mFDR at .10.

3.2 The Oracle Procedure Based on z Values

We write the Lfdr statistic as $T_{OR}(Z_i) = (1-p)f_0(Z_i)/f(Z_i)$ and call it the *oracle test statistic*. We assume that $T_{OR}(Z_i)$ is distributed with marginal cdf $G_{OR} = (1-p)G^0_{OR} + pG^1_{OR}$, where G^0_{OR} and G^1_{OR} are the cdf's of $T_{OR}(Z_i)$ under the null and the alternative. Set $\tilde{G}_{OR}(t) = 1 - G_{OR}(t)$. The mFDR of the oracle rule $\delta(T_{OR}, \lambda)$ is $Q_{OR}(\lambda) = (1-p)G^0_{OR}(\lambda)/G_{OR}(\lambda)$. We assume that $T_{OR} \in \mathcal{T}$ satisfies the SMLR assumption; then (b) of Proposition 1 implies that $Q_{OR}(\lambda)$ is increasing in the threshold λ. Therefore, the oracle procedure thresholds $T_{OR}(z)$ at $\lambda_{OR} = \sup\{t \in (0,1) : Q_{OR}(t) \le \alpha\} \doteq Q^{-1}_{OR}(\alpha)$. Thus the oracle testing rule based on z values is

$$\boldsymbol{\delta}(T_{OR}, \lambda_{OR}) = \big\{I[T_{OR}(z_1) < \lambda_{OR}], \ldots, I[T_{OR}(z_m) < \lambda_{OR}]\big\}. \tag{6}$$

The corresponding mFDR$= Q_{OR}(\lambda_{OR})$ is just α, and the mFNR is

$$\tilde{Q}_{OR}(\lambda_{OR}) = p\tilde{G}^1_{OR}(\lambda_{OR})/\tilde{G}_{OR}(\lambda_{OR}). \tag{7}$$

Example 1. We consider a random sample $z_1, \ldots, z_m$ from a normal mixture model,

$$f(z) = (1-p_1-p_2)\phi(z) + p_1\phi(z-\mu_1) + p_2\phi(z-\mu_2), \tag{8}$$

along with corresponding m tests $H^i_0 : \mu = 0$ versus $H^i_1 : \mu \ne 0$, $i = 1, \ldots, m$. Let $p = p_1 + p_2$ be the total proportion of nonnulls. We assume that $\mu_1 < 0$ and $\mu_2 > 0$, so that the rejection region will be two-sided.

The p value distribution under the alternative can be derived as $G_1(t) = (p_1/p)\{\tilde{\Phi}(Z_{t/2}+\mu_1) + \tilde{\Phi}(Z_{t/2}-\mu_1)\} + (p_2/p)\{\tilde{\Phi}(Z_{t/2}+\mu_2) + \tilde{\Phi}(Z_{t/2}-\mu_2)\}$. Thus the optimal p value cutoff u^* is the solution to equation $G_1(u)/u = (1/p-1)\times(1/\alpha - 1)$, with corresponding mFNR of $p\tilde{G}(u^*)/[p\tilde{G}_1(u^*) + (1-p)(1-u^*)]$.

It is easy to verify that in normal mixture model (8), the oracle testing rule (6) is equivalent to rejecting the null when $z_i < c_l$ or $z_i > c_u$. The threshold λ_{OR} can be obtained using the bisection method. For a given λ, c_l and c_u can be solved numerically from the equation $\lambda[p_1\exp(-\mu_1 z - \frac{1}{2}\mu_1^2) + p_2\exp(\mu_2 z - \frac{1}{2}\mu_2^2)] - (1-\lambda)(1-p) = 0$. The corresponding mFDR and mFNR can be calculated as

$$\begin{aligned}\text{mFDR} = {} & (1-p)[\Phi(c_l) + \tilde{\Phi}(c_u)] \\ & /\big((1-p)[\Phi(c_l) + \tilde{\Phi}(c_u)] + p_1[\Phi(c_l+\mu_1) \\ & + \tilde{\Phi}(c_u+\mu_1)] + p_2[\Phi(c_l-\mu_2) + \tilde{\Phi}(c_u-\mu_2)]\big)\end{aligned}$$

and

$$\begin{aligned}\text{mFNR} = {} & \big(p_1[\Phi(c_u+\mu_1) - \Phi(c_l+\mu_1)] \\ & + p_2[\Phi(c_u-\mu_2) - \Phi(c_l-\mu_2)]\big) \\ & /\big((1-p)[\Phi(c_u) - \Phi(c_l)] + p_1[\Phi(c_u+\mu_1) \\ & - \Phi(c_l+\mu_1)] + p_2[\Phi(c_u-\mu_2) - \Phi(c_l-\mu_2)]\big),\end{aligned}$$

where $\Phi(\cdot)$ is the cdf of a standard normal and $\tilde{\Phi}(\cdot) = 1 - \Phi(\cdot)$.

In this example we choose the mixture model $.8\text{N}(0,1) + p_1\text{N}(-3,1) + (.2-p_1)\text{N}(4,1)$. This is different from the model shown in Figure 1, which has alternative means -3 and 3; therefore, different patterns are seen. Both oracle procedures control the mFDR at level .10, and we plot the mFNRs of the two procedures as a function of p_1 in Figure 2(a). We can see that again, the p value oracle procedure is dominated by the z value oracle procedure. We set $p = .15$ and characterize the rejection regions of the two oracle procedures in Figure 2(b). Some calculations show that the rejection region of the p value oracle procedure is $|z_i| > 2.27$ with mFNR $= .046$, whereas the rejection region of the z value oracle procedure is $z_i < -1.97$ and $z_i > 3.41$ with mFNR $= .038$. It is interesting to note that the z value oracle procedure rejects observation $x = -2$ (p value $= .046$) but accepts observation $x = 3$ (p value $= .003$). We provide more numerical comparisons in Section 5, and show that the z value oracle procedure yields a lower mFNR level than the p value oracle procedure on all possible configurations of alternative hypotheses, with the difference significant in many cases.

3.3 Connection to Work of Spjøtvoll (1972) and Storey (2007)

An "optimal" multiple-testing procedure was introduced by Spjøtvoll (1972). Let $f_{01}, \ldots, f_{0N}$ and $f_1, \ldots, f_N$ be integrable functions. Let $S'(\gamma)$ denote the set of all tests $(\psi_1, \ldots, \psi_N)$ that satisfy $\sum_{t=1}^N \int \psi_t(\mathbf{x}) f_{0t}(\mathbf{x})\, d\mu(\mathbf{x}) = \gamma$, where $\gamma > 0$ is prespecified. Spjøtvoll (1972) showed that the test $(\phi_1, \ldots, \phi_N) = \{I[f_t(\mathbf{x}) > c f_{0t}(\mathbf{x})] : t = 1, \ldots, N\}$ maximizes

$$\sum_{t=1}^N \int \phi_t(\mathbf{x}) f_t(\mathbf{x})\, d\mu(\mathbf{x}) \tag{9}$$

among all tests $(\psi_1, \ldots, \psi_N) \in S'(\gamma)$. Storey (2007) proposed the optimal discovery procedure (ODP) based on a shrinkage test statistic that maximizes the expected number of true positives (ETP) for a fixed level of the expected number of false positives (EFP). The result of Storey (2007) on the ODP follows directly from the theorem of Spjøtvoll (1972) by choosing appropriate f_{0t} and f_t.

In the setting of the present article, both Spjøtvoll's optimal procedure and Storey's ODP procedure depend on unknown functions that are not estimable from the data. Moreover, the optimal cutoffs for a given test level are not specified in both procedures. These limitations make the two methods inapplicable in terms of the goal of FDR control.

The formulation of our testing procedure (6) has two advantages over the procedures of Spjøtvoll (1972) and Storey (2007). First, we can form good estimates (Jin and Cai 2007) from the data for the unknown functions in the oracle test statistic that we propose. Second, we specify the optimal cutoff for each mFDR level, which was not discussed by Spjøtvoll (1972) or Storey (2007). These advantages greatly facilitate the development of the adaptive procedure (10), where a consistent cutoff (relative to the oracle cutoff) is suggested for each test level based on a simple step-up procedure. This task is impossible based on the formulation of the work of Spjøtvoll or Storey, because good estimates of the unknown quantities in their test "statistics" are not available, and obtaining appropriate cutoffs is difficult.

4. ADAPTIVE PROCEDURES FOR mFDR CONTROL

Genovese and Wasserman (2004) discussed a class of plug-in procedures for the purpose of practical implementations of the oracle procedure based on the p values. However, the idea of plug-in is difficult to apply to the z value oracle procedure, because it essentially involves estimating the distribution of oracle test statistic $T_{OR}(z)$, which is usually very difficult. Instead, we develop an adaptive procedure that requires only estimation of the distribution of the z values, so that the difficulty of estimating the distribution of $T_{OR}(z)$ is avoided. The adaptive procedure can be easily implemented by noting that the z values are usually distributed as a normal mixture after appropriate transformations are applied, and several methods for consistently estimating the normal nulls have been developed in the literature (see, e.g., Efron 2004b; Jin and Cai 2007).

In this section we first introduce an adaptive p value–based procedure proposed by Benjamini and Hochberg (2000). We then turn to the development of an adaptive procedure based on the z values. We show that the adaptive z value–based procedure asymptotically attains the performance of the oracle procedure (6). We report simulation studies in Section 5 demonstrating that the z value–based adaptive procedure is more efficient than the traditional p value–based testing approaches.

4.1 Adaptive Procedure Based on p Values

Suppose that $P_{(1)}, \ldots, P_{(m)}$ are the ranked p values from m independent tests, and let $k = \max\{i : P_{(i)} \le \alpha i/[m(1-\hat{p})]\}$. Then the adaptive procedure of Benjamini and Hochberg (2000), designated as BH hereinafter, rejects all $H^{(i)}$, $i \le k$. Genovese and Wasserman (2004) proposed the *plug-in threshold*, $t(\hat{p}, \hat{G})$, for p values, where $t(p, G) = \sup\{t : (1-p)t/G(t) \le \alpha\}$ is the threshold by the oracle p value procedure and $\hat{p}$ and $\hat{G}$ are estimates of p and G. The next theorem shows that the GW plug-in procedure and the BH adaptive procedure are equivalent when the empirical distribution for p values $\mathbb{G}_m$ is used to estimate G.

Theorem 3. Let $\hat{p}$ be an estimate of p and let $\mathbb{G}_m$ be the empirical cdf of the p values. Then the GW plug-in procedure is equivalent to the BH adaptive procedure.

It follows from our Theorem 3 and theorem 5 of Genovese and Wasserman (2004) that the BH adaptive procedure controls the mFDR at level α asymptotically.

4.2 Adaptive Procedure Based on z Values

Here 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 $z_1, \ldots, z_m$ be a random sample from the mixture model (1) with 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 have been provided by for example, Jin and Cai (2007). 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 $\boldsymbol{\delta}(T_{OR}, \lambda) = \{I[T_{OR}(z_i) < \lambda] : i = 1, \ldots, m\}$ 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 as defined in Section 3. 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

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

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

$$\begin{aligned} &\text{Let } k = \max\{i : \tfrac{1}{i}\textstyle\sum_{j=1}^i \hat{\mathrm{Lfdr}}_{(j)} \le \alpha\}; \\ &\text{then reject all } H^{(i)}, i = 1, \ldots, k. \end{aligned} \tag{10}$$

Similar to the discussion in the proof of Theorem 3, it is sufficient to consider only the discrete cutoffs in R and the adaptive procedure (10) is equivalent to the plug-in procedure

$$\boldsymbol{\delta}(\hat{T}_{OR}, \hat{\lambda}_{OR}) = \left\{I[\hat{T}_{OR}(z_1) < \hat{\lambda}_{OR}], \ldots, I[\hat{T}_{OR}(z_m) < \hat{\lambda}_{OR}]\right\},$$

which is very difficult to implement because obtaining $\hat{\lambda}_{OR}$ is difficult.

Remark 4. The procedure (10) is more adaptive than the BH adaptive procedure in the sense that it adapts to both the global feature p and local feature f_0/f. In contrast, the BH method adapts only to the global feature p. Suppose that we use the theoretical null N(0, 1) in the expression of $\hat{T}_{OR} = (1-\hat{p})f_0/\hat{f}$. 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 nonnulls around $-z$ [i.e., $\hat{f}(-z)$ is larger], then observation $-z$ will be correspondingly ranked higher than observation z.

Remark 5. In the FDR literature, z value–based methods such as the Lfdr procedure (Efron 2004a,b) are used only to calculate individual significance levels, whereas the p value–based procedures are used for global FDR control to identify nonnull cases. It is also notable that the goals of global error control and individual case interpretation are naturally unified in the adaptive procedure (10).

The next two theorems show that the adaptive procedure (10) 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 (6) are also asymptotically achieved by the adaptive z value–based procedure (10).

Theorem 4 (Asymptotic validity of the adaptive procedure). Let θ_i, $i = 1, \ldots, m$, be iid Bernoulli(p) variables and let $x_i|\theta_i$, $i = 1, \ldots, m$, be independent observations from the mixture model (1) with PDF $f = (1-p)f_0 + pf_1$. Suppose that

f is continuous and positive on the real line. Assume that $T_{OR}(z_i) = (1-p)f_0(z_i)/f(z_i)$ is distributed with the marginal pdf $g = (1-p)g_0 + pg_1$ and that $T_{OR} \in \mathcal{T}$ satisfies the SMLR assumption. Let $\hat{p}$, $\hat{f}_0$, and $\hat{f}$ be estimates of p, f_0, and f such that $\hat{p} \xrightarrow{p} p$, $E\|\hat{f} - f\|^2 \to 0$, and $E\|\hat{f}_0 - f_0\|^2 \to 0$. Define the test statistic $\hat{T}_{OR}(z_i) = (1-\hat{p})\hat{f}_0(z_i)/\hat{f}(z_i)$. Let $\widehat{\text{Lfdr}}_{(1)}, \ldots, \widehat{\text{Lfdr}}_{(m)}$ be the ranked values of $\hat{T}_{OR}(z_i)$; then the adaptive procedure (10) controls the mFDR at level α asymptotically.

Theorem 5 (Asymptotic attainment of adaptive procedure). Assume that random sample $z_1, \ldots, z_m$ and test statistics $T_{OR}(z_i)$, $\hat{T}_{OR}(z_i)$ are the same as in Theorem 4. Then the mFNR level of the adaptive procedure (10) is $\tilde{Q}_{OR}(\lambda_{OR}) + o(1)$, where $\tilde{Q}_{OR}(\lambda_{OR})$, given in (7), is the mFNR level achieved by the z value oracle procedure (6).

5. NUMERICAL RESULTS

We now turn to the numerical performance of our adaptive z value–based procedure (10). The procedure is easy to implement; the R code for the procedure is available at *http://stat.wharton.upenn.edu/~tcai/paper/html/FDR.html*. The goal of this section is to provide numerical examples to show that the conventional p value–based procedures are inefficient. We also explore the situations where these p value–based procedures can be substantially improved by the z value–based procedures. We describe the application of our procedure to the analysis of microarray data from an HIV study in Section 6.

In our numerical studies, we consider the following normal mixture model:

$$Z_i \sim (1-p)\text{N}(\mu_0, \sigma_0^2) + p\text{N}(\mu_i, \sigma_i^2), \tag{11}$$

where (μ_i, σ_i^2) follows some bivariate distribution $F(\mu, \sigma^2)$. This model can be used to approximate many mixture distributions and is found in a wide range of applications (see, e.g., Magder and Zeger 1996; Efron 2004b). We compare the p value–based and z value–based procedures in two numerical examples.

Numerical Study 1. We compare the p value and z value oracle procedures in the normal mixture model (8), a special case of (11). The algorithm for calculating the oracle cutoffs and the corresponding mFNRs is given in Example 1 at the end of Section 3. Figure 3 compares the performance of these two oracle procedures. Panel (a) plots the mFNR of the two oracle procedures as a function of p_1, where the mFDR level is set at .10, the alternative means are $\mu_1 = -3$ and $\mu_2 = 3$, and the total proportion of nonnulls is $p = .2$. Panel (b) plots the mFNR as a function of p_1 in the same setting except with alternative means $\mu_1 = -3$ and $\mu_2 = 6$. In panel (c), mFDR $= .10$, $p_1 = .18$, $p_2 = .02$, and $\mu_1 = -3$, and we plot the mFNR as a function of μ_2. Panel (d) plots the mFNR as a function of the mFDR level while holding $\mu_1 = -3$, $\mu_2 = 1$, $p_1 = .02$, and $p_2 = .18$ fixed. We discuss these results at the end of this section.

Numerical Study 2. We compare the step-up procedure (Benjamini and Hochberg 1995), the adaptive p value–based procedure (Benjamini and Hochberg 2000; Genovese and Wasserman 2004), and the adaptive z value–based procedure (10), designated by BH95, BHGW, and AdaptZ hereinafter. Note that the BH step-up procedure is easier to implement than either BHGW or AdaptZ. The BHGW procedure requires estimating the proportion of nonnulls, and the AdaptZ procedure also requires an additional density estimation step.

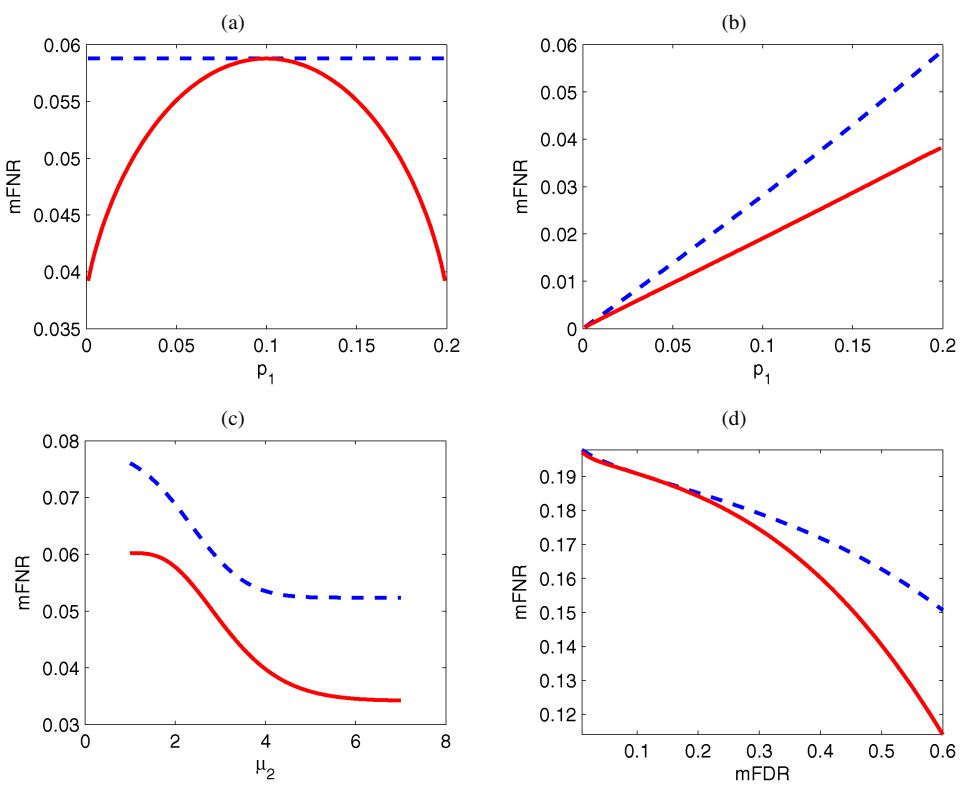

Figure 3. Comparison of the p value (- - -) and z value (——) oracle rules. (a) mFDR $= .10$, $\mu = (-3; 3)$; (b) mFDR $= .10$, $\mu = (-3; 6)$; (c) mFDR $= .10$, $p_1 = .18$, $p_2 = .02$, $\mu_1 = -3$; (d) $\mu = (-3; 1)$, $p = (.02, .18)$.

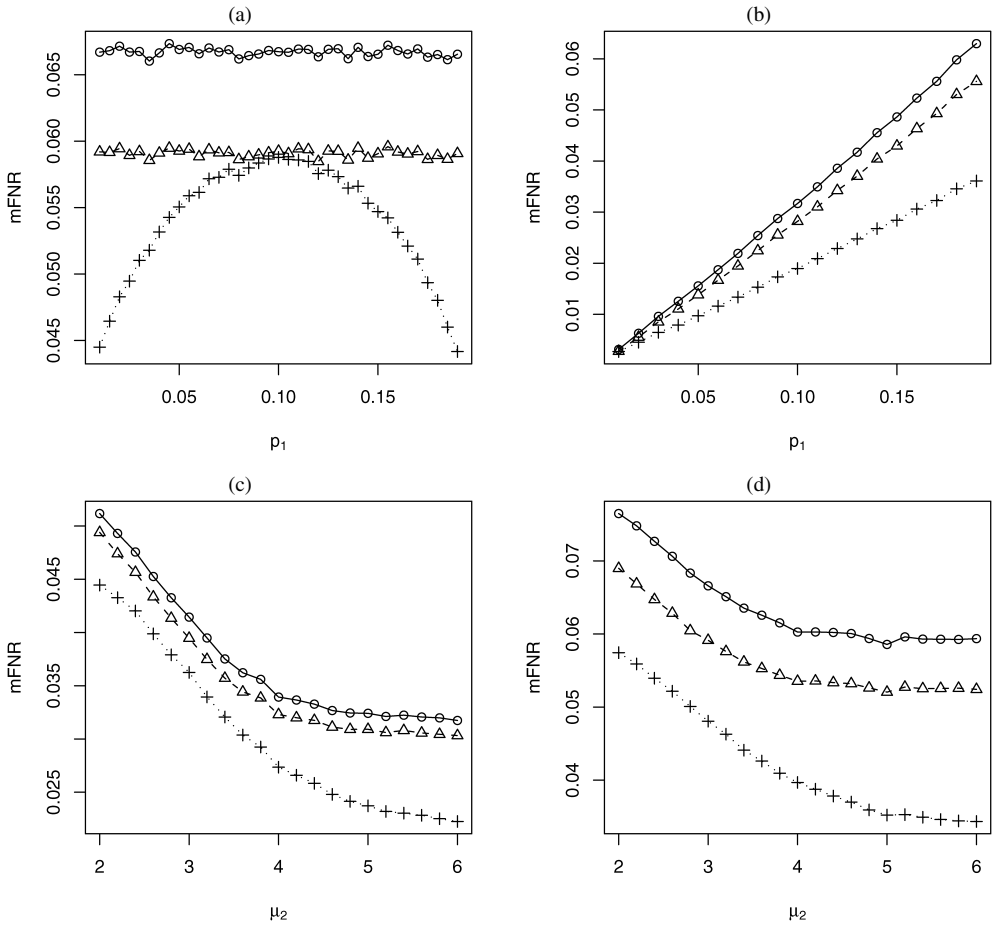

Figure 4. Comparison of procedures for FDR control. (a) $\mu = (-3, 3)$, $p = .20$; (b) $\mu = (-3, 6)$, $p = .20$; (c) $\mu_1 = -3$, $p = (.08, .02)$; (d) $\mu_1 = -3$, $p = (.18, .02)$. ($\circ$, the BH step-up procedure based on the p values; $\triangle$, the BHGW adaptive procedure based on the p values; $+$, the adaptive procedure based on the z-values.) The mFDR is controlled at level .10.

The alternative means μ_{1i} and μ_{2i} are generated from uniform distributions $\mathrm{U}(\theta_1 - \delta_1, \theta_1 + \delta_1)$ and $\mathrm{U}(\theta_2 - \delta_2, \theta_2 + \delta_2)$, after which z_i is generated from the normal mixture model (8) based on μ_{1i} and μ_{2i}, $i = 1, \ldots, m$. This hierarchical model also can be viewed as a special case of the mixture model (11). In estimating the Lfdr statistic $\mathrm{Lfdr}(z_i) = (1 - p) f_0(z_i)/f(z_i)$, f_0 is chosen to be the theoretical null density $\mathrm{N}(0, 1)$, p is estimated consistently using the approach of Jin and Cai, and f is estimated using a kernel density estimator with bandwidth chosen by cross-validation. The comparison results are displayed in Figure 4, where the top row gives plots of the mFNR as a function of p_1 and the bottom row gives plots of the mFNR as a function of θ_2, with other quantities held fixed. All points in the plots are calculated based on a sample size of $m = 5{,}000$ and 200 replications.

The following observations can be made based on the results from these two numerical studies:

a. When $|\mu_1| = |\mu_2|$, the mFNR remains a constant for the p value oracle procedure. In contrast, the mFNR for the z value oracle procedure increases first and then decreases [Fig. 3(a)]. The p value and z value oracle procedures yield the same mFNR levels only when the alternative is symmetric about the null. This reveals the fact that the z value procedure adapts to the asymmetry in the alternative distribution but the p value procedure does not. Similar phenomena are shown in Figure 4 for adaptive procedures.

b. The p value oracle procedure is dominated by the z value oracle procedure. The largest difference occurs when $|\mu_1| < \mu_2$ and $p_1 > p_2$, where the alternative distribution is highly asymmetric about the null [Figs. 3(b) and 3(c)]. Similarly, the BH95 is dominated by BHGW, which is again dominated by AdaptZ (Fig. 4).

c. The mFNRs for both p value and z value procedures decrease as μ_2 moves away from 0.

d. Within a reasonable range (mFDR $< .6$), the mFNR is decreasing in the mFDR [Fig. 3(d)], which verifies part (d) of Proposition 1, one consequence of our SMLR assumption.

Additional simulation results show that the difference in the mFNRs of the p value and z value procedures is increasing in the proportion of nonnulls and that the adaptive procedures (BHGW and AdaptZ) control the mFDR at the nominal level α approximately, whereas the BH95 procedure controls the mFDR at a lower level.

6. APPLICATION TO MICROARRAY DATA

We now 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

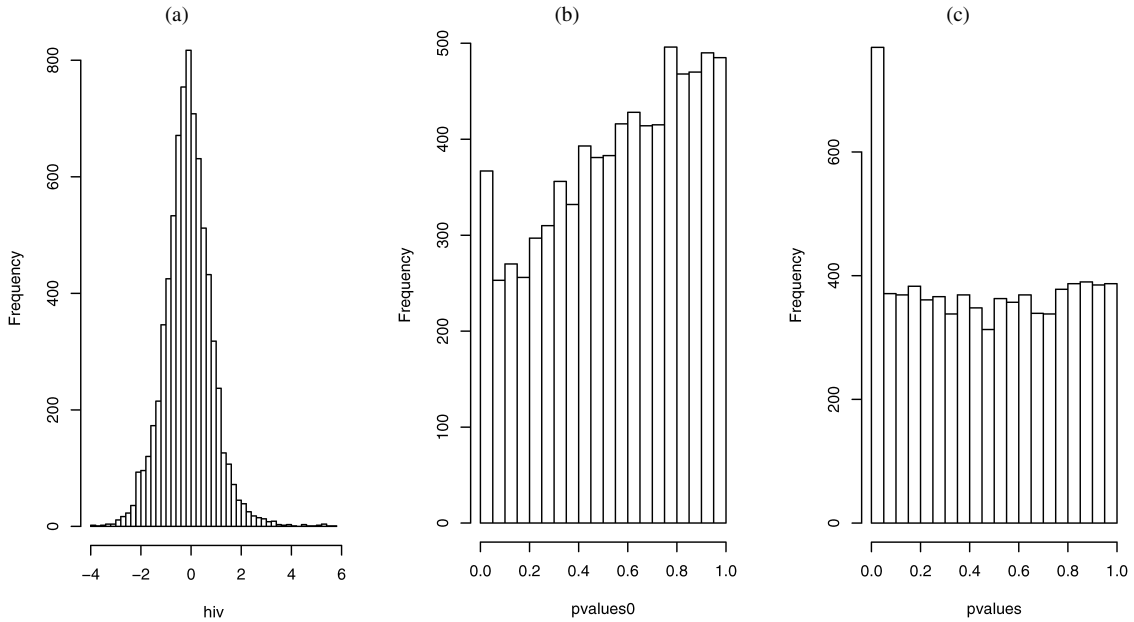

Figure 5. Histograms of the HIV data. (a) z values; (b) p values; transformed p values, approximately distributed as uniform $(0, 1)$ for the null cases.

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 $N = 7{,}680$ genes. A total of N two-sample t-tests were performed, and N two-sided p values were obtained. The z values were converted from the t statistics using the transformation $z_i = \Phi^{-1}[G_0(t_i)]$, where Φ and G_0 are the cdf's of a standard normal and a t variable with 6 degrees of freedom. The histograms of the z values and p values were presented in Figure 5. 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. But the theoretical nulls are usually quite different from the empirical nulls for the data arising from microarray experiments (see Efron 2004b for more discussion on the choice of the null in multiple testing). We take the approach of Jin and Cai (2007) to estimate the null distribution as $\mathrm{N}(\hat{\mu}_0, \hat{\sigma}_0^2)$. The estimates $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ are consistent. We then proceed to estimate the proportion of the nonnulls $\hat{p}$ 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 $\widehat{\mathrm{Lfdr}}(z_i) = (1 - \hat{p})\hat{f}_0(z_i)/\hat{f}(z_i)$. 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 shown in Figure 5(b), after transformation, the distribution of the transformed p values is approximately uniform when the null is true.

We compare the BH95, BHGW, and AdaptZ procedures using both the theoretical nulls and estimated nulls. We calculate the number of rejections for each mFDR level; the results are shown in Figure 6. For Figure 6(a), f_0 is chosen to be the theoretical null $\mathrm{N}(0, 1)$, and the estimate for the proportion of nulls is 1. Therefore, the BH and BHGW procedures yield the same number of rejections. For Figure 6(b), the estimated null distribution is $\mathrm{N}(-.08, .77^2)$, with estimated proportion of nulls $\hat{p}_0 = .94$. 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:

a. The number of rejections is increasing as a function of the mFDR.

b. For both the p value–based and z value–based approaches, more hypotheses are rejected by using the estimated null.

c. Both comparisons show that AdaptZ is more powerful than BH95 and BHGW, which are based on p values.

7. DISCUSSION

We have developed a compound decision theory framework for inference about the mixture model (1) and showed that the oracle procedure $\boldsymbol{\delta}(T_{OR}, \lambda_{OR})$, given in (6), is optimal in the multiple-testing problems for mFDR control. We have proposed an adaptive procedure based on the z values that mimics the oracle procedure (6). The adaptive z value–based procedure attains the performance of the oracle procedure asymptotically and outperforms the traditional p value–based approaches. The decision-theoretic framework provides insight into the superiority of the adaptive z value–based procedure. Each multiple-testing procedure involves two steps, ranking and thresholding. The process of ranking the relative importance of the m tests can be viewed as a compound decision problem, $\mathbf{P} = (P_{ij}, 1 \le i < j \le m)$, with $m(m-1)/2$ components. Each component problem P_{ij} is a pairwise comparison with its own data, $\mathbf{y}_{ij} = (x_i, x_j)$. Then the solution to $\mathbf{P}$ based on p values is *simple* (note that $\delta_{ij} = I[p(x_i) < p(x_j)]$ depends on $\mathbf{y}_{ij}$ alone), whereas the rule based on the estimated Lfdr is *compound*. The gain in efficiency of the adaptive z value–based procedure is due to the fact that the scope of attention is extended from the

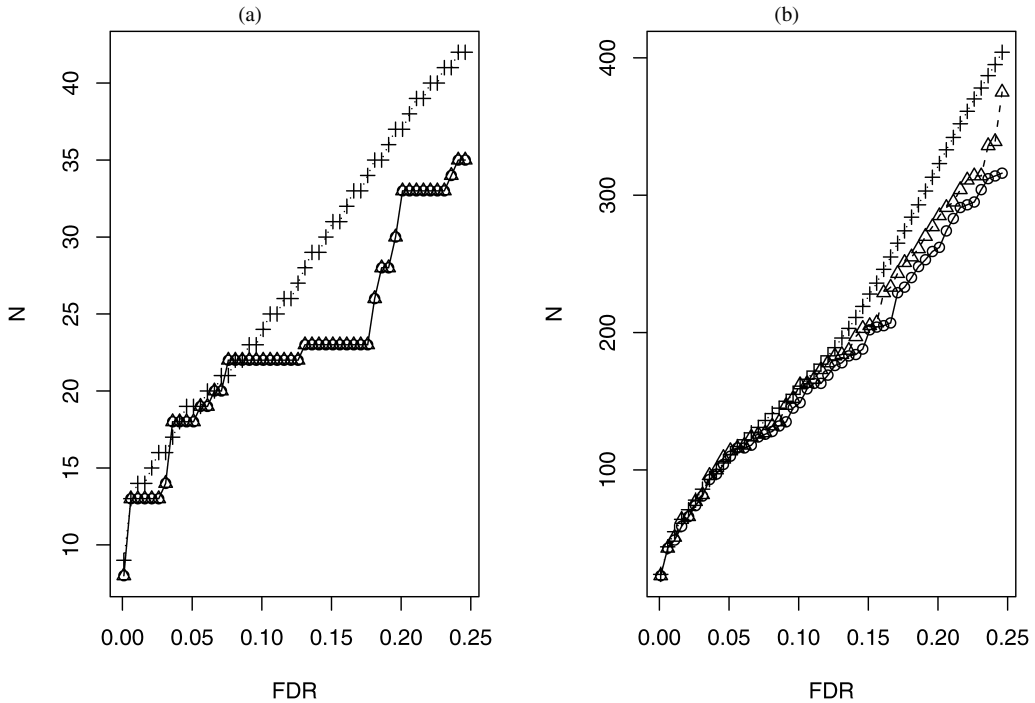

Figure 6. Analysis of the HIV data: Number of rejections versus FDR levels for the theoretical null (a) and the estimated null (b) ∘, the BH step-up procedure based on the p values; △, the BHGW adaptive procedure based on the p values; +, the adaptive procedure based on the z values.

class of simple rules to the class of compound rules in the ranking process.

The compound decision-theoretic approach to multiple testing also suggests several future research directions. First, the oracle statistic (Lfdr) may no longer be optimal in ranking the significance of tests when the observations are correlated. In addition, different cutoffs should be chosen under dependency because the multiple-testing procedures developed for independent tests can be too liberal or too conservative for dependent tests (Efron 2006). Second, the oracle procedure is only optimal in the class of symmetric decision rules. We expect to develop more general models in the compound decision-theoretic framework and extend our attention to asymmetric rules to further increase the efficiency of the multiple-testing procedures. Finally, Genovese, Roeder, and Wasserman (2005), Wasserman and Roeder (2006), Rubin, Dudoit, and van der Laan (2006), and Foster and Stine (2006) have discussed methods that use different cutoffs for p values for different tests when prior information is available or some correlation structure is known. It would be interesting to construct z value–based testing procedures that also exploit external information. We leave this topic for future research.

APPENDIX: PROOFS

Proof of Proposition 1

To prove part (a), note the MLR assumption (4) implies that

$$\int_0^c g_0(t)\,dt \Big/ \int_0^c g_1(t)\,dt = \int_0^c g_0(t)\,dt \Big/ \int_0^c \frac{g_1(t)}{g_0(t)} g_0(t)\,dt < \int_0^c g_0(t)\,dt \Big/ \int_0^c \frac{g_1(c)}{g_0(c)} g_0(t)\,dt = \frac{g_0(c)}{g_1(c)},$$

which is equivalent to $g_1(c)G_0(c) < g_0(c)G_1(c)$. Thus the derivative of $P(\theta_i = 1|T(X_i) \le c) = pG_1(c)/G(c)$ is $(1-p)p\{g_1(c)G_0(c) - g_0(c)G_1(c)\}/\{G(c)\}^2 < 0$. Therefore, $P(\theta_i = 1|T(X_i) \le c)$ is decreasing in c. Parts (b) and (c) can be proved similarly to part (a) by noting that $\mathrm{mFDR} = p_0G_0(c)/G(c)$ and $\mathrm{mFNR} = p_1[1 - G_1(c)]/[1 - G(c)]$. Part (d) follows from parts (b) and (c). For part (e), the classification risk of δ is $R_\lambda = E[L_\lambda(\theta, \delta)] = \lambda(1-p)G_0(c) + p\{1 - G_1(c)\}$. The optimal cutoff c^* that minimizes R_λ satisfies $[g_0(c^*)/g_1(c^*)] = [p/\lambda(1-p)]$. Equivalently, $\lambda = pg_1(c^*)/[(1-p)g_0(c^*)]$. Therefore, λ (c^*) is monotone decreasing in c^* (λ) if (4) holds. Note that $r = G(c^*)$ and G is increasing; we conclude that r is decreasing in λ.

Proof of Theorem 1

According to part (b) of Proposition 1, for any $T \in \mathcal{T}$ and mFDR level α, there exists a unique threshold c such that the mFDR is controlled at level α by $\boldsymbol{\delta}(T, c) = \{I[T(x_1) < c], \ldots, I[T(x_m) < c]\}$. Let r be the expected number of rejections of $\delta(T, c)$. Then, again by part (b) of Proposition 1, there exists a unique threshold c^* such that the expected number of rejections of $\boldsymbol{\delta}(\Lambda, c^*)$ is also r. Next, according to part (e) of Proposition 1, there exists a unique $\lambda(\alpha)$ with respect to the choice of c^* such that the classification risk with $\lambda(\alpha)$ as the weight is minimized by $\boldsymbol{\delta}(\Lambda, c^*)$, and the expected number of subjects classified to the nonnull population is r.

Suppose that among the r rejected hypotheses, there are v_L from the null and k_L from the nonnull when $\boldsymbol{\delta}(L, c_L)$ is applied, where $(L, c_L) = (T, c)$ or (Λ, c^*). Then $r = v_T + k_T = v_\Lambda + k_\Lambda$. Now we argue that $v_T \ge v_\Lambda$ and $k_T \le k_\Lambda$. If not, then suppose that $v_T < v_\Lambda$ and $k_T > k_\Lambda$. Note that the classification risk can be expressed as $R_{\lambda(\alpha)} = p + \frac{1}{m}\{\lambda(\alpha)v_L - k_L\}$, and it is easy to see that $\delta(T, c)$ yields a lower classification risk than $\boldsymbol{\delta}(\Lambda, c^*)$. This contradicts the fact that $\boldsymbol{\delta}(\Lambda, c^*)$ minimizes the classification risk for the choice of $\lambda(\alpha)$. Therefore, we must have that $v_T \ge v_\Lambda$ and $k_T \le k_\Lambda$.

Let mFDR_L and mFNR_L be the mFDR and mFNR of $\boldsymbol{\delta}(L, c_L)$, where $L = T, \Lambda$. Now we apply both $\boldsymbol{\delta} = \boldsymbol{\delta}(T, c)$ and $\boldsymbol{\delta}^* = \boldsymbol{\delta}(\Lambda, c^*)$ in the multiple-testing problem. Then the mFDR using rule $\boldsymbol{\delta}^*$ is $\mathrm{mFDR}_\Lambda = v_\Lambda/r \le v_T/r = \mathrm{mFDR}_T = \alpha$, and the mFNR using $\boldsymbol{\delta}^*$ is $\mathrm{mFNR}_\Lambda = [(m_1 - k_\Lambda)/(m-r)] \le [(m_1 - k_T)/(m-r)] = \mathrm{mFNR}_T$. Therefore, $\boldsymbol{\delta}^*$ is the testing rule in class $\mathcal{D}_S$ that controls the mFDR at level α with the smallest mFNR.

Proof of Theorem 2

The joint distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is $\pi(\boldsymbol{\theta}) = \prod_i (1-p)^{1-\theta_i} p^{\theta_i}$. The posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{x}$ can be calculated as $P_{\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x}) = \prod_i P_{\theta_i|X_i}(\theta_i|x_i)$, where

$$P_{\theta_i|X_i}(\theta_i|x_i) = \frac{I(\theta_i = 0)(1-p) f_0(x_i) + I(\theta_i = 1) p f_1(x_i)}{(1-p) f_0(x_i) + p f_1(x_i)}. \quad \text{(A.1)}$$

Let $I_i = I(\theta_i = 1)$ and $\mathcal{I} = \{(I_1, \ldots, I_m) : I_i = 0 \text{ or } 1\}$ be collections of all m-binary vectors; then $|\mathcal{I}| = 2^m$. The elements in $\mathcal{I}$ are designated by superscripts, $\mathbf{I}^{(j)}$, $j = 1, \ldots, 2^m$. Note that $\sum_{\theta_i=0}^{1} [I(\theta_i = 0)(1-p) f_0(x_i) + I(\theta_i = 1) p f_1(x_i)][\lambda I(\theta_i = 0)\delta_i + I(\theta_i = 1)(1-\delta_i)] = \lambda(1-p) f_0(x_i)\delta_i + p f_1(x_i)(1-\delta_i)$; then the posterior risk is

$$\begin{aligned} & E_{\boldsymbol{\theta}|\mathbf{X}} L(\boldsymbol{\theta}, \boldsymbol{\delta}) \\ &= \sum_{j=1}^{2^m} L\big(\boldsymbol{\theta} = \mathbf{I}^{(j)}, \delta\big) \prod_i P_{\theta_i = I_i^{(j)}|X_i}\big(\theta_i = I_i^{(j)}|x_i\big) \\ &= \frac{1}{m} \sum_i \sum_{\theta_i=0}^{1} P_{\theta_i|X_i}(\theta_i|x_i)\{\lambda I(\theta_i = 0)\delta_i + I(\theta_i = 1)(1-\delta_i)\} \\ &= \frac{1}{m} \sum_i \frac{p f_1(x_i)}{f(x_i)} + \frac{1}{m} \sum_i \frac{\lambda(1-p) f_0(x_i) - p f_1(x_i)}{f(x_i)} \delta_i. \end{aligned}$$

The Bayes rule is the simple rule $\boldsymbol{\delta}^\pi(\mathbf{x}) = [\delta_1^\pi(x_i), \ldots, \delta_m^\pi(x_m)]$, where $\delta_i^\pi = I[\lambda(1-p) f_0(x_i) < p f_1(x_i)]$. The expected misclassification risk of $\boldsymbol{\delta}^\pi$ is

$$\begin{aligned} E E_{\boldsymbol{\theta}|\mathbf{X}} L(\boldsymbol{\theta}, \boldsymbol{\delta}^\pi) &= E \frac{p f_1(X)}{f(X)} - E\Bigg[\frac{\lambda(1-p) f_0(X) - p f_1(X)}{f(X)} \\ &\qquad \times I\{\lambda(1-p) f_0(X) < p f_1(X)\}\Bigg] \\ &= p + \int_{\lambda(1-p) f_0 < p f_1} [\lambda(1-p) f_0(x) - p f_1(x)]\, dx. \end{aligned}$$

Proof of Theorem 3

Observe that for each t, if $p_{(i)} \le t < p_{(i+1)}$, then the number of rejections is i. In addition, the ratio $(1-\hat{p})t/\mathbb{G}_m(t)$ increases in t over the range on which the number of rejections is a constant, which implies that if $t = p_{(i)}$ does not satisfy the constraint, then neither does any choice of t between $p_{(i)}$ and $p_{(i+1)}$. Thus it is sufficient to investigate the thresholds that are equal to one of the $p_{(i)}$'s. Then the ratio $(1-\hat{p})t/\mathbb{G}_m(t)$ becomes $(m/i)(1-\hat{p})p_{(i)}$. Therefore, the plug-in threshold is given by $\sup\{p_{(i)} : (m/i)(1-\hat{p})p_{(i)} \le \alpha\}$, which is equivalent to choosing the largest i such that $p_{(i)} \le i\alpha/[m(1-\hat{p})]$.

Proof of Theorem 4

To prove this theorem, we need to state the following lemmas.

Lemma A.1. Let $\hat{p}$, $\hat{f}$, and $\hat{f}_0$ be estimates such that $\hat{p} \xrightarrow{p}$, $E\|\hat{f} - f\|^2 \to 0$, $E\|\hat{f}_0 - f_0\|^2 \to 0$, and then $E\|\hat{T}_{OR} - T_{OR}\|^2 \to 0$.

Proof. Note that f is continuous and positive on the real line, then there exists $K_1 = [-M, M]$ such that $\Pr(z \in K_1^c) \to 0$ as $M \to \infty$. Let $\inf_{z \in K_1} f(z) = l_0$ and $A_\varepsilon^f = \{z : |\hat{f}(z) - f(z)| \ge l_0/2\}$. Note that $E\|\hat{f} - f\|^2 \ge (l_0/2)^2 P(A_\varepsilon^f)$; then $\Pr(A_\varepsilon^f) \to 0$. Thus f and $\hat{f}$ are bounded below by a positive number for large m except for an event that has a low probability. Similar arguments can be applied to the upper bound of $\hat{f}$ and f, as well as to the upper and lower bounds for $\hat{f}_0$ and f_0. Therefore, we conclude that f_0, $\hat{f}_0$, f, and $\hat{f}$ are all bounded in the interval $[l_a, l_b]$, $0 < l_a < l_b < \infty$ for large m except for an event, say A_ε, that has algebraically low probability. Therefore, $0 < l_a < \inf_{z \in A_\varepsilon} \min\{f_0, \hat{f}_0, f, \hat{f}\} < \sup_{z \in A_\varepsilon^c} \max\{f_0, \hat{f}_0, f, \hat{f}\} < l_b < \infty$.

Noting that $\hat{T}_{OR} - T_{OR} = [\hat{f}_0 f(p - \hat{p}) + (1-p) f(\hat{f}_0 - f_0) + (1-p) f_0(f - \hat{f})]/(\hat{f} f)$, we conclude that $(\hat{T}_{OR} - T_{OR})^2 \le c_1(p - \hat{p})^2 + c_2(\hat{f}_0 - f_0)^2 + c_3(\hat{f} - f)^2$ in A_ε^c. It is easy to see that $\|\hat{T}_{OR} - T_{OR}\|^2$ is bounded by L. Then $E\|\hat{T}_{OR} - T_{OR}\|^2 \le L \Pr(A_\varepsilon) + c_1 E(\hat{p}_0 - p)^2 + c_2 E\|\hat{f} - f\|^2 + c_3 E\|\hat{f}_0 - f_0\|^2$. Note that $E(\hat{p}_0 - p)^2 \to 0$ by lemma 2.2 of van der Vaart (1998), and $E\|\hat{f} - f\|^2 \to 0$, $E\|\hat{f}_0 - f_0\|^2 \to 0$ by assumption; then we have that for a given $\varepsilon > 0$, there exists $M \in \mathbb{Z}^+$ such that we can find A_ε, $\Pr(A_\varepsilon) < \varepsilon/(4L)$, and at the same time $E(\hat{p}_0 - p)^2 < \varepsilon/(4c_1)$, $E\|\hat{f} - f\|^2 < \varepsilon/(4c_2)$, and $E\|\hat{f}_0 - f_0\|^2 < \varepsilon/(4c_3)$ for all $m \ge M$. Consequently, $E\|\hat{T}_{OR} - T_{OR}\|^2 < \varepsilon$ for $m \ge M$, and the result follows.

Lemma A.2. $E\|\hat{T}_{OR} - T_{OR}\|^2 \to 0$ implies that $\hat{T}_{OR}(Z_i) \xrightarrow{p} T_{OR}(Z_i)$.

Proof. Let $A_\varepsilon = \{z : |\hat{T}_{OR}(z) - T_{OR}(z)| \ge \varepsilon\}$. Then $\varepsilon^2 \Pr(A_\varepsilon) \le E\|\hat{T}_{OR} - T_{OR}\|^2 \to 0$. Consequently, $\Pr(A_\varepsilon) \to 0$. Therefore, $\Pr(|\hat{T}_{OR}(Z_i) - T_{OR}(Z_i)| \ge \varepsilon) \le \Pr(A_\varepsilon) + \Pr(\{|\hat{T}_{OR}(Z_i) - T_{OR}(Z_i)| \ge \varepsilon\} \cap A_\varepsilon^c) = \Pr(A_\varepsilon) \to 0$, and the result follows.

Lemma A.3. For $\alpha < t < 1$, $\mathrm{E}[1\{\hat{T}_{OR}(Z_i) < t\}\hat{T}_{OR}(Z_i)] \to \mathrm{E}[1\{T_{OR}(Z_i) < t\}T_{OR}(Z_i)]$.

Proof. $\hat{T}_{OR}(Z_i) \xrightarrow{p} T_{OR}(Z_i)$ (Lemma A.2) implies that $\hat{T}_{OR}(Z_i) \xrightarrow{d} T_{OR}(Z_i)$. Let $h(x) = 1\{x < t\}x$; then $h(x)$ is bounded and continuous for $x < t$. By lemma 2.2 of van der Vaart (1998), $\mathrm{E}[h(\hat{T}_{OR}(Z_i))] \to \mathrm{E}[h(T_{OR}(Z_i))]$, and the result follows.

Lemma A.4. Construct the empirical distributions $\hat{G}_{OR}(t) = \frac{1}{m} \times \sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) \le t\}$ and $\hat{G}_{PI}^0(t) = \frac{1}{m} \sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) \le t\} \times \hat{T}_{OR}(z_i)$. Define $\hat{Q}_{OR}(t) = \hat{G}_{OR}^0(t)/\hat{G}_{OR}(t)$, the estimated mFDR. Then for $\alpha < t < 1$, $\hat{Q}_{OR}(t) \xrightarrow{p} Q_{OR}(t)$.

Proof. Let $\rho_m = \mathrm{cov}[\hat{T}_{OR}(Z_i), \hat{T}_{OR}(Z_j)]$, where Z_i and Z_j are two independent random variables from the mixture distribution F. $\hat{T}_{OR}(Z_i) \xrightarrow{p} T_{OR}(Z_i)$ (Lemma A.2) implies that $E[\hat{T}_{OR}(Z_i) \times \hat{T}_{OR}(Z_j)] \to E[T_{OR}(Z_i)T_{OR}(Z_j)]$. Therefore, $\rho_m = \mathrm{cov}[\hat{T}_{OR}(Z_i), \hat{T}_{OR}(Z_j)] \to \mathrm{cov}[T_{OR}(Z_i), T_{OR}(Z_j)] = 0$. Let $\sigma_m^2 = \mathrm{var}(\hat{T}_{OR}(Z_1))$; then $\sigma_m^2 \le E[\hat{T}_{OR}(Z_1)]^2 \le 1$.

Let $\mu_m = \Pr\{\hat{T}_{OR}(Z_i) < t\}$ and $S_m = \sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) \le t\}$. Note that $\mathrm{var}(S_m)/m^2 = (1/m)\sigma_m^2 + [(m-1)/m]\rho_m \le 1/m + \rho_m \to 0$. According to the weak law of large numbers for triangular arrays, we have that $(1/m)\sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) \le t\} - \mu_m \xrightarrow{p} 0$. Also note that $\mu_m = \Pr\{\hat{T}_{OR}(Z_i) < t\} \to \Pr\{T_{OR}(Z_i) < t\} = G_{OR}(t)$, we conclude that $\hat{G}_{OR}(t) \xrightarrow{p} G_{OR}(t)$. Next, we let $v_m = \mathrm{E}[1\{\hat{T}_{OR}(Z_i) \le t\}\hat{T}_{OR}(Z_i)]$. Similarly, we can prove that $\hat{G}_{OR}^0(t) = \frac{1}{m} \times \sum_{i=1}^m 1\{\hat{T}_{OR}(z_i) \le t\}\hat{T}_{OR}(z_i) - v_m \xrightarrow{p} 0$. Note that by Lemma A.2, we have $\mathrm{E}[1\{\hat{T}_{OR}(Z_i) < t\}\hat{T}_{OR}(Z_i)] \to \mathrm{E}[1\{T_{OR}(Z_i) < t\}T_{OR}(Z_i)]$; thus $v_m \to \mathrm{E}(1\{T_{OR}(Z_i) < t\}T_{OR}(Z_i)) = (1-p)\int 1\{T_{OR}(Z_i) < t\} f_0 dt = (1-p)G_{OR}^0(t)$ and $\hat{G}_{OR}^0(t) \xrightarrow{p} (1-p)G_{OR}^0(t)$. Finally, note that for $t > \alpha$, $G_{OR}(t)$ is bounded away from 0, and we obtain $\hat{Q}_{OR}(t) \xrightarrow{p} (1-p)G_{OR}^0(t)/G_{OR}(t) = Q_{OR}(t)$.

Lemma A.5. Define the estimate of the plug-in threshold $\hat{\lambda}_{OR} = \sup\{t \in (0,1) : \hat{Q}_{OR}(t) \le \alpha\}$. If $\hat{Q}_{OR}(t) \xrightarrow{p} Q_{OR}(t)$, then $\hat{\lambda}_{OR} \xrightarrow{p} \lambda_{OR}$.

Proof. Note that $\hat{Q}_{OR}(t)$ is not continuous, and the consistency of $\hat{Q}_{OR}(t)$ does not necessarily imply the consistency of $\hat{\lambda}_{OR}$. Thus we first construct an envelope for $\hat{Q}_{OR}(t)$ using two continuous random functions $\hat{Q}^-_{OR}(t)$ and $\hat{Q}^+_{OR}(t)$ such that for $\mathrm{Lfdr}_{(k)} < t < \mathrm{Lfdr}_{(k+1)}$,

$$\hat{Q}^-_{OR}(t) = \hat{Q}_{OR}\big(\mathrm{Lfdr}_{(k-1)}\big)\frac{t - \mathrm{Lfdr}_{(k)}}{\mathrm{Lfdr}_{(k+1)} - \mathrm{Lfdr}_{(k)}} + \hat{Q}_{OR}\big(\mathrm{Lfdr}_{(k)}\big)\frac{\mathrm{Lfdr}_{(k+1)} - t}{\mathrm{Lfdr}_{(k+1)} - \mathrm{Lfdr}_{(k)}}$$

and

$$\hat{Q}^+_{OR}(t) = \hat{Q}_{OR}\big(\mathrm{Lfdr}_{(k)}\big)\frac{t - \mathrm{Lfdr}_{(k)}}{\mathrm{Lfdr}_{(k+1)} - \mathrm{Lfdr}_{(k)}} + \hat{Q}_{OR}\big(\mathrm{Lfdr}_{(k+1)}\big)\frac{\mathrm{Lfdr}_{(k+1)} - t}{\mathrm{Lfdr}_{(k+1)} - \mathrm{Lfdr}_{(k)}}.$$

Noting that $\hat{Q}_{OR}(\mathrm{Lfdr}_{(k+1)}) - \hat{Q}_{OR}(\mathrm{Lfdr}_{(k)}) = [k\mathrm{Lfdr}_{(k+1)} - \sum_{i=1}^{k}\mathrm{Lfdr}_{(i)}]/[k(k+1)] > 0$, we have that $\hat{Q}^-_{OR}(t) \le \hat{Q}_{OR}(t) \le \hat{Q}^+_{OR}(t)$, with both $\hat{Q}^-_{OR}(t)$ and $\hat{Q}^+_{OR}(t)$ strictly increasing in t. Let $\hat{\lambda}^-_{OR} = \sup\{t \in (0,1) : \hat{Q}^-_{OR}(t) \le \alpha\}$ and $\hat{\lambda}^+_{OR} = \sup\{t \in (0,1) : \hat{Q}^+_{OR}(t) \le \alpha\}$; then $\hat{\lambda}^+_{OR} \le \hat{\lambda}_{OR} \le \hat{\lambda}^-_{OR}$.

We claim that $\hat{\lambda}^-_{OR} \xrightarrow{p} \lambda_{OR}$. If not, then there must exist ε_0 and η_0 such that for all $m \in \mathbb{Z}^+$, $\Pr(|\hat{\lambda}^-_{OR} - \lambda_{OR}| > \varepsilon_0) > 2\eta_0$ holds for some $m \ge M$. Suppose that $\Pr(\hat{\lambda}^-_{OR} > \lambda_{OR} + \varepsilon_0) > 2\eta_0$. It is easy to see that $\hat{Q}^-_{OR} \xrightarrow{p} Q_{OR}$, because $\hat{Q}^-_{OR}(t) - \hat{Q}^+_{OR}(t) \xrightarrow{\text{a.s.}} 0$, $\hat{Q}^-_{OR}(t) \le \hat{Q}_{OR}(t) \le \hat{Q}^+_{OR}(t)$, and $\hat{Q}_{OR}(t) \xrightarrow{p} Q_{OR}(t)$. Thus there exists $M \in \mathbb{Z}^+$ such that $\Pr(|\hat{Q}^-_{OR}(\lambda_{OR} + \varepsilon_0) - Q_{OR}(\lambda_{OR} + \varepsilon_0)| < \delta_0) > 1 - \eta_0$ for all $m \ge M$, where we let $2\delta_0 = Q_{OR}(\lambda_{OR} + \varepsilon_0) - \alpha$. Now for the choice of m, there exists event K_{1m} such that $\Pr(K_{1m}) \ge 1 - \eta_0$ and for all outcomes $\omega \in K_{1m}$, $|\hat{Q}^-_{OR}(\lambda_{OR} + \varepsilon_0) - Q_{OR}(\lambda_{OR} + \varepsilon_0)| < \delta_0$. At the same time, there exists event K_{2m} such that $\Pr(K_{2m}) \ge 2\eta_0$ and for all outcomes $\omega \in K_{2m}$, $\hat{\lambda}^-_{OR} > \lambda_{OR} + \varepsilon_0$. Let $K_m = K_{1m} \cap K_{2m}$; then $\Pr(K_m) = \Pr(K_{1m}) + \Pr(K_{2m}) - \Pr(K_{1m} \cup K_{2m}) \ge \Pr(K_{1m}) + \Pr(K_{2m}) - 1 \ge \eta_0$. That is, K_m has positive measure. Then for all outcomes in K_m, $\hat{Q}^-_{OR}(t)$ is continuous and strictly increasing, and we have $\alpha = \hat{Q}^-_{OR}(\hat{\lambda}^-_{OR}) > \hat{Q}^-_{OR}(\lambda_{OR} + \varepsilon_0) > Q_{OR}(\lambda_{OR} + \varepsilon_0) - \delta_0 > \alpha + \delta_0$. This is a contradiction. Therefore, we must have $\hat{\lambda}^-_{OR} \xrightarrow{p} \lambda_{OR}$. Similarly, we can prove $\hat{\lambda}^+_{OR} \xrightarrow{p} \lambda_{OR}$. Note that $\hat{Q}^-_{OR}(t) - \hat{Q}^+_{OR}(t) \xrightarrow{\text{a.s.}} 0$ implies that $\hat{\lambda}^-_{OR} - \hat{\lambda}^+_{OR} \xrightarrow{\text{a.s.}} 0$, the result follows by noting that $\hat{\lambda}^+_{OR} \le \hat{\lambda}_{OR} \le \hat{\lambda}^-_{OR}$.

Proof of Theorem 4. Implementing the adaptive procedure (10) is equivalent to choosing $\hat{T}_{OR}(Z_i)$ as the test statistic and $\hat{\lambda}_{OR}$ as the threshold. The mFDR of decision rule $\delta(\hat{T}_{OR}, \hat{\lambda}_{OR})$ is $\mathrm{mFDR} = E(N_{10})/E(R) = [(1-p)P_{H_0}(\hat{T}_{OR} < \hat{\lambda}_{OR})]/[\mathrm{P}(\hat{T}_{OR} < \hat{\lambda}_{OR})]$. Noting that $\hat{T}_{OR} \xrightarrow{p} T_{OR}$ by Lemma A.2 and $\hat{\lambda}_{OR} \xrightarrow{p} \lambda_{OR}$ by Lemma A.5, it follows that $(1-p)P_{H_0}(\hat{T}_{OR} < \hat{\lambda}_{OR}) \to (1-p)P_{H_0}(T_{OR} < \lambda_{OR}) = (1-p)G^0_{OR}(\lambda_{OR})$ and $P(\hat{T}_{OR} < \hat{\lambda}_{OR}) \to P(T_{OR} < \lambda_{OR}) = G_{OR}(\lambda_{OR})$. Noting that $P(T_{OR} < \lambda_{OR})$ is bounded away from 0, we have that $\mathrm{mFDR} \to (1-p)G^0_{OR}(\lambda_{OR})/G_{OR}(\lambda_{OR}) = Q_{OR}(\lambda_{OR}) = \alpha$.

Proof of Theorem 5

Similar to the proof of Theorem 4, the mFNR of the adaptive procedure (10) is $\mathrm{mFNR} = E(N_{01})/E(S) = P_{H_1}(\hat{T}_{OR} > \hat{\lambda}_{OR})/P(\hat{T}_{OR} > \hat{\lambda}_{OR})$. By Lemmas A.2 and A.5, we have that $P_{H_1}(\hat{T}_{OR} > \hat{\lambda}_{OR}) \to P_{H_1}(T_{OR} > \lambda_{OR}) = \tilde{G}^1_{OR}(\lambda_{OR})$ and $P(\hat{T}_{OR} > \hat{\lambda}_{OR}) \to P(T_{OR} > \lambda_{OR}) = \tilde{G}_{OR}(\lambda_{OR})$. Noting that $\tilde{G}_{OR}(\lambda_{OR})$ is bounded away from 0, we obtain $\mathrm{mFNR} \to p\tilde{G}^1_{OR}(\lambda_{OR})/\tilde{G}_{OR}(\lambda_{OR}) = \tilde{Q}_{OR}(\lambda_{OR})$.

[Received November 2006. Revised March 2007.]

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

——— (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.

Copas, J. (1974), "On Symmetric Compound Decision Rules for Dichotomies," *The Annals of Statistics*, 2, 199–204.

Efron, B. (2004a), "Local False Discovery Rate," technical report, Stanford University, Dept. of Statistics, available at *http://www-stat.stanford.edu/~brad/papers/False.pdf*.

——— (2004b), "Large-Scale Simultaneous Hypothesis Testing: the Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104.

——— (2006), "Correlation and Large-Scale Simultaneous Testing," technical report, Stanford University, Dept. of Statistics, available at *http://www-stat.stanford.edu/~brad/papers/Correlation-2006.pdf*.

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.

Foster, D., and Stine, R. (2006), "Alpha Investing: A New Multiple Hypothesis Testing Procedure That Controls mFDR," technical report, University of Pennsylvania, Dept. of Statistics.

Genovese, C., and Wasserman, L. (2002), "Operating Characteristic and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society*, Ser. B, 64, 499–517.

——— (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061.

Genovese, C., Roeder, K., and Wasserman, L. (2005), "False Discovery Control With p-Value Weighting," *Biometrika*, 93, 509–524.

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.

Magder, L., and Zeger, S. (1996), "A Smooth Nonparametric Estimate of a Mixing Distribution Using Mixtures of Gaussians," *Journal of the American Statistical Association*, 91, 1141–1151.

Robbins, H. (1951), "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," in *Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA: University of California Press, pp. 131–148.

Rubin, D., Dudoit, S., and van der Laan, M. (2006), "A Method to Increase the Power of Multiple Testing Procedures Through Sample Splitting," working paper, University of California, Berkeley, Dept. of Biostatistics, available at *http://www.bepress.com/ucbbiostat/paper171/*.

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

Storey, J. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 64, 479–498.

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the Q-Value," *The Annals of Statistics*, 31, 2012–2035.

——— (2007), "The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing," *Journal of the Royal Statistical Society*, Ser. B, 69, 347–368.

van der Laan, M., Dudoit, S., and Pollard, K. (2004), "Multiple Testing, Part III: Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives," University of California Berkeley, Dept. of Biostatistics.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

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.

Wasserman, L., and Roeder, K. (2006), "Weighted Hypothesis Testing," technical report, Carnegie Mellon University, Dept. of Statistics, available at *http://www.arxiv.org/abs/math.ST/0604172*.