Harvard University
Harvard University Biostatistics Working Paper Series

Year 2007 *Paper* 63

Conservative Estimation of Optimal Multiple Testing Procedures

James E. Signorovitch*

*Harvard University, James_Signorovitch@hms.harvard.edu

Conservative Estimation of Optimal Multiple Testing Procedures

James E. Signorovitch
Department of Biostatistics
Harvard University, Boston, MA 02115
email: jsignoro@hsph.harvard.edu

March 20, 2007

Abstract

Multiple hypothesis testing is studied under a two-level hierarchical model. The parameters of interest follow an unknown distribution in the lower level of the model and govern the distribution of the observed data at the top level. Multiple testing is viewed as the joint problem of (1) estimating a rejection region in the possibly high-dimensional space occupied by the observed data and (2) estimating the false discovery rate in the estimated rejection region. Optimal rejection regions, that maximize power for a given rate of false discoveries, depend on the unknown data-generating distribution and are generally not identifiable. By expressing optimal rejection regions as functions of certain sufficient statistics we define conservatively optimal rejection regions that are identifiable. A simple algorithm is described for conservative estimation of optimal rejection regions under the general hierarchical model, and implemented in detail for the case in which observed data follow a general linear model. Proposed testing procedures are evaluated through simulations and applications to gene expression data and are found to outperform the estimated 'Optimal Discovery Procedure' (Storey 2005, Storey *et. al* 2006) and the 'Empirical Alternative Hypothesis' (Signorovitch 2006).

1 Introduction

The statistical theory of multiple hypothesis testing has provided a valuable framework for the discovery of interesting genes in large-scale microarray experiments (Tusher *et al.* 2001, Storey and Tibshirani 2003). In a typical experiment aimed at identifying differentially expressed genes across several tissue types, the null hypothesis of constant mean expression across tissues is tested for every gene. If a gene's true differential expression leads to the rejection of this null hypothesis we have a *true positive*, whereas if the null is rejected for a gene that truly has constant mean expression across tissues we have a *false positive*.

An ongoing statistical challenge has been to extract as much relevant information as possible from gene expression data so as to increase the rate of true positives while controlling the rate of false positives.

A fundamental insight is that multiple testing for differential expression can benefit from the combination of information across genes (Efron *et al.* 2001, Tusher *et al.* 2001, Storey 2005, Signorovitch 2006). In this paper, information is combined across genes by building on the following observation.

Consider reducing the data from each gene to a p-value and a statistic S having the property that conditional on any value of S the p-value is uniformly distributed on [0,1] under the null hypothesis. For simplicity, assume that S is binary, taking the value 0 or 1.

Now suppose that when the p-values from thousands of genes are divided into two groups according to S, the p-values tend to be smaller in the $S = 1$ group than in the $S = 0$ group (Figure 1). We can say that the $S = 1$ group provides more *evidence* of differential expression in the following sense. Suppose the null hypothesis is rejected for all p-values less than some threshold a, regardless of S. If the method of Storey *et al.* (2004) is used to estimate upper bounds on the FDR separately in the $S = 0$ and $S = 1$ groups, the estimated bound should be smaller in the $S = 1$ group, as can be seen by applying Storey *et al.*'s (2004) FDR estimate

$$\widehat{FDR} = \frac{a}{1-\xi} \times \frac{\#\{p_i > \xi\}}{\#\{p_i < a\}},$$

with tuning parameter $0 \leq \xi < 1$, to the illustrative data in Figure 1.

In practice S could indicate, for example, the sign of a t-statistic, with the p-value corresponding to a two-sided t-test. S need not be binary. If the

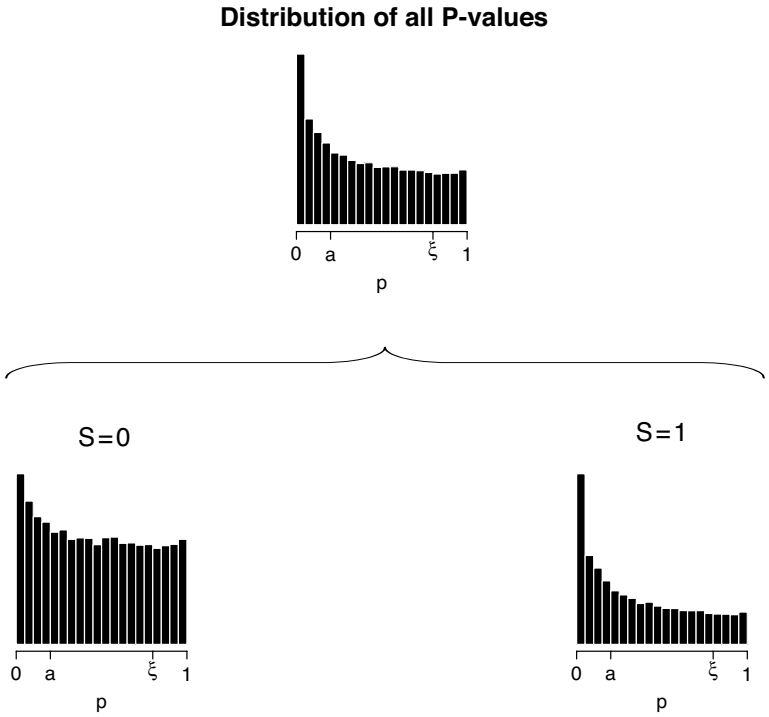

Figure 1: The false discovery rate in the rejection region $[0, a]$ can be bounded below a smaller value when $S = 1$ than when $S = 0$.

p-value corresponds to an F-test, S could depend on the possibly multivariate direction of departure from the null (Signorovitch 2006). Furthermore, since the joint distribution of S and P need not be known under the null, S could also depend on estimates of so-called nuisance parameters. For example when testing for a treatment effect, S could depend on the overall mean response, information that is ignored by the Optimal Discovery Procedure (ODP) (Storey 2005, Storey *et al.* 2006) and the Empirical Alternative Hypothesis (EAH) (Signorovitch 2006). In these examples, the value of S for a single gene in isolation often contains no information regarding differential expression. Only after combining realizations of S and P across many genes do we have enough information to influence our assessment of significance under an FDR criterion.

This paper lays out a framework for combining all relevant information across related hypothesis tests. The sharing of information is justified under

a two-level hierarchical model described in Section 2. A simple FDR-based optimality criterion given in Section 3 leads to a view of multiple testing as the joint problem of using data from all tests to (1) estimate optimal rejection regions and (2) estimate the FDR in the estimated rejection regions. Section 4 shows that optimal rejection regions are in general not identifiable. By making use of certain sufficient statistics, we define in Section 4 near-optimal rejection regions that are identifiable and maximize power under a localized upper bound on the rate of false discoveries. Given the data in Figure 1, a near-optimal rejection region would allow different significance thresholds in the $S = 0$ and $S = 1$ groups to maximize the expected number of rejected hypotheses while controlling the overall FDR.

Consistent estimates of the near-optimal rejection regions provide *conservatively-estimated optimal* (CEO) rejection regions. A general program for obtaining CEO rejection regions under the hierarchical model is described in Section 5. The program is then applied in detail to a two-sample Gaussian setting in Section 6 and extended to the general linear model in Section 7. CEO multiple testing procedures are evaluated in Section 8 through simulations and in Section 9 through application to gene expression data. Connections to other multiple testing procedures are explored in Section 10 and possible extensions of CEO multiple testing are discussed in Section 11.

2 The Probabilistic Setting

Consider the random triple (Y, Θ, H) generated by a semiparametric hierarchical model with

$$\begin{aligned} H &\sim \text{Bernoulli}(1 - p_0) \\ \Theta | H = h &\sim G_h \\ Y | \Theta = \theta &\sim \mathcal{P}_\theta. \end{aligned} \tag{1}$$

The indicator H represents a null hypothesis that is true ($H = 0$) with probability p_0 and false ($H = 1$) otherwise. The state of H in turn governs the distribution of a d-dimensional parameter Θ with $\Theta \sim G_1$ supported on $\mathbb{R}^d$ when $H = 1$ and $\Theta \sim G_0$ supported on a linear subspace $\mathcal{V}_0 \subset \mathbb{R}^d$ when $H = 0$. We suppose that H and Θ are unobserved and that p_0, G_0 and G_1 are unknown. Each realization of Θ specifies a particular instance of the known parametric model $\{\mathcal{P}_\theta : \theta \in \mathbb{R}^d\}$ and we do observe the value $Y \in \mathbb{R}^n$

sampled from $\mathcal{P}_\theta$. We assume that the conditional distributions of Y given $H = 0$ and $H = 1$ have continuous densities f_0 and f_1, respectively, on $\mathbb{R}^n$.

Suppose an experiment generates m independent realizations (y_i, θ_i, h_i), $i = 1, \ldots, m$, of (Y, Θ, H). Given the m observations y_i, $i = 1, \ldots, m$, our objective is to determine as well as possible the values of the unobserved indicators h_i, $i = 1, \ldots, m$, or, equivalently, to test the hypotheses

$$\theta_i \in \mathcal{V}_0, i = 1, \ldots, m,$$

for a known subspace $\mathcal{V}_0 \subset \mathbb{R}^p$.

In most experimental settings the observed data are generated according to fixed parameter values $\theta_1, \ldots, \theta_m$. Modeling these values as realizations of a random variable captures the idea that the observed data $Y_1, \ldots, Y_m$ are related by some underlying phenomena. Inference under the frequentist setting in which the θ_i's and h_i's are fixed is considered in Section 5.

The probabilistic setting described above provides a simple model for a gene expression microarray experiment with m genes on n arrays. Each $n \times 1$ observation vector y_i can represent the expression measurements for the ith gene across the n arrays. Assuming a Gaussian model for gene expression, the gene-specific parameter $\theta_i = (\boldsymbol{\beta}_i, \sigma_i)$ can specify the distribution of Y_i such that

$$Y_i \sim N(\mathbf{X}'\boldsymbol{\beta}_i, \sigma_i^2 \mathbf{I}_n),$$

where the rows of the $n \times d$ matrix $\mathbf{X}$ contain array-specific covariates, for example the tissue type, patient's age, gender etc., whose effects on the mean expression level of the ith gene are given by the $d \times 1$ coefficient vector $\boldsymbol{\beta}_i$. The gene-specific hypothesis h_i may indicate whether or not certain elements of $\boldsymbol{\beta}_i$ are zero. Dependence across genes is considered in the Appendix.

3 Optimal Rejection Regions

We restrict our attention to multiple testing procedures that produce multiple decision rules of the form

$$\text{reject } h_i \text{ if } y_i \in \Gamma, i = 1, \ldots, m,$$

for some rejection region $\Gamma \subseteq \mathbb{R}^n$. These rules ensure that once Γ is specified, acceptance or rejection of h_i depends only on the value of y_i, though of course the selection of a decision rule Γ may depend jointly on all y_i's.

A rejection region Γ is considered optimal at level α if it maximizes, over all subsets of $\mathbb{R}^n$, the probability of rejecting a false null hypothesis given that the frequency of true nulls among the rejected hypotheses is less than α. That is, letting $\mathcal{P}$ denote the joint distribution of (Y, Θ, H), the optimal level-α rejection region is the solution to the constrained optimization problem

$$\begin{aligned} \text{Maximize} \quad &: \quad \mathcal{P}(Y \in \Gamma | H = 1) \\ \text{subject to} \quad &: \quad \mathcal{P}(H = 0 | Y \in \Gamma) \leq \alpha, \end{aligned} \tag{2}$$

where the constrained quantity

$$FDR(\Gamma) \equiv \mathcal{P}(H = 0 | Y \in \Gamma)$$

is the false discovery rate in Γ under the data-generating distribution $\mathcal{P}$. Similarly to the Neyman-Pearson setting for testing a single hypothesis, where the constrained quantity is $\mathcal{P}(Y \in \Gamma | H = 0)$ instead of FDR, optimal rejection regions are likelihood ratio (LR) level sets for the conditional densities of Y given $H = 0$ and $H = 1$.

Theorem 1 *A rejection region $\Gamma^* \subseteq \mathbb{R}$ solves (2) for some $0 \leq \alpha \leq 1$ if and only if*

$$\Gamma^* = \left\{ y \in \mathbb{R}^n : \frac{f_1(y)}{f_0(y)} > \gamma \right\} \ (a.e.) \tag{3}$$

for some $\gamma \geq 0$.

This theorem complements that of Storey (2005), where different optimality criteria led to the same collection of optimal rejection regions. The argument used to prove Theorem 1 in the Appendix also provides the following.

Corollary 1 *A rejection region Γ^* maximizes $\mathcal{P}(Y \in \Gamma)$ over all $\Gamma \subseteq \mathbb{R}^n$ such that $FDR(\Gamma) \leq \alpha$ for some $0 \leq \alpha \leq 1$ if and only if Γ^* is an LR level set.*

Since LR level sets provide optimal rejection regions, asymptotically optimal multiple testing procedures could be based on consistent estimates of LR level sets.

4 Optimal Conservative Rejection Regions

In general, LR level sets are not identifiable since the ratio of f_1 to f_0 is not identifiable, even up to monotone transformations, given only observations of Y. In this section we define identifiable rejection regions that are near-optimal and conservative. References to identifiablilty will always assume that only Y is observed.

We exploit the representation of LR level sets as maximizers of the function

$$R_\gamma(\Gamma) = \mathcal{P}(Y \in \Gamma | H = 1) - \gamma \mathcal{P}(Y \in \Gamma | H = 0). \tag{4}$$

The idea is that if R_γ were identifiable at every Γ, LR level sets could be identified as

$$\Gamma_\gamma = \arg\max_{\Gamma \subseteq \mathbb{R}^n} R_\gamma(\Gamma)$$

and estimated as maximizers of a consistent estimate of R_γ. In this section we find approximations to R_γ that are identifiable.

As in Efron *et al.* (2001) it is convenient to absorb the non-identifiable quantity $\mathcal{P}(Y \in \Gamma | H = 1)$ into the identifiable quantity $\mathcal{P}(Y \in \Gamma)$ and rewrite (4) as

$$Q_\lambda(\Gamma) = \mathcal{P}(Y \in \Gamma) - \lambda \mathcal{P}(Y \in \Gamma, H = 0) \tag{5}$$

with $\lambda = 1 + \gamma(1 - p_0)/p_0 \geq 1$. Still, Q_λ is generally not identifiable because any set Γ's 'false content,'

$$\mathcal{P}(Y \in \Gamma, H = 0) = p_0 \int_{\mathcal{V}_0} \int_\Gamma f_0(y|\theta) dy dG_0(\theta),$$

is not identifiable, even up to a multiplicative constant.

Since we can not consistently estimate Q_λ or its maximizer, a reasonable compromise is to replace the false content of Γ in Q_λ with an identifiable upper bound, since upper bounds on the false content will yield upper bounds on the FDR. The main idea of this paper is that identifiable upper bounds on the false content of any rejection region can be substantially tightened by conditioning on certain sufficient statistics.

Suppose there exists a statistic $S = S(Y)$, supported on $\mathcal{A}$, such that the conditional distribution of Y given S is free of θ for all $\theta \in \mathcal{V}_0$, that is

$$\mathcal{P}(Y | S = s, \Theta = \theta, H = 0) = P_0(Y | S = s), \qquad \text{for all } \theta \in \mathcal{V}_0,$$

where $P_0(Y|S = s)$ is a known distribution for any $s \in \mathcal{A}$. Since S is a sufficient statistic for Θ under the null hypothesis, we call S a *null-sufficient statistic*. Let $P_h(s)$ be the distribution of S conditional on $H = h$, $h = 0, 1$, and let $P(s)$ be the marginal distribution of S. Also define the conditional null frequency as $\pi_0(s) = \mathcal{P}(H = 0|S = s)$.

The objective function Q_λ (5) can now be written as

$$Q_\lambda(\Gamma) = \mathcal{P}(Y \in \Gamma) - \lambda \int P_0(Y \in \Gamma|S = s)\pi_0(s)dP(s). \tag{6}$$

Considering (6) with an eye towards estimation, we notice that $\mathcal{P}(Y \in \Gamma)$ and $P(s)$ can be replaced by their empirical counterparts and that $P_0(Y \in \Gamma|S = s)$ is known since S is null-sufficient. Only the conditional null frequency $\pi_0(s)$ is not identifiable, and it is for this quantity that we will derive identifiable upper bounds.

To see that identifiable upper bounds $\pi_0^u(\cdot) \geq \pi_0(\cdot)$ exist, consider any set-valued function $\Lambda : \mathcal{A} \to \mathbb{R}^n$. Since

$$\mathcal{P}(Y \in \Lambda(s)|S = s) \geq P_0(Y \in \Lambda(s)|S = s)\pi_0(s)$$

for all $s \in \mathcal{A}$ we have, similarly to Storey *et al.* (2004),

$$\pi_0^u(s) \equiv 1 \wedge \frac{\mathcal{P}(Y \in \Lambda(s)|S = s)}{P_0(Y \in \Lambda(s)|S = s)} \geq \pi_0(s), \quad s \in \mathcal{A},$$

in which the numerator is an identifiable function of $s \in \mathcal{A}$ and the denominator is a known function.

An identifiable upper bound on π_0 leads to identifiable upper bounds on FDR since, letting

$$FDR(\Gamma, \eta) = \frac{\int P_0(Y \in \Gamma|S = s)\eta(s)dP(s)}{\mathcal{P}(Y \in \Gamma)} \tag{7}$$

for some function $\eta : \mathcal{A} \to [0, 1]$, we have

$$FDR(\Gamma, \pi_0^u) \geq FDR(\Gamma, \pi_0) = FDR(\Gamma), \quad \Gamma \subseteq \mathbb{R},$$

with $FDR(\Gamma, \pi_0^u)$ identifiable.

An identifiable upper bound on π_0 also leads to identifiable objective functions that approximate Q_λ. Let

$$Q_\lambda(\Gamma, \eta) \equiv \mathcal{P}(Y \in \Gamma) - \lambda \int P_0(Y \in \Gamma|S = s)\eta(s)dP(s) \tag{8}$$

so that $Q_\lambda(\Gamma, \pi_0) = Q_\lambda(\Gamma)$ as defined in (5) and define *conservatively optimal* rejection regions under π_0^u as

$$\Gamma_\lambda^u = \arg\max_{\Gamma \subseteq \mathbb{R}} Q_\lambda(\Gamma, \pi_0^u), \quad \lambda \geq 1. \tag{9}$$

In principle, the maximizer Γ_λ^u is identifiable since $Q_\lambda(\Gamma, \pi_0^u)$ is identifiable.

If $\pi_0^u(s) = \pi_0(s)$ for $\mathcal{P}$-almost all s, the rejection regions $\{\Gamma_\lambda^u\}_{\lambda \geq 0}$ are the optimal LR level sets of Theorem 1 for the true data-generating distribution. However if $\pi_0^u(s)$ and $\pi_0(s)$ differ, as they will in practice, Γ_λ^u is only guaranteed to be an LR level set under the conditional distributions for Y given $H = 0$ and $H = 1$ determined by the true marginal distribution $\mathcal{P}(Y)$ and the conditional null frequency $\pi_0^u(\cdot)$. Thus Γ_λ^u may not satisfy (2) for any α under the true data-generating distribution. However Corollary 1 can be applied to prove the following.

Theorem 2 *A rejection region Γ^* maximizes $\mathcal{P}(Y \in \Gamma)$ over all $\Gamma \subseteq \mathbb{R}^n$ such that $FDR(\Gamma, \pi_0^u) \leq \alpha$ for some $0 \leq \alpha \leq 1$ if and only if $\Gamma^* = \Gamma_\lambda^u$ for some $\lambda \geq 1$.*

Unlike Theorem 1 and Corollary 1, Theorem 2 provides solutions to an optimization problem involving only identifiable quantities. The identifiable region Γ_λ^u may be called *conservatively optimal* in that by Theorem 2 it maximizes the expected number of rejected hypotheses among all rejection regions with equal or smaller upper bounds on the FDR under $\pi_0^u(\cdot)$.

5 Conservative Estimation of Optimal Rejection Regions

This section defines conservatively-estimated optimal (CEO) rejection regions as estimates of conservatively optimal rejection regions and gives theorems for FDR control. We will use the notation $\mathcal{P}Z \equiv \int Z(y) d\mathcal{P}(y)$ for functions Z of y and $\mathcal{P}\Gamma \equiv \int I(y \in \Gamma) d\mathcal{P}(y)$ for sets $\Gamma \subseteq \mathbb{R}^n$. Also define

$$v(\Gamma, \eta)(y) \equiv P_0(Y \in \Gamma | S = S(y)) \eta(S(y))$$

so that (8) can be written as

$$Q_\lambda(\Gamma, \eta) = \mathcal{P}\{\Gamma - \lambda v(\Gamma, \eta)\}.$$

The theoretical developments of the preceding sections involved maximizations over all subsets of $\mathbb{R}^n$. The practical CEO tests described in this paper restrict attention to a small class $\mathcal{G}$ of potential rejection regions. Consequently, we do not need an upper bound on $\pi_0(s)$ for all $s \in \mathcal{A}$. A function $\pi_0^u(\cdot)$ that weakly dominates $\pi_0(\cdot)$ over $\mathcal{G}$ such that

$$\sup_{\Gamma \in \mathcal{G}} \mathcal{P}\{v(\Gamma, \pi_0^u) - v(\Gamma, \pi_0)\} \geq 0 \tag{10}$$

provides an upper bound on the false content of any potential rejection region in $\mathcal{G}$.

To define CEO tests, begin by letting $\mathcal{G}$ be a Vapnik-Chervonenkis (VC) class of potential rejection regions such that $(\mathcal{G}, d)$ is a complete, pathwise connected pseudometric space containing the empty set $\emptyset$ and $\mathbb{R}^n$, with $d(A, B)$ giving the Lebesgue measure of the symmetric difference between sets $A, B \in \mathcal{G}$. In principle $\mathcal{G}$ could be allowed to grow with m, but here we provide asymptotic theory for the simpler case of fixed $\mathcal{G}$. Applications and simulations will show that even for modestly sized classes $\mathcal{G}$, CEO testing offers substantial improvement over other methods.

Suppose we have an estimator $\widehat{\pi}_0^u(\cdot)$ of the conditional null frequency such that

$$\sup_{s \in \mathcal{A}} |\widehat{\pi}_0^u(s) - \pi_0^u(s)| \xrightarrow{a.s.} 0$$

with $\pi_0^u(\cdot)$ weakly dominating $\pi_0(\cdot)$ over $\mathcal{G}$ in the sense of (10). Given $\mathcal{G}$ and $\pi_0^u(\cdot)$, conservatively optimal rejection regions are defined for $\lambda \geq 1$ as

$$\Gamma_\lambda^u = \arg\max_{\Gamma \in \mathcal{G}} \mathcal{P}\{\Gamma - \lambda v(\Gamma, \pi_0^u)\}, \tag{11}$$

and CEO rejection regions are defined as their empirical counterparts,

$$\widehat{\Gamma}_{\lambda,m}^u = \arg\max_{\Gamma \in \mathcal{G}} \mathbb{P}_m\{\Gamma - \lambda v(\Gamma, \widehat{\pi}_0^u)\}, \tag{12}$$

where $\mathbb{P}_m$ is the empirical measure based on $\{y_i\}_{i=1}^m$.

FDR is estimated for any non-empty $\Gamma \in \mathcal{G}$ as the empirical counterpart of (7),

$$\widehat{FDR}_m(\Gamma, \widehat{\pi}_0^u) = \frac{\mathbb{P}_m v(\Gamma, \widehat{\pi}_0^u)}{\mathbb{P}_m \Gamma}. \tag{13}$$

If $\mathbb{P}_m \Gamma = 0$ we set $\widehat{FDR}_m(\Gamma, \widehat{\pi}_0^u) = 0$.

When evaluating rejection regions, we may also be interested in the realized rate of false discoveries

$$rFDR_m(\Gamma) \equiv \frac{\sum_{i=1}^m I(y_i \in \Gamma)(1-h_i)}{1 \vee \sum_{i=1}^m I(y_i \in \Gamma)}.$$

Under technical conditions given in the Appendix, we have the following.

Theorem 3 *For any fixed λ^* such that $\mathcal{P}\Gamma^u_{\lambda^*} = \delta > 0$ we have as $m \to \infty$*

a. $\sup_{\lambda \geq 1} d(\widehat{\Gamma}^u_{\lambda,m}, \Gamma^u_\lambda) \xrightarrow{a.s.} 0$,

b. $\sup_{1 \leq \lambda \leq \lambda^*} FDR(\widehat{\Gamma}^u_{\lambda,m}, \pi_0) - \widehat{FDR}(\widehat{\Gamma}^u_{\lambda,m}, \widehat{\pi}_0) \leq 0$ *w.p.1 and*

c. $\sup_{1 \leq \lambda \leq \lambda^*} rFDR(\widehat{\Gamma}^u_{\lambda,m}) - \widehat{FDR}(\widehat{\Gamma}^u_{\lambda,m}, \widehat{\pi}_0) \leq 0$ *w.p.1.*

The proof of this theorem is given in the Appendix.

Theorems 3a and 3b ensure that the collection of CEO rejection regions and their estimated FDRs can be interpreted simultaneously for $1 \leq \lambda \leq \lambda^*$ as estimated optimal conservative rejection regions. Theorem 3c ensures large-m control of the true proportion of false discoveries. Since the convergence in Theorem 3c occurs for almost every sequence of observations, a frequentist interpretation follows. So long as the fixed sequence $\{(\theta_i, h_i)\}_{i=0}^\infty$ can be thought of as a typical realization from some underlying distribution, rFDR is controlled asymptotically with probability 1. Further discussion of the connection between the Bayesian and Frequentist views of multiple testing can be found in Genovese and Wasserman (2002) and Storey (2002, 2003).

Once the parametric model $\{\mathcal{P}_\theta : \theta \in \Theta\}$ has been specified, the above framework for CEO testing can be applied in five steps:

(i) choose a null-sufficient statistic S,

(ii) choose a class $\mathcal{G}$ of potential rejection regions,

(iii) estimate an identifiable $\pi_0^u(\cdot)$ that weakly dominates $\pi_0(\cdot)$ over $\mathcal{G}$,

(iv) obtain CEO rejection regions via (12) and

(v) conservatively estimate the FDR in the CEO rejection regions via (13).

In the following section we illustrate in detail the application of this program to testing for a difference of means in a two-sample Gaussian model. In Section 7 we extend this application to testing the mean parameter in a general linear model.

6 The Two-Sample Gaussian Problem

Suppose $Y = (Y_1, \ldots, Y_{2n})$ contains n independent normally-distributed observations from each of two groups with common variance σ^2 and possibly different means β_1 and β_2. The unobserved random parameter $\Theta = (\beta_1, \beta_2, \sigma^2)$ follows an unknown distribution as in Section 2. An experiment generates m independent realizations of (Y, H, Θ) yielding the observations $\{y_i\}_{i=1}^m$. The goal is to test for each y_i the null hypothesis that the underlying realizations of β_1 and β_2 are equal.

We summarize Y via the usual two-sample statistics. Let $\widehat{\beta}_1(Y) = n^{-1}\sum_{i=1}^n Y_i$ denote the mean in the first group, let $\widehat{\beta}_2(Y) = n^{-1}\sum_{i=n+1}^{2n} Y_i$ denote the mean in the second group and let $\widehat{\beta}_0(Y) = \{\widehat{\beta}_1(Y) + \widehat{\beta}_2(Y)\}/2$ denote the pooled mean. The usual variance estimate under the null is $s_0^2(Y) = (2n-1)^{-1}\sum_{i=1}^{2n}\{Y_i - \widehat{\beta}_0(Y)\}^2$ and the variance estimate under the alternative is

$$s_1^2(Y) = (2n-2)^{-1}\Big[\sum_{i=1}^{n}\{Y_i - \widehat{\beta}_1(Y)\}^2 + \sum_{i=n+1}^{2n}\{Y_i - \widehat{\beta}_2(Y)\}^2\Big]$$

This setting models a gene expression microarray experiment in which expression levels for m genes are measured on n samples from each of two tissue types. The goal of such experiments is often to identify genes with different mean expression levels across the two tissues.

Step (i): Choose a Null-Sufficient Statistic

We choose the null-sufficient statistic

$$S = \Big[s_0(Y), \widehat{\beta}_0(Y), \mathrm{sign}\{\widehat{\beta}_1(Y) - \widehat{\beta}_2(Y)\}\Big].$$

This is not a minimal sufficient statistic under the null, but this choice will be seen to provide computational advantages without altering the limiting performance. The remaining information in Y regarding θ is captured by the usual F-statistic

$$T = 2^{-1}n\{\widehat{\beta}_1(Y) - \widehat{\beta}_2(Y)\}^2/s_1^2(Y),$$

in the sense that (S, T) is sufficient for θ even under the alternative. For convenience we replace T by its corresponding p-value, P, under an F-distribution with 1 and $2n-2$ degrees of freedom. Without loss of information, we may convert the observations $\{y_i\}_{i=1}^m$ to the sufficient statistics $\{(s_i, p_i)\}_{i=1}^m$

Step (ii): Choose a Class of Potential Rejection Regions

By sufficiency, LR level sets for Y can be defined in terms of (S, P). Since departures from the null hypothesis can only make P stochastically smaller than its Uniform[0,1] distribution under the null, any LR level set Γ can be expressed as a function of S having the form

$$r : \mathbb{R}^+ \times \mathbb{R} \times \{-1, 1\} \to [0, 1]$$

such that $P < r(S)$ if and only if $Y \in \Gamma$. A class $\mathcal{G}$ of potential rejection regions that can approximate LR level sets can now be defined through constraints on the function r. In this paper we use a simple piecewise constant model for r, letting

$$r(S) = \sum_{k=1}^{K} r_k I(S \in B_k)$$

for scalars $r_1, \ldots, r_K$ corresponding to sets $B_1, \ldots, B_K$ partitioning the support of S.

Step (iii): Conservatively Estimate the Conditional Null Frequency

For fixed $0 \leq \xi < 1$, the function

$$\pi_0^u(s) = \sum_{k=1}^{K} I(s \in B_k)\mathcal{P}(P > \xi | S \in B_k)/(1-\xi)$$

weakly dominates $\pi_0(s)$ over $\mathcal{G}$ and can be uniformly consistently estimated by

$$\widehat{\pi}_0^u(s) = \sum_{k=1}^{K} I(s \in B_k)\widehat{\pi}_{k,0}^u$$

with

$$\widehat{\pi}^u_{k,0} = 1 \wedge m_k^{-1} \sum_{i=1}^{m} I(s_i \in B_k) I(p_i > \xi)/(1-\xi)$$

where $m_k = \sum_{i=1}^{m} I(s_i \in B_k)$. As in Storey *et al.* (2004) setting $\xi = 0$ yields the most conservative estimate, while increasing ξ makes the bound tighter but increases its estimation variance.

Step (iv): Estimate Optimal Conservative Rejection Regions

For any value of $\lambda \geq 1$, the optimal conservative rejection region is

$$\Gamma^u_\lambda = \{Y : P < r_\lambda(S)\}$$

with $r_\lambda(S) = \sum_{k=1}^{K} r_{k,\lambda} I(S \in B_k)$ and, simplifying from (9),

$$(r_{1,\lambda}, \ldots, r_{K,\lambda})' = \arg \max_{(r_1,\ldots,r_K)} \mathcal{P}\Big[I\{P < r(S)\} - \lambda r(S) \pi^u_0(S) \Big].$$

Following Theorem 3, this set can be consistently estimated by

$$\widehat{\Gamma}^u_\lambda = \{Y : P < \widehat{r}_\lambda(S)\} \tag{14}$$

with $\widehat{r}_\lambda(S) = \sum_{k=1}^{K} \widehat{r}_{k,\lambda} I(S \in B_k)$ and, simplifying from (12),

$$(\widehat{r}_{1,\lambda}, \ldots, \widehat{r}_{K,\lambda})' = \arg \max_{(r_1,\ldots,r_K)} m^{-1} \sum_{i=1}^{m} \left\{ I\{p_i < r(s_i)\} - \lambda r(s_i) \widehat{\pi}^u_0(s_i) \right\}$$

This optimization problem is solved by a simple algorithm. Let $p(j;k)$ be the jth largest p-value among the group of p_i's having $s_i \in B_k$, $j = 0, \ldots, m_k$, $k = 1, \ldots, K$ and set $p(0;k) = 0$ for all k. The estimated near-optimal rejection region is given by

$$\widehat{r}_{k,\lambda} = p(\widehat{j}_k; k), k = 1, \ldots, K$$

with

$$\widehat{j}_k = \arg\max_{0 \leq j \leq m_k} \{ j - \lambda m_k p(j;k) \widehat{\pi}^u_{0,k} \}. \tag{15}$$

In practice there is no need to repeatedly solve (15) for different values of λ since values of λ that change the solution correspond to vertices of the lower

convex majorant of the empirical distribution of p-values, as can be seen in the R implementation of this method available from the author.

Step (v): Conservatively estimate the FDR for Estimated Rejection Regions

The conservative FDR estimator (13) evaluated at the estimated piecewise constant rejection region (14) simplifies to

$$\widehat{FDR}(\widehat{\Gamma}^u_\lambda, \widehat{\pi}^u_0) = \frac{\sum_{i=1}^K m_k \widehat{r}_{k,\lambda} \widehat{\pi}^u_{0,k}}{1 \vee \sum_{i=1}^m I\{p_i < \widehat{r}_\lambda(s_i)\}}. \tag{16}$$

7 The General Linear Model

Consider the probabilistic setting of Section 2 with $\Theta = (\boldsymbol{\beta}, \sigma)$ and $Y \sim \mathcal{P}_\theta = \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ where $\boldsymbol{\beta}$ is $d_1 \times 1$ and $\sigma > 0$ for a fixed and known $n \times d_1$ covariate matrix $\mathbf{X}$. Without loss of generality, we assume that $\mathbf{X}$ is orthonormal. The null hypothesis has the form $\boldsymbol{\beta} \in \mathcal{V}_0$ for some subspace $\mathcal{V}_0 \subset \mathbb{R}^{d_1}$ with dimension d_0.

Each observation of Y can be summarized without loss of information by the least squares estimate of $\boldsymbol{\beta}$ under the alternative $\widehat{\boldsymbol{\beta}}_1(Y) = \mathbf{X}'Y$, the estimate of $\boldsymbol{\beta}$ under the null, $\widehat{\boldsymbol{\beta}}_0(Y)$, which by orthonormality of $\mathbf{X}$ is the projection of $\widehat{\boldsymbol{\beta}}_1(Y)$ into $\mathcal{V}_0$, the estimated variances under the null and alternative

$$s_i^2(Y) = \|Y - \mathbf{X}\widehat{\boldsymbol{\beta}}_i(Y)\|^2/(n - d_i), \quad i = 0, 1,$$

and the 'direction of departure from the null' (Signorovitch 2006)

$$\phi(Y) = \frac{\widehat{\boldsymbol{\beta}}_1(Y) - \widehat{\boldsymbol{\beta}}_0(Y)}{\|\widehat{\boldsymbol{\beta}}_1(Y) - \widehat{\boldsymbol{\beta}}_0(Y)\|}.$$

Analogously to the two-sample setting of Section 6, we choose the null-sufficient statistic to be

$$S = \left\{ s_0(Y), \widehat{\boldsymbol{\beta}}_0(Y), \phi(Y) \right\}$$

and we summarize the remaining information in Y regarding θ with the usual F-statistic,

$$T = \frac{\|\widehat{\boldsymbol{\beta}}_1(Y) - \widehat{\boldsymbol{\beta}}_0(Y)\|^2/(d_1 - d_0)}{s_1^2(Y)}$$

which is then converted to its p-value, P, under the F-distribution with $d_1 - d_0$ and $n - d_1$ degrees of freedom. By sufficiency, any LR level set contains exactly those values of Y for which $P < r(S)$ for some

$$r : \mathbb{R}^+ \times \mathbb{R}^{d_0} \times \mathcal{S}_{d_1 - d_0} \to [0, 1],$$

where $\mathcal{S}_{d_1-d_0}$ is the surface of the unit ball in $d_1 - d_0$ dimensions supporting $\phi(Y)$. Once the support of S is partitioned into K regions over which r is modeled as constant, as in step (ii) of Section 6, CEO testing procedures can be implemented as in steps (iii) through (v).

8 Simulation Study

This section compares CEO multiple testing to EAH tests (Signorovitch 2006) and the ODP (Storey 2005, Storey *et al.* 2006) under four simulation scenarios described by Storey *et al.* (2006). Briefly, paraphrasing from Storey *et al.* (2006), scenario (a) generates expression data from two tissues with symmetric patterns of differential expression and variances simulated from a unimodal distribution. Scenario (b) introduces some asymmetry in differential expression between the two tissues and simulates variances from a bimodal distribution. Scenario (c) generates data for three tissues with slight asymmetry in differential expression and variances simulated from a unimodal distribution. Scenario (d) introduces stronger asymmetry in differential expression across the three groups and samples variances from a bimodal distribution. Under each scenario, data were simulated for 1000 differentially expressed genes and 2000 non-differentially expressed genes in 8 samples from each tissue.

The ODP was applied using the EDGE software (Leek *et al.* 2006). CEO tests for scenarios (a) and (b) were implemented as described in Section 6 with the support of $S(Y) = [s_0(Y), \widehat{\beta}_0(Y), \text{sign}\{\widehat{\beta}_1(Y) - \widehat{\beta}_2(Y)\}]$ partitioned by first splitting the data into two groups according to $\text{sign}\{\widehat{\beta}_1(Y) - \widehat{\beta}_2(Y)\}$ and then further splitting within each group according to quintiles of $\widehat{\beta}_0(Y)$, ignoring $s_0(Y)$. For the three-tissue comparisons in simulations (c) and (d), CEO tests were implemented as described in Section 7 with the support of $S = \{s_0(Y), \widehat{\boldsymbol{\beta}}_0(Y), \phi(Y)\}$ partitioned first into two groups according to the median of $\widehat{\beta}_0(Y)$ and further partitioning within each group according to quintiles of $\phi(Y)$, ignoring $s_0(Y)$. Discretized EAH tests were performed

under the CEO framework by partitioning only by $\text{sign}\{\widehat{\beta}_1(Y) - \widehat{\beta}_2(Y)\}$ in the two-sample setting and partitioning only by deciles of $\phi(Y)$ in the three-sample setting.

Operating characteristics of the multiple testing procedures were evaluated by averaging across 100 simulated data sets under each scenario. Figure 2 compares the average number of false null hypotheses rejected by each multiple testing procedure as a function of the estimated FDR. In all scenarios, the CEO tests were found to reject on average more false nulls at each level of estimated FDR. The improvement offered by the CEO tests is especially notable in scenario (a) where the EAH, ODP and ANOVA tests have similar performance. Figure 3 illustrates FDR control for CEO tests by showing that at any level of estimated FDR the rFDR is expected to be smaller. FDR control was also achieved by the discretized EAH tests.

Example CEO rejection regions given in Figure 4 illustrate how CEO tests adapt to information contained in null-sufficient statistics. In scenarios (a) and (b), the estimated rejection regions, expressed in terms of signed t-statistics, efficiently capture false nulls by varying with $\widehat{\beta}_0(Y)$. In scenario (c), separating genes with high and low values of $\widehat{\beta}_0(Y)$ concentrates evidence for differential expression in the high-$\widehat{\beta}_0(Y)$ group, allowing CEO rejection regions to outperform the other procedures which ignore $\widehat{\beta}_0(Y)$. Rejection regions for scenario (d) were similar to those for (c).

9 Application to Gene Expression Data

Multiple testing procedures based on CEO rejection regions, EAH tests, the ODP and ANOVA tests were applied to the three-tissue microarray experiment of Spira *et al.* (2004). This experiment measured the expression of 22,214 genes in human airway epithelial cells from 34 Current Smokers, 23 Never Smokers and 18 Former Smokers. Normalized data obtained from `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE994` were log-transformed prior to analysis.

For a three-sample analysis of the Spira *et al.* (2004) data we aimed to detect genes with differential expression across the three tissue groups. ANOVA tests for each gene were performed as for a standard three-sample comparison. ODP tests were implemented using EDGE (Leek *et al.* 2006) and CEO tests were implemented as described in Section 7 with the support of S partitioned by splitting at the median of $\widehat{\beta}_0(Y)$ and then splitting by

deciles of $\phi(Y)$. Discretized EAH tests were implemented under the CEO framework by splitting the genes into 20 groups based on quantiles of $\phi(Y)$.

Figure 5a shows that the CEO testing procedure rejected more genes than the other procedures at each level of estimated FDR. From the CEO rejection region illustrated in Figure 6, we see that the CEO testing procedure adapts to variation in the evidence for differential expression across partitions of S. Among genes with low values of $\widehat{\beta}_0(Y)$ there is strong evidence for differential expression among genes with a decreasing trend in expression from never to former to current smokers. This pattern also exists among genes with high values of $\widehat{\beta}_0(Y)$, but here we also have strong evidence of differential expression among genes with increasing expression from never to former to current smokers.

Since CEO tests use information in the null-sufficient statistic $\widehat{\beta}_0(Y)$ that is ignored by other procedures, the potential value of CEO testing increases with the dimension of $\widehat{\beta}_0(Y)$. To explore this idea in the Spira *et al.* (2004) data, we tested for differential expression between never and current smokers while adjusting for age, sex and race (Caucasian, African, Hispanic or other). For each gene, the estimated effects of these potential confounders on expression is included in $\widehat{\beta}_0(Y)$.

CEO testing in this setting was implemented as in Section 7 with the support of S partitioned by first splitting according to the sign of $\widehat{\beta}_1(Y)$ and then splitting by tertiles of the first component of $\widehat{\beta}_0(Y)$ and then by tertiles of the second component of $\widehat{\beta}_0(Y)$. Figure 5b shows that the CEO tests based on this partition detected many more genes as differentially expressed than the ANOVA or EAH tests, which were also adjusted for age, sex and race but ignore the information in $\widehat{\beta}_0(Y)$ when assessing significance. The value of the information in $\widehat{\beta}_0(Y)$ is evident in Figure 7, which shows how the distribution of p-values changes across the nine partitions of S corresponding to negative values of $\widehat{\beta}_1(Y)$. We have the most evidence of differential expression when the first component of $\widehat{\beta}_0(Y)$ is high and the second component is low. When both the first and second components are high there appears to be no evidence for differential expression.

10 Connections to Other Testing Procedures

If data are reduced to statistics $Z = Z(Y)$ with known null distributions, CEO testing coincides with existing methods. The key connection is that

when Z has a known null distribution, LR level sets for Z are identifiable since the objective function Q_λ (5) can be written as

$$Q_\psi(\Gamma) = \mathcal{P}(Z \in \Gamma) - \psi \mathcal{P}(Z \in \Gamma | H = 0),$$

with $\psi = \lambda p_0$, which is an identifiable function of Γ for each ψ. Even though the frequency of false nulls p_0 is unknown and non-identifiable, the maximizers of Q_ψ for any fixed ψ are LR level sets, just at some unknown level. It follows that when Z has a known null distribution, the class of optimal rejection regions can be consistently estimated. However choosing a rejection region to control FDR will require conservative estimation of the *unconditional* null frequency p_0.

Consider the extreme data reduction in which Y is converted to an unbiased p-value. In this case there is no need to estimate the optimal rejection regions on $[0, 1]$ since every region $[0, a)$, $a \in (0, 1]$, is an LR level set. The multiple testing problem is therefore reduced to choosing a threshold a to control FDR. From (15) it is clear that when given only p-values, CEO testing coincides with the methods of Storey *et al.* (2004), controlling FDR through conservative estimation of p_0.

Under the less extreme data reduction that converts each Y to a t-statistic, LR level sets have the form $t \notin (a, b)$, $a < b$, and can be estimated from level cuts of the consistently estimated ratio $f(t)/f_0(t)$ as described by Efron et al. (2001). The EAH procedure (Signorovitch 2006) generalizes this idea to multivariate statistics $Z(Y)$ with known null distributions. Again, even though optimal rejection regions can be identified in these cases, the choice of a rejection region to control FDR requires conservative estimation of the unconditional frequency of true nulls.

Wasserman and Roeder (2006) and Rubin *et al.* (2006) study 'variable threshold procedures' in which evidence against the null hypotheses is summarized by a univariate test statistic for each test. The optimal rejection region is then defined by possibly different significance thresholds for each test that depend on the true data-generating distributions. From the perspective of p-values, CEO testing can be viewed as a variable threshold procedure in which we share information across tests to conservatively estimate the optimal thresholds as functions of the null-sufficient statistic S.

The popular SAM procedure (Tusher *et al.* 2001) accepts the null hypothesis of equal mean expression in two groups when

$$l < \frac{\widehat{\Delta}}{\widehat{s}_1 + c} < u,$$

for upper and lower bounds u and l where $\widehat{\Delta}$ is the estimated mean difference between groups with standard error $\widehat{s}_1$ and c is some positive constant. These acceptance regions can be written as

$$(l^{-1} - c\widehat{\Delta}^{-1})^{-1} < t < (u^{-1} - c\widehat{\Delta}^{-1})^{-1},$$

where t is the usual two-sample t-statistic. Since SAM rejection regions depend on $\widehat{\Delta}$, which is not a null-sufficient statistic, they do not fall within the CEO testing framework.

11 Discussion

This paper has shown that null-sufficient statistics can contain valuable information for multiple hypothesis testing.

The piecewise constant model for CEO rejection region boundaries used in this paper has the advantage of computational simplicity. However smoother models for CEO rejection region boundaries could more efficiently capture the information in null-sufficient statistics, leading to more powerful testing procedures. For example, the tightest identifiable upper bound on $\pi_0(s)$ is provided by $g(1|s)$ where $g(\cdot|s)$ is the conditional density for P given that $S = s$. Plugging this upper bound into (9) and maximizing over all subsets of $\mathbb{R}^n$ leads to identifiable rejection regions of the form

$$\text{reject } h_i \text{ if } \frac{g(p_i|s_i)}{g(1|s_i)} > c$$

for some constant c. Notice that even for the simple two-sample problem, consistent estimation of these rejection regions would require nonparametric density estimation in essentially three dimensions.

A promising way to improve upon CEO testing for gene expression experiments is through the incorporation of information external the gene expression measurements. For example gene-level information such as GO terms (The Gene Ontology Consortium 2000), locations of genes in pathways or networks, or the presence of specific cis-regulatory elements could all be used to augment the null-sufficient statistic S. If such external data are related to differential expression, their incorporation into CEO tests could facilitate the statistical detection of differential expression.

With or without external information, it is only practical to obtain CEO rejection regions in a limited number of dimensions. An interesting direction

for future research is the development of data-driven dimension reduction for multiple testing.

12 Appendix

Proof of Theorem 1. Let

$$\Gamma(\gamma) = \{y : f_1(y) > \gamma f_0(y)\}, \quad \gamma \geq 0$$

denote the LR level set at level γ. Since f_0 and f_1 are continuous the function

$$z(\gamma) \equiv \int_{\Gamma(\gamma)} f_0(y)dy$$

is a continuous and decreasing map from $[0, \gamma^*]$ to $[0, 1]$ for $\gamma^* = \sup_y f_1(y)/f_0(y)$ and the function

$$C(z) = \sup_{\Gamma \subseteq \mathbb{R}} \Big\{ \int_\Gamma f_1(y)dy : \int_\Gamma f_0(y)dy \leq z \Big\} \tag{17}$$

is a continuous map from $[0, 1]$ to $[0, 1]$. By the Neyman-Pearson Lemma (Lehmann 1986, pp. 74-76) the supremum in (17) given each $0 \leq z \leq 1$ is achieved at a unique LR-level set and we can write

$$C(z(\gamma)) = \int_{\Gamma(\gamma)} f_1(y)dy.$$

Note that $C(z)$ is concave and increasing on $[0, 1]$. Consider approximating the derivative of $C(z)$ between the points $z(\gamma_1)$ and $z(\gamma_2)$ with $0 \leq \gamma_1 < \gamma_2 \leq \gamma^*$. Since $\Gamma(\gamma_2) \subset \Gamma(\gamma_1)$ and $\gamma_1 f_0(y) < f_1(y) \leq \gamma_2 f_0(y)$ for $y \in \mathcal{D} \equiv \Gamma(\gamma_1) \cap \Gamma(\gamma_2)^c$ we have

$$\gamma_1 \leq \frac{C(z(\gamma_1)) - C(z(\gamma_2))}{z(\gamma_1) - z(\gamma_2)} = \frac{\int_\mathcal{D} f_1(y)dy}{\int_\mathcal{D} f_0(y)dy} \leq \gamma_2,$$

and taking the limit as $\gamma_2 \downarrow \gamma_1$ we see that the derivative of $C(z)$ evaluated at $z(\gamma_1)$ is γ_1. Since the inverse function of $z(\gamma)$ decreases with z and has non-negative range, it follows that $C(z)$ is increasing and concave. Note that if the continuity assumptions of this theorem are violated so that $C(z)$ is

not concave or continuous, it can always be made concave and continuous by permitting randomized rejection regions.

The remainder of the proof can be accomplished graphically. Imagine plotting for every $\Gamma \in \mathcal{G}$ the point in $[0,1] \times [0,1]$ given by $\mathcal{P}(Y \in \Gamma | H = 0)$ on the horizontal axis and $\mathcal{P}(Y \in \Gamma | H = 1)$ on the vertical. The function $C(z)$ defines the the concave, non-decreasing upper bound on this set of points. The FDR constraint in (2) can be written as

$$\mathcal{P}(Y \in \Gamma | H = 1) \geq \frac{p_0(1-\alpha)}{\alpha(1-p_0)} \mathcal{P}(Y \in \Gamma | H = 0),$$

which defines a region lying above a line from the origin. If this line intersects the concave, non-decreasing curve $C(z)$, the unique point of intersection corresponds to an LR level set that maximizes $\mathcal{P}(Y \in \Gamma | H = 1)$ under the FDR constraint. Furthermore, since $C(z)$ is concave and increasing, every LR level set will have the highest value of $\mathcal{P}(Y \in \Gamma | H = 1)$ for some FDR constraint $0 < \alpha < 1$. If the FDR constraint line falls entirely above $C(z)$ the rejection region is the empty set; if the FDR line falls entirely below $C(z)$ the rejection region is the whole space.

Note that Corollary 1 follows from this proof since the concavity and increasingness of the map $z \to zp_0 + C(z)(1-p_0)$ follows from that of $C(z)$.

The following property of $\widehat{\Gamma}^u_\lambda$ is worth noting for practical applications and is used in the proof of Theorem 3.

Proposition 1 *For any $\mathcal{G}$ and $\lambda_1 < \lambda_2$, $\mathbb{P}_m \widehat{\Gamma}_{\lambda_1} \geq \mathbb{P}_m \widehat{\Gamma}_{\lambda_2}$.*

Proof of Proposition 1 For any fixed Γ,

$$\widehat{Q}_\lambda(\Gamma) = \mathbb{P}_m \Gamma - \lambda \mathbb{P}_m v(\Gamma, \widehat{\pi}^u_0)$$

is an affine function of λ that is either decreasing or constant. As the supremum of such functions,

$$\sup_{\Gamma \in \mathcal{G}} \widehat{Q}_\lambda(\Gamma)$$

is continuous, decreasing and convex in λ. For any $\lambda^* \geq 1$, the affine function $\widehat{Q}_\lambda(\widehat{\Gamma}^u_{\lambda^*})$ must therefore be tangent to $\sup_{\Gamma \in \mathcal{G}} \widehat{Q}_\lambda(\Gamma)$ at λ^* which implies

$$\mathbb{P}_m v(\widehat{\Gamma}^u_{\lambda_1}, \widehat{\pi}^u_0) \geq \mathbb{P}_m v(\widehat{\Gamma}^u_{\lambda_2}, \widehat{\pi}^u_0), \qquad \text{for all } \lambda_1 < \lambda_2. \tag{18}$$

Combined with $\widehat{Q}_{\lambda_1}(\widehat{\Gamma}^u_{\lambda_1}) \geq \widehat{Q}_{\lambda_1}(\widehat{\Gamma}^u_{\lambda_2})$, (18) implies the desired result:

$$\mathbb{P}_m\{\Gamma^u_{\lambda_1} - \Gamma^u_{\lambda_2}\} \geq \lambda_1 \mathbb{P}_m\{v(\widehat{\Gamma}^u_{\lambda_1}, \widehat{\pi}^u_0) - v(\widehat{\Gamma}^u_{\lambda_2}, \widehat{\pi}^u_0)\} \geq 0.$$

Proof of Theorem 3. Suppose Q_λ is continuous on $(\mathcal{G}, d)$ with a unique maximum over $\mathcal{G}$ for each $\lambda \geq 1$. Also suppose that $\widehat{Q}_{\lambda,m}(\Gamma) = \mathbb{P}_m\{\Gamma - \lambda v(\Gamma, \widehat{\pi}^u_0)\}$ always has a unique maximum over $\mathcal{G}$ for each $\lambda \geq 1$.

Letting $\|\cdot\|_{\mathcal{G}}$ denote the supremum norm over $\mathcal{G}$, note that

$$\|\mathbb{P}_m - \mathcal{P}\|_{\mathcal{G}} \xrightarrow{a.s.} 0 \tag{19}$$

since $\mathcal{G}$ is a VC class and

$$\|\mathbb{P}_m v(\Gamma, \widehat{\pi}^u_0) - \mathcal{P} v(\Gamma, \pi^u_0)\|_{\mathcal{G}} \xrightarrow{a.s.} 0 \tag{20}$$

since the left side of (20) is bounded above by

$$\sup_{s \in \mathcal{A}} |\widehat{\pi}^u_0(s) - \pi^u_0(s)| + \|\{\mathbb{P}_m - \mathcal{P}\} v(\Gamma, \pi^u_0)\|_{\mathcal{G}}$$

with the first term converging $a.s.$ to 0 by assumption and the second term converging $a.s.$ to zero since $v(\Gamma, \pi^u_0)(y)$ is continuous on $(\mathcal{G}, d)$ uniformly in y and Theorem 2.7.11 (van der Vaart and Wellner 1996, p.164).

By (19) and (20) we have for any $1 \leq \lambda^* < \infty$

$$\sup_{1 \leq \lambda \leq \lambda^*} \|\widehat{Q}_{\lambda,m} - Q_\lambda\|_{\mathcal{G}} \xrightarrow{a.s.} 0$$

and Theorem 3a follows from the argmax theorem. Note that for large enough λ^* we have $\Gamma^u_\lambda = \emptyset$ for all $\lambda \geq \lambda^*$ so that convergence at λ^* and Property 1 together imply $\widehat{\Gamma}^u_{\lambda,m} \xrightarrow{a.s.} \emptyset$ for all $\lambda \geq \lambda^*$, so the upper bound λ^* need not appear in the statement of Theorem 3a.

For Theorems 3b and 3c the restriction to $1 \leq \lambda < \lambda^*$ with $\mathcal{P}(\Gamma^u_\lambda) = \delta > 0$, together with Proposition 1, ensures that the denominators of FDR, $\widehat{FDR}$

and $rFDR$ do not converge to zero. To prove Theorem 3b we can write

$$\begin{aligned}
&\sup_{1\le\lambda\le\lambda^*} FDR(\widehat{\Gamma}^u_\lambda,\pi_0) - \widehat{FDR}(\widehat{\Gamma}^u_\lambda,\widehat{\pi}^u_0) \\
&\quad\le \sup_{1\le\lambda\le\lambda^*} FDR(\widehat{\Gamma}^u_\lambda,\pi^u_0) - \widehat{FDR}(\widehat{\Gamma}^u_\lambda,\widehat{\pi}^u_0) \\
&\quad\le \sup_{1\le\lambda\le\lambda^*}\left|\frac{\mathcal{P}v(\widehat{\Gamma}^u_\lambda,\pi^u_0)}{\mathcal{P}\widehat{\Gamma}^u_\lambda} - \frac{\mathbb{P}_m v(\widehat{\Gamma}^u_\lambda,\widehat{\pi}^u_0)}{\mathcal{P}\widehat{\Gamma}^u_\lambda}\right| + \sup_{1\le\lambda\le\lambda^*}\left|\frac{\mathbb{P}_m v(\widehat{\Gamma}^u_\lambda,\widehat{\pi}^u_0)}{\mathcal{P}\widehat{\Gamma}^u_\lambda} - \frac{\mathbb{P}_m v(\widehat{\Gamma}^u_\lambda,\widehat{\pi}^u_0)}{\mathbb{P}_m\widehat{\Gamma}^u_\lambda}\right| \\
&\quad\le \frac{\|\mathcal{P}v(\Gamma,\pi^u_0) - \mathbb{P}_m v(\Gamma,\widehat{\pi}^u_0)\|_{\mathcal{G}}}{\inf_{1\le\lambda\le\lambda^*}\mathcal{P}\widehat{\Gamma}^u_\lambda} + \frac{\|\mathbb{P}_m - \mathcal{P}\|_{\mathcal{G}}}{\inf_{1\le\lambda\le\lambda^*}\mathcal{P}(\widehat{\Gamma}^u_\lambda)\mathbb{P}_m(\widehat{\Gamma}^u_\lambda)},
\end{aligned}$$

with (19), (20) and Theorem 3a, ensuring that the final two terms converge $a.s.$ to zero by the continuous mapping theorem. Theorem 3c can be proved by a similar argument. Theorem 3 remains valid under weak dependence across realizations of the hierarchical model, so long as the underlying empirical processes $I(y\in\Gamma)$ and $v(\Gamma,\pi^u_0)$ converge uniformly to their expectations over $\mathcal{G}$. Convergence would occur for example if genes were dependent only within finite blocks. The assumptions of Theorem 3 are easily verified for application to the piecewise constant model used in this paper.

Acknowledgements

This work was supported by an NIH pre-doctoral interdisciplinary training grant in biostatistics. The author thanks Tianxi Cai, Jamie Robins, Armin Schwartzman and L.J. Wei for helpful discussions.

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.

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

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. (2000) 25: 25-29.

Genovese, C. and Wasserman, L. (2003) Bayesian and Frequentist Multiple Testing, Bayesian Statistics, Oxford University Press, pp. 145-162.

Leek, J.T., Monsen, E., Dabney, A.R. and Storey, J.D. (2006) EDGE: Extraction and Analysis of Differential Gene Expression, Bioinformatics 22(4):507-508.

Lehmann, E. L., (1986). *Testing Statistical Hypotheses*, Second Edition, Springer-Verlag, New York.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Rubin, D., Dudoit, S., and van der Laan, M. (2006). A Method to Increase the Power of Multiple Testing Procedures Through Sample Splitting, *Statistical Applications in Genetics and Molecular Biology*, 5,1,Article 19.

Signorovitch, J.E. (2006). Multiple Testing With an Empirical Alternative Hypothesis, Harvard University Biostatistics Working Paper Series, Working Paper 60, http://www.bepress.com/harvardbiostat/paper60

Spira A., Beane J., Shah V., Liu G., Schembri F., Yang X., Palma J. and Brody J.S. (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci U S A*,101(27):10143-8.

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

Storey, J.D. (2003). The Positive False Discovery Rate: A Bayesian Interpretation and the q-value, *Annals of Statistics*, 31, 2013-2035.

Storey, J. D. (2005). The optimal discovery procedure: A new approach to simultaneous significance testing. UW Biostatistics Working Paper Series, Working Paper 259. http://www.bepress. com/uwbiostat/paper259/

Storey, J.D., and Tibshirani, R. (2003). Statistical Significance for Genomewide Studies, *Proceedings of the National Academy of Sciences*, 100, 9440-9445.

Storey, J.D., Taylor, J.E., and Siegmund, D. (2004). Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach, *Journal of the Royal Statistical Society, Series B*, 66,187-205.

Storey, J.D., Dai, J.Y. and Leek, J.T. (23 August 2006). The Optimal Discovery Procedure for Large-Scale Significance Testing, With Applications to Comparative Microarray Experiments. *Biostatistics* doi:10.1093/biostatistics/kxl019.

Tusher, V.G., Tibshirani, R, and Chi, G. (2001) Significance Analysis of Microarrays Applied to the Ionizing Radiation Response, *Proceedings of the National Academy of Sciences*, 98(9):5116-5121.

Wasserman, L., and Roeder, K. (2006). Weighted Hypothesis Testing, http://arxiv.org/abs/math.ST/0604172.

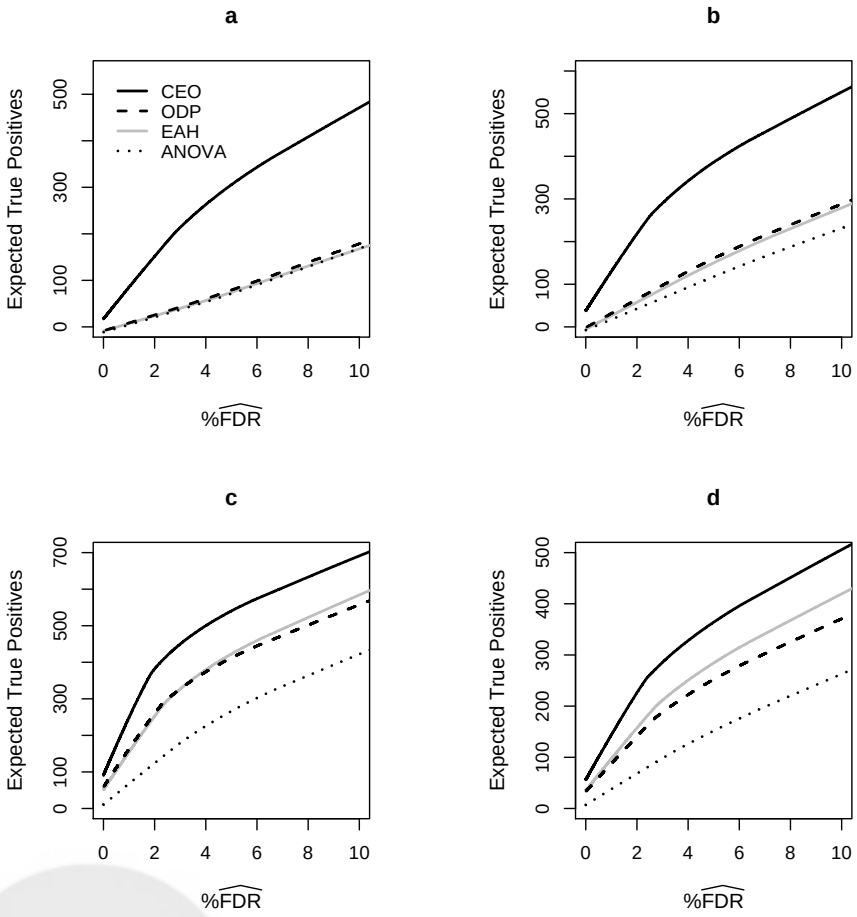

Figure 2: Comparison of the expected number of true positives rejected as a function of estimated FDR for CEO, EAH, ODP and ANOVA tests, as estimated from 100 simulated data sets under each of the scenarios (a) through (d) of Storey *et al.* (2006).

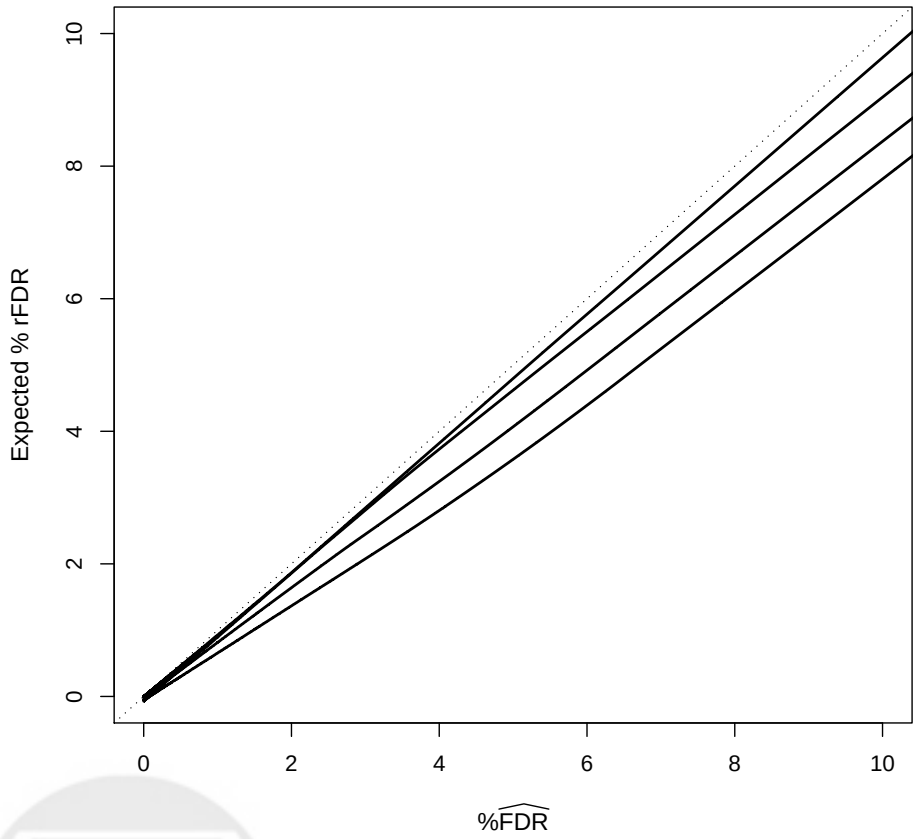

Figure 3: Assessment of FDR control for CEO tests in simulation scenarios (c), (d), (a), and (b) (ordered from highest to lowest expected % rFDR at $\widehat{FDR} = 10$ %).

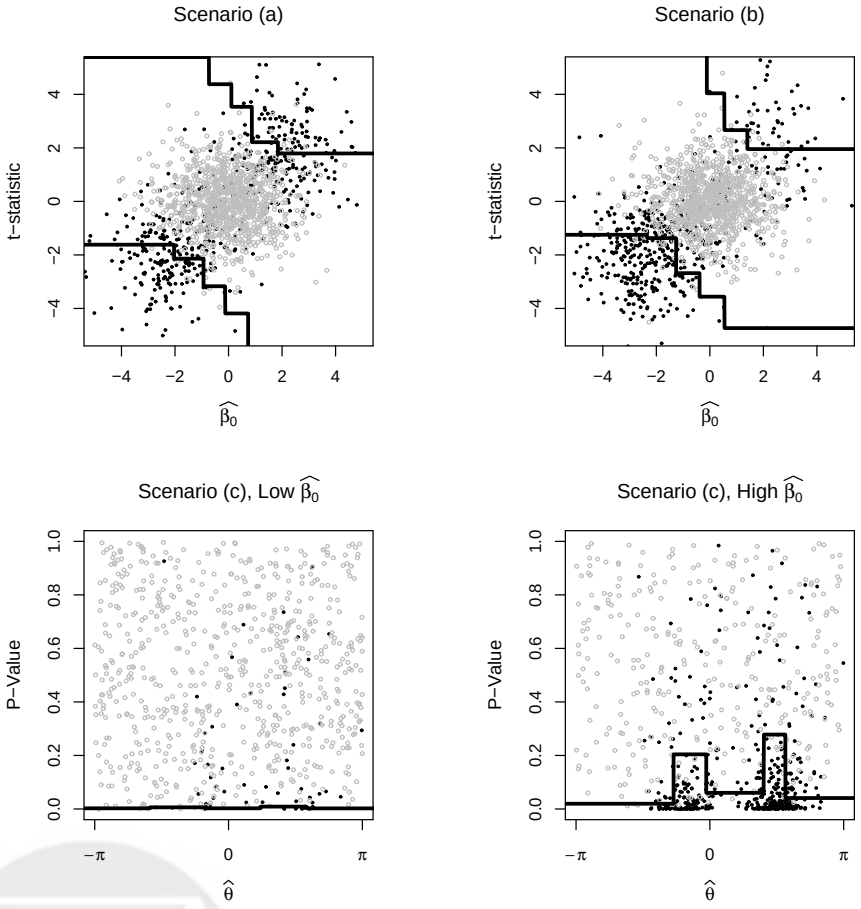

Figure 4: Example CEO rejection regions under simulation scenarios (a), (b) and (c). Open gray circles correspond to true null hypotheses and solid black circles correspond to false nulls.

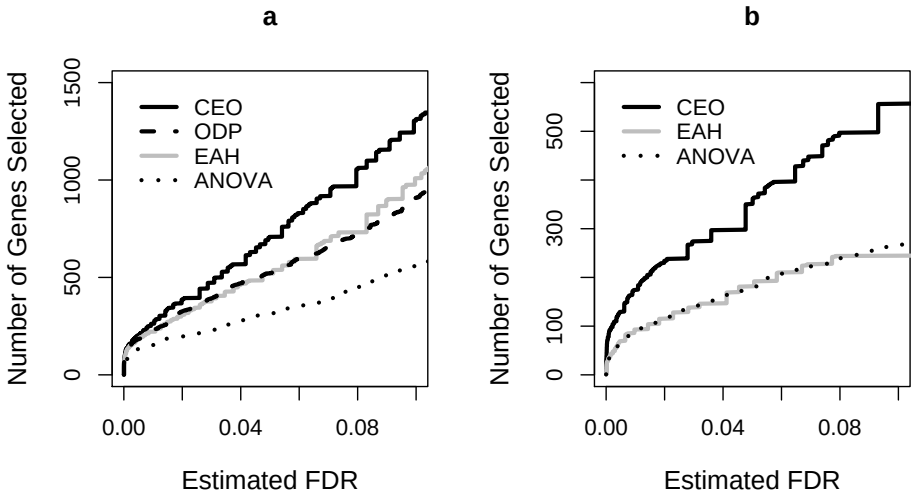

Figure 5: Comparison of multiple testing procedures applied to the detection of differentially expressed genes in (a) the three-tissue and (b) the two-tissue analysis of the the Spira *et al.* (2004) data.

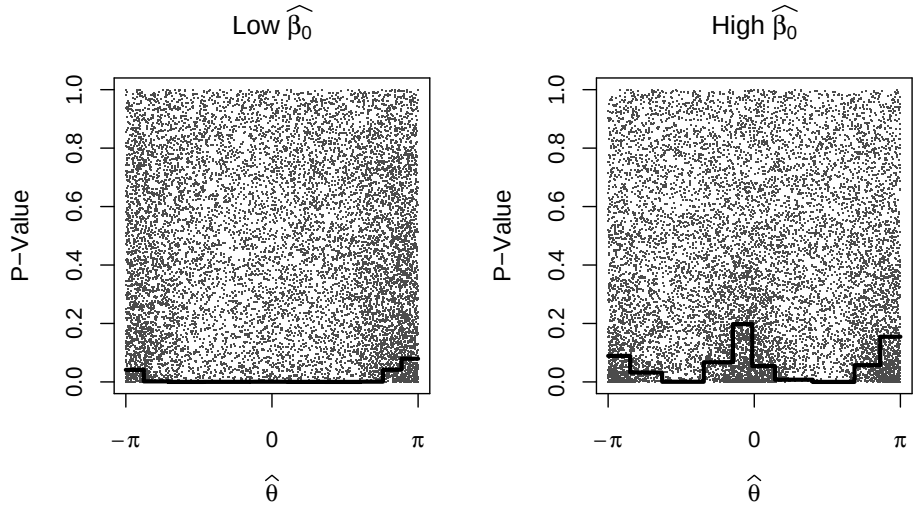

Figure 6: CEO rejection region for the smoking data of Spira *et al.* (2004). $\widehat{\theta} = 0$ corresponds to genes with increasing expression from never to former to current smokers. Genes with $|\widehat{\theta}| = \pi$ have the reverse trend in expression across groups. $\widehat{\beta}_0$ is the mean expression level across all patients.

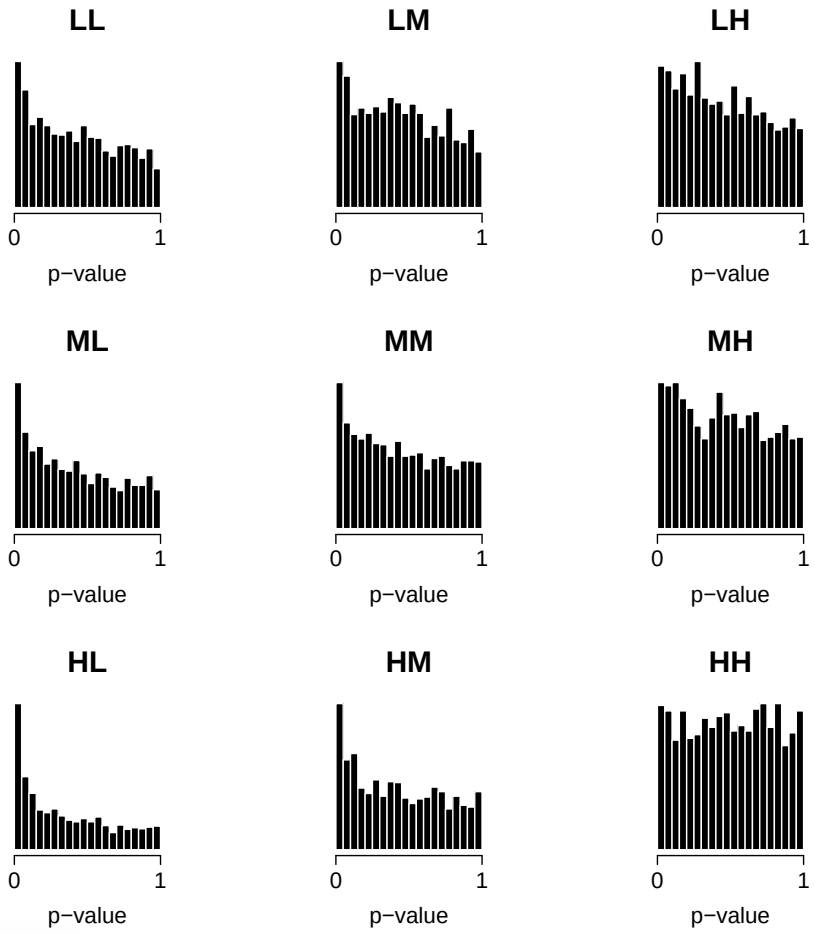

Figure 7: Distributions of adjusted p-values for differential expression between current and never smokers in the Spira *et al.* (2004) data. Each panel corresponds to one of the nine partitions based on the null-sufficient statistic with negative values of $\widehat{\beta}_1(Y)$. The first letter of the panel label indicates whether the first component of $\widehat{\beta}_0(Y)$ is high (H), middle (M) or low (L) and the second letter indicates the value of $\widehat{\beta}_0(Y)$. Each histogram is based on approximately 1,400 p-values.