Harvard University
Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 60

Multiple Testing With an Empirical Alternative Hypothesis

James E. Signorovitch*

*Harvard University, James_Signorovitch@hms.harvard.edu

Multiple Testing With an Empirical Alternative Hypothesis

James E. Signorovitch

Abstract

An optimal multiple testing procedure is identified for linear hypotheses under the general linear model, maximizing the expected number of false null hypotheses rejected at any significance level. The optimal procedure depends on the unknown data-generating distribution, but can be consistently estimated. Drawing information together across many hypotheses, the estimated optimal procedure provides an empirical alternative hypothesis by adapting to underlying patterns of departure from the null. Proposed multiple testing procedures based on the empirical alternative are evaluated through simulations and an application to gene expression microarray data. Compared to a standard multiple testing procedure, it is not unusual for use of an empirical alternative hypothesis to increase by 50% or more the number of true positives identified at a given significance level.

Multiple Testing With an Empirical Alternative Hypothesis

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

November 15, 2006

Abstract

An optimal multiple testing procedure is identified for linear hypotheses under the general linear model, maximizing the expected number of false null hypotheses rejected at any significance level. The optimal procedure depends on the unknown data-generating distribution, but can be consistently estimated. Drawing information together across many hypotheses, the estimated optimal procedure provides an *empirical alternative hypothesis* by adapting to underlying patterns of departure from the null. Proposed multiple testing procedures based on the empirical alternative are evaluated through simulations and an application to gene expression microarray data. Compared to a standard multiple testing procedure, it is not unusual for use of an empirical alternative hypothesis to increase by 50% or more the number of true positives identified at a given significance level.

Keywords: Empirical Bayes; False discovery rate; Clustering; Density estimation.

Author's Footnote:

James E. Signorovitch is Doctoral Candidate, Department of Biostatistics, Harvard School of Public Health, Boston MA 02115 (E-mail: jsignoro@hsph.harvard.edu).
This work was supported by an NIH pre-doctoral interdisciplinary training grant in biostatistics. The author thanks L.J. Wei, Tianxi Cai, Jun Liu, Armin Schwartzman, Lihua Zou, Chuck Weitz, Florian Storch and Carlos Paz for helpful comments and suggestions.

1. INTRODUCTION

Multiple hypothesis testing plays an increasingly prominent role in applied statistics. New data gathering technologies, especially in the biological sciences, allow researchers to study thousands of related items, such as genes or cell types, in a single experiment. Such experiments are often aimed at identifying a subset of items that behave in a specified fashion. This goal can be addressed statistically though multiple hypothesis testing. For example, to identify genes with expression levels that vary across several tissue types one could test for every gene the null hypothesis that mean expression is constant across tissues. When a test rejects this null hypothesis for a gene that truly has constant mean expression we have a *false positive* result for that gene. Failure to reject this null hypothesis for a gene with non-constant mean expression constitutes a *false negative*. A good multiple testing procedure should minimize as much as possible the rate of false negatives while controlling the rate of false positives.

Following the landmark introduction of false discovery rate (FDR) controlling procedures by Benjamini and Hochberg (1995), many practical methods have been introduced for controlling the rate of false positives in large-scale multiple testing (Storey and Tibshirani 2003; Storey *et al.* 2004; Efron 2004; Dudoit *et al.* 2004). This paper focuses on decreasing the rate of false negatives by increasing the average power of multiple testing procedures under the general linear model. Power and optimality for multiple testing has recently been studied by Storey (2005), Rubin *et al.* (2006) and Wasserman and Roeder (2006).

Consider an experimental setting in which repeated observations are made from a general linear model with a fixed design matrix and random parame-

ters. That is, the experiment generates independent realizations of $(\mathbf{y}, \boldsymbol{\beta}, \sigma)$ where $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\sigma > 0$ are unobserved parameters drawn from some unknown distribution and $\mathbf{y}$ is an observed $n \times 1$ response vector distributed as

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

for a fixed and known $n \times d$ design matrix $\mathbf{X}$. Given m observations $\mathbf{y}_1, \ldots, \mathbf{y}_m$, corresponding to the unobserved realized parameters $\sigma_i, \boldsymbol{\beta}_i$, $i = 1, \ldots, m$, our objective is to test the m null hypotheses,

$$H_i : \boldsymbol{\beta}_i \in \mathcal{V}_0, \quad i = 1, \ldots, m,$$

determined by a linear space $\mathcal{V}_0 \subset \mathbb{R}^d$. Dependence among observations is considered in the Appendix.

This setting provides a simple model for gene expression data from m genes on n arrays, where each array is associated with a $d \times 1$ vector of covariates such as time, tissue type or treatment. An experimenter may, for example, be interested in testing for each of the m genes the null hypothesis that mean expression does not depend on some covariates while controlling for others.

Given a design matrix $\mathbf{X}$ and a null hypothesis $\boldsymbol{\beta} \in \mathcal{V}_0$, we consider the class $\mathcal{T}$ of all statistics $T : \mathbb{R}^n \to [0, 1]$ such that

i. $T(\mathbf{y})$ is invariant under the group $\mathcal{G}$ of transformations $\mathbf{y} \to c\mathbf{y} + \mathbf{X}\boldsymbol{\beta}_0$, with $c \neq 0$ and $\boldsymbol{\beta}_0 \in \mathcal{V}_0$, and

ii. $T(\mathbf{y}) \sim \text{Uniform}[0, 1]$ given any fixed $\sigma > 0$ and $\boldsymbol{\beta} \in \mathcal{V}_0$.

Requirement (i) is a weakening of the invariance requirements under which the standard F-test for linear hypotheses is uniformly most powerful (Lehmann

1986, pp. 365-370). The suitability of $\mathcal{G}$-invariance for multiple testing under the general linear model is examined in the Discussion and Remark E. Any $\mathcal{G}$-invariant statistic with a known null distribution has representatives in $\mathcal{T}$ that satisfy requirement (ii). For example, the p-value from a standard F-test is in $\mathcal{T}$. The requirement that T has a known null distribution simplifies the problem of assessing and controlling the rate of false positives.

We consider multiple testing procedures that employ a single $T \in \mathcal{T}$ and reject all H_i such that $T(\mathbf{y}_i) < \alpha$ for some threshold α. Such procedures are invariant to the ordering of individual hypotheses and preclude dependence on information external to $\mathbf{X}$, $\mathcal{V}_0$ and the $\mathbf{y}_i$'s. The important case in which α is a function of the data, chosen to control the rate of false positives, is considered in Remark A.

For any data-generating distribution, specified by the distribution of $(\boldsymbol{\beta}, \sigma)$, the worth of a statistic $T \in \mathcal{T}$ for multiple testing can be measured by its average power function,

$$\pi(T; \alpha) \equiv Pr(T(\mathbf{y}) < \alpha | \boldsymbol{\beta} \notin \mathcal{V}_0),$$

where the probability is computed with respect to the random variables $\mathbf{y}, \sigma$ and $\boldsymbol{\beta}$. The average power function is a reasonable performance criterion for multiple testing since, with any number of hypotheses, $\pi(T; \alpha)$ is the expected fraction of false nulls rejected by T (the sensitivity) when the probability of rejecting a true null (1 - specificity) is α. Storey (2005), Rubin *et al.* (2006) and Wasserman and Roeder (2006) also measure the performance of multiple testing procedures using average power.

For a given data-generating distribution, a test $T' \in \mathcal{T}$ is preferred over

T if its average power function is dominant, that is if

$$\pi(T';\alpha) \geq \pi(T;\alpha), \quad 0 \leq \alpha \leq 1$$

and the inequality is strict for some α. Equivalently, if two elements of $\mathcal{T}$ can be stochastically ordered under the data-generating distribution, the stochastically smaller one is preferred.

This paper identifies in Section 3 the stochastically minimal element $T^* \in \mathcal{T}$ as a function of the data-generating distribution. If the data-generating distribution were known, the corresponding T^* would provide a multiple testing procedure with the maximal average power function

$$\pi^*(\alpha) \equiv \pi(T^*;\alpha) \geq \pi(T;\alpha), \;\; 0 \leq \alpha \leq 1, \text{ for all } T \in \mathcal{T},$$

maximizing the sensitivity over all significance thresholds α. In this sense, T^* is optimal in $\mathcal{T}$.

In practice the data-generating distribution is unknown and T^* can not be computed directly. In Section 4 we provide an estimator $\widehat{T}^*_m(\cdot) = \widehat{T}^*_m(\cdot;\mathbf{y}_1,\ldots,\mathbf{y}_m)$, depending on all the data, such that $\sup_{\mathbf{y}\in\mathbb{R}^n} |\widehat{T}^*_m(\mathbf{y}) - T^*(\mathbf{y})| \xrightarrow{a.s.} 0$ as $m \to \infty$. Since $\widehat{T}^*_m$ consistently estimates T^*, which depends on the data-generating distribution for false nulls, $\widehat{T}^*$ is referred to as employing an *empirical alternative hypothesis (EAH)*. Furthermore, we show in Section 5 that *a priori* knowledge about the data-generating distribution for false nulls can not improve the limiting performance of $\widehat{T}^*_m$ for an increasing number of hypotheses.

The proposed multiple testing procedure based on $\widehat{T}^*_m$ is applied to simulated data in Section 6 and to a search for rhythmically expressed genes in the mouse eye in Section 7. The discussion in Section 8 concludes with some

technical remarks that are referenced throughout the paper. Selected proofs are in an Appendix. The following section provides two examples that motivate our search for T^* and illustrate the practical value of using an estimate $\widehat{T}^*_m$.

2. EXAMPLES

Our first example illustrates how an optimal test for a single hypothesis may not capture all the information available for multiple testing.

2.1 Two-sample problem

Consider a microarray experiment comparing gene expression across two tissue types, with expression levels for 1000 genes following the general linear model described above on $n = 2k$ arrays with $k \geq 2$ samples from each tissue type. The statistical goal is to detect genes that are differentially expressed in one tissue type relative to the other.

If we were testing only a single gene for differential expression, a two-sided, two-sample t-test would provide the uniformly most powerful unbiased test. But suppose we observe, say, 900 positive t-statistics and 100 negative. Certainly the overabundance of positive signs suggests that many genes are differentially expressed. Two-sided t-tests would ignore this information.

The information in the signs of the t-statistics could be captured if we knew the probability density f_1 for t-statistics corresponding to differentially expressed genes. A likelihood ratio test for the ith gene, based on the t-statistic t_i, would yield the p-value

$$p_i \equiv Pr\left\{\frac{f_1(t_0)}{f_0(t_0)} > \frac{f_1(t_i)}{f_0(t_i)}\right\},$$

where the probability is computed for t_0 following a central Student's t-

distribution with the appropriate degrees of freedom and density function f_0. Efron *et al.* (2001) show that even though we observe t-statistics sampled from the mixture density $f = p_0 f_0 + (1 - p_0) f_1$, with unknown proportion p_0 of true null hypotheses, we can write

$$p_i = Pr\left\{ \frac{f(t_0)}{f_0(t_0)} > \frac{f(t_i)}{f_0(t_i)} \right\},$$

and these p-values can be estimated by plugging-in an estimate of f based on the empirical distribution of t-statistics across all genes. Efron *et al.* (2001) note that this likelihood ratio procedure could, in principle, be applied for any data reduction with a known or estimable null distribution.

In this paper we restrict our attention to $\mathcal{G}$-invariant data reductions with known null distributions and extend this empirical approach from the two-tissue comparison to the general linear model. Some of the key ideas involved in making this extension are illustrated by a three-tissue comparison.

2.2 Three-sample problem

Suppose that $m = 2000$ genes are tested for differential expression across three tissue types with six arrays for each type. For each gene $i = 1, \ldots, m$, let F_i be the usual ANOVA F-statistic, with $(3 - 1)$ and $(18 - 3)$ degrees of freedom, for testing the null hypothesis H_i that gene i's mean expression level is constant across the three tissue types. Consider the statistic

$$\hat{\theta}_i \equiv \text{the directed angle between } (\hat{\beta}_{2i} - \hat{\beta}_{1i}, \hat{\beta}_{3i} - \hat{\beta}_{1i})' \text{ and } (0, 1),'$$

where $\hat{\beta}_{ki}$ is the estimated mean expression level of gene i in tissue k, $i = 1, \ldots, m$, $k = 1, 2, 3$.

Note that $\hat{\theta}_i$ is uniformly distributed over its support and independent of F_i when H_i is true. For this reason, and others described in Section 3,

the statistic $(F_i, \hat{\theta}_i)$ is the three-sample analog of $(t_i^2, \text{sgn}(t_i))$ in the previous two-tissue example. Letting a_i be the p-value corresponding to F_i we obtain the statistics $(a_i, \hat{\theta}_i)$, $i = 1, \ldots, m$, taking values on $[0, 1] \times (-\pi, \pi)$.

Figure 1 shows a plot of the $(a_i, \hat{\theta}_i)$'s distributed over $[0, 0.1] \times (-\pi, \pi)$ for a data set simulated with the six expression measurements for gene i in tissue k distributed independently as $\mathcal{N}(\beta_{ki}, 1)$, $i = 1, \ldots, m$, $k = 1, 2, 3$. Points corresponding to 1000 non-differentially expressed genes (i.e. $\beta_{1i} = \beta_{2i} = \beta_{3i}$) are shown as open circles and should be uniformly distributed over the plot region. The filled circles correspond to 1000 differentially expressed genes, with mean expression values generated using $(\beta_{1i}, \beta_{2i}, \beta_{3i})' = (0.325, 0, -0.325)'$ with probability 1/2 and $(\beta_{1i}, \beta_{2i}, \beta_{3i})' = -(0.325, 0, -0.325)'$ with probability 1/2 for each i indexing a differentially expressed gene. With these parameter values, the standard F-test for individual hypotheses has only 25% power to detect true alternatives when $\alpha = 0.05$.

Using only the F-statistics, a multiple testing procedure controlling the false discovery rate (FDR) at 10% using the method of Storey *et al.* (2004) produced the rejection region below the horizontal line at $a = 0.004$ in Figure 1. The realized FDR in this region, the actual proportion of true nulls among the rejected hypotheses, is about 9%, with 4 false positives among 45 rejected hypotheses. The solid lines in Figure 1 define a different rejection region that has adapted to the empirical distribution of $(a, \widehat{\boldsymbol{\theta}})$ using methods described in the remainder of this paper. This rejection region has captured about five times as many true alternatives, containing 224 rejected hypotheses with a realized FDR of 8% (18/224). Simulations in Section 6 will show that this result is not out of the ordinary.

3. THE OPTIMAL TEST T^*

This section identifies an optimal T^* in $\mathcal{T}$ with the maximal average power function $\pi^*(\alpha)$ for any particular data-generating distribution. We begin by finding a $\mathcal{G}$-invariant function of $\mathbf{y}$ that contains as much information as possible about $\boldsymbol{\beta}$ and σ. Assume without loss of generality that $\mathbf{X}^t\mathbf{X} = \mathbf{I}$. The ordinary least-squares estimate of $\boldsymbol{\beta}$ based on $\mathbf{y}$ is denoted by $\widehat{\boldsymbol{\beta}}(\mathbf{y})$. Orthogonal projection into a linear space $\mathcal{V}$ is denoted by $\mathbf{P}_{\mathcal{V}}$.

Lemma 1 *Under the general linear model, all $\mathcal{G}$-invariant statistics follow distributions that depend on $\boldsymbol{\beta}$ and σ only through $\boldsymbol{\eta} = \sigma^{-1}\mathbf{P}_{\mathcal{V}_o^\perp}\boldsymbol{\beta}$ and the statistic*

$$\mathcal{M}(\mathbf{y}) = \frac{\mathbf{P}_{\mathcal{V}_0^\perp}\widehat{\boldsymbol{\beta}}(\mathbf{y})}{\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y})\|}$$

is minimal sufficient for $\boldsymbol{\eta}$ among all $\mathcal{G}$-invariant statistics.

The function $\mathcal{M}(\mathbf{y})$ describes the magnitude and direction of $\mathbf{y}$'s apparent deviation from the null hypothesis. Letting $\widehat{\boldsymbol{\beta}}_0(\mathbf{y})$ denote the least squares estimate of $\boldsymbol{\beta}$ constrained to $\mathcal{V}_0$, the magnitude of deviation is measured by

$$\|\mathcal{M}(\mathbf{y})\|^2 = \frac{\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y}) - \mathbf{X}\widehat{\boldsymbol{\beta}}_0(\mathbf{y})\|^2}{\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y})\|^2}, \tag{1}$$

which is proportional to the standard F-statistic for testing $\boldsymbol{\beta} \in \mathcal{V}_0$ against the unrestricted alternative. The direction of apparent deviation from the null, illustrated in Figure 2, is measured by

$$\widehat{\boldsymbol{\theta}}(\mathbf{y}) \equiv \frac{\mathcal{M}(\mathbf{y})}{\|\mathcal{M}(\mathbf{y})\|} = \frac{\mathbf{P}_{\mathcal{V}_0^\perp}\widehat{\boldsymbol{\beta}}(\mathbf{y})}{\|\mathbf{P}_{\mathcal{V}_0^\perp}\widehat{\boldsymbol{\beta}}(\mathbf{y})\|}. \tag{2}$$

In a two-tissue comparison, $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ reduces to the sign of the t-statistic. In the three-tissue example of Section 2, $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ lies on the edge of a 2-dimensional unit circle and can be identified with a scalar directed angle.

Since the null hypothesis is true exactly when $\boldsymbol{\eta} = 0$, it follows from Lemma 1 that any $\mathcal{G}$-invariant statistic follows a null distribution free of $\boldsymbol{\beta}$ and σ. The null distribution of $\mathcal{M}(\mathbf{y}) = F(\mathbf{y})\widehat{\boldsymbol{\theta}}(\mathbf{y})\frac{d-d_0}{n-d}$ is provided as follows.

Proposition 1 *Suppose $\boldsymbol{\beta} \in \mathcal{V}_0$. Then $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is independent of $F(\mathbf{y})$ and uniformly distributed on the surface of the unit $(d-d_0)$-sphere centered at the origin in $\mathcal{V}_0^{\perp}$.*

It is convenient to convert the F-statistic to its corresponding p-value $a(\mathbf{y}) = 1 - F_{(d-d_0),(n-d)}\{F(\mathbf{y})\}$ and then represent $\mathcal{M}(\mathbf{y})$ by an algebraic equivalent,

$$(a, \widehat{\boldsymbol{\theta}}) \equiv \big(a(\mathbf{y}), \widehat{\boldsymbol{\theta}}(\mathbf{y})\big),$$

which is uniformly distributed over its support under the null hypothesis. To be precise, let $\mathcal{S}_{\mathcal{V}_0^{\perp}}$ denote the surface of the unit $(d-d_0)$-sphere centered at the origin in $\mathcal{V}_0^{\perp}$ with surface area $s_{d-d_0} = 2\pi^{(d-d_0)/2}/\Gamma\{(d-d_0)/2\}$. The null density for $(a, \widehat{\boldsymbol{\theta}})$ is

$$g_0(a, \widehat{\boldsymbol{\theta}}) = s_{d-d_0}^{-1}, \qquad \text{for } (a, \widehat{\boldsymbol{\theta}}) \in [0,1] \times \mathcal{S}_{\mathcal{V}_0^{\perp}}.$$

Since $(a, \widehat{\boldsymbol{\theta}})$ contains all the information regarding $\boldsymbol{\beta}$ and σ available from any $\mathcal{G}$-invariant statistic, T^* can be identified by employing $(a, \widehat{\boldsymbol{\theta}})$ in an optimal fashion. Suppose that when the null hypothesis is false, $(a, \widehat{\boldsymbol{\theta}})$ is distributed with a known density g_1 that is absolutely continuous with respect to g_0. Hypothesis testing could then be based on the likelihood ratio g_1/g_0, leading to

$$T^*(\mathbf{y}) \equiv \int\int_{[0,1]\times\mathcal{S}_{\mathcal{V}^{\perp}}} I[g_1(a,\theta) > g_1\{a(\mathbf{y}), \widehat{\boldsymbol{\theta}}(\mathbf{y})\}] dG_0(a,\theta), \tag{3}$$

where G_0 is the null distribution of (a, θ) with density g_0. We assume that g_1 has no flat parts (see Remark B). Given Lemma 1, the following is an immediate consequence of the Neyman-Pearson Lemma.

Theorem 1 *Given any data-generating distribution,* $T^*(\mathbf{y})$ *as defined in* (3) *is stochastically minimal in* $\mathcal{T}$.

As described in the Introduction, the stochastically minimal statistic T^* provides a multiple testing procedure with the maximal average power function $\pi^*(\alpha)$ achievable with elements of $\mathcal{T}$. Of course in practice $T^*(\cdot)$ is unknown because g_1 is unknown. However when faced with multiple testing we observe m realizations of $(a, \widehat{\boldsymbol{\theta}})$ which, as shown in the following section, can provide a uniformly consistent estimate of $T^*(\cdot)$ as m increases to infinity.

We will use the notation

$$G_0\left\{z > z(a_0, \widehat{\boldsymbol{\theta}}_0)\right\} \equiv \int\int_{[0,1]\times\mathcal{S}_{\mathcal{V}^\perp}} I\{z(a,\theta) > z(a_0, \widehat{\boldsymbol{\theta}}_0)\} dG_0(a,\theta), \qquad (4)$$

for $z : [0,1] \times \mathcal{S}_{\mathcal{V}^\perp} \to \mathbb{R}$ so that we may write

$$T^*(\mathbf{y}) = G_0\left\{g_1 > g_1(a, \widehat{\boldsymbol{\theta}})\right\},$$

where it is implied that $a = a(\mathbf{y})$ and $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{y})$.

4. ESTIMATING T^*

In practice we will observe independent samples $(a_i, \widehat{\boldsymbol{\theta}}_i)$, $i = 1, \ldots, m$, from the mixture density

$$g = p_0 g_0 + (1 - p_0) g_1$$

with an unknown proportion $0 \le p_0 \le 1$ of true null hypotheses (see the Appendix regarding dependence among observations). Due to Proposition 1, this mixing causes no additional difficulty since, as in Efron *et al.* (2001), the statistic T^* defined in (3) can be expressed as a function of g only, with

$$T^*(\mathbf{y}) \equiv G_0\left\{g_1 > g_1(a, \widehat{\boldsymbol{\theta}})\right\} = G_0\left\{g > g(a, \widehat{\boldsymbol{\theta}})\right\}. \qquad (5)$$

An estimate of T^* can therefore be obtained by plugging an estimate of g into (5).

Our proposed density estimate for g at the point $(a_0, \boldsymbol{\theta}_0)$ begins by assigning kernel weights to the observations $(a_i, \widehat{\boldsymbol{\theta}}_i)$, $i = 1, \ldots, m$, based on their angle from $\boldsymbol{\theta}_0$, with

$$w_i(\boldsymbol{\theta}_0, \kappa) = m^{-1} c_0(\kappa) \exp(\kappa \widehat{\boldsymbol{\theta}}_i^t \boldsymbol{\theta}_0), \quad i = 1, \ldots, m,$$

where κ is a smoothing parameter and the normalization constant $c_0(\kappa)$ is given by Hall *et al.* (1987). To estimate the density along $0 \leq a < 1$ for fixed θ_0 we use these kernel weights to construct a histogram estimator with b bins of equal widths $1/b$ by assigning masses

$$h_k(\boldsymbol{\theta}_0) = bm^{-1} \sum_{i=1}^{m} I\left\{(k-1)b^{-1} \leq a_i < kb^{-1}\right\} w_i(\boldsymbol{\theta}_0, \kappa), \quad k = 1, \ldots, b,$$

to the corresponding intervals $[(k-1)b^{-1}, kb^{-1})$, $k = 1, \ldots, b$. With $F(\mathbf{y})$ and $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ defined in (1) and (2) for the unrestricted alternative hypothesis $\boldsymbol{\beta} \in \mathcal{V}_0^{\perp} \cap \mathbb{R}^d$, $g(a, \theta)$ is non-increasing in a. We therefore sort the bin masses to obtain a non-increasing histogram estimator at $\boldsymbol{\theta}_0$. That is, letting $h_{(k)}(\boldsymbol{\theta}_0)$ denote the kth largest of the bin masses $h_j(\boldsymbol{\theta}_0)$, $j = 1, \ldots, b$, the estimate of g at the point $(a_0, \boldsymbol{\theta}_0)$ becomes

$$\hat{g}_m(a_0, \boldsymbol{\theta}_0) = h_{(k)}(\boldsymbol{\theta}_0), \quad \text{with } k \text{ chosen such that } (k-1)b^{-1} \leq a_0 < kb^{-1}.$$

This sorting operation can only improve the estimate or leave it unchanged (Remark C).

Define

$$\widehat{T}_m^*(\mathbf{y}) \equiv G_0\left\{\hat{g}_m > \hat{g}_m(a, \widehat{\boldsymbol{\theta}})\right\}.$$

Under conditions described in the Appendix, we have

Theorem 2 $\sup_{\mathbf{y}\in\mathbb{R}^n} \left|\widehat{T}^*_m(\mathbf{y}) - T^*(\mathbf{y})\right| \xrightarrow{a.s.} 0.$

In the Appendix, we prove Theorem 2 and show that it ensures (i) convergence to the maximal power function, $\pi(\widehat{T}^*_m;\alpha) \xrightarrow{a.s.} \pi^*(\alpha)$ as $m \to \infty$, and (ii) asymptotic control of FDR through the methods of Storey *et al.* (2004) .

5. CONSTRAINED ALTERNATIVE HYPOTHESES

Tests of a single null hypothesis often require a tradeoff in the choice of an alternative. Under the general linear model, if true alternatives are thought to lie in some linear subspace $\mathcal{V}_s$, with $\mathcal{V}_0 \subset \mathcal{V}_s \subset \mathcal{V}$, we could use the corresponding F-statistic, $F^{(s)}$, to test the hypothesis that $\boldsymbol{\beta}$ is in $\mathcal{V}_s$ against the null hypothesis that $\boldsymbol{\beta}$ is in $\mathcal{V}_0$. If we are correct in supposing that $\boldsymbol{\beta} \in \mathcal{V}_s$ for true alternatives, then the test based on $F^{(s)}$ will have more power to detect departures from the null than a test based on F. However if the true alternatives do not lie in $\mathcal{V}_s$, a test based on $F^{(s)}$ could perform miserably compared to F.

This tradeoff disappears asymptotically when an empirical alternative hypothesis is used for multiple testing. If we suppose that true alternatives lie in $\mathcal{V}_s$ we could define $\widehat{\boldsymbol{\theta}}^{(s)}$ for $\mathcal{V}_s$ analogously to the definition of $\widehat{\boldsymbol{\theta}}$ for $\mathcal{V}$. The optimal test $T^{(s)}$ for alternatives confined to $\mathcal{V}_s$ could then be defined analogously to T^* in (3) using the likelihood ratio for $(F^{(s)}, \widehat{\boldsymbol{\theta}}^{(s)})$. But $T^{(s)}$ can have no more power than T^*.

Proposition 2 *For any linear spaces $\mathcal{V}_0 \subset \mathcal{V}_s \subset \mathcal{V}$, $\pi(T^*;\alpha) \geq \pi(T^{(s)};\alpha)$ for $0 \leq \alpha \leq 1$. When true alternatives are constrained to $\mathcal{V}_s$ we have equality and, furthermore, $T^{(s)}(\mathbf{y}) = T^*(\mathbf{y})$ for almost every $\mathbf{y} \in \mathbb{R}^n$.*

Property 2 follows from the Neyman-Pearson Lemma and the fact that $F^{(s)}$

and $\widehat{\boldsymbol{\theta}}^{(s)}$ are functions of F and $\widehat{\boldsymbol{\theta}}$, with

$$\widehat{\boldsymbol{\theta}}^{(s)} = \frac{\mathbf{P}_{\mathcal{V}_s}\widehat{\boldsymbol{\theta}}}{\|\mathbf{P}_{\mathcal{V}_s}\widehat{\boldsymbol{\theta}}\|}$$

and

$$F^{(s)} = \frac{c_s\|\mathbf{P}_{\mathcal{V}_s}\widehat{\boldsymbol{\theta}}\|^2}{\|\mathbf{P}_{\mathcal{V}_s^\perp}\widehat{\boldsymbol{\theta}}\|^2 + c_1F^{-1}}$$

where $c_1 = (n-d)/(d-d_0)$ and $c_s = (n-d_s)/(d_s-d_0)$ with $d_s = dim(\mathcal{V}_s)$.

It follows from Proposition 2 that when true alternatives lie in $\mathcal{V}_s$, a consistent estimate of T^* is simultaneously consistent for $T^{(s)}$ even though $\mathcal{V}_s$ is unspecified. No price is paid asymptotically, in terms of average power, for using the larger alternative hypothesis $\mathcal{V}$ when true alternatives are confined to a subspace $\mathcal{V}_s$. Of course if true alternatives were constrained to some known $\mathcal{V}_s$ we could estimate $T^{(s)}(\cdot)$ at a faster rate from realizations of $(F^{(s)}, \widehat{\boldsymbol{\theta}}^{(s)})$, as this would require density estimation over fewer dimensions.

6. SIMULATION STUDY

Data sets were simulated from the three-tissue model described in Section 2. Each data set consisted of $m = 2000$ genes with either 50% or 80% of the genes expected to follow the null hypothesis of constant mean expression across the three tissues. The differentially expressed genes were equally likely to follow a linear increase or decrease in mean expression across tissue types 1 though 3, with $\boldsymbol{\beta}$ and σ scaled to control the power of the standard F-test when $\alpha = 0.05$. Under each simulation regime, 100 data sets were generated and analyzed using standard three-sample ANOVA F-tests, F-tests for trend and EAH-tests with FDR controlled asymptotically in all cases using the method of Storey *et al.* (2004). Realized false discovery rates and sensitivities are summarized in Table 1.

Since genes with increasing and decreasing linear trends in expression were simulated with equal probability, independently of the magnitude of $\boldsymbol{\beta}$ and σ, the F-test for trend is in fact the optimal test in $\mathcal{T}$ for these simulated data sets. Due to Proposition 2, the F-test for trend provides a benchmark for the EAH-tests which should converge to the trend tests with increasing m. In general the EAH-tests summarized in Table 1 provided sensitivities close to those of the optimal tests for trend, despite the relatively small sample size of 2000 genes. The benefit of using EAH tests was greatest when F-tests yielded the lowest sensitivities. For example in the setting used to generate the three-tissue example of Section 2, with 25% power and 50% true nulls, the expected sensitivity is more than quadrupled upon moving from F-tests to EAH-tests.

The EAH tests generally controlled FDR near the target level of 10%, with the exception of simulations where 80% of the genes followed the null hypotheses and differential expression was detectable with 25% power. Poor control of FDR in this case was likely due to under-smoothing of the density estimate used to construct the EAH tests. For all simulations we used $b = 100$ bins and a data-dependent kernel bandwidth of $\kappa^{-1} = \widehat{m}_1^{-1/6}$, with $\widehat{m}_1 = m - \sum_{i=1}^{m} I(a_i > 0.8)/0.2$ approximating the number of false nulls as in Storey *et al.* (2004). These choices are somewhat arbitrary, but the dependence on $\widehat{m}_1$ should provide more smoothing when there are few false nulls and the true density is flatter. When the bandwidth κ^{-1} given above was doubled to provide more smoothing the EAH-tests with $p_0 = 0.8$ and 25% power had an average realized FDR of 11.6%, with 25%- and 75%-quantiles (0, 12.5) and average sensitivity of 2%, (.5, 2.7).

7. RHYTHMIC GENE EXPRESSION IN THE MOUSE EYE

A gene expression experiment, fully described in Storch *et al.* (2006), was conducted to identify rhythmically expressed genes in the mouse eye. A population of synchronized mice was reared in a controlled environment with alternating 12-hour periods of light and dark. Every four hours, three mice were randomly sampled and a pooled extraction of mRNA was obtained from their eyes. During periods of scheduled darkness, sampled mice were captured and sacrificed with the aid of night vision goggles. After three days of sampling every four hours, 18 mRNA samples were available for microarray analysis. After preprocessing the microarray data as described in Storch *et al.* (2006), expression levels for 33,377 probe sets were available for statistical analysis.

A design matrix for this experiment was constructed using periodic basis functions as follows. Let $t_1, \ldots t_{18}$ be the sampling times, in cumulative hours, for the $n = 18$ arrays and let $s_j = \pi t_j/24$, $j = 1, \ldots, 18$. Define the 18×6 design matrix $\mathbf{D}$ with jth row

$$\mathbf{D}_j = [1, \sin(2s_j), \cos(2s_j), \sin(4s_j), \cos(4s_j), \cos(6s_j)]'.$$

A matrix $\mathbf{X}$ with orthonormal columns can be obtained by standardizing the columns of $\mathbf{D}$. A gene's mean expression level across the 18 arrays can then be modeled as $\mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5)'$.

Only rhythmic expression with a period of 24 hours was of interest, so the null hypothesis $\beta_1 = \beta_2 = 0$ was chosen to allow higher frequency variation in mean expression over time. Note that since every sixth sampling occurred at the same time of day, the full model with $\boldsymbol{\beta} \in \mathbb{R}^6$ is equivalent to allowing an unrestricted mean for each of the six unique sampling times.

The null hypothesis that $\beta_1 = \beta_2 = 0$ was tested for each gene using standard F-tests (with 2 and 12 degrees of freedom) and using the EAH procedure described above with $\kappa = \widehat{m}_1^{1/6}$ and $b = 400$. The distributional assumptions of the general linear model were checked using the approach described in Remark D.

At any level of FDR, controlled asymptotically as in Storey *et al.* (2004), more genes were detected as rhythmic by the EAH procedure than by F-tests (Figure 3). With FDR controlled asymptotically at 10%, F-tests detected $1,975$ genes as rhythmic and the EAH procedure detected $3,018$ genes as rhythmic.

The distribution of $(a, \widehat{\boldsymbol{\theta}})$ for this data set is shown in Figure 4. Each gene's $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2)'$ can be represented by the time t^* at which

$$\widehat{\theta}_1 \sin(2\pi t/24) + \widehat{\theta}_2 \cos(2\pi t/24)$$

achieves its maximum over $t \in (0, 24]$, which we call the estimated phase. The EAH rejection region in Figure 4 detects more genes as rhythmic by adapting to clustering among the estimated phases. The dense clustering of estimated phases just prior to 0/24 hours corresponds to genes achieving their peak expression levels just before the mice entered a 12-hour period of illumination beginning at time 0. A smaller group of genes have estimated phases clustered prior to hour 12, when the lights were turned off. Noticeably few genes have estimated phases during the initial 4 hours of darkness. The EAH analysis suggests that many of the genes detected as rhythmic by the F-tests with estimated phases between 12 and 16 hours are likely to be false positives.

Source code, in the R language, for all analyses in this paper, is available from the author.

8. DISCUSSION

The asymptotically optimal procedure identified in this paper augments the classical F-statistic by measuring the direction of apparent deviation from the null hypothesis, $\widehat{\boldsymbol{\theta}}$. When testing a single hypothesis, $\widehat{\boldsymbol{\theta}}$ generally provides no useful information. We have shown that when testing multiple hypotheses, the collection of $\widehat{\boldsymbol{\theta}}$'s taken together provides information about the data-generating distribution that can greatly increase the sensitivity of multiple testing procedures.

In the case of gene expression data, patterns of association among the $\widehat{\boldsymbol{\theta}}$'s describe clusters of genes with similar mean expression profiles. Often the identification of such clustering is a final goal of gene expression analysis, with clustering algorithms applied to a list of genes deemed significantly differentially expressed by a multiple testing procedure. It is therefore appealing that the EAH procedure provides a principled method for incorporating apparent clustering into the detection of differential expression. Furthermore, the prevalence of clustering in real data sets indicates that EAH tests will often lead to substantial increases in sensitivity over procedures that ignore clustering.

Given that EAH tests and F-tests produce different significance rankings, experimenters will wonder which method better meets their needs. It may seem that EAH tests are 'unfair,' in that differentially expressed genes with unique directions of deviation from the null are disadvantaged relative to genes sharing their direction of deviation with many others. Would it not be more fair to individually evaluate each gene's significance level? We disagree with this point of view. As can be seen in Figure 1, and inferred from Figure 4, clusters of differentially expressed genes expand the F-tests'

rejection region uniformly in $\widehat{\boldsymbol{\theta}}$, admitting primarily false positives outside the strong clusters. Testing procedures based only on F-statistics are therefore unfair in that the conditional false positive rate can vary dramatically across directions of deviation $\widehat{\boldsymbol{\theta}}$.

The EAH procedure is asymptotically optimal among a class of tests satisfying two requirements. The requirement that tests have known null distributions avoids the difficult problem of separately estimating both the null and alternative distributions from mixed observations. The invariance requirement is not necessary in all experimental settings. In oligonucleotide arrays, for example, all expression measurements are normalized to the same scale and $\mathcal{G}$-invariant tests could miss information relevant to multiple testing. For example false nulls could be more prevalent among genes with high levels of mean expression. Optimal testing without the invariance requirement is an interesting direction for future research.

If the Gaussian model assumed in this paper does not hold, simple transformations, in the Box-Cox family for example, together with the model-checking techniques described in Remark D may make the data amenable to an EAH procedure. Efron *et al.* (2001) avoid parametric assumptions by relying on a cleverly designed experiment that allows the null distribution to be estimated from contrasts between arrays that eliminate the treatment effects under investigation. Generalizations of this experimental design that facilitate estimation of the null distribution for $(a, \widehat{\boldsymbol{\theta}})$ would have great practical value. Gao (2006) has studied the estimation of null distributions for more general F-statistics.

This paper has illustrated the EAH procedure using null and alternative hypotheses that differ by two dimensions. EAH procedures can easily be ap-

plied to higher-dimensional alternatives as long as enough data are available to estimate the EAH. If the data are too sparse, some dimension reduction could be practical. For example, the empirical alternative could be restricted to the leading principle components of the empirical distribution of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$.

Remark A: Stochastic ordering and FDR control

Suppose T^* is stochastically minimal in $\mathcal{T}$, with a uniformly consistent estimator $\widehat{T}^*_m$, and let T be any other test in $\mathcal{T}$. For a sample $\mathbf{y}_i$, $i = 1, \ldots, m$, let $\widehat{F}^*_m(t) = m^{-1}\sum_{i=1}^m I\{\widehat{T}^*_m(\mathbf{y}_i) \le t\}$, $\widehat{F}_m(t) = m^{-1}\sum_{i=1}^m I\{T(\mathbf{y}_i) \le t\}$, $F^*(t) = Pr\{T^*(\mathbf{y}) \le t\}$ and $F(t) = Pr\{T(\mathbf{y}) \le t\}$.

Given any test $T \in \mathcal{T}$, Storey *et al.* (2004) show that FDR is controlled asymptotically at level α by rejecting all H_i such that $T(\mathbf{y}_i) \le t_\alpha(\widehat{F}_m)$ with the data-dependent threshold

$$t_\alpha(\widehat{F}_m) = \sup\left\{0 \le t \le 1 : \frac{\widehat{F}_m(t)}{1 - \widehat{F}_m(\lambda)} \ge \frac{t}{\alpha(1-\lambda)}\right\},$$

depending on a tuning parameter $0 \le \lambda \le 1$.

We have $t_\alpha(\widehat{F}_m) \xrightarrow{a.s.} t_\alpha(F)$ and, due to Theorem 2, $t_\alpha(\widehat{F}^*_m) \xrightarrow{a.s.} t_\alpha(F^*)$. Assuming that T is reasonable, $F(t) \ge t$ and by the minimality of T^* we have $F^*(t) \ge F(t)$, $0 \le t \le 1$. From

$$\frac{F^*(t)}{1 - F^*(\lambda)} \ge \frac{F(t)}{1 - F(\lambda)}, 0 \le t \le 1,$$

it follows that

$$t_\alpha(F^*) \ge t_\alpha(F)$$

and therefore

$$F^*\{t_\alpha(F^*)\} \ge F\{t_\alpha(F)\}.$$

In the limit, $\widehat{T}_m^*$ will reject a larger fraction of false nulls than any other $T \in \mathcal{T}$ when FDR is controlled using a data-dependent threshold as in Storey *et al.* (2004).

Remark B: Flat Parts in g

We assume that g_1 has no flat parts, i.e. sets of the form $\{(a,\theta) : g_1(a,\theta) = c\}$ with non-zero measure. This is guaranteed under the general linear model with an unrestricted alternative hypothesis, and therefore g will have no flat parts if the proportion of false nulls is non-zero.

To make T^* well-defined when g has flat parts we could adopt the following convention. Since c will always correspond to some $c = g(a_0, \widehat{\boldsymbol{\theta}}_0)$, we could redefine T^* in (3), and the notation in (4), by replacing the integration over $[0,1] \times \mathcal{S}_{\mathcal{V}^\perp}$ with integration over

$$\{(a,\theta) \in [0,1] \times \mathcal{S}_{\mathcal{V}^\perp} : g(a,\theta) > g(a_0, \widehat{\boldsymbol{\theta}}_0) \text{ or } g(a,\theta) = g(a_0, \widehat{\boldsymbol{\theta}}_0), a < a_0\}.$$

Remark C: Sorted Histograms

This application required a density estimator guaranteed to be non-increasing in a over a bounded support. The Grenander estimator (e.g., Van der vaart 1998, pp. 349-353), while necessarily monotone, was found to be unsuitably sensitive to small perturbations of the data, leading to rejection regions with erratic boundaries for small samples. The sorted histogram estimator described in Section 4 was more stable and can only improve on the unsorted histogram when the true density is monotone. For example, consider two true frequencies $f_1 \geq f_2$ and estimates $\hat{f}_1 < \hat{f}_2$ and note that

$$|f_1 - \hat{f}_1| \vee |f_2 - \hat{f}_2| \geq |f_1 - \hat{f}_2| \vee |f_2 - \hat{f}_1|.$$

Exchanging the order, using $\hat{f}_1$ to estimate f_2 and $\hat{f}_2$ to estimate f_1, can't increase the maximum absolute deviation of the estimates from the truth.

Sorting a histogram with many bins can be accomplished by a sequence of such pairwise exchanges.

Remark D: Model Checking

Let $\mathbf{A}$ be an $(n-d)\times n$ matrix with orthonormal rows spanning the nullspace of $\mathbf{X}'$. Then since

$$\mathbf{z} \equiv \mathbf{A}\mathbf{y} \sim N(0, \sigma^2 \mathbf{I}_{n-d}),$$

regardless of $\boldsymbol{\beta}$, $\mathbf{z}$ may be used for model checking. For example, after partitioning $\mathbf{z} = (\mathbf{z}_1', \mathbf{z}_2')'$ the statistic $\|\mathbf{z}_1\|^2/\|\mathbf{z}_2\|^2$ should, after rescaling, follow a central F distribution with the appropriate degrees of freedom and be independent of $\mathbf{z}_1/\|\mathbf{z}_1\|$, which is uniformly distributed over its support.

Applying this approach to the mouse data of Section 7, we choose $\mathbf{z}_1$ to be 2×1 and $\mathbf{z}_2$ to be 10×1. By making a plot similar to Figure 4, but using $F \propto \|\mathbf{z}_1\|^2/\|\mathbf{z}_2\|^2$ and $\widehat{\boldsymbol{\theta}} = \mathbf{z}_1/\|\mathbf{z}_1\|$, we visually detected no departure from the uniform distribution expected under the general linear model.

Remark E: Invariance for Multiple Testing

Note that $\widehat{T}_m^*$ depends on the $\widehat{\boldsymbol{\theta}}$'s only through their pairwise inner products. This ensures that a multiple testing procedure based on $\widehat{T}_m^*$ is invariant not just to the product group $\mathcal{G}^m$, but also to certain orthogonal transformations that preserve inner products among the $\widehat{\boldsymbol{\theta}}_i$'s. This Remark describes this desirable invariance property in more detail.

The general F-statistic defined in (1) is invariant to linear transformations of the form

$$\mathbf{y} \to a\mathbf{Q}\mathbf{y} + \mathbf{v} \tag{6}$$

where $a \in \mathbb{R}$, $\mathbf{v} \in \mathbf{X}\mathcal{V}_0$ and $\mathbf{Q}$ is a member of a special subgroup of orthogonal matrices as described by Lehmann (1986, pp. 365-368). If we let $\mathcal{G}_F$ denote

this group of linear transformations, where each $g \in \mathcal{G}_F$ specifies an a, $\mathbf{Q}$ and $\mathbf{v}$ to be applied to a single data vector $\mathbf{y}$ as in (6), then the multiple testing procedure based only on F-statistics will be invariant under the product group $\mathcal{G}_F^m$ containing transformations of the form $\mathbf{g} = (g_1, \ldots, g_m)$ that act on the entire data set with $\mathbf{g}(\mathbf{y}_1, \ldots, \mathbf{y}_m) = (g_1\mathbf{y}_1, \ldots, g_m\mathbf{y}_m)$ where each $g_i \in \mathcal{G}_F$ specifies an a_i, $\mathbf{Q}_i$ and $\mathbf{v}_i$ to be applied to $\mathbf{y}_i$ such that

$$g_i\mathbf{y}_i = a_i\mathbf{Q}_i\mathbf{y}_i + \mathbf{v}_i, \quad i = 1, \ldots, m.$$

The EAH procedure based on $\widehat{T}_m^*$ is invariant under the smaller group of transformations

$$\mathcal{G}_0^m = \{\mathbf{g} \in \mathcal{G}^m : \widehat{\boldsymbol{\theta}}(\mathbf{y}_1)^t\widehat{\boldsymbol{\theta}}(\mathbf{y}_2) = \widehat{\boldsymbol{\theta}}(g_i\mathbf{y}_1)^t\widehat{\boldsymbol{\theta}}(g_j\mathbf{y}_2) \text{ for all } 1 \le i < j \le m \text{ and any } \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n\}.$$

It is easily verified that $\mathcal{G}_0^m$ is a subgroup of $\mathcal{G}_F^m$ that essentially ensures that all $\mathbf{y}_i$'s are subjected to the same orthogonal transformation affecting the $\widehat{\boldsymbol{\theta}}_i$'s. It can also be shown that the collection of inner-products $\widehat{\boldsymbol{\theta}}(\mathbf{y}_i)^t\widehat{\boldsymbol{\theta}}(\mathbf{y}_j)$, $1 \le i < j \le m$, together with the F-statistics $F(\mathbf{y}_i)$, $i = 1, \ldots, m$, comprise a maximal invariant under $\mathcal{G}_0^m$ (proof available from author upon request).

Invariance of an entire multiple testing procedure under $\mathcal{G}_0^m$ is desirable in that the procedure is sensitive to the relative directions of deviation from the null without regard to any prior reference point. In the two tissue example, the EAH test is sensitive to any imbalance in the frequency of over- versus under-expression, without any prior bias towards either. In the three-sample problem of Section 2, the EAH test is invariant to any relabeling of the tissue groups as long as they are relabeled in the same way for all genes. More generally, in the case of differential expression across any number of tissue types,

with balanced sample allocation, the inner product $\widehat{\boldsymbol{\theta}}(\mathbf{y}_i)^t\widehat{\boldsymbol{\theta}}(\mathbf{y}_j)$ is simply the Pearson correlation between $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\boldsymbol{\beta}}_j$. In this case $\mathcal{G}_0^m$-invariant multiple testing procedures are sensitive to the pattern of pairwise correlation among estimated mean expression levels.

APPENDIX: PROOFS

To prove Theorem 2, note that

$$\begin{aligned}\sup_{\mathbf{y}\in\mathbb{R}^n}|\widehat{T}_m^*(\mathbf{y})-T^*(\mathbf{y})| &\leq \sup_{a,\theta}|G_0\{\widehat{g}_m>\widehat{g}_m(a,\theta)\}-G_0\{g>g(a,\theta)\}|\\ &\leq \sup_{c\geq 0}G_0\{|g-c|<2\phi_m\},\end{aligned}$$

with

$$\phi_m=\sup_{a,\theta}|\widehat{g}_m(a,\theta)-g(a,\theta)|.$$

For simplicity we continue to assume that g has no flat parts. The proof is therefore completed by showing that $\widehat{g}_m$ is strongly and uniformly consistent for g. Following Remark C, if the unsorted predecessor of $\widehat{g}_m$ is uniformly consistent, then $\widehat{g}_m$ is uniformly consistent for the same function. Without the sorting operation, $\widehat{g}_m$ employs the kernel

$$K(u_1,u_2)=\exp(u_1)I\{|u_2|<1\}.$$

with $u_1=\kappa\boldsymbol{\theta}'\boldsymbol{\theta}_0$ and $u_2=2b(a_0-a)$ for the bin centers $a_0\in\{(2k+1)/(2b):k=0,1,\ldots,b-1\}$. Functions of this form, indexed by $(a_0,\boldsymbol{\theta}_0)\in[0,1]\times\mathcal{S}_{\mathcal{V}_0^\perp}$, $\kappa>0$ and $b\in\mathbb{N}$, constitute a Vapnik-Červonenkis class of measurable functions on $[0,1]\times\mathcal{S}_{\mathcal{V}_0^\perp}$, satisfying the conditions of Giné *et al.* (2004). Under mild conditions on g and for appropriate sequences $\kappa_m\to\infty$ and $b_m\to\infty$, the results of Giné *et al.* (2004) can be used to show almost sure uniform convergence of $\widehat{g}_m$ to g.

Since gene expression data often exhibit dependency across genes, we note that Theorem 2 may be extended to sequences $\{\mathbf{y}_i\}_{i=1}^{\infty}$ of dependent random variables using the results of Nobel and Dembo (1993) for general empirical processes. For example, Theorem 2 will still hold if genes are dependent only within finite blocks.

Uniform convergence of the power function $\pi(\widehat{T}_m^*;\alpha)$ to $\pi^*(\alpha)$ follows from the weak convergence of $\widehat{T}_m^*(\mathbf{y})$ to $T^*(\mathbf{y})$ implied by Theorem 2.

We can satisfy the sufficient conditions for asymptotic FDR control given by Storey *et al.* (2004) by showing that

$$|\widehat{F}_m^*(t) - F^*(t)| \xrightarrow{a.s.} 0 \text{ as } m \to \infty$$

for almost every t, where $\widehat{F}_m^*(t)$ and $F^*(t)$ are defined in Remark A. Let $\widetilde{F}_m^* = m^{-1}\sum_{i=1}^m I\{T^*(\mathbf{y}_i) \le t\}$. For any t we have

$$|\widehat{F}_m^*(t) - F^*(t)| \le |\widehat{F}_m^*(t) - \widetilde{F}_m^*(t)| + \sup_t \|\widetilde{F}_m^*(t) - F^*(t)\|$$

with the second term on the right converging almost surely to 0. Theorem 2 ensures that the first term on the right converges almost surely to zero for almost every t since

$$|\widehat{F}_m^*(t) - \widetilde{F}_m^*(t)| \le m^{-1}\sum_{i=1}^m I(|T^*(\mathbf{y}_i) - t| \le \varepsilon_m)$$

with $\varepsilon_m = \sup_{\mathbf{y}\in\mathbb{R}^n} |\widehat{T}_m^*(\mathbf{y}) - T^*(\mathbf{y})|$. To completely satisfy the conditions of Storey *et al.* (2004), the above convergence of empirical distribution functions must be shown to occur separately for sequences of true and false nulls. This is easily verified in the current setting.

REFERENCES

[1] 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.

[2] Dudoit, S., van der Laan, M.J., and Pollard, K.S. (2004), "Multiple Testing Part I: Single-Step Procedures for Control of General Type I Error Rates," *Statistical Applications in Genetics and Molecular Biology*, 3,1, Article 13.

[3] 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.

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

[5] Gao, X. (2006), "Construction of Null Statistics in Permutation-Based Multiple Testing for Multi-factorial Microarray Experiments," *Bioinformatics*, 22,1486-1494.

[6] Giné, E., Koltchinskii, V., and Zinn, J. (2004), "Weighted Uniform Consistency of Kernel Density Estimators," *The Annals of Probability*, 32,2570-2605.

[7] Hall, P., Watson, G.S., and Cabrera, J. (1987), "Kernel Density Estimation With Spherical Data," *Biometrika*, 74,751-762.

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

[9] Nobel, A., and Dembo, A. (1993), "A Note on Uniform Laws of Averages for Dependent Processes," *Statistics & Probability Letters*, 17,169-172.

[10] 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.

[11] 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.

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

[13] 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.

[14] Storey, J.D. (2005), "The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing," *UW Biostatistics Working Paper Series*,Paper 259, https://www.bepress.com/uwbiostat/paper259/.

[15] Storch, K-F., Paz, C., Signorovitch, J., Raviola, E., Pawlyk, B., Li, T., and Weitz, C. (2006), "Circadian Clock in The Mammalian Retina: Importance for Retinal Responses to Light," *under review*.

[16] van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.

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

Table 1: Means and quantiles (25%, 75%) summarizing the performance of multiple testing procedures applied to simulated data sets with FDR controlled asymptotically at 10%.

		Realized FDR (%)			Realized Sensitivity (%)		
p_0	Power	F-test	Trend	EAH	F-test	Trend	EAH
0.5	0.25	8.8	8	8.6	3.9	18.8	18.3
		(5, 12)	(6.1, 9.7)	(6.8, 10.1)	(1.9, 5.7)	(16.3, 21.3)	(14.9, 21.9)
	0.5	9.6	9.5	10	51.6	71.3	70.5
		(9, 10)	(8.7, 10.3)	(8.8, 10.8)	(49.7, 53.3)	(69.3, 73.1)	(68.6, 72.8)
	0.75	10	10	10.6	84.4	92.9	91.6
		(9, 11)	(9.2, 10.6)	(9.5, 11.7)	(83.2, 85.6)	(92.1, 93.6)	(90.7, 92.5)
	Random	8.6	8.7	9.3	52.3	64.7	63.8
		(8, 10)	(7.7, 9.5)	(8, 10.3)	(50.8, 54.2)	(63.1, 66.5)	(62.4, 65.6)
0.8	0.25	7.1	9.6	16	0.3	2.1	2.3
		(0, 0)	(0, 17.8)	(0, 24.2)	(0, 0.5)	(0.5, 3.3)	(0.5, 3.2)
	0.5	9.6	9.6	10.2	13.8	35.5	32.5
		(7, 12)	(8, 11.4)	(8.2, 12.3)	(10, 17.5)	(31.4, 39.7)	(29.3, 38.7)
	0.75	10.3	9.8	11	53.8	74.5	72.3
		(9, 12)	(8.3, 11.5)	(9.6, 12.9)	(51.6, 56.2)	(72.7, 76)	(69.7, 75.5)
	Random	9.8	9.6	10.7	28.7	43.8	41.3
		(8, 12)	(8, 10.8)	(8.3, 13)	(26.4, 30.8)	(41.6, 46)	(37.5, 44.9)

NOTE: p_0 gives the expected fraction of true null hypotheses. Random powers were uniformly distributed over [0,1].

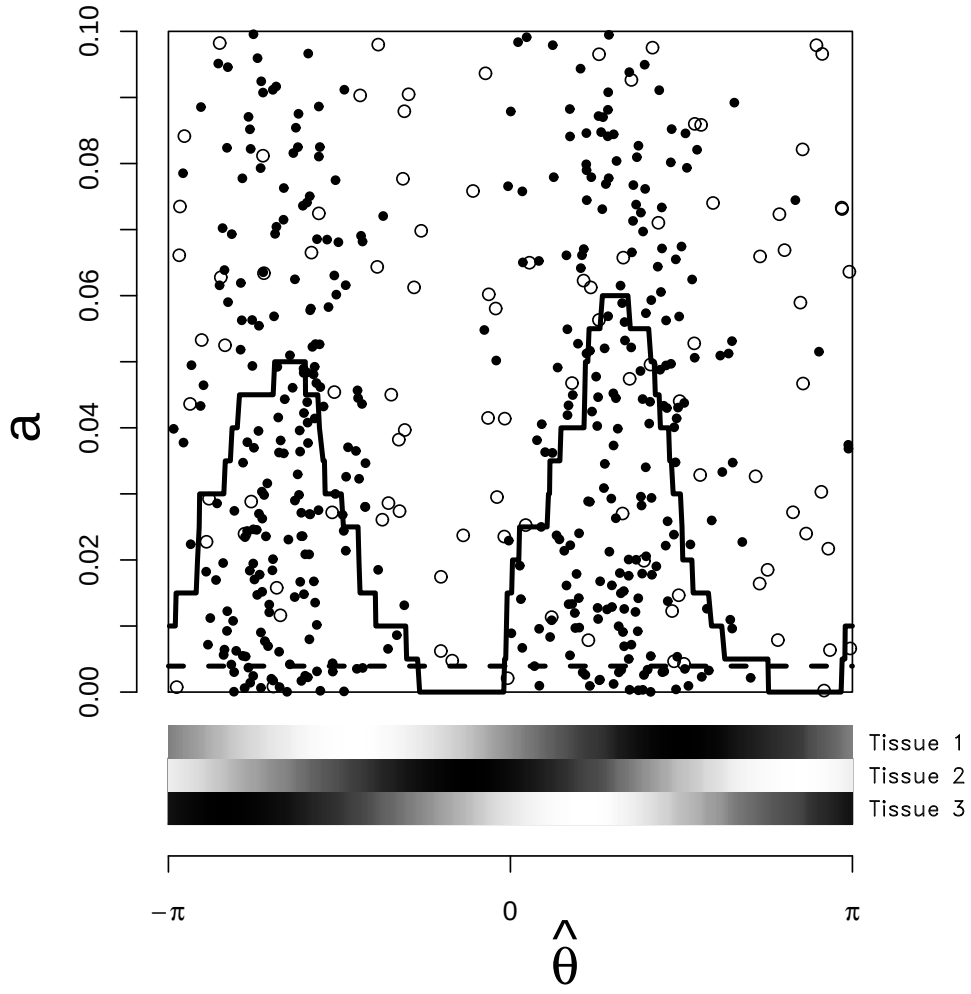

Figure 1: Simulated statistics $(a, \hat{\theta})$ corresponding to true null hypotheses (open circles) and false null hypotheses (filled circles). Statistics falling below the dashed line at $a = 0.004$ correspond to hypotheses rejected by F-tests. The solid line defines the rejection region generated by the procedure proposed in this paper. Both rejection regions were obtained with FDR controlled asymptotically at 10%. Shaded bars illustrate for each value of $\hat{\theta}$ the relative mean expression levels across tissues.

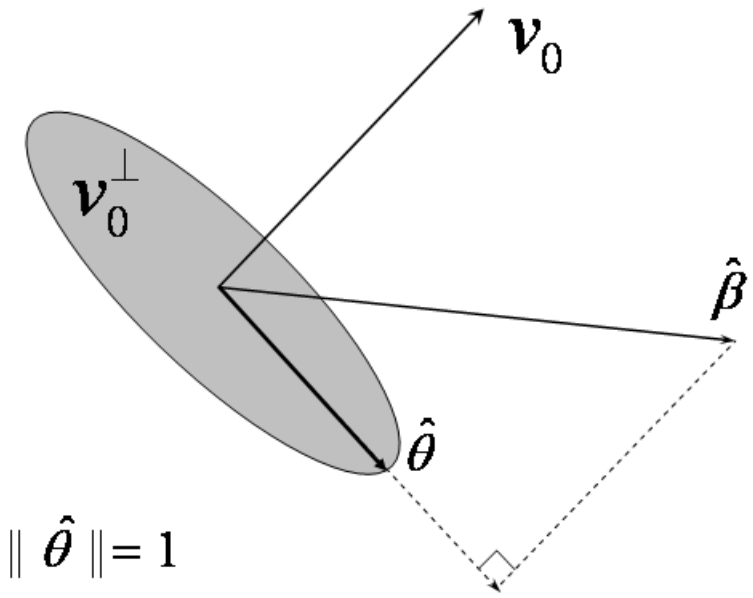

Figure 2: $\widehat{\boldsymbol{\theta}}(\mathbf{y}) \in \mathcal{V}_0^{\perp}$ measures the direction of apparent deviation of $\widehat{\boldsymbol{\beta}}(\mathbf{y})$ from the null hypothesis $\boldsymbol{\beta} \in \mathcal{V}_0$.

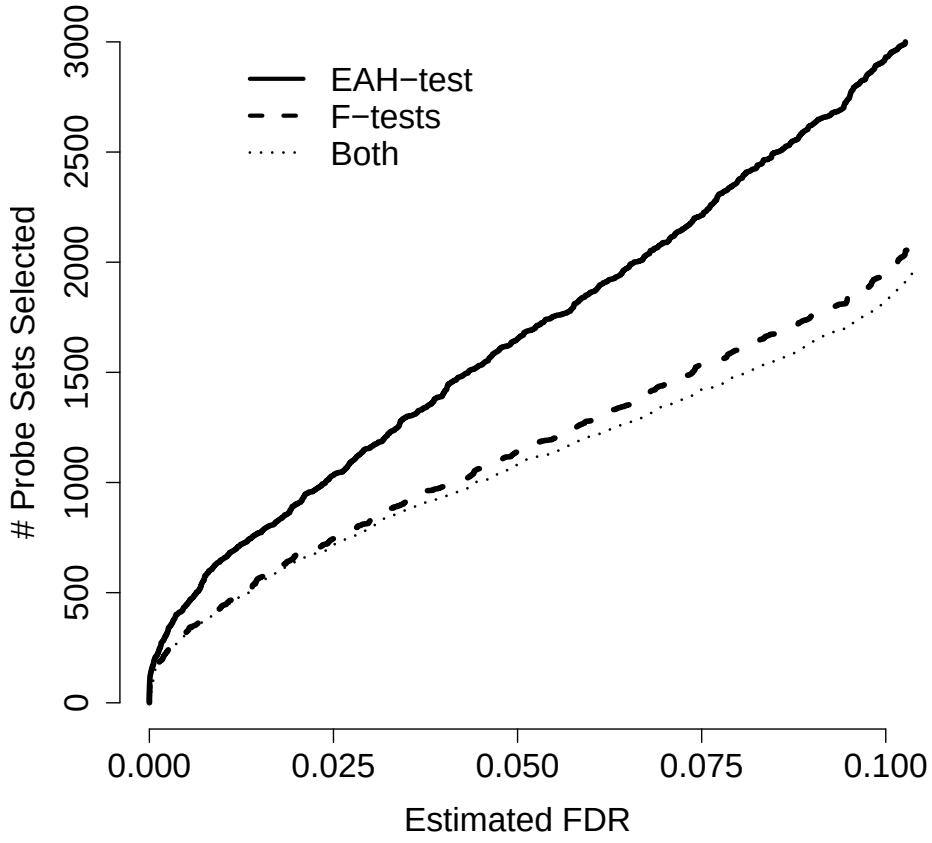

Figure 3: Numbers of rejected hypotheses as a function of estimated FDR in the mouse expression data.

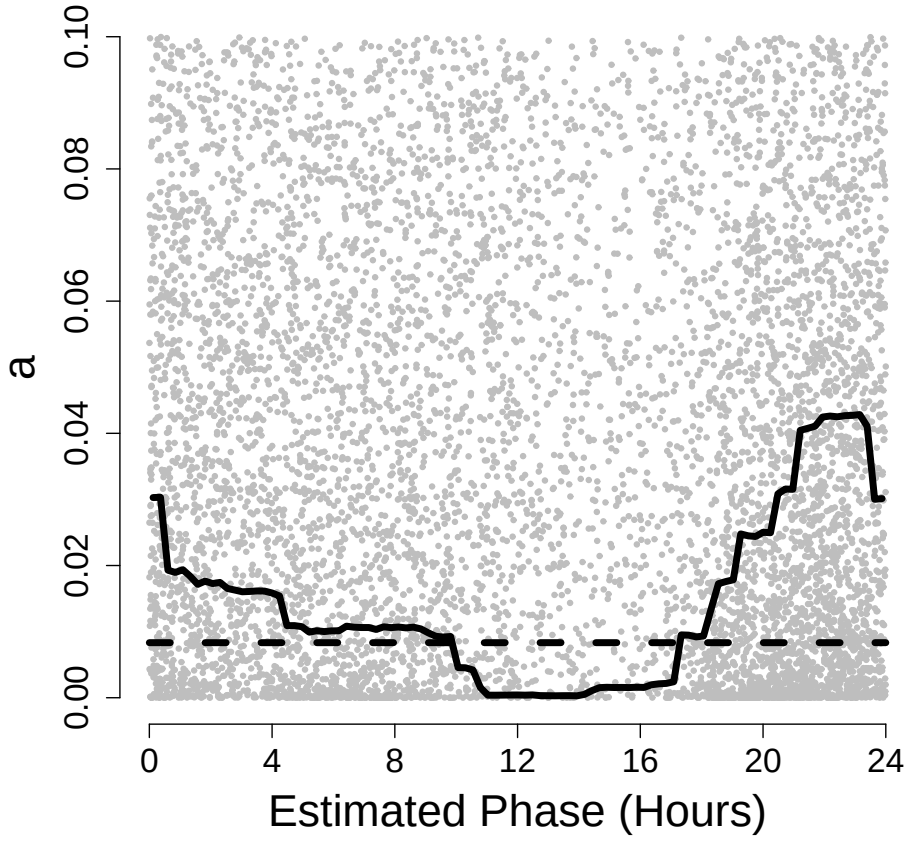

Figure 4: Rejection regions for the mouse gene expression data. Grey circles correspond to the values of a and $\widehat{\boldsymbol{\theta}}$ for each gene, with $\widehat{\boldsymbol{\theta}}$ represented by the estimated phase. Genes detected as rhythmic by the F-tests fall below the dashed line and genes detected by the EAH-tests fall below the solid line. Both rejection regions were obtained with FDR controlled asymptotically at 10%.