11-2005

Exploring the Information in P-values

David Ruppert
Cornell University

Dan Nettleton
Iowa State University, dnett@iastate.edu

J.T. Gene Hwang
Cornell University

Exploring the Information in P-values

Abstract
A new methodology is proposed for estimating the proportion of true null hypotheses in a large collection of tests. The proportion of true null hypotheses is needed, for example, when controlling the false discovery rate in the analysis of microarray data. We assume that each test concerns a single parameter δ whose value is specified by the null hypothesis. Our methodology combines a parametric model for the conditional CDF of the p-value given δ with a nonparametric spline model for the density $g(\delta)$ of δ under the alternative hypothesis. The proportion of true null hypotheses and the coefficients in the spline model are estimated by penalized least-squares subject to constraints that guarantee that the spline is a density. The constrained, penalized leastsquares estimator is computed efficiently using quadratic programming. Our procedure gives estimators with less bias compared to other estimators in the literature, which are positively biased because they are based upon an estimate of the marginal density of the p-values at 1. We define three estimators with different degrees of positive bias. In a simulation study, we compare our estimators to the convex, decreasing estimator of Langaas, Ferkingstad, and Lindqvist. The most biased of our estimators is very similar in performance to the convex, decreasing estimator. Our methodology produces an estimate $gb(\delta)$ of the density of δ when the null is false and can address such questions as "when the null is false, is the parameter usually close to the null or far away?" This leads us to define a "falsely interesting discovery rate" (FIDR), a generalization of the false discovery rate. We contrast the FIDR approach to Efron's "empirical null hypothesis" technique. We discuss the use of gb in sample size calculations based on the expected discovery rate (EDR). As an illustration, we analyze differences in gene expression between resistant and susceptible strains of barley after the plants have been exposed to a pathogen.

Keywords
expected discovery rate, false discovery rate, inverse problem, microarray, penalty, power and sample size, quadratic programming, simultaneouse tests, splines

Disciplines
Statistics and Probability

Comments
This preprint was published as David Ruppert, Dan Nettleton, and J.T. Gen Hwang, "Exploring the Information in p-Values for the Analysis and Planning of Multiple-Test Experiments" *Biometrics* (2007): 483-495, doi: 10.1111/j.1541-0420.2006.00704.x.

Exploring the Information in P-values

David Ruppert* Dan Nettleton† J. T. Gene Hwang‡

November 9, 2005

Abstract

A new methodology is proposed for estimating the proportion of true null hypotheses in a large collection of tests. The proportion of true null hypotheses is needed, for example, when controlling the false discovery rate in the analysis of microarray data. We assume that each test concerns a single parameter δ whose value is specified by the null hypothesis. Our methodology combines a parametric model for the conditional CDF of the p-value given δ with a nonparametric spline model for the density $g(\delta)$ of δ under the alternative hypothesis. The proportion of true null hypotheses and the coefficients in the spline model are estimated by penalized least-squares subject to constraints that guarantee that the spline is a density. The constrained, penalized least-squares estimator is computed efficiently using quadratic programming. Our procedure gives estimators with less bias compared to other estimators in the literature, which are positively biased because they are based upon an estimate of the marginal density of the p-values at 1. We define three estimators with different degrees of positive bias. In a simulation study, we compare our estimators to the convex, decreasing estimator of Langaas, Ferkingstad, and Lindqvist. The most biased of our estimators is very similar in performance to the convex, decreasing estimator. Our methodology produces an estimate $\widehat{g}(\delta)$ of the density of δ when the null is false and can address such questions as "when the null is false, is the parameter usually close to the null or far away?" This leads us to define a "falsely interesting discovery rate" (FIDR), a generalization of the false discovery rate. We contrast the FIDR approach to Efron's "empirical null hypothesis" technique. We discuss the use of $\widehat{g}$ in sample size calculations based on the expected discovery rate (EDR). As an illustration, we analyze differences in gene expression between resistant and susceptible strains of barley after the plants have been exposed to a pathogen.

1 Introduction

We consider the problem of testing $H_{0i} : \delta_i = 0$ for $i = 1, \ldots, n$. Let π_0 be the proportion of $\delta_1, \ldots, \delta_n$ equal to the null value 0. We assume that the remaining proportion $(1 - \pi_0)$ have

*Andrew Schultz Jr. Professor of Engineering and Professor of Statistical Science, School of Operations Research and Industrial Engineering, Cornell University, Rhodes Hall, Ithaca, NY 14853, USA. E-mail: dr24@cornell.edu. Research supported by NSF Grant DMS 04-538 and NIH Grant CA57030.

†Associate Professor, Department of Statistics, Iowa State University, Ames, Iowa 50011-1210, USA. Email: dnett@iastate.edu. D. Nettleton acknowledges support of the Cooperative State Research, Education, and Extension Service, U.S. Department of Agriculture, under Agreement No. 2002-35300-12619.

‡Professor, Department of Mathematics, Cornell University, Malott Hall, Ithaca, NY 14853, USA. E-mail: hwang@math.cornell.edu.

an empirical distribution G_n that can be well approximated by a continuous distribution G with density g. We estimate both π_0 (the proportion of tests with true null hypotheses) and G (the distribution of δ_i for tests with true alternatives). Estimates of π_0 are useful for several purposes such as estimation of the false discovery rate (FDR). An estimate of G can be used for sample size calculations based on the expected discovery rate (EDR) and to determine the proportion null hypotheses that are "false but uninteresting" meaning that the null is false but δ is close to the null value.

Suppose that the parameters $\delta_1, \ldots, \delta_n$ have associated p-values, $p_1, \ldots, p_n$, with p_i coming from a test of $H_{0i} : \delta_i = 0$ versus either a one or two-sided alternative. The conditional CDF of p_i given δ_i will be denoted by $F_{p|\delta}(p\,;\delta_i)$, e.g., for a t-test $F_{p|\delta}$ would be derived from a non-central t-distribution with non-centrality parameter δ_i. Since the p-value is uniformly distributed under H_0, the marginal CDF of p_i is

$$F_p(p\,;\pi_0) = \pi_0 p + (1-\pi_0)\mathrm{EDR}(p) \tag{1}$$

where

$$\mathrm{EDR}(p) = \int_{-\infty}^{\infty} F_{p|\delta}(p\,;\delta)dG(\delta) \tag{2}$$

is the Expected Discovery Rate (Gadbury et al., 2004). If one fixes α and varies δ, then $F_{p|\delta}(\alpha)$ is the power curve of a level-α test and $\mathrm{EDR}(\alpha)$ is the expected power.

Much of the recent interest in estimation of π_0 is due to applications to false discovery rates. However, there are other interesting applications, e.g., Meinshausen and Rice (2005) discuss the problem of estimating the number of objects in the Kuiper Belt. These objects are detected by a reduction in light when they pass between a star and an observer. The null hypothesis is that there is no reduction, and the number of false null hypotheses gives information about the number of objects.

Currently, the most popular estimators of π_0 are equal to some estimator of the p-value density evaluating at 1, i.e., of $f_p(1;\pi_0) = F_p'(1;\pi_0)$. The underlying assumption is that p-values near 1 come from the null. However, this need not be true, especially if G has considerable probability near 0. In one of the cases of the simulation study of Section 8, π_0 is 0.7 but estimates of $f_p(1;\pi_0)$ are near 0.85; thus, the probability of a false null is twice what one of the currently available estimators would report. The semiparametric estimators proposed in this paper are designed to reduce this positive bias and our simulations show that this can be done successfully.

Although our original and still primary goal is to estimate π_0, a secondary goal is to learn something about G, the distribution of δ under the alternative. We believe that it is an advantage of our approach that we estimate G along with π_0. An important question is "when is a false null interesting?" The answer will be determined mostly by the application,

but knowing something about g will also influence the answer. We consider the possibility of "false but uninteresting null hypotheses" and generalize the false discovery rate to a "falsely interesting discovery rate" (FIDR). We also contrast our approach with the "empirical null hypothesis" of Efron (2004).

Our methodology is related to methods for estimation of mixing distributions, deconvolution, and other inverse problems, e.g., O'Sullivan (1986), Carroll and Hall (1988), Fan (1991), and Lesperance and Kalbfleisch (1992). Estimation of (π_0, g) is an inverse problem, which we approach in the same way as O'Sullivan (1986) by using B-splines with a roughness penalty, one slight difference being that we have a mixing distribution with a discrete component coming from the null hypothesis and another being the constraints we use to make $\widehat{g}$ a density.

Our estimation methodology is described in Section 2–6. The promising convex, decreasing estimator of Langaas, Ferkingstad, and Lindqvist (2005) and two additional estimators based on our semiparametric approach are introduced in Section 7. A simulation study is represented in Section 8 compares our estimators to the convex, decreasing estimator. Section 9 discusses estimation of the false discovery rate. In Section 10 the "falsely interesting discovery rate" is defined. Power and sample size calculations are discussed in Section 11, where we show how to estimate the expected discovery rate (EDR), the true positive (TP) probability, and the true negative (TN) probability for differernt sample sizes. A example using gene expression data is in Section 12 and a summary in Section 13.

2 The Semiparametric Estimator

We will model g as $g(\delta\,;\boldsymbol{\beta})$ where $g(\cdot\,;\cdot)$ is a known function and $\boldsymbol{\beta}$ is a vector of parameters. There is a wide variety of possible choices for $g(\delta\,;\boldsymbol{\beta})$ ranging from "classical" parametric models such as the lognormal, Gamma, and Beta families to "flexible" parametric models such as mixtures of Gamma or Beta densities or splines. We will use splines so that $\boldsymbol{\beta}$ is vector of coefficients of the basis functions. Let $F_p(\cdot\,;\pi_0,\boldsymbol{\beta})$ be given by (1) with $g(\delta)$ replaced by $g(\delta\,;\boldsymbol{\beta})$. Flexible parametric families are for practical purposes nonparametric, since the number of parameters can be arbitrarily large and some type of "regularization" is needed for stable estimation. Regularization can be achieved by penalized least squares or penalized likelihood, by Bayes or empirical Bayes estimation, or by constraining the parameters. We use penalized least squares for regularization and use constraints only to guarantee that the estimate of g is non-negative and integrates to 1.

In many applications, n is very large, often 10,000–20,000 and sometimes even larger, and for computationally efficiency it is useful to bin the p-values into, say, 2000 bins. Binning reduces computation in two ways. Obviously, it reduces the number of data points, but,

more importantly, it allows us to estimate the parameters by constrained, weighted least-squares, which turns out to be a quadratic programming problem. Quadratic programs can be solved more quickly than general constrained optimization problems such as maximum likelihood using the non-binned data. Let $N_{\rm bin}$ be the number of bins, let l_i, c_i, r_i, and $w_i = r_i - l_i$ be the left edge, center, right edge, and width of the ith bin, $i = 1, \ldots, N_{\rm bin}$, and let $M_1, \ldots, M_{N_{\rm bin}}$ be the bin counts. Then

$$y_i = \frac{M_i}{n w_i}$$

is an unbiased estimate of

$$m_i(\pi_0, \boldsymbol{\beta}) = \frac{F_p(r_i\,;\pi_0, \boldsymbol{\beta}) - F_p(l_i\,;\pi_0, \boldsymbol{\beta})}{w_i} \approx f_p(c_i\,;\pi_0). \tag{3}$$

We will estimate $(\pi_0, \boldsymbol{\beta})$ by minimizing the penalized weighted sum of squares,

$$SS(\pi_0, \boldsymbol{\beta}\,;\lambda) = \sum_{i=1}^{N_{\rm bin}} \omega_i^2 \left\{y_i - m_i(\pi_0, \boldsymbol{\beta})\right\}^2 + \lambda Q(\boldsymbol{\beta}) \tag{4}$$

where ω_i^2 is a weight, $Q(\boldsymbol{\beta})$ is a penalty to be discussed later, and $\lambda \geq 0$ is a penalty parameter. This estimator of π_0 will be called the "semiparametric" estimator since it combines a parametric model for $f_{p|\delta}(p\,;\delta)$ with an essentially nonparametric model for $g(\delta\,;\boldsymbol{\beta})$; the model for $g(\delta\,;\boldsymbol{\beta})$ will be a spline and is nonparametric in the sense that the dimension of $\boldsymbol{\beta}$ is large and can depend on the data and in an asymptotic analysis might increase with n. The weights could be $\omega_i^2 \equiv 1$ or they could be the reciprocals of the estimated variances of the y_i. In the latter case, the weighted least-squares estimator is an approximate minimum chi-squared statistic since the weighted sum-of-squares is a sum of terms of the form (observed − expected)^2/expected.

3 The Conditional CDF $F_{p|\delta}$

To use (1) to evaluate $m_i(\pi_0, \boldsymbol{\beta})$ in (3), we need to know the form of $F_{p|\delta}$. Suppose we observe iid $X_1, \ldots, X_n$ with conditional CDF $F_x(x\,;\delta)$ and the rejection regions for the ith test are of the form $X_i > \kappa$ for some κ. Then the ith p-value is $1 - F_x(X_i\,;0)$. The CDF of the p-value under δ is

$$F_{p|\delta}(p\,;\delta) = 1 - F_x\{F_x^{-1}(1-p\,;0)\,;\delta\},\ 0 < p < 1 \tag{5}$$

and the PDF is

$$f_{p|\delta}(p\,;\delta) = \frac{F_x'\{F_x^{-1}(1-p\,;0)\,;\delta\}}{F_x'\{F_x^{-1}(1-p\,;0)\,;0\}}.$$

3.1 Location Problems

As an example, in this section we consider a location parameter model where $X_1, \ldots, X_n$ are iid $\mathcal{F}(x - \delta)$, where $\mathcal{F}$ is a known CDF with PDF $\mathcal{F}'$.

3.1.1 One-Sided

The p-value for testing $H_0 : \delta_i = 0$ versus $H_1 : \delta_i > 0$ is $1 - \mathcal{F}(X_i)$, and by (5) the conditional CDF of the p-value given δ_i is

$$F_{p|\delta}(p\,;\delta_i) = P\left\{1 - \mathcal{F}(X_i) \le p \,|\delta_i\right\} = P\left\{\mathcal{F}(X_i) \ge 1 - p\right\} = 1 - \mathcal{F}\left\{\mathcal{F}^{-1}(1-p) - \delta_i\right\}.$$

Also, the conditional PDF is

$$f_{p|\delta}(p\,;\delta_i) = \frac{\mathcal{F}'\left\{\mathcal{F}^{-1}(1-p) - \delta_i\right\}}{\mathcal{F}'\left\{\mathcal{F}^{-1}(1-p)\right\}}.$$

In the normal case where $\mathcal{F} = \Phi$ and $\mathcal{F}' = \phi$, where Φ and ϕ are the standard normal CDF and PDF, respectively, we have

$$f_{p|\delta}(p\,;\delta_i) = \exp\left\{\delta_i \Phi^{-1}(1-p) - \delta_i^2/2\right\}.$$

This density is shown in Figure 1 for the cases $\delta_i = 0.5$, 1, 1.5, and 2. This is an example where $f_{p|\delta}(1\,;\delta) = 0$ but $f_{p|\delta}$ is not convex near 1.

3.1.2 Symmetric Two-Sided

Suppose that $\mathcal{F}$ is symmetric about 0 and we are testing H_0: $\delta = 0$ against H_1': $|\delta| > 0$. Taking the test statistic to be $|X_i|$ rather than X_i and using rejection regions $|X_i| > \kappa$ for some κ, the p-value is $2\{1 - \mathcal{F}(|X_i|)\}$,

$$F_{p|\delta}(p\,;\delta_i) = 1 - \mathcal{F}\left\{\mathcal{F}^{-1}(1-p/2) - \delta_i\right\} + \mathcal{F}\left\{-\mathcal{F}^{-1}(1-p/2) - \delta_i\right\}, \tag{6}$$

and

$$f_{p|\delta}(p\,;\delta_i) = \frac{1}{2}\left[\frac{\mathcal{F}'\left\{\mathcal{F}^{-1}(1-p/2) - \delta_i\right\}}{\mathcal{F}'\left\{\mathcal{F}^{-1}(1-p/2)\right\}} + \frac{\mathcal{F}'\left\{-\mathcal{F}^{-1}(1-p/2) - \delta_i\right\}}{\mathcal{F}'\left\{\mathcal{F}^{-1}(1-p/2)\right\}}\right]. \tag{7}$$

If $\mathcal{F} = \Phi$, then

$$f_{p|\delta}(p\,;\delta_i) = \frac{1}{2}\left[\exp\left\{\delta_i \Phi^{-1}(1-p/2) - \delta_i^2/2\right\} + \exp\left\{-\delta_i \Phi^{-1}(1-p/2) - \delta_i^2/2\right\}\right].$$

This is an example, where the density under alternatives is positive at $p = 1$ since

$$f_{p|\delta}(1\,;\delta_i) = \exp\left\{-\delta_i^2/2\right\}.$$

Because $\mathcal{F}$ is symmetric about 0, (6) and (7) depend on δ_i only through $|\delta_i|$. In the following, we will be interested only in one-sided and symmetric two-sided tests, so there is no loss in generality by assuming that

$$\delta \geq 0, \tag{8}$$

or, alternatively, of viewing $|\delta|$ rather than δ as the parameter. Assumption (8) is especially convenient for modeling g and will be made throughout this paper.

3.2 t-tests

Let T be a statistic whose CDF is $F_t(\cdot\,;\nu,\delta)$, the non-central-t CDF with ν degrees of freedom and non-centrality parameter δ, that is, $T = (\delta + Z)/\sqrt{Y/\nu}$ where Z is standard normal, Y is χ^2_ν, and Z and Y are independent. Many tests, e.g., about coefficients in a linear model, are based on such a statistic with $\delta = 0$ under the null hypothesis.

3.2.1 One-Sided

Suppose that the p-value for a one-sided test of $\delta = 0$ is $1 - F_t(T\,;\nu,0)$, i.e., large values of T are significant. By (5), the CDF of the p-value is

$$F_{p|\delta}(p\,;\delta_i) = 1 - F_t\left\{F_t^{-1}(1-p\,;\nu,0)\,;\nu,\delta\right\}.$$

3.2.2 Two-Sided

Let $F_{|t|}(\cdot\,;\nu,\delta) = F_t(\cdot\,;\nu,\delta) - F_t(-\cdot\,;\nu,\delta)$ be the CDF of $|T|$. The p-value for testing that $\delta = 0$ versus a two-sided alternative is $1 - F_{|t|}(|T|\,;\nu,0)$. Note that $F_{|t|}(t\,;\nu,0) = 2F_t(t\,;\nu,0) - 1$ and $F^{-1}_{|t|}(p\,;\nu,0) = F_t^{-1}\{(p+1)/2\,;\nu,0\}$. Therefore, by (5), the CDF of the p-value is

$$F_{p|\delta}(p\,;\delta_i) = 1 - \{F_t(t\,;\nu,\delta) - F_t(-t\,;\nu,\delta)\}\Big|_{t=F_t^{-1}(1-p/2\,;\nu,0)}.$$

4 The Spline Model for $g(\cdot)$

The density g will be modeled as a linear spline and estimated using the B-spline basis. We will be using assumption (8). Let δ^* be an upper bound for δ so that g is assumed to have support contained in $[0,\delta^*]$. The spline will have K knots, $0 = \kappa_1,\ldots,\kappa_K = \delta^*$, equally spaced between 0 and δ^*, so that the distances between adjacent knots are all equal to $d = \delta^*/(K-1)$. The choice of K is not critical as long as it is large enough. Because the spline is penalized, the "effective" number of parameters is controlled by the penalty parameter and K only provides an upper bound. We have experimented with

$K = 8$ and 16 and found that in many situations both choices work well, because data-driven methods for choosing the effective number of parameters choose a value less than the upper bound of 8 when $K = 8$. For example, in an experiment with 5000 p-values, the approximate generalized cross validation method we introduce in Section 6 chose between 4 and 5 effective parameters when using either $K = 8$ or $K = 16$. In our numerical examples of Sections 8 and 12, we use $K = 12$.

Another issue is the choice of δ^*, the upper bound for δ. We have used $\delta^* = 6$ in our empirical studies and this choice proved satisfactory. The explanation for this is that the tests we studied were z-tests or t-tests with δ the non-centrality parameter. Thus, δ is the deviation of a parameter from its null value expressed in standard deviation units, so that 6 is a reasonable upper bound for δ. If we bin the p-values into 2000 bins, say, then there is virtually no information about the exact value of δ once it exceeds 6, for any δ above 6 is almost certain to produce a p-value in the [0, 1/2000] bin.

The B-splines are plotted in Figure 2 for the case $\delta^* = 6$ and $K = 7$. The first B-spline, B_1, decreases linearly from $2/d$ to 0 on the interval $[0, \kappa_2] = [\kappa_1, \kappa_2]$ and is zero elsewhere. The remaining B-splines $B_2, \ldots, B_{K-1}$ are such that B_k increases linearly from 0 to $1/d$ on $[\kappa_{k-1}, \kappa_k]$ and then decreases linearly from $1/d$ to 0 on $[\kappa_k, \kappa_{k+1}]$ and is 0 elsewhere. The B-splines span the space of linear splines with knots $\kappa_1, \ldots, \kappa_K$ and constrained to be zero at the last knot. This constraint forces the splines to be continuous on $[0, \infty)$, which seems reasonable. The constraint could be removed by adding an additional B-spline that increases linearly from κ_{K-1} to κ_K and is zero elsewhere. This B-spline is shown as a dashed line in Figure 2.

Each B-spline has been normalized so that it is a density, and therefore any convex combination of the B-splines is also a density. Thus, our model for g will be

$$g(\delta, \boldsymbol{\beta}) = \sum_{k=1}^{K-1} \beta_k B_k(\delta), \tag{9}$$

where $\beta_k \geq 0$ for all k and $\sum_{k=1}^{K-1} \beta_k = 1$.

5 The Penalized Least-Squares Estimator

To find an more explicit expression for the right hand side of (3), we now write $F_p(\cdot\,; \pi_0, \boldsymbol{\beta})$ in terms of the B-splines. It is convenient to reparameterize to a parameter vector $\boldsymbol{\theta}$ as follows. Define $\theta_1 = \pi_0$ and $\theta_{k+1} = (1 - \pi_0)\beta_k$ for $k = 1, \ldots, K - 1$, and define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\mathsf{T}$. Let $Z_1(p) = p$ be the (uniform) CDF of the p-values under H_0, and for $k = 1, \ldots, K - 1$ let

$$Z_{k+1}(p) = \int F_{p|\delta}(p; \delta) B_k(\delta) d\delta \tag{10}$$

be the marginal CDF of a p-value if the density of δ is B_k. Then the marginal CDF of a p-value is modeled as

$$F_p(p\,;\boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_k Z_k(p), \tag{11}$$

where

$$\theta_k \geq 0,\ \forall\ k,\ \text{and}\ \sum_{k=1}^{K} \theta_k = 1. \tag{12}$$

The roughness penalty we will use penalizes deviations of $\widehat{g}$ from a linear function using a finite difference approximation to the second derivative of g. It is convenient if the roughness penalty is expressed in terms of $\boldsymbol{\theta}$. The value of g at the knots is $g(\kappa_1) = g(0) = 2\beta_1/d = 2(1-\pi_0)\theta_2/d$, $g(\kappa_k) = \beta_k/d = (1-\pi_0)\theta_{k+1}/d$ for $k = 2, \ldots, K-1$, and $g(\kappa_K) = g(\delta^*) = 0$. The roughness penalty is

$$\begin{aligned} Q(\boldsymbol{\theta}) &= (2\theta_2 - 2\theta_3 + \theta_4)^2 + \sum_{k=3}^{K-2} (\theta_k - 2\theta_{k+1} - \theta_{k+2})^2 \\ &= \{d(1-\pi_0)\}^2 \sum_{k=1}^{K-3} \{g(\kappa_k) - 2g(\kappa_{k+1}) - g(\kappa_{k+2}))\}^2. \end{aligned} \tag{13}$$

Now define $\boldsymbol{y} = (y_1, \ldots, y_{N_{\text{bin}}})^\mathsf{T}$ and let $\boldsymbol{Z}$ be the $N_{\text{bin}} \times K$ matrix whose i,jth element is

$$Z_{i,j} = \{Z_j(r_i) - Z_j(l_i)\}/w_i. \tag{14}$$

Then, by (3), (4), (11), and (13) the sum of squares is

$$\begin{aligned} SS(\boldsymbol{\theta};\lambda) &= \sum_{i=1}^{N_{\text{bin}}} \omega_i^2 \left\{ y_i - \sum_{k=1}^{K} \theta_k Z_{i,k} \right\}^2 \\ &+ \lambda \left\{ (2\theta_2 - 2\theta_3 + \theta_4)^2 + \sum_{k=3}^{K-2} (\theta_k - 2\theta_{k+1} - \theta_{k+2})^2 \right\} \\ &= \|\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta})\|^2 + \lambda \boldsymbol{\theta}^\mathsf{T} \left\{ (\boldsymbol{D}\boldsymbol{A})^\mathsf{T} \boldsymbol{D}\boldsymbol{A} \right\} \boldsymbol{\theta}, \end{aligned} \tag{15}$$

where $\boldsymbol{W} = \text{diag}(\omega_1, \ldots, \omega_n)$, $\boldsymbol{A} = \text{diag}(0, 2, 1, \ldots, 1)$ and $\boldsymbol{D}$ is a $(K-3) \times K$ "differencing matrix" whose ith row has $+1$ in the columns $i+1$ and $i+3$, -2 in column $i+2$ and zeros elsewhere. Minimizing (15) is equivalent to minimizing

$$\boldsymbol{f}^\mathsf{T}\boldsymbol{\theta} + 0.5\, \boldsymbol{\theta}^\mathsf{T} \boldsymbol{H} \boldsymbol{\theta} \tag{16}$$

where $\boldsymbol{f}^\mathsf{T} = -\boldsymbol{y}^\mathsf{T}\boldsymbol{W}\boldsymbol{Z}$ and $\boldsymbol{H} = \boldsymbol{Z}^\mathsf{T}\boldsymbol{W}\boldsymbol{Z} + \lambda \boldsymbol{A}^\mathsf{T}\boldsymbol{D}^\mathsf{T}\boldsymbol{D}\boldsymbol{A}$, and in either case the constraints are

$$\boldsymbol{\theta} \geq 0 \text{ and } \mathbf{1}^\mathsf{T}\boldsymbol{\theta} = 1, \tag{17}$$

where $\mathbf{1}$ is a K-dimensional vector of ones. Objective function (16) and the constraints (17) are in the form used by the quadratic programming algorithm `quadprog` of MATLAB, which we used to find the constrained penalized least-squares estimator. The penalty is somewhat complicated because the maximum height of B_1 is twice the maximum height of the other B-splines. We considered redefining B_1 so that all the B-splines would have the same maximum height, but then B_1, having mass 1/2, not 1, would not be a density. This would make the form the constraints slightly more complex and would not give an overall simpler form to the optimization problem.

If λ is chosen by cross-validation, then the quadratic program must be solved for each value of λ on some grid. However, much of the computational effort is devoted to computing $\boldsymbol{f}$ and $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{Z}$ since $\boldsymbol{y}$ is $N_{\text{bin}} \times 1$ and $\boldsymbol{Z}$ is $N_{\text{bin}} \times K$ and N_{bin} is reasonably large; we used $N_{\text{bin}} = 2000$. However, these matrices can be computed once and then used repeatedly to minimize (16) for a grid of λ values. The dimensions of the matrices in (16)–(17) are $K \times 1$ or $K \times K$ and K, which is the number of basis functions in our model, will be reasonably small, e.g., 12.

Computation of the $Z_{i,j}$'s defined by (14) requires that we compute $Z_j(p)$ given by (10) with p equal to each of the bin edges. We computed the integral in (10) numerically using 500 values of δ. Doing this required that $F_{p|\delta}(p\,;\delta)$ be valued at $N_{\text{bin}} \times 500$ combinations of p and δ. For the t-tests this takes several minutes since computation of the non-central-t CDF is somewhat computationally expensive. To speed up computations, we computed these values of $F_{p|\delta}(p\,;\delta)$ and saved them in files, one file for each value of DF that we used, and then loaded the appropriate file into memory when needed.

The fitted value

$$\widehat{f}_p(c_i) = \widehat{y}_i = \sum_{k=1}^{K} \widehat{\theta}_k Z_{i,k} \tag{18}$$

estimates $m_i(\pi_0, \boldsymbol{\beta})$ given by (3), which is an approximation to $f_p(c_i)$, the marginal density of the p-values at the center of the ith bin. Also, $F_p(p)$ can be estimated by

$$\widehat{F}_p(r_i) = \sum_{i'=1}^{i} w_{i'} \left\{ \sum_{k=1}^{K} \widehat{\theta}_k Z_{i',k} \right\} \tag{19}$$

when p is some right bin edge r_i and then interpolated to other values of p. The estimator of π_0 is

$$\text{“Semi, } \theta_1\text{”} = \widehat{\theta}_1. \tag{20}$$

The notation “Semi, θ_1” is intended to remind the reader that this is a semiparametric estimator based only on $\widehat{\theta}_1$. Two other semiparametric estimators based on $\widehat{\boldsymbol{\theta}}$ will be introduced in Section 7.

Also, let $\widehat{g}(\delta) = \sum_{k=1}^{K-1} \widehat{\beta}_k B_k(\delta)$ and $\widehat{G}(\delta) = \int_0^\delta \widehat{g}(u)du$.

6 Approximate Cross-Validation

An obvious method for choosing λ is cross-validation (CV). However, exact cross-validation would be slow to compute, so instead we used an approximation to the generalized cross-validation (GCV) statistic. The GCV statistic itself is not defined for our estimator because the constraints make the estimator nonlinear in $\boldsymbol{y}$. Thus, there is no hat matrix and the usual method of defined the degrees of freedom of the fit (DF) does not apply—see Chapter 3 and Section 5.3 of Ruppert, Wand, and Carroll (2003) for an introduction to GCV, linear estimators, the hat matrix, GCV, and DF for penalized least-squares estimators. Therefore, we use the DF parameter from estimating $\boldsymbol{\theta}$ by minimizing (15) without constraint—this is a poor estimator of $\boldsymbol{\theta}$ but gives a DF value that worked well in our simulations when put into the GCV formula.

The unconstrained minimizer of (15) is $\left\{\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{Z} + \lambda(\boldsymbol{D}\boldsymbol{A})^{\mathsf{T}}(\boldsymbol{D}\boldsymbol{A})\right\}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{y}$, and has hat matrix $\boldsymbol{H}(\lambda) = \boldsymbol{Z}\left\{\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{Z} + \lambda(\boldsymbol{D}\boldsymbol{A})^{\mathsf{T}}(\boldsymbol{D}\boldsymbol{A})\right\}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}$. Then $\mathrm{DF}(\lambda) = \mathrm{trace}\{\boldsymbol{H}(\lambda)\}$ $= \mathrm{trace}\Big[\left\{\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{Z} + \lambda(\boldsymbol{D}\boldsymbol{A})^{\mathsf{T}}(\boldsymbol{D}\boldsymbol{A})\right\}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{Z}\Big]$, and the approximate GCV statistic we use is

$$\mathrm{GCV}(\lambda) = \frac{\|\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\theta}}(\lambda)\|^2}{\{N_{\mathrm{bin}} - \mathrm{DF}(\lambda)\}^2},$$

where $\boldsymbol{\theta}(\lambda)$ is the estimator of Section 5 that minimizes (15) *with* constraints (17). The smoothing parameter λ is chosen by computing $\mathrm{GCV}(\lambda)$ on a grid on λ values and choosing the value that minimizes $\mathrm{GCV}(\lambda)$.

7 Alternative Estimators of π_0

Alternative estimators of π_0 can be obtained by minimizing estimators of f_p, the marginal density of the p-values. In this section, we describe two additional estimators based on penalized least-squares as well as the convex, decreasing estimator proposed by Langaas et al. (2005).

7.1 Estimator Based On The Penalized Least-Squared Fit

7.1.1 "Semi, min$\{\widehat{f}\}$"

An alternative semiparametric estimator, denoted by "Semi, min$\{\widehat{f}\}$" is the minimum over i of (18), i.e.,

$$\text{"Semi, min}\{\widehat{f}\}\text{"} = \min_i \widehat{f}_p(c_i) = \widehat{f}_p(c_{N_{\mathrm{bin}}}) = \sum_{k=1}^{K}\widehat{\theta}_k Z_{N_{\mathrm{bin}},k}. \tag{21}$$

The minimum occurs at $i = N_{\mathrm{bin}}$ because the estimated density is decreasing.

7.1.2 "Semi, compromise"

We found that "Semi, θ_1" can biased downward and is somewhat more variable than "Semi, $\min\{\widehat{f}\}$". However, when a substantial proportion of the p-values near 1 come from the alternative hypothesis, then "Semi, $\min\{\widehat{f}\}$" can be biased upwards to such an extend that nearly 100% of the MSE (mean squared error) is attributable to squared bias; see Section 8. These results motivated us to find an estimator that is a compromise between "Semi, θ_1" and "Semi, $\min\{\widehat{f}\}$". The former attempts to separate $f_p(1)$ into a component from the null and another component from the alternative and uses only the component from the null to estimate π_0. The latter uses both components. The problem with "Semi, θ_1" is that it is difficult to separate p-values coming from the null from those coming from alternative values of δ near the null. To circumvent this problem, we defined a new estimator, "Semi, compromise", which decomposes $f_p(1)$ in three components, one from the null, one from the alternative near the null, and the third from the alternative away from the null. Then "Semi, compromise" uses the first two components. This induces a slight upward bias which provides a margin of safety. It also decreases variability. More precisely, we define

$$\text{"Semi, compromise"} = \sum_{k=1}^{2} \widehat{\theta}_k Z_{N_{\text{bin}},k} \approx \text{"near null part" of } \widehat{f}_p(1) \tag{22}$$

One can see from (20), (21), and (22) and the fact that $Z_{i,k} \geq 0$ for all i, k, that "Semi, θ_1" $\leq$ "Semi, compromise"$\leq$ "Semi, $\min\{\widehat{f}\}$".

7.1.3 Weighting

We studied two versions each of "Semi, θ_1", "Semi, $\min\{\widehat{f}\}$", and "Semi, compromise", an unweighted version where $\omega_i \equiv 1$ and a weighted version where $\omega_i = 1/\sqrt{\widehat{f}_p(c_i)}$, where $\widehat{f}_p$ is some preliminary estimator, e.g., the unweighted version of "Semi, θ_1" or "Semi, $\min\{\widehat{f}\}$". The latter weights are based on the facts that the bin counts are approximately Poisson distributed and, by (3), $f_p(c_i)$ is approximately proportional to the expected count for the ith bin. We found that weighting did not have a consistent effect on "Semi, θ_1", "Semi, $\min\{\widehat{f}\}$", and "Semi, compromise", but that the weighted versions of these estimators often had a somewhat smaller mean squared error compared to the unweighted versions.

7.2 The Convex, Decreasing Estimator

Langaas, Ferkingstad, and Lindqvist (2005) considered a number of different estimators of π_0. The best performing of these is convex, decreasing density estimator applied to the p-values and evaluated at 1. These authors show that any twice differentiable, convex, and

decreasing density f on $[0,1]$ has a representation as

$$f(x) = \int_0^1 f_\theta(x)\gamma(\theta)\,d\theta + f_0(x)a_0 + f_1(x)a_1, \tag{23}$$

where f_0 is the uniform$(0,1)$ density,

$$f_\theta(x) = \frac{2(\theta - x)_+}{\theta^2},\ 0 \le x \le 1,\ -0 < \theta \le 1, \tag{24}$$

$a_0 = f(1)$, $a_1 = -1/2f'(1)$, and $\gamma = (1/2)\,\theta^2/f''(\theta)$. The nonparametric MLE (NPMLE) of a convex, decreasing density maximizes the likelihood over this class of densities. Langaas et al. suggest an iterative algorithm for approximating the nonparametric MLE by a discrete mixture, using only values of θ contained in some fine grid, e.g., $\{0, 0.01, 0.02, \ldots, 1\}$.

We developed a algorithm for approximating the NPMLE, that differed in a few ways from the Langaas et al. algorithm. First, we minimized a chi-squared statistic rather than maximizing the likelihood. Second, we used all θ on the grid $\{0, 0.01, 0.02, \ldots, 1\}$. Finally, we used quadratic programming to optimize. Our estimators were of the form

$$\sum_{i=0}^{100} b_i f_{i/100}(x), b_i \ge 0\ \forall\ i,\ \sum_{i=0}^{100} b_i = 1. \tag{25}$$

The minimum chi-squared statistics is a quadratic function of $(b_0, \ldots, b_{100})$ and the constraints in (25) are linear. Thus, our estimator can be calculated by quadratic programming in the same way that (15) was minimized. Computation is very fast using this algorithm. Following Langaas et al., we denote this estimator evaluated at 1 by "Convex". It is well known that minimum chi-square estimators are asymptotically equivalent to MLE based on grouped data, e.g., see Holland (1967) or Rao (1973), and there should be little loss of information in grouping data into a large number, e.g., 2000, bins, so the estimate computed by our algorithm is expected to be nearly equal to the MLE.

8 Simulation Studies

We performed simulations to compare "Convex", "Semi, θ_1", "Semi, $\min\{\widehat{f}\}$", and "Semi, compromise".

8.1 z-tests

We started our simulation study with z-tests, that is, tests where the test statistic has either an exact or an approximate normal distribution with a known variance. The one- and two-sample tests about the means of normal populations with known variances are, of course, exact z-tests. Approximate rather than exact z-tests are most common, since a

test statistic often can be transformed so as to have an asymptotically normal distribution under the null hypothesis as well as under local alternatives.

We simulated exact z-tests where, conditional on $\delta_1, \ldots, \delta_n$, $X_1, \ldots, X_n$ are iid $N(\delta_i, 1)$, the null hypotheses is $H_0 : \delta = 0$, and under the alternative the δ_i were generated from a Beta(b_1, b_2) density on $[\delta_{\min}, \delta_{\max}]$ where $(\delta_{\min}, \delta_{\max}, b_1, b_2, \pi_0)$ was varied during the study. In each simulation, we generated 10,000 p-values and exactly $10{,}000\pi_0$ came from the null, that is, π_0 is the actual sample proportion of true nulls. Here X_i represents a normalized sufficient statistic for the ith test, e.g., a sample mean or the difference between two sample means divided by its standard deviation.

The simulations used six cases of (π_0, g). In Cases 1–3 $\pi_0 = 0.95$ and in Cases 4–6 $\pi_0 = 0.7$. Three densities were used for g and their parameters were, respectively, $(\delta_{\min}, \delta_{\max}, b_1, b_2) = (0, 4, 3, 2)$, $(0, 4, 2, 2)$, and $(0.5, 4.5, 1, 2)$. The first density, which is used in Cases 1 and 4, has support [0, 4] and is concentrated around 0, its mode, making it difficult to distinguish p-values that come from the null from those coming from the alternative. This density is similar to the estimates of g in the gene expression study discussed in Section 12, which suggests that difficulty distinguishing p-values from the null from those from the alternative might be common in practice, at least in gene expression studies. The second density, used in Cases 2 and 5, also has support [0, 4], but has a mode away from 0. The third density, used in Cases 3 and 6, has support [0.5, 4.5] which is separated from the null hypothesis. The three densities are labelled, respectively, "near," "moderately near," and "far" from the null in the Tables. Two of these three densities are shown as dotted-and-dashed curves on the tops of Figures 3 and 4.

We studied both one- and two-sided tests. For a two-sided z-test, the distribution of the p-value depends only on $|\delta|$, so there is no loss in generality in having the alternative density of δ supported on a positive interval. For the semiparametric estimator, N_{bin} was fixed at 2000 and K was 12. We experimented with $K = 8$ and 16 and found that the value of K had little effect. This was expected since λ controls the amount of smoothing and K is only an upper bound on the effective degrees of freedom.

We also found that the weighted versions of "Semi, θ_1" and "Semi, $\min\{\widehat{f}\}$" did not dominate unweighted versions, but generally the weighted estimators were somewhat better. To save space, only results for the weighted estimators will be presented. The estimated root mean squared errors (RMSE) and biases are in Table 1 for one-sided z-tests and Table 2 for two-sided z-tests. Our main conclusions from these results are that:

1. In Cases 1, 2, 3, and 6, "Semi, $\min\{\widehat{f}\}$", "Semi, compromise", and "Convex" all perform well, especially for one-sided z-tests. The RMSE of "Semi, compromise" is less than 0.02 for all four of these cases for both one- and two-sided z-tests.

2. For two-sided z-tests, "Semi, compromise" has a smaller RMSE than "Convex" in five of the six cases. The exception is Case 3 where π_0 is large and g is far from the null, a situation where the assumption that all p-values near 1 are from the null is nearly true. Therefore, estimators based on this assumption, e.g., "Semi, min$\{\widehat{f}\}$" and "Convex", do very well. Interestingly, "Semi, min$\{\widehat{f}\}$" has approximately half the RMSE of "Convex" in the case.

3. The "Semi, min$\{\widehat{f}\}$" and "Convex" estimators have a substantial positive bias in Cases 4 and 5, especially in Case 4 and especially for two-sided z-tests.

4. "Semi, θ_1" tends to have a negative bias and is more variable than the other estimators. This was the motivation for introducing "Semi, compromise".

8.2 t-tests

We performed simulations of the t-test where $t_i = X_i/\sqrt{\chi_i^2(DF)/DF}$, $X_i, \ldots, X_n$ had the same distribution as in the z-test simulations, and $\chi_i^2(DF)$ was an independent chi-squared random variate with DF degrees of freedom. We only considered two-sided tests with DF = 4. If DF is large, then the results for t-tests were expected to be similar to those for z-tests, so we intentionally used a small value of DF. The results are in Table 3. From these results, we conclude that:

- In Cases 1, 2, 4, and 5 where g is near or moderately near the null, "Semi, compromise" has the smallest RMSE of the four estimators.

- In Case 4, "Semi, min$\{\widehat{f}\}$" and "Convex" have severe positive bias because many of the p-values near 1 are from the alternative. In this case, "Semi, compromise" is far superior to "Semi, min$\{\widehat{f}\}$" and "Convex".

- In Cases 3 and 6 where g is far from the null, "Semi, min$\{\widehat{f}\}$" has the smallest RMSE of the four estimators.

- "Semi, θ_1" has large RMSE values, as seen also for z-tests.

8.3 Autocorrelated Tests

We also simulated autoregressive two-sided t-tests with DF=4 and n = 10,000. The ith p-value was based on $t_i = (\delta_i + e_i)/\sqrt{s_i^2/DF}$ where e_i is an AR(1) process and s_i^2 is independent of e_i and χ_{DF}^2 distributed with $DF = 4$. Specifically, $e_i = \rho e_{i-1} + u_i$ where the u_i are independent $N(0, 1-\rho^2)$, so that the e_i are $N(0,1)$ and e_i and e_j have correlation $\rho^{|i-j|}$. The s_i^2 were mutually independent.

When $\rho \neq 0$, the joint distribution of the p-values will depend on how the δ_i are ordered. We considered two orderings of the δ_i, "permute" where the δ_i were randomly permuted and "sort" where the δ_i were sorted from smallest to largest, so, in particular, all the p-values from true nulls came first. Under "sort" p-values with similar values of δ will be more highly correlated.

The proportion of true nulls, π_0, was fixed at 0.9 and $(\delta_{\min}, \delta_{\max}, b_1, b_2)$ was fixed at (0, 4, 2, 2). There were 600 Monte Carlo simulated data sets per case.

The results are in Table 4. We were at first surprised to see that the RMSE's of "Semi, $\min\{\widehat{f}\}$", "Semi, compromise", and "Convex" were nearly independent of ρ and also of whether the δ_i were permuted or sorted. However, there is a simple explanation. For these estimators, the largest component of RMSE is squared bias, not variance, and bias should depend little, if at all, on the amount of autocorrelation. In contrast, "Semi, θ_1" has a larger component due to variance and its RMSE is larger when ρ is larger. However, the RMSE of "Semi, θ_1" also depends very little upon whether the δ_i were permuted or sorted.

Note that "Semi, compromise" had a smaller RMSE than "Convex" and "Semi, $\min\{\widehat{f}\}$" in all five cases in Table 4.

8.4 Discussion of the Simulation Results

"Convex", by definition, uses a convex density estimator while "Semi, $\min\{\widehat{f}\}$" uses a density estimator whose shape comes from the shape of $f_{p|\delta}$ and, in particular, will be convex for two-sided tests but will be concave near 1 for one-sided tests. Thus, it is not too surprising that "Convex" and "Semi, $\min\{\widehat{f}\}$" have similar performances, but that "Semi, $\min\{\widehat{f}\}$" is less biased in some situations such as Case 4 for one-sided z-tests.

How much one is willing to tolerate bias will influence the choice of estimator. When using an estimate of π_0 for determining the false discovery rate, a positive bias is often considered to be less serious than a negative bias, because it leads to conservative false discovery rates. However, a bias of 0.15 seen in "Semi, $\min\{\widehat{f}\}$" and "Convex" may be too conservative, and it is unnecessary now that "Semi, compromise" is available. In other applications, such as determining the number of Kuiper Belt objects (Meinshausen and Rice, 2005), bias in either direction is undesirable.

In general, bias is a major component of the RMSE of the estimators. The amount of bias depends on both π_0 and g. Obviously, there can be little positive bias if π_0 is close to 1, but if π_0 is 0.7 then bias can be severe. Fortunately, our semiparametric methodology provides estimates of both of π_0 and g, so one can be alerted to situations where bias may be severe. Figures 3 and 4 show estimates of f_p and g from 10 independent simulated data sets for Cases 4 and 6, respectively. The estimates of g are rather close to g itself, showing

that it is possible to determine whether values of δ under the alternative are mostly close to or far from the null; it is only when they are mostly close to the null that severe bias should be expected. In Cases 4–6, $\pi_0 = 0.7$ so there is more information about g than in cases 1–3, where g is not estimated quite so well. However, when π_0 is large as in Cases 3–6, the estimate of π_0 will indicate both that g may not be estimated accurately and that, fortunately, positive bias will not be severe. Of course, the accuracy of $\widehat{g}$ depends also on n. In Figures 3 and 4, $n = 10,000$ which is not an extreme case and is less than half the value of n in the application in Section 12.

9 Estimating the False Discovery Rate

Suppose that for some fixed α, all null hypotheses with a p-value below α will be rejected. We wish to estimate $P(H_0 \text{ is true}|p\text{-value} < \alpha)$. Morton (1955) called this probability the *posterior error rate*. Storey (2002, 2003) provided conditions under which the positive False Discovery Rate (pFDR) and the posterior error rate are identical. With some slight abuse of terminology, we will refer to $P(H_0 \text{ is true}|p\text{-value} < \alpha)$ as the False Discovery Rate (FDR) and estimate it by

$$\widehat{\text{FDR}} = \frac{\alpha\,\widehat{\pi}_0}{\widehat{F}_p(\alpha)}, \tag{26}$$

where $\widehat{F}_p(\alpha)$ is estimated by (19) and $\widehat{\pi}_0$ is some estimator of π_0, e.g., "Semi, compromise" or "Convex". Note that (26) is identical to the expression denoted $\widehat{\text{FDR}}$ by Storey and Tibshirani (2003) except that we use a different estimator of π_0 and our denominator is not simply the proportion of observed p-values that fall below α. From (26) we see that upwards bias in $\widehat{\pi}_0$ will cause upward bias in $\widehat{\text{FDR}}$.

10 When is a p-value Interesting?

Efron (2004) discusses a problem that may occur when one has a large number of tests: the number of false nulls is often very large and we do not necessarily want to "discover" every one of them. This problem does not always occur. In the astronomy example of Meinshausen and Rice (2005) mentioned in Section 1, we are not really interested in which nulls are false, only in how many nulls are false, so there is no danger of discovering too many false nulls. However, in gene expression studies one is primarily interested in finding nulls that are both false and "important" or "interesting" biologically. For example, in the microarray experiment described in Section 12, the biologists were looking for barley genes that are involved in resistance to a fungal pathogen. Many genes are likely to change expression during attack by a pathogen, but some changes may be quite small and play

only a minor role in a plants defense response. While all changes, regardless of size, are potentially of interest, researchers may wish to focus attention initially on the genes that exhibit the largest and most consistent changes in expression. In some cases this may provide a clearer picture of the biology than attempting to simultaneously interpret the meaning of small changes in thousands of genes.

10.1 The Falsely Interesting Discovery Rate

If "interesting" is interpreted as an especially unusual p-value as in Efron (2004), then we are assuming that the δ values farthest from the null are of greatest interest. This assumption is debatable, of course, and we do not think that making this assumption is always a good idea. However, if we are willing to make it, then our estimates of π_0 and g can be useful for determining the number of interesting δ values. Suppose we want to know what proportion of the p-values come either from a true null or from a null that is false but with a δ value that is "uninteresting," where "uninteresting" is defined by subject-matter considerations to mean that $\delta < \delta'$ for some fixed $\delta' > 0$. Then this is proportion can be estimated by

$$\widehat{\pi}_0 + (1 - \widehat{\pi}_0) \int_0^{\delta'} \widehat{g}(\delta)\, d\delta. \tag{27}$$

We define the "falsely interesting discovery rate" (FIDR) as the conditional probability that a null hypothesis is either true or false but with an uninteresting value of δ, given that it has been rejected, i.e., again assuming that a null hypothesis is rejected if the p-value is its less than α,

$$\text{FIDR}(\alpha, \delta') = \frac{P(\delta < \delta' \text{ and } p\text{-value} < \alpha)}{P(p\text{-value} < \alpha)}. \tag{28}$$

The denominator of (28) can be estimated by $\widehat{F}_p(\alpha)$. If δ' is one of the knots, say the k'th, then the numerator of (28) can be estimated when α is a right bin edge, say r_i, by

$$\sum_{i'=1}^{i} w_{i'} \left\{ \sum_{k=1}^{k'} \widehat{\theta}_k Z_{i',k} + (1/2)\widehat{\theta}_{k'+1} Z_{i',k'+1} \right\} \tag{29}$$

and then interpolated for other values of α. Here we use the facts that θ_1 is the probability that the null is true, that θ_{k+1} is the coefficient of the kth B-spline, and that the kth B-spline peaks at the kth knot and has half of its probability to the left of that knot.

Efron (2004) has a rather different approach to the problem of rejecting too many nulls. He replaces the "theoretical null hypothesis" by an "empirical null hypothesis." Efron applies the inverse probit transformation to the p-values so that those "z-values" coming from the null will have an exact $N(0,1)$ distribution. He then finds an estimate, $\widehat{f}_z$, of the density of the z-values. The "empirical null" is that the z-value is $N(\delta_z, \sigma_z^2)$ where δ_z is the mode of the estimated density and σ_z^2 is the $-1/\{\log(\widehat{f}_z)\}''(\delta_z)$.

Often the theoretical null hypothesis has a clear scientific meaning, e.g., that the expression level of a gene is the same in two populations, but the empirical null does not have a similar interpretation. Subject-matter specialists might find it confusing to use to a null hypothesis that is estimated from the data, that can vary across tests within a study, and that does not have as clear an interpretation as the theoretical null. In contrast, the idea that "the null is false but not by much" seems natural.

11 Power and Sample Sizes

Gadbury et al. (2004) define the Expected Discovery Rate (EDR) to be the probability of a "discovery," given that the effect is real, i.e., the probability that an effect is declared significant, given that the null hypothesis is false. The EDR in our notation was given by (2). We assume that g has been estimated and we are now contemplating a repetition of the same experiment, or perhaps a similar experiment, with new sample sizes that differ from the old by a factor η. We assume that δ represents the non-centrality parameter of a test that changes from δ to $\delta^* = \sqrt{\eta}\delta$ with the new sample size. This would be the case, for example, if we were considering two-sample t-tests with n observations per sample.

It is of interest to see how EDR changes with η. Since G is the conditional distribution of δ given that the null is false, the EDR for any η is defined by

$$\mathrm{EDR}(\alpha, \eta) = \int_0^\infty F_{p|\delta}(\alpha\,;\sqrt{\eta}\delta)\,dG(\delta). \tag{30}$$

Let $\widehat{\mathrm{EDR}}(\alpha, \eta)$ be (30) with G replaced by $\widehat{G}$. Gadbury et al. also define TN (True Negative) as the probability an effect is not real given that it is declared not significant and TP (True Positive) as the probability an effect is real given that it is declared significant. In our notation

$$\mathrm{TN}(\alpha, \eta) = \frac{(1-\alpha)\pi_0}{(1-\alpha)\pi_0 + (1-\pi_0)\{1 - \mathrm{EDR}(\alpha, \eta)\}} \tag{31}$$

and

$$\mathrm{TP}(\alpha, \eta) = \frac{(1-\pi_0)\mathrm{EDR}(\alpha, \eta)}{\alpha\pi_0 + (1-\pi_0)\mathrm{EDR}(\alpha, \eta)}. \tag{32}$$

$\mathrm{TN}(\alpha, \eta)$ and $\mathrm{TP}(\alpha, \eta)$ can be estimated by plugging $\widehat{\mathrm{EDR}}(\alpha, \eta)$ and "Semi, $\min\{\widehat{f}\}$" or "Semi, compromise" into (31) and (32).

We also define an Expected Interesting Discovery Rate (EIDR) as the probability of a "discovery" given that the null is false and interesting meaning that $\delta > \delta'$. Thus,

$$\mathrm{EIDR}(\alpha, \eta, \delta') = \frac{\int_{\delta'}^\infty F_{p|\delta}(\alpha\,;\sqrt{\eta}\delta) g(\delta) d\delta}{\int_{\delta'}^\infty g(\delta) d\delta}. \tag{33}$$

Examining estimates of EDR, EIDR, TP, and TN as a function of α for varying choices or η will help researchers determine appropriate sample sizes for future microarray experiments. For example, a researcher may have the goal of identifying 90% of all "interesting" gene expression differences, where "interesting" is defined by specifying a value for δ'. Furthermore, suppose this level of discovery is to be achieved while maintaining a true positive rate (TP) in excess of 0.95. By estimating EIDR and TP from pilot data, we can estimate the sample size relative to that in the pilot experiment (η) that will be required to meet the desired performance criteria. Such information will prevent researchers from wasting effort and resources on experiments that are likely to fall far short of their performance goals, or from using more resources than necessary to achieve their performance goals. These calculations are particularly valuable for microarray experiments where labor and supply costs are quite high.

12 Example: Gene Expression in Barley

Caldo, Nettleton, and Wise (2004) conducted a microarray experiment to identify barley genes that play a role in resistance to a fungal pathogen. To illustrate our methods, in this section we describe the analysis of a subset of the data they considered.

Two genotypes of barley seedlings, one resistant and one susceptible to a fungal pathogen, were grown in separate trays randomly positioned in a growth chamber. Each tray contained six rows of 15 seedlings each. The six rows in each tray were randomly assigned to six tissue collection times: 0, 8, 16, 20, 24, and 32 hours after fungal inoculation. After simultaneously inoculating plants with the pathogen, each row of plants was harvested at its randomly assigned time. One Affymetrix GeneChip was used to measure gene expression in the plant material from each row of seedlings. The entire process was independently repeated a total of three times, yielding data on 22,840 probe sets (corresponding to barley genes) for each of 36 GeneChips (2 genotypes $\times$ 6 time points $\times$ 3 replications). This can be viewed as a split-plot experimental design with replications as blocks, trays as whole plots, and rows of seedlings as split plots.

A mixed linear model corresponding to the split-plot design was separately fit to the 36 log-scale measures of expression for each gene. Specifically, each mixed linear model included fixed effects for genotypes, times, and genotype-by-time interaction along with random effects for replications, replication-by-genotype terms (i.e., trays), and residuals corresponding to rows of seedlings. The usual assumptions regarding independence, normality, and constant variance were assumed for the random effects within a gene.

Genes that exhibit different patterns of expression over the time course following inoculation are of primary interest because this type of differential gene activity may help to

explain why the one genotype is resistant to the fungus while the other is susceptible. Thus interaction between genotype and time is of primary interest in this experiment. We focus here on t-tests intended to detect specific sub-interactions within the overall genotype-by-time interaction. In particular, for each gene indexed by i and for each time $t = 8$, 16, 20, 24, and 32 hours after inoculation, we test

$$H_{0i}^{(t)} : \mu_{irt} - \mu_{ist} = \mu_{ir0} - \mu_{is0}$$

where μ_{irt} and μ_{ist} denote the mean expression of gene i in resistant and susceptible barley genotypes, respectively, at t hours after inoculation. Note that rejection of $H_{0i}^{(t)}$ suggests that the expression difference between genotypes at time t during fungal attack has changed from the baseline difference between the genotypes at the initial time point.

According to our mixed-linear model, the test statistic for $H_{0i}^{(t)}$ will have a non-central t distribution with 20 degrees of freedom and non-centrality parameter

$$\delta_i^{(t)} = \frac{\sqrt{n}(\mu_{irt} - \mu_{ist} - \mu_{ir0} + \mu_{is0})}{\sqrt{4\sigma_e^2}},$$

where n denotes the number of replications ($n = 3$ in this case) and σ_e^2 denotes the residual variance component. Clearly $H_{0i}^{(t)}$ is equivalent to $\delta_i^{(t)} = 0$. We now present results for the five sets of p-values obtained by testing $H_{0i}^{(t)} : \delta_i^{(t)} = 0$ for all $i = 1, \ldots, 22,840$ at each time $t = 8$, 16, 20, 24, and 32 hours after inoculation.

Because we are only showing how our methods can be used in practice, not providing a complete analysis of the data, we will focus on the 0-8 hour and 0-32 hour interactions, especially the latter. This does not imply that the other interactions are of less importance.

12.1 Estimating π_0 and g

Table 5 contains the "Semi, θ_1", "Semi, compromise", "Semi, $\min\{\widehat{f}\}$", and "Convex" estimates of π_0 for tests of the 0-8, 0-16, 0-20, 0-24, and 0-32 interactions. The "Semi, compromise" estimates for the 0-t interaction decreases as t increases from 8 to 32, indicating that more genes are being differentially expressed as the time since exposure increases. The "Convex" and "Semi, $\min\{\widehat{f}\}$" estimates are similar to each other and both are larger than the "Semi, compromise" estimates. Moreover, the differences between the "Convex" or "Semi, $\min\{\widehat{f}\}$" estimate and the "Semi, compromise" estimate increases as the estimates get smaller (t gets larger). This is the same pattern we saw in the simulation study—"Convex" and "Semi, $\min\{\widehat{f}\}$" are more upwardly biased as π_0 decreases.

Figure 5 shows the estimates of f_p and g for each set of tests. Note that the estimates of g peak at 0, indicating that most values of the non-centrality parameters are near the

null. This is another reason why the "Convex" and "Semi, min$\{\widehat{f}\}$" estimates of π_0 have a large positive bias.

Table 6 has the results from bootstrapping "Semi, compromise". Figure 6 contain 30 bootstrap estimates of g and histograms of 250 bootstrap "Semi, compromise" estimates from resampling p-values. The bootstrap results suggest that g and π_0 can be estimated with reasonably good accuracy. However, the bootstrap may overestimate accuracy if the p-values are not conditionally independent, given $\delta_1, \ldots, \delta_n$, so the bootstrap results should be interpreted with caution.

Figure 7 is a plot of (27), the estimated proportion of δ_i less than δ', versus δ' for the 0-32 hour interaction. The estimate of π_0 is "Semi, compromise", which is 0.57 in this example. If $\delta' = 1$, then one can see in Figure 7 that about 82% of the null hypotheses are either true or "false but with δ uninteresting." Since "Semi, compromise" = 0.57, it appears that about 25% of the null hypotheses are "false but with δ uninteresting" and about 18% are "false and δ is interesting."

12.2 Estimating the FDR and FIDR

Figures 8 shows estimates of FDR as functions of the critical value α for the p-value. That figure has estimates using both "Semi, compromise" and "Convex" for the 0-8 and 0-32 hour interactions. For the 0-8 hour interaction, "Semi, θ_1" and "Convex" are very close to each other and therefore give similar FDR estimates. For the 0-32 hour interaction, the upward bias of "Convex" causes a some overestimation of the FDR; if $\alpha = 0.002$, then the estimated FDR is about 0.038 using "Semi, compromise" but 0.046, about 21% higher, using "Convex".

Figure 9 shows the estimate of the FIDR(α, δ') for the 0-32 hour interaction data. Here $\delta' = 0.55$, 1.09, and 2.18 which are the second, third, and fifth of 12 knots. Suppose we use $\alpha = 0.002$, that is, we reject the null if p-value < 0.002. Then one can see from 9 that the FIDR(0.002,1.09) is about 0.13, more than three times the FDR of 0.038. However, FIDR(0.002,0.55) is only 0.05, much closer to the FDR.

If we wanted to have the FIDR$(\alpha, 1.09)$ close to 0.1, for example, then Figure 9 suggests $\alpha = 0.001$. For the 0-32 hour interaction, about 2.3% (534 of 22,840) of the p-values are below 0.001.

12.3 Estimating EDR, TN, and TP

Estimates of the EDR(α, η), EIDR$(\alpha, \eta, 1)$, EIDR$(\alpha, \eta, 2)$, TP(α, η), and TN(α, η) for the 0-32 hour interaction are shown in Figure 10 for $0.001 \leq \alpha \leq .05$ and $\eta = 1$, 2, and 4. There are vertical lines in these plots through $\alpha = 0.01$. Suppose we use this value of α. Then from

the top plot in Figure 10 we see that the EDR is 0.1 if the current number of replicates, three, is maintained. If six replicates are used, then the EDR rises to about 0.18, and if twelve replicates are used then the EDR is about 0.3. These numbers somewhat discouraging—even with twelve replicates only about 30% of the genes with a 0-32 interaction will be discovered. The problem here is that most of these expressed genes are difficult to discover because they have a non-centrality parameter for the 0-32 interaction that is near 0. If we only consider "interesting" genes with $\delta > 1$ then the value of $\text{EIDR}(0.01, \eta, 1)$ is nearly double the value of $\text{EDR}(0.02, \eta)$, i.e., about 0.2, 0.38, and 0.6 for three, six, and twelve replicates, respectively. Moreover, $\text{EIDR}(0.01, \eta, 2)$ is even larger, approximately 0.4, 0.7, and 0.95 for three, six, and twelve replicates, respectively. Thus, with twelve replicates, we can expect to discover 95% of the genes with a 0-32 hour interaction so large that the non-centrality parameter is 2 to larger.

13 Summary

The barley gene expression data suggests that g is close to the null for these data. In such situations, the simulation results show that "Convex" and "Semi, $\min\{\widehat{f}\}$" are positively biased and "Semi, compromise" is the best of these three estimators, especially when π_0 is not close to 1.

In other studies, g may be far from the null and then the simulation results suggest that "Convex" and "Semi, $\min\{\widehat{f}\}$" will outperform "Semi, compromise". The simulation studies also suggest that "Semi, $\min\{\widehat{f}\}$" will be somewhat superior to "Convex" is such cases.

Since our semiparametric methodology produces an estimator of g, in any application we can assess whether g is near the null or not. This assessment will provide guidance as to whether "Semi, compromise" or "Semi, $\min\{\widehat{f}\}$" should be used.

In our example, we found that in all five cases the alternative was poorly separated from the null in that g peaked at 0 and the probability was at least 1/2 that the non-centrality parameter was less than 1. This is different from the types of alternatives used in simulation studies by other investigators, e.g., Broberg (2005) and Langaas et al. (2005). In our Monte Carlo study, we used three rather different g which range from being poorly to well separated from H_0. We found that bias depends strongly on g. We suggest that other investigators use our methods to estimate g in their studies and that future simulation studies investigate g that are poorly separated from the null. Langaas et al. state they use a g separated from the null "to make the estimable upper bound $\overline{\pi}_0$ close to the true π_0," that is, to ameliorate the positive bias of "Convex" and the other estimators they consider and they also state that "does not mean that we imply that smaller changes are biologically

uninteresting." Our results suggest that one can target π_0 itself as the quantity to estimate rather than the upper bound of $\overline{\pi}_0 = f_p(1)$ and then there is no need to restrict g as they have done.

If one must choose a single estimator among those studied, we recommend "Semi, compromise" since it had generally good performances in all cases in our simulations study, for both one- and two-sided tests and for z-tests as well as t-tests. No other estimator in our study performed well across all cases.

There are many other estimators of π_0 beside those we have studied. Broberg (2005) describes and compares eight of them in a simulation study. However, the estimators in Langaas et al. (2005) are not included in Broberg's study. A full comparison of all available estimators is beyond the scope of this paper. As can be seen in the Tables 1–4 as well as Table 3 in Broberg, bias is often the major component of RMSE and the size and direction of bias depends heavily upon π_0 and g. Finding an estimator with a consistently small bias should be the goal of further research in this area, but whether this goal is attainable or not is unclear.

No other estimator that we are aware of also provides an estimate of g. We believe that this is an important advantage of our methodology, since $\widehat{g}$ can be used to assess the possible size and direction of bias, to estimate how many false nulls are close to the null, and to determine sample sizes appropriate for future studies.

Genovese and Wasserman (2004) discuss the question of when π_0 is identified. They mention that π_0 will be identified under parametric assumptions. We make a parametric assumption about f_p and this seems enough to identify π_0, though we know of no proof. We intend to study in the future the robustness of our methodology to this parametric assumption. Robustness is an issue even for more nonparametric estimators such as "Convex" that assume that under the null the p-value is uniformly distributed. For the t-test, for example, this assumption will hold only under the ideal circumstance that all the parametric assumptions behind the t-test hold.

References

Broberg, P. (2005) A comparative review of estimates of the proportion unchanged genes and the false discovery rate, *BMC Bioinformatics*, 6:199 doi10.1186/1471-2105-6-199 (available at http://www.biomedcentral.com/1471-2105/6/199)

Carroll, R. J., and Hall, P. (1988) Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83, 1184–1186.

Caldo, R. A., Nettleton, D., and Wise, R. P. (2004) Interaction-dependent gene expression

in *Mla*-specified response to barley powdery mildew, *The Plant Cell*, 16, 2514–2528.

Efron, B. (2004) Large-scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis, *Journal of the American Statistical Association*, 99, 96–104.

Fan, J. (1991) On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19, 1257–1272.

Gadbury, G. L., Page, G. P., Edwards, J., Kayo, T., Prolla, T. A., Weindruch, R., Permana, P. A., Mountz, J. D., and Allison, D. B. (2004) Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research*, 13, 325–338.

Genovese, C., and Wasserman, L. (2004) A stochastic process approach to false discovery control, *The Annals of Statistics*, 32, 1035–1061.

Hájek, J. and Šidák, Z. (1967) *Theory of Rank Tests*, Academia, Prague.

Holland, P. (1967) A variation on the minimum chi-square test. *Journal of Mathematical Psychology*, 4, 377–413.

Lesperance, M. L., and Kalbfleisch, J. D. (1992) An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, 87, 120–126.

Meinshausen, N., and Rice, J. (2005) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses, *The Annals of Statistics*, to appear.

Morton, N. (1955) Sequential tests for the detection of linkage. *American Journal of Human Genetics* 7, 277-318.

O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems, *Statistical Science*, 1, 502–518.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications, 2nd edition*, John Wiley & Sons, NY.

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. *The Annals of Statistics*, 31, 2013-2035.

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

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
π_0	0.95	0.95	0.95	0.7	0.7	0.7
nearest of g from null	near	moderate	far	near	moderate	far
RMSE						
"Semi, θ_1"	0.0153	0.0155	0.0163	0.0326	0.0404	0.0244
"Semi, min$\{\widehat{f}\}$"	0.0133	0.0102	0.0125	0.0391	0.0195	0.0127
"Semi, compromise"	0.0132	0.0103	0.0126	0.0381	0.0194	0.0128
"Convex"	0.0163	0.0119	0.0103	0.0674	0.0259	0.0128
Bias						
"Semi, θ_1"	0.0001	−0.0099	−0.0103	−0.0042	−0.0131	−0.0086
"Semi, min$\{\widehat{f}\}$"	0.0068	−0.0048	−0.0080	0.0351	0.0022	−0.0040
"Semi, compromise"	0.0066	−0.0050	−0.0081	0.0339	0.0013	−0.0042
"Convex"	0.0092	0.0008	−0.0039	0.0642	0.0187	−0.0006

Table 1: One-sided z-tests. 600 Monte Carlo simulated data sets, each with 10,000 p-values. RMSE is the root mean squared error. $(\delta_{\min}, \delta_{\max}, b_1, b_2)$ is (0, 4, 1, 2) in Cases 1 and 4, (0, 4, 2, 2) in Cases 2 and 5, and (0.5, 4.5, 3, 2) in Cases 3 and 6.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
π_0	0.95	0.95	0.95	0.7	0.7	0.7
nearest of g from null	near	moderate	far	near	moderate	far
RMSE						
"Semi, θ_1"	0.0222	0.0265	0.0288	0.0540	0.1008	0.0947
"Semi, min$\{\widehat{f}\}$"	0.0265	0.0127	0.0057	0.1507	0.0766	0.0188
"Semi, compromise"	0.0197	0.0090	0.0135	0.1004	0.0319	0.0137
"Convex"	0.0232	0.0147	0.0107	0.1476	0.0759	0.0205
Bias						
"Semi, θ_1"	−0.0019	−0.0183	−0.0207	0.0020	−0.0350	−0.0306
"Semi, min$\{\widehat{f}\}$"	0.0259	0.0115	−0.0012	0.1505	0.0762	0.0176
"Semi, compromise"	0.0171	0.0007	−0.0093	0.0994	0.0260	−0.0028
"Convex"	0.0204	0.0090	−0.0019	0.1467	0.0742	0.0157

Table 2: Two-sided z-tests. 600 Monte Carlo simulated data sets, each with 10,000 p-values. RMSE is the root mean squared error. $(\delta_{\min}, \delta_{\max}, b_1, b_2)$ is (0, 4, 1, 2) in Cases 1 and 4, (0, 4, 2, 2) in Cases 2 and 5, and (0.5, 4.5, 3, 2) in Cases 3 and 6.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
π_0	0.95	0.95	0.95	0.7	0.7	0.7
nearest of g from null	near	moderate	far	near	moderate	far
RMSE						
"Semi, θ_1"	0.0251	0.0348	0.0346	0.0503	0.1035	0.0732
"Semi, min$\{\widehat{f}\}$"	0.0276	0.0145	0.0066	0.1519	0.0748	0.0187
"Semi, compromise"	0.0206	0.0124	0.0269	0.0492	0.0268	0.0458
"Convex"	0.0238	0.0148	0.0121	0.1485	0.0767	0.0214
Bias						
"Semi, θ_1"	0.0001	−0.0228	−0.0286	−0.0063	−0.0703	−0.0233
"Semi, min$\{\widehat{f}\}$"	0.0266	0.0123	−0.0014	0.1517	0.0743	0.0166
"Semi, compromise"	0.0076	0.0007	−0.0221	0.0367	0.0158	−0.0169
"Convex"	0.0208	0.0089	−0.0020	0.1476	0.0749	0.0167

Table 3: Two-sided t-tests with $DF = 4$ and $n = 10{,}000$ p-values per data set. RMSE and bias for 4 estimators and 4 cases. 600 Monte Carlo simulated data sets per case. RMSE is the root mean squared error. $(\delta_{\min}, \delta_{\max}, b_1, b_2)$ is (0, 4, 1, 2) in Cases 1 and 4, (0, 4, 2, 2) in Cases 2 and 5, and (0.5, 4.5, 3, 2) in Cases 3 and 6.

ρ	0	0.5	0.75	0.5	0.75
δ	–	"permute"	"permute"	"sort"	"sort"
RMSE					
"Semi, θ_1"	0.0333	0.0359	0.0582	0.0395	0.0523
"Semi, min$\{\widehat{f}\}$"	0.0519	0.0515	0.0518	0.0519	0.0525
"Semi, compromise"	0.0364	0.0361	0.0378	0.0368	0.0389
"Convex"	0.0491	0.0491	0.0488	0.0490	0.0498
Bias					
"Semi, θ_1"	−0.0044	−0.0061	−0.0102	−0.0065	−0.0083
"Semi, min$\{\widehat{f}\}$"	0.0512	0.0507	0.0504	0.0509	0.0510
"Semi, compromise"	0.0337	0.0330	0.0324	0.0332	0.0332
"Convex"	0.0472	0.0468	0.0456	0.0462	0.0464

Table 4: Two-sided t-tests with DF=4, $n = 10{,}000$, and AR(1) data. ρ is the AR parameter. The ith p-value is based on $t_i = (\delta_i + e_i)/\sqrt{s_i^2/4}$ where e_i is an AR process and s_i^2 is independent of e_i and χ^2_4 distributed. "permute" means that the δ_i are randomly permuted. "sort" means that the δ_i are sorted. π_0 is fixed at 0.9 and $(\delta_{\min}, \delta_{\max}, b_1, b_2)$ is fixed at (0, 4, 2, 2). 600 Monte Carlo simulated data sets per case.

Interaction	0-8	0-16	0-20	0-24	0-32
"Semi, θ_1"	0.8639	0.6497	0.2734	0.1644	0.2307
"Semi, compromise"	0.9195	0.8519	0.8075	0.7884	0.5728
"Semi, min$\{\widehat{f}\}$"	0.9435	0.9074	0.8743	0.8605	0.7097
"Convex"	0.9324	0.9087	0.8668	0.8468	0.7032

Table 5: Estimates of π_0 for barley gene expression interaction tests. "0-t" is the interaction between resistant/susceptible and time at 0 and t hours after exposure.

Interaction	0-8	0-16	0-20	0-24	0-32
estimate for barley data	0.9195	0.8519	0.8075	0.7884	0.5728
bootstrap mean	0.9197	0.8522	0.8133	0.7885	0.5721
bootstrap std dev	0.0067	0.0101	0.0100	0.0114	0.0116
bootstrap 2.5 %	0.8885	0.8277	0.7907	0.7592	0.5395
bootstrap 97.5 %	0.9412	0.9003	0.8592	0.8185	0.6042

Table 6: Bootstrapping "Semi, compromise" using the barley gene expression interaction tests. The top row gives the value of "Semi, compromise" for the original data. The remaining rows are the bootstrap mean, bootstrap standard deviation, bootstrap 2.5 percentile, and bootstrap 97.5 percentile from 250 bootstrap samples.

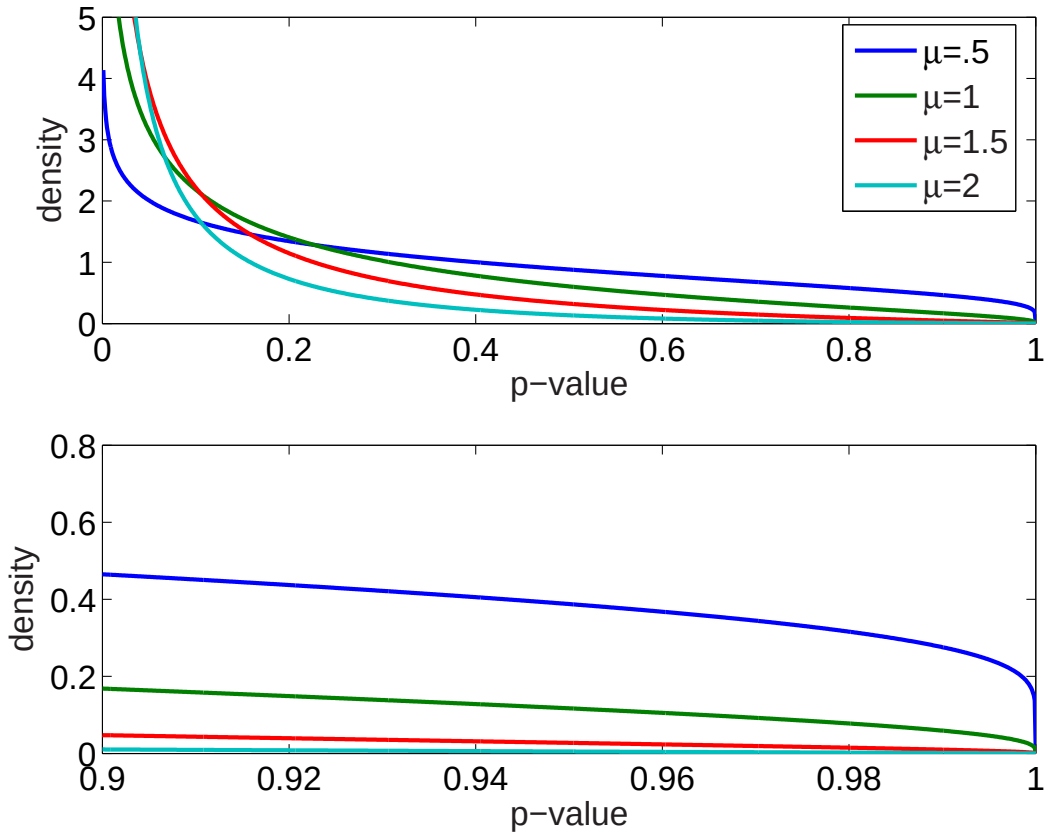

Figure 1: Density of the p-value from a z-test of $H_0 : \delta = 0$ versus $H_1 : \delta > 0$ when $\delta = .5$, 1, 1.5, and 2. The lower plot zooms in on the region where the density is concave.

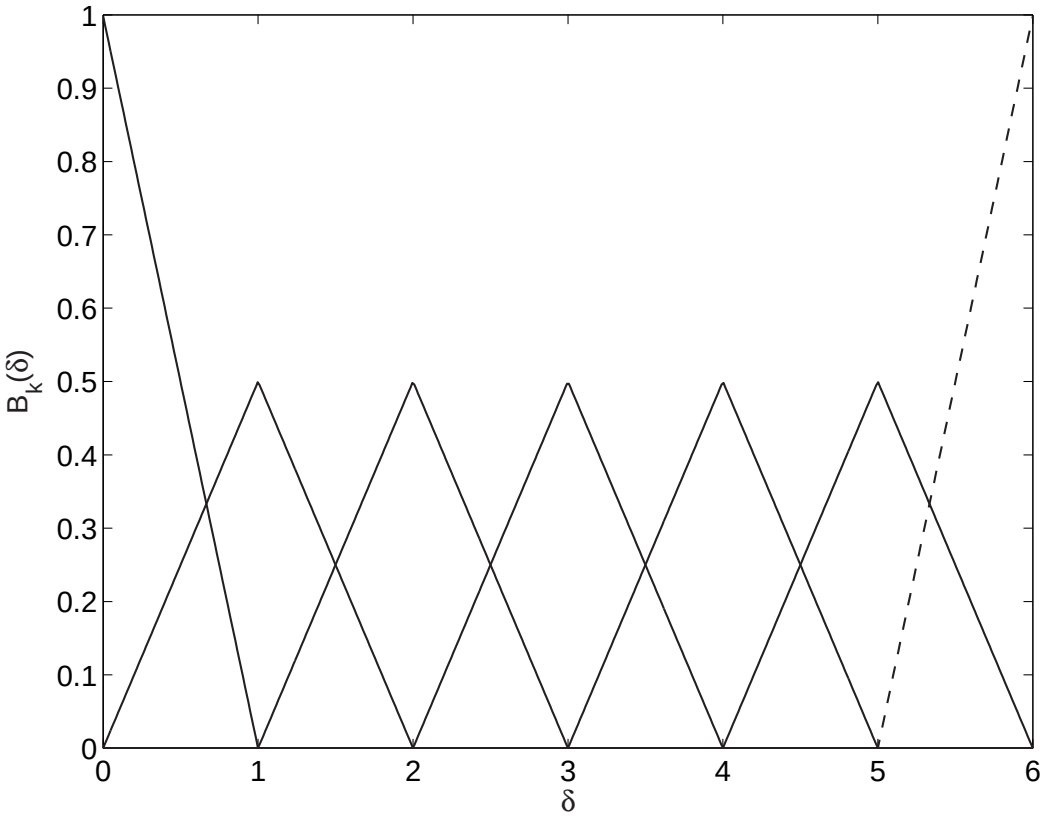

Figure 2: B-splines with 7 knots and $\delta^* = 6$ used to model g. Each B-spline is normalized to be a density. The B-spline with support [5, 6] is shown as a dashed line and is not used in the model for g because it is discontinuous at 6.

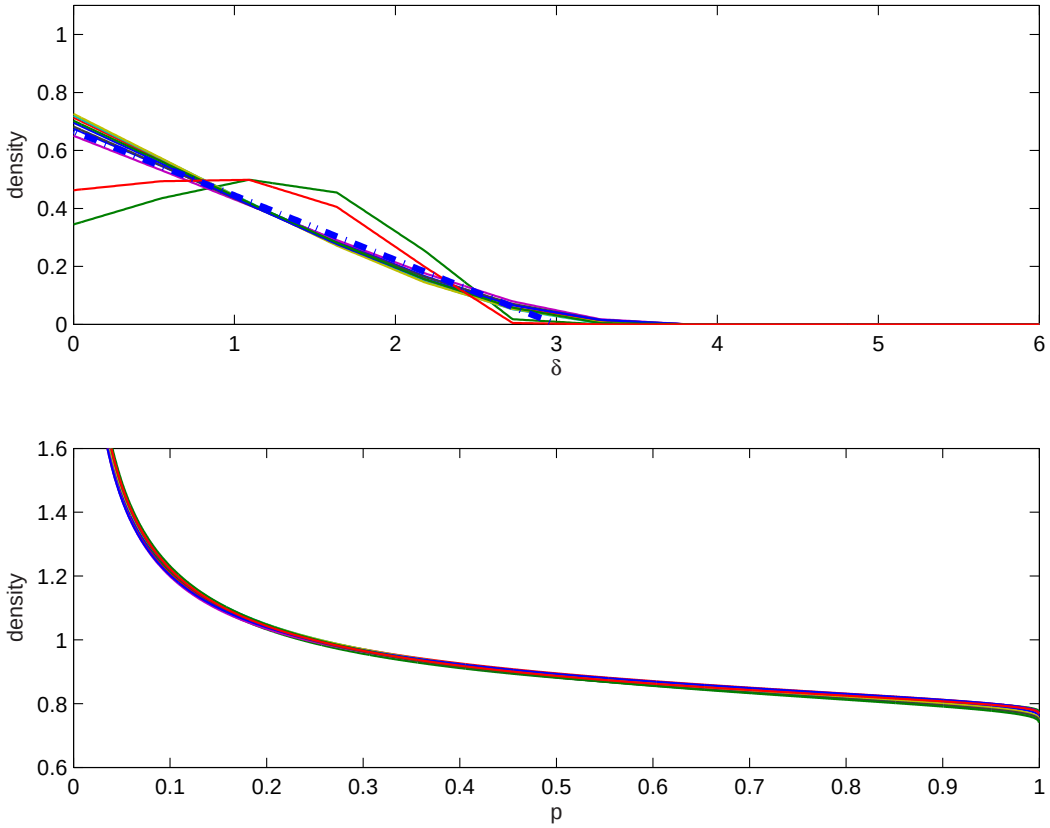

Figure 3: One-sided z-tests. Case 4 where $\pi_0 = 0.7$ and the alternative values of δ are close to the null. Estimates of g (top) and f_p (bottom) from 10 independently simulated data sets, each with $n = 10,000$. The true density g is a thick dotted-and-dashed line.

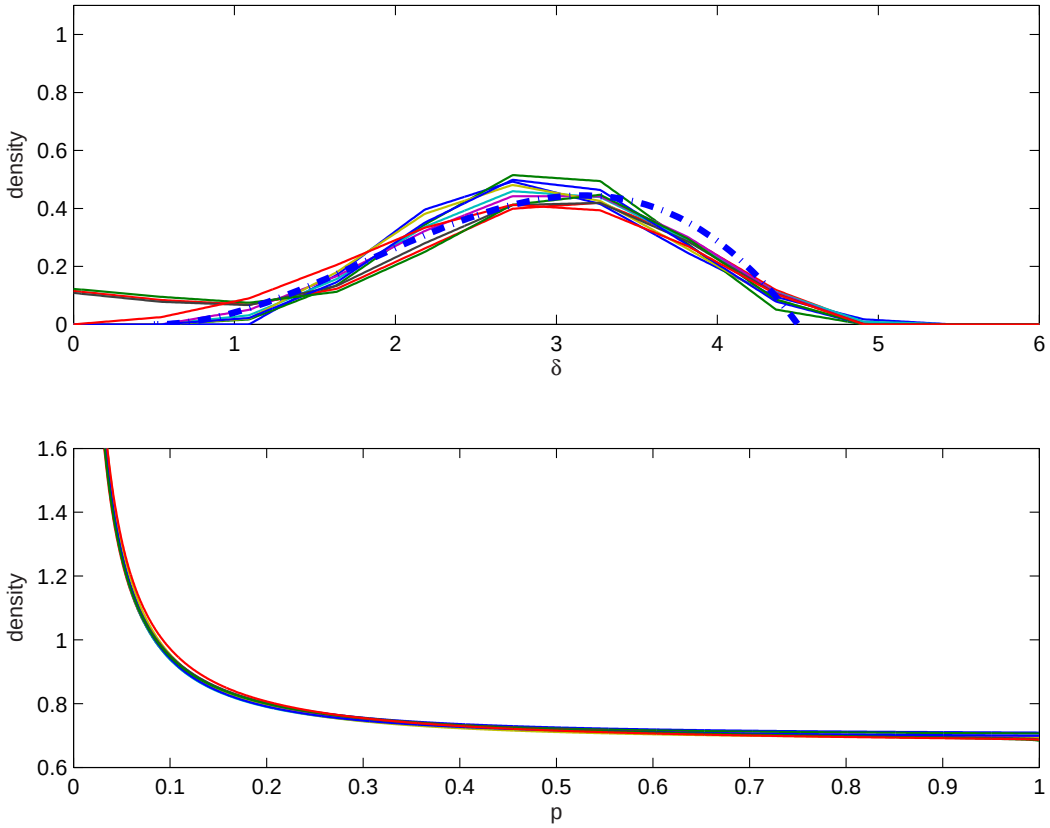

Figure 4: One-sided z-tests. Case 6 where $\pi_0 = 0.7$ and the alternative values of δ are far from the null. Estimates of g (top) and f_p (bottom) from 10 independently simulated data sets, each with $n = 10,000$. The true density g is a thick dotted-and-dashed line.

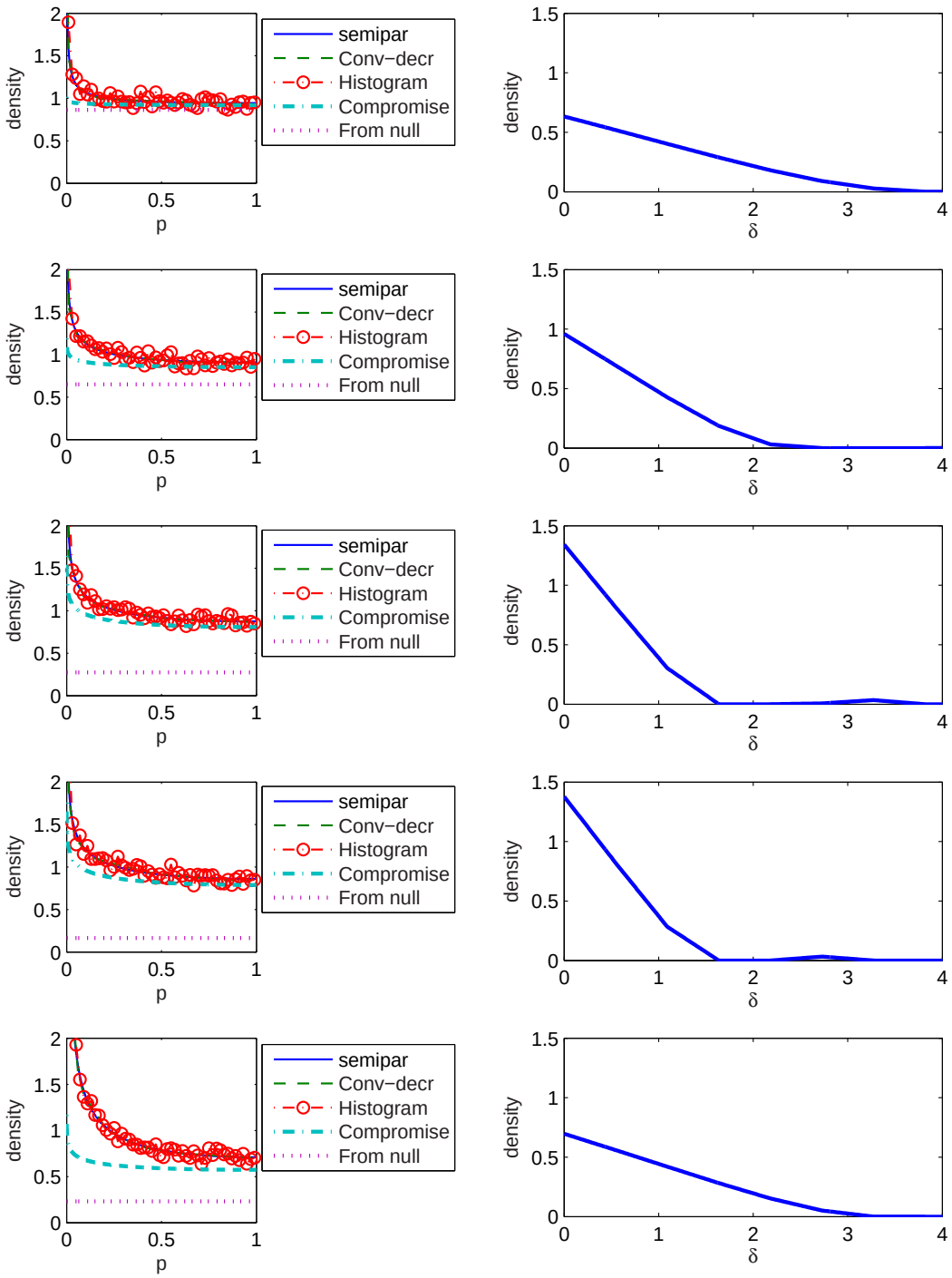

Figure 5: Barley gene expression data. Top to bottom rows: 0-8, 0-16, 0-20, 0-24, and 0-32 hour interactions. Left plots show the semiparametric (semipar) and convex, decreasing (conv-decr) estimates of f_p and a histogram of the p-values—the "o" are at the tops of the 50 bins. "From null" shows the estimated component of f_p coming from the null hypotheses—it is the uniform (0,1) density multiplied by "Semi, θ_1". "Compromise" shows the estimated component of f_p coming from the null hypotheses or δ close to the null value—see text. The height of the "compromise" estimate of f_p at 1 is the "Semi, compromise" estimate of π_0. The right plots are the estimates of g.

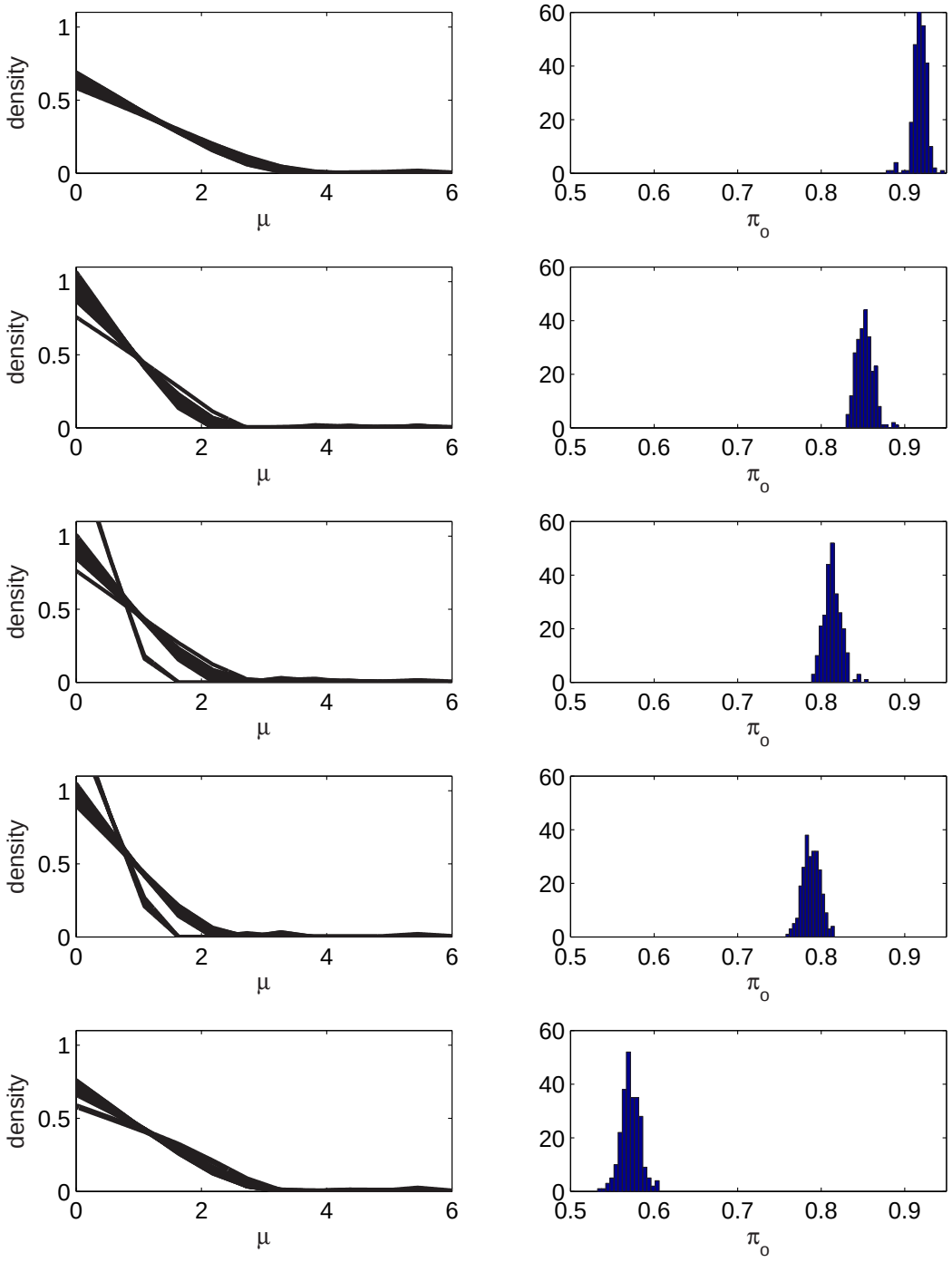

Figure 6: Plot of 30 bootstrap estimates of g (left) and histogram of 250 bootstrap estimates of π_0 (right). Top to bottom: 0-8, 0-16, 0-20, 0-24, and 0-32 hour interactions.

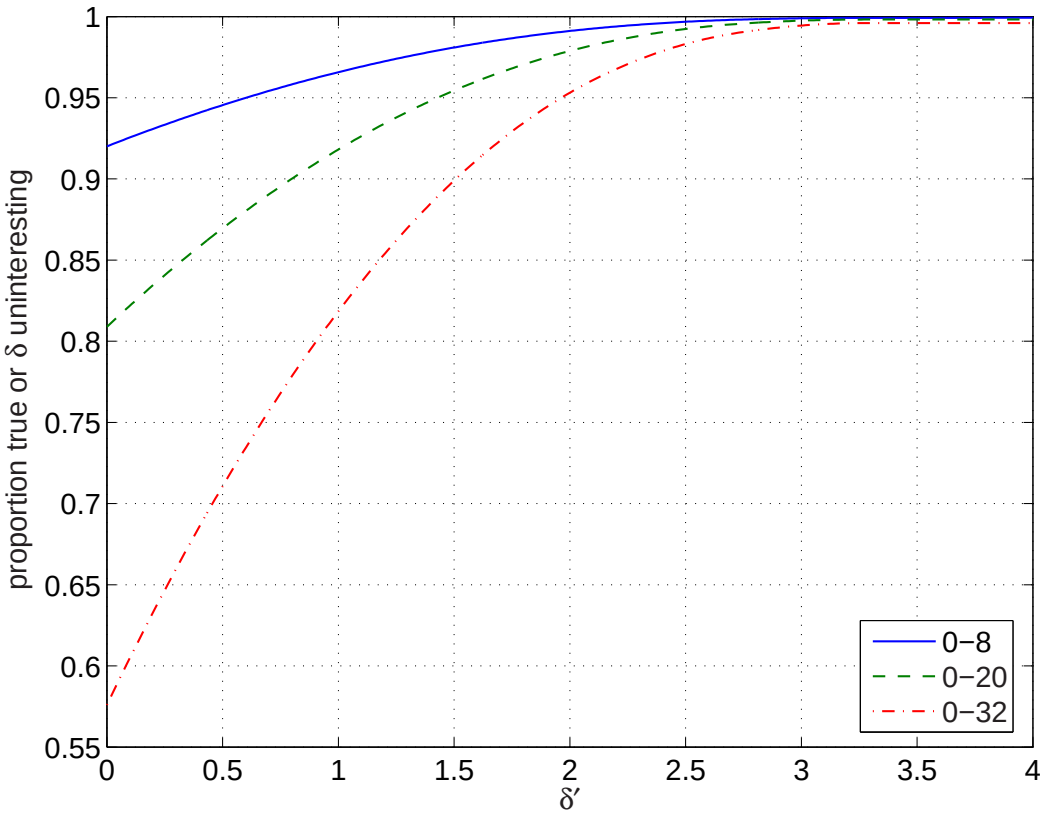

Figure 7: Barley data, 0-8, 0-20, and 0-32 hour interactions. Estimated proportion of hypotheses that are either true or "false but δ uninteresting," where the latter means that $0 < \delta < \delta'$.

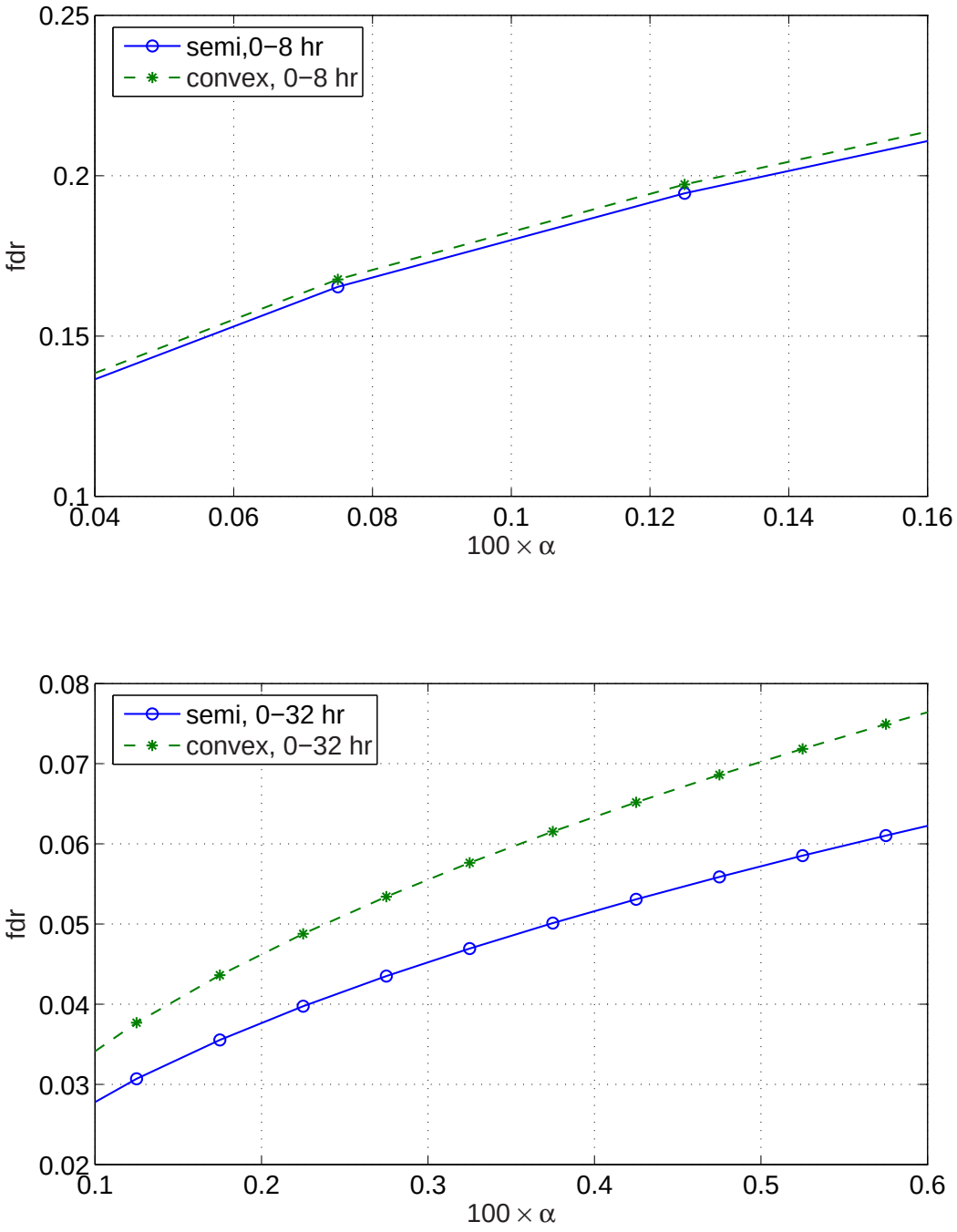

Figure 8: Estimating the false discovery rate when a null hypothesis is rejected if the p-value is less than α.

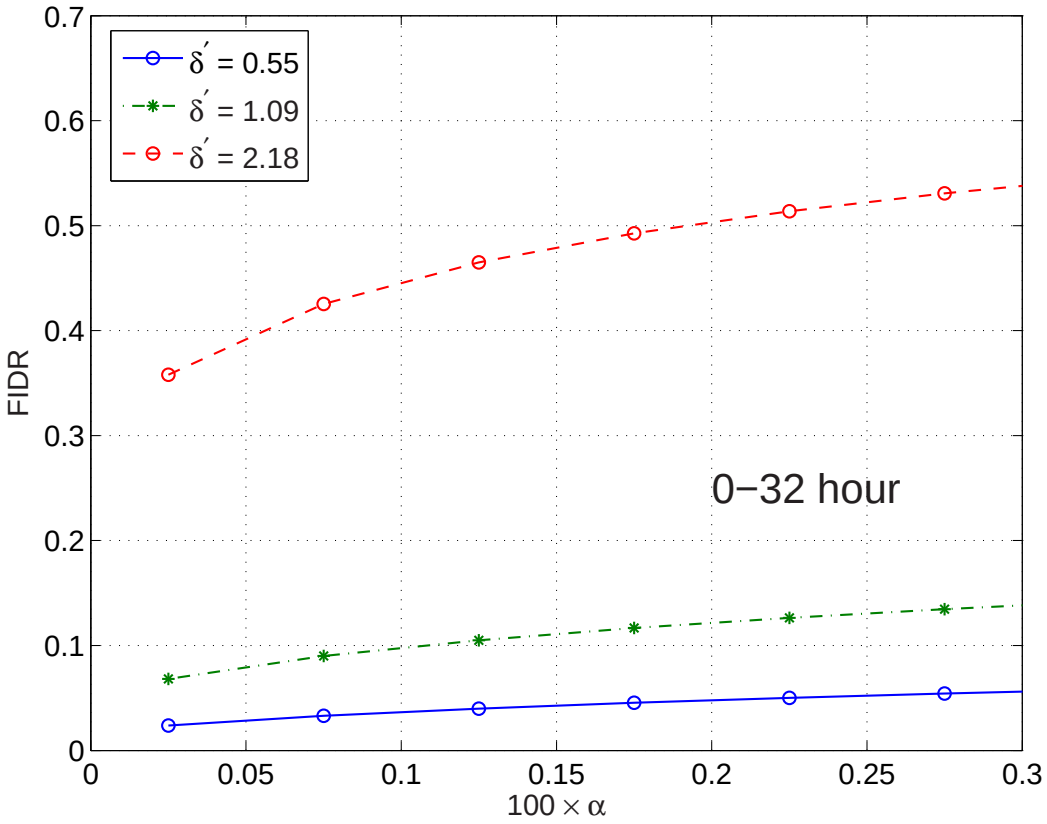

Figure 9: Estimating the falsely interesting discovery rate (FIDR), 0-32 hour interaction, when a null hypothesis is rejected if the p-value is less than α. A null hypothesis is "false but δ is uninteresting" if $0 < \delta < \delta'$. The FIDR is the proportion of rejected nulls that are either true or false but with δ uninteresting. The FIDR increases as δ' increases and as α increases.

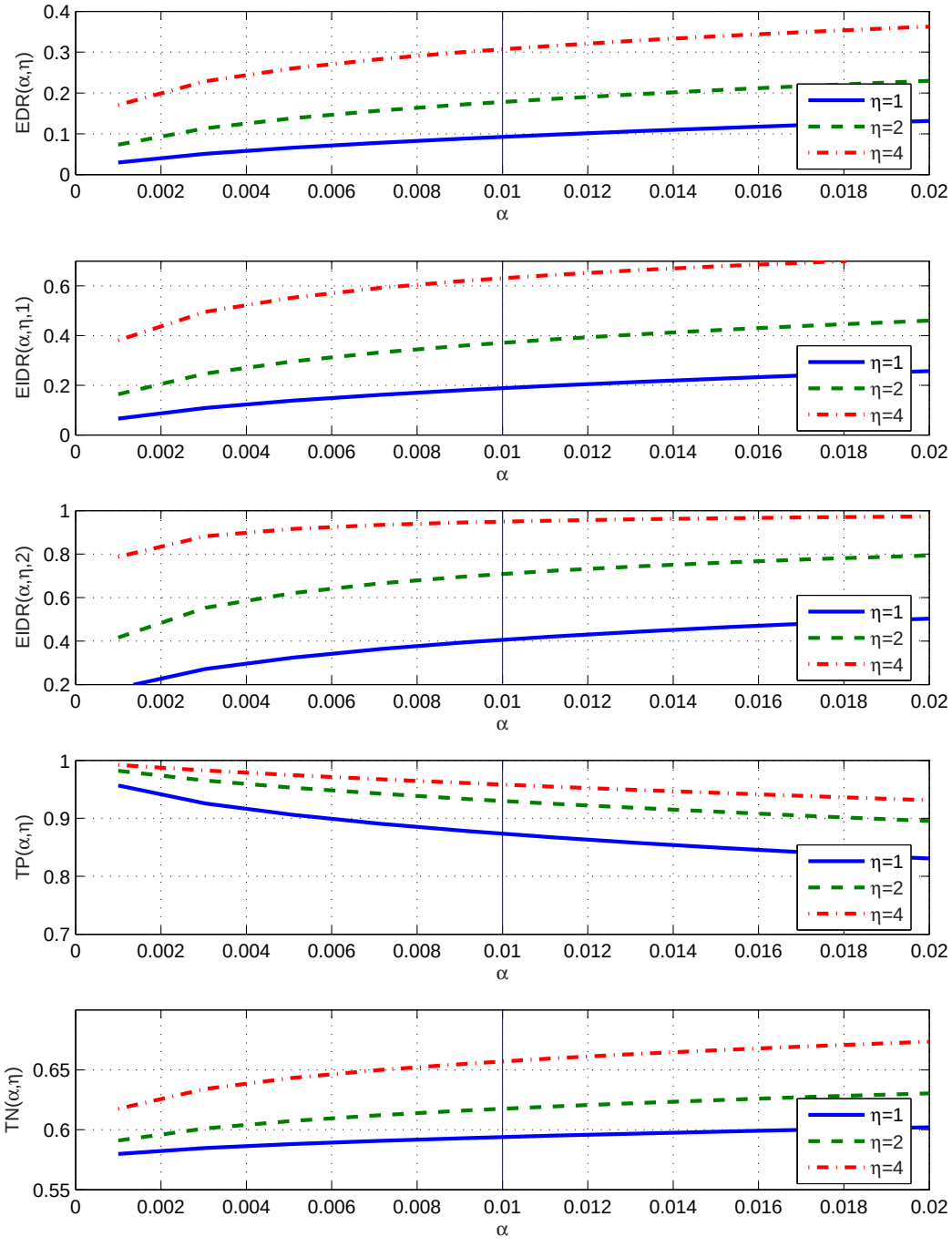

Figure 10: Barley data. 0-32 hour interaction. Estimates of $\mathrm{EDR}(\alpha, \eta)$ (Expected Discovery Rate), $\mathrm{EIDR}(\alpha, \eta, 1)$ (Expected Interesting Discovery Rate with $\delta' = 1$), $\mathrm{TP}(\alpha, \eta)$ (True Positive), and $\mathrm{TN}(\alpha, \eta)$ (True Negative) curves for $\eta = 1$, 2, 4.