Harvard University
Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 77

Empirical Null and False Discovery Rate Inference for Exponential Families

Armin Schwartzman*

*Harvard School of Public Health, armins@hsph.harvard.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/harvardbiostat/paper77

Copyright ©2008 by the author.

Empirical null and false discovery rate inference for exponential families

Armin Schwartzman

January 25, 2008

Authors' footnote

Armin Schwartzman is Assistant Professor, Department of Biostatistics, Harvard School of Public Health, and Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, both in Boston, MA 02115 (E-mail: *armins@hsph.harvard.edu*).
This research was partially supported by a William R. and Sara Hart Kimball Stanford Graduate Fellowship. The author thanks Bradley Efron and Jonathan Taylor for their guidance, Christoph Lange for providing the FBAT-data analysis results, and an associate editor for helpful suggestions.

From the Framingham Heart Study of the National Heart, Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. This work was supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the National Heart, Lung and Blood Institute.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Abstract

In large scale multiple testing, the use of an empirical null distribution estimated from the data rather than the theoretical null distribution can be critical for correct inference. While previous empirical null methods have focused on situations where the theoretical null is $N(0,1)$ or t, other theoretical null distributions such as χ^2 and F are also common in genomics and imaging applications. This paper proposes a 'mode matching' method as an extension to the central matching method for fitting an empirical null, applicable in more general situations where the theoretical null belongs to any exponential family. Mode matching estimates the null density by fitting an appropriately defined exponential family to the histogram of the test statistics by Poisson regression in a region surrounding the mode. The empirical null estimate is then used to estimate local and tail false discovery rate (FDR) for inference. Delta-method covariance formulas and approximate asymptotic bias formulas are provided, as well as simulation studies of the effect of the tuning parameters of the procedure on the bias-variance trade-off. The standard FDR estimates are found to be biased down at the far tails. Correlation between test statistics is taken into account in the covariance estimates, providing a generalization of Efron's 'wing function' for exponential families. An example using χ^2 statistics is shown in a family-based genome-wide association study from the Framingham Heart Study.

Keywords: multiple testing, multiple comparisons, mixture models, Poisson regression, genome-wide association.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

1 Introduction

In large-scale multiple testing problems, the observed distribution of the test statistics often does not accurately match the theoretical null distribution (Efron et al. 2001; Efron 2004, 2005). In such cases, the use of an empirical null distribution, estimated from the data itself, can be critical for making correct inferences. Previous empirical null methods (Efron 2004, 2007b; Jin and Cai 2007) have focused on situations where the theoretical distribution of the test statistics is $N(0,1)$ or t, typically found, for example, in two-group microarray gene expression studies. Many other multiple testing problems, however, present theoretical null distributions that are not normal or t. For example, χ^2 tests are commonplace in genome-wide association studies based on single nucleotide polymorphisms (SNPs) (Van Steen et al. 2005; Kong et al. 2006), while multivariate F tests appear in brain imaging studies based on magnetic resonance imaging (MRI) and voxel-based mophometry (Everitt and Bullmore 1999; Schwartzman et al. 2005; Lee et al. 2007; Schwartzman et al. 2008).

This paper extends the scope of the empirical null to distributions that belong to general exponential families, treating the normal and χ^2, as well as their counterparts t and F, as special cases. This extension allows the empirical null to be flexibly chosen as a parametric exponential family version of the theoretical null. For example, where the theoretical null $N(0,1)$ may be replaced by an empirical null $N(\mu,\sigma^2)$ with arbitrary mean μ and variance σ^2, a theoretical null $\chi^2(\nu_0)$ with fixed ν_0 degrees of freedom may be replaced by a scaled χ^2 density (i.e. gamma) with arbitrary scaling factor a and arbitrary number of degrees of freedom ν (Schwartzman et al. 2008).

The proposed method for fitting the empirical null, which we call 'mode matching', is a generalization of the central matching method for z-scores (Efron et al. 2001; Efron 2004, 2007b). Mode matching consists of fitting the empirical null to a region of the histogram of the test statistics surrounding the mode, which for the normal

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

distribution coincides with matching the center. We present mode matching here with a one-step approach, fitting the empirical null to the histogram directly by Poisson regression. This contrasts with the two-step scheme previously proposed, where a nonparametric density is first fitted to the histogram by Poisson regression and then the empirical null is fitted to the nonparametric density estimate by least squares. The one-step fit not only simplifies the theoretical analysis of bias and variance and avoids the need to tune additional parameters for nonparametric density estimation. It also highlights why mode matching is effective for exponential families: for these the log-link function of the Poisson regression becomes linear in the regression parameters.

The empirical null may be used with any multiple testing procedure. Nonetheless, mode matching is particularly suited for estimating the false discovery rate (FDR), a commonly used error measure in multiple testing problems (Benjamini and Hochberg 1995; Genovese and Wasserman 2004; Storey et al. 2004). We present formulas for calculating the local and tail FDR estimates and show that, as with central matching (Efron 2005, 2007b), these estimates follow easily from mode matching calculations for general exponential families.

We derive delta method covariance formulas for both the empirical null and FDR estimates and show that these formulas produce variance estimates similar to those obtained by the bootstrap method. Further, we derive approximate formulas for the bias of both the empirical null and FDR estimates and show that the bias in the empirical null is driven mainly by the likelihood ratio between the alternative and null distributions. Simulations are used to inform the choice of the two tuning parameters of mode matching (histogram bin width and fitting interval) in terms of the bias-variance trade-off. For example, in agreement with Efron (2007b), we find that in the normal case mode matching is fairly insensitive to the choice of bin width, but in the χ^2 case, the choice of bin width is affected by the curvature of the

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

density, which sharply increases when the number of degrees of freedom is less than 2. The fitting interval is even more important, as it controls the bias introduced by the alternative distribution. In terms of FDR estimation, the bias formulas reveal that both the local and tail FDR estimates can be deceptively biased down for very high thresholds (low p-values), where the number of observed test statistics is low. We argue that this effect should be carefully taken into account when making inferences in real data sets. The effect of correlation between test statistics is taken into account in the covariance estimates. We show that Efron's enigmatic "wing function" Efron (2007a) is a special case of the large family of Lancaster polynomials of bivariate exponential families, which reduces to the Hermite polynomials in the normal case and to the Laguerre polynomials in the χ^2 case.

The above methods are both computationally efficient and easy to implement because they are based on Poisson regression, for which software is widely available. The analysis is demonstrated in a family-based genome-wide association study based on SNPs, where the theoretical null distribution of the test statistics is χ^2. For an illustration of how mode matching is applied to F-statistics, the reader is referred to Schwartzman et al. (2008), which shows an analysis example in the context of brain imaging.

2 Mode matching for exponential families

2.1 Setup

Let $T_1, \ldots, T_N$ be a large collection of N independent test statistics (the dependent case is considered in Section 5). The two-class mixture model (Efron et al. 2001; Storey 2003; Efron 2004; Sun and Cai 2007)

$$f(t) = p_0 f_0(t) + (1 - p_0) f_A(t) \tag{1}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

specifies that a fixed fraction p_0 of the test statistics behave according to a common null distribution with density $f_0(t)$, which is assumed unimodal. The zero assumption, needed for identifiability of the model, mantains that most of the probability mass near the mode of $f(t)$ is due to the null term $p_0 f_0(t)$ (e.g. $p_0 > 0.9$). The alternative density $f_A(t)$ may itself be a mixture but its form is irrelevant as long as its contribution $(1-p_0)f_A(t)$ in (1) is small near the mode of $f(t)$. The objective of the empirical null methodology is to estimate p_0 and f_0 from $T_1, \ldots, T_N$.

Mode matching begins by summarizing the data into a vector of histogram counts $\boldsymbol{y} = (y_1, \ldots, y_K)'$ with $y_k = \sum_{i=1}^{N} \mathbf{1}\{T_i \in B_k\}$, $k = 1, \ldots, K$, for K bins B_k centered at $\boldsymbol{t} = (t_1, \ldots, t_K)'$. For simplicity, we assume all bins have the same width Δ. Given N, the counts $\boldsymbol{y}$ follow a multinomial distribution with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, $\pi_k = P(T_i \in B_k)$. By Taylor expansion around t_k,

$$\pi_k = \int_{B_k} f(t)\, dt = \Delta f(t_k) + \frac{\Delta^3}{24} f''(t_k) + \cdots \approx \Delta f(t_k). \tag{2}$$

The approximation is valid if the bin width Δ is small and $f(t)$ is smooth (the effect of curvature is discussed in Section 4). Thus, for large N, the scaled histogram

$$\hat{f}(\boldsymbol{t}) = \frac{\boldsymbol{y}}{N\Delta} \tag{3}$$

is a nearly unbiased estimate of the mixture density $f(\boldsymbol{t})$ at the bin centers $\boldsymbol{t}$.

The next step is to choose a closed interval S_0 where the zero assumption may hold. S_0 is the union of $K_0 < K$ consecutive bins containing the mode of $f(t)$. For example, for a two-sided test with theoretical null $N(0,1)$, S_0 may be of the form $S_0 = [t_{\min}, t_{\max}]$, while for a one-sided test with theoretical null χ^2, S_0 may be of the form $S_0 = [0, t_{\max}]$. Within S_0, the zero assumption makes (3) an estimate of the scaled null $p_0 f_0(t)$ in (1), with additional bias $(1-p_0)f_A(t)$.

Suppose $f_0(t)$ is a parametric density. Instead of maximizing the multinomial likelihood given $\boldsymbol{y}$, mode matching uses, almost equivalently, Poisson regression. The

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

idea, also called Lindsey's method (Efron and Tibshirani 1996; Efron 2007b), is to consider the number of tests N as a Poisson variable $N \sim Po(\gamma)$. This makes the histogram counts independent Poisson variables $y_k \sim Po(\lambda_k)$ with $\lambda_k = \gamma\pi_k$. If N is large, this is essentially the same as the usual Poisson approximation to the multinomial. Using (2), we have $\lambda_k = \gamma\pi_k \approx \gamma\Delta f(t_k)$. Thus within S_0, the zero assumption leads to the general Poisson regression model $y_k \sim Po(\lambda_k)$ with

$$\lambda_k \approx \gamma\Delta p_0 f_0(t_k), \quad t_k \in S_0, \tag{4}$$

where γ is replaced by its MLE, the observed count N.

2.2 Exponential families

Since the link function for Poisson regression is logarithmic, the precise parametric form of $f_0(t)$ needed to make $\log(\lambda_k)$ in (4) linear in the parameters is an exponential family. Let

$$f_0(t) = g_0(t)\exp\big(\boldsymbol{x}(t)'\boldsymbol{\eta} - \psi(\boldsymbol{\eta})\big) \tag{5}$$

where $g_0(t)$ is the carrier density, $\boldsymbol{\eta}$ is the vector of canonical parameters, $\boldsymbol{x}(t)$ is the sufficient vector and $\psi(\boldsymbol{\eta})$ is the cumulant generating function. Replacing in (4) gives the linear Poisson regression model $y_k \sim Po(\lambda_k)$ with

$$\log(\lambda_k) = \boldsymbol{x}(t_k)'\boldsymbol{\eta} + C + h_k, \tag{6}$$

where the entries of $\boldsymbol{x}(t_k)$ play the role of predictors,

$$C = C(\boldsymbol{\eta}) = \log p_0 - \psi(\boldsymbol{\eta}) \tag{7}$$

is a constant intercept, and $h_k = \log(N\Delta g_0(t_k))$ is an offset. It is convenient to write model (6) in vector form as

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X}\boldsymbol{\eta}^{+} + \boldsymbol{h} \tag{8}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)'$, $\boldsymbol{\eta}^+ = (C, \boldsymbol{\eta}')'$ is the augmented parameter vector, the design matrix $\boldsymbol{X}$ has rows $(1, \boldsymbol{x}(t_k)')$ for $k = 1, \ldots, K$, and $\boldsymbol{h} = (h_1, \ldots, h_K)'$. The fit is restricted to the interval S_0 by providing the Poisson regression algorithm with an external set of weights $\boldsymbol{w} = (w_1, \ldots, w_K)'$ where w_k is equal to 1 or 0 according to whether t_k is in S_0 or not. For later use, we define the diagonal matrix $\boldsymbol{W}$ with diagonal equal to $\boldsymbol{w}$ (not to be confused with the weighting matrix used internally in the iterative solving of the Poisson regression).

Solving (8) gives estimates $\hat{\boldsymbol{\eta}}^+ = (\hat{C}, \hat{\boldsymbol{\eta}})'$, which include the empirical null parameter estimates $\hat{\boldsymbol{\eta}}$. From these, an estimate of the null probability p_0 is also obtained using (7) as $\hat{p}_0 = \exp(\hat{C} + \psi(\hat{\boldsymbol{\eta}}))$. Notice that $\hat{p}_0$ is not constrained to be less or equal to 1. The predicted histogram counts $\hat{\boldsymbol{\lambda}} = N\Delta\hat{f}_0(\boldsymbol{t}) = \hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_K)'$ corresponding to the empirical null for all bins (not just within S_0) are

$$\hat{\boldsymbol{y}} = \exp(\boldsymbol{X}\hat{\boldsymbol{\eta}}^+ + \boldsymbol{h}). \tag{9}$$

As a result, the predicted histogram counts corresponding to the alternative component in (1) are

$$N\Delta(1 - \hat{p}_0)\hat{f}_A(\boldsymbol{t}) = N\Delta\big(\hat{f}(\boldsymbol{t}) - \hat{p}_0\hat{f}_0(\boldsymbol{t})\big) = \boldsymbol{y} - \hat{\boldsymbol{y}}. \tag{10}$$

Empirical null densities are more naturally specified using natural parameters rather than canonical parameters. When the theoretical null is $N(0, 1)$, the empirical null is $N(\mu, \sigma^2)$ with $\boldsymbol{\theta} = (\mu, \sigma^2)'$ (Efron 2004, 2007b) (t-statistics are handled by a quantile transformation to $N(0, 1)$). When the theoretical null is χ^2 with ν_0 d.f., an appropriate empirical null is a scaled χ^2 with ν d.f. and scaling factor a, denoted $a\chi^2(\nu)$, with density

$$f_0(t) = \frac{1}{(2a)^{\nu/2}\Gamma(\nu/2)} e^{-t/(2a)} t^{\nu/2-1} \tag{11}$$

where $\boldsymbol{\theta} = (a, \nu)'$ (Schwartzman et al. 2008). This is the same as a gamma density with shape parameter $\nu/2$ and scaling parameter $2a$, but using the χ^2 notation helps

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

keep the connection to the theoretical null. F-statistics are handled by a quantile transformation to χ^2 with the same numerator number of degrees of freedom.

Let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$ denote the vector of natural parameters of the exponential family empirical null (5), a one-to-one function of $\boldsymbol{\eta}$. Let $\boldsymbol{\theta}^+ = (\log p_0, \boldsymbol{\theta}')'$ be the augmented parameter vector. The MLE of $\boldsymbol{\theta}^+$ is $\hat{\boldsymbol{\theta}}^+ = (\log \hat{p}_0, \boldsymbol{\theta}(\hat{\boldsymbol{\eta}})')'$. The derivation of natural parameter estimates from the canonical parameter estimates for both the normal and χ^2 cases is worked out in the appendix.

Other distributions are treated in a similar way. For p-values, whose theoretical null is uniform, the empirical null may be a beta distribution with fitting interval $S_0 = [t_{\min}, 1]$. If the theoretical null is a discrete exponential family (e.g. binomial, Poisson, negative binomial), the mode matching procedure is the same as above except that the bins width $\Delta = 1$ is automatically set by the discrete nature of the distribution, making equations (2) and (4) exact rather than approximate.

2.3 Exponential subfamilies

In some cases, one may want to adjust only some of the parameters in (5) and leave the others fixed as prescribed by the theoretical null. For instance, the microarray analysis examples in Efron (2007b) suggest the empirical null $N(0, \sigma^2)$, while in some fMRI studies involving z-scores, an appropriate empirical null may be $N(\mu, 1)$ (Ghahremani and Taylor 2005). If fixing some natural parameters results in another lower dimensional exponential family, then the procedure is similar to the one above after the canonical parameters have been redefined. Let $\boldsymbol{\eta}^+ = (C, \boldsymbol{\eta}_1', \boldsymbol{\eta}_2')'$, where $\boldsymbol{\eta}_1$ is the vector of canonical parameters to be estimated and $\boldsymbol{\eta}_2$ is the vector of parameters whose values are fixed. Let $\boldsymbol{X} = (\mathbf{1}_K, \boldsymbol{X}_1, \boldsymbol{X}_2)$ be the corresponding split of the design matrix where $\mathbf{1}_K$ indicates a column of K ones. The regression equation (8)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

becomes

$$\log(\boldsymbol{\lambda}) = \mathbf{1}_K C + \boldsymbol{X}_1\boldsymbol{\eta}_1 + (\boldsymbol{X}_2\boldsymbol{\eta}_2 + \boldsymbol{h}) \tag{12}$$

and is solved as before, except that the fixed term $\boldsymbol{X}_2\boldsymbol{\eta}_2$ is absorbed into the offset in parenthesis. The specific exponential subfamilies of the normal and χ^2 cases are worked out in detail in the appendix.

The simplest restricted case is that where we believe the theoretical null and no adjustment of parameters is necessary, except for p_0 (Efron 2004). In that case, only the intercept C needs to be estimated in (12), treating all the other terms as offset. The estimate of p_0 is then given by $\hat{p}_0 = \exp(\hat{C}+\psi(\boldsymbol{\eta}))$. Notice that, for the regression (12) to remain linear, p_0 cannot be fixed apriori.

2.4 Covariance estimates

Covariance estimates for the empirical null parameter estimates $\hat{\boldsymbol{\eta}}^+$ can be obtained by the delta method in a way similar to Efron (2005). For this we first need an estimate of the covariance of $\boldsymbol{y}$. As noted by Efron and Tibshirani (1996), there are two such estimates. The Poisson regression fit regards the observations y_k as independent, estimating their covariance $\widehat{\mathrm{cov}}(\boldsymbol{y})$ as $\hat{\boldsymbol{V}} = \mathrm{diag}(\hat{\boldsymbol{y}})$, a $K \times K$ diagonal matrix with diagonal entries $\hat{y}_k$. On the other hand, recall that conditional on N the y_k are multinomial, for which a more appropriate estimate is $\hat{\boldsymbol{V}}_N = \mathrm{diag}(\hat{\boldsymbol{y}}) - \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}'/N$.

Proposition 1. *Let $\dot{\psi}(\hat{\boldsymbol{\eta}})$ and $\dot{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}})$ denote the derivatives of ψ and $\boldsymbol{\theta}$ with respect to $\boldsymbol{\eta}$ evaluated at $\hat{\boldsymbol{\eta}}$. The delta method covariance estimates of $\hat{\boldsymbol{\eta}}^+$ and $\hat{\boldsymbol{\theta}}^+$ are respectively*

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\eta}}^+) = \left(\boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{V}}_N\boldsymbol{W}\boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{X}\right)^{-1} \tag{13}$$

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\theta}}^+) = \hat{\boldsymbol{D}}\ \widehat{\mathrm{cov}}(\hat{\boldsymbol{\eta}}^+)\hat{\boldsymbol{D}}', \quad \hat{\boldsymbol{D}} = \begin{pmatrix} 1 & \dot{\psi}(\hat{\boldsymbol{\eta}})' \\ 0 & \dot{\boldsymbol{\theta}}(\hat{\boldsymbol{\eta}})' \end{pmatrix}. \tag{14}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

Proof. The score equation for the Poisson regression (8) including the external weigths $\boldsymbol{W}$ is

$$\boldsymbol{X}'\boldsymbol{W}\big[\boldsymbol{y} - \exp(\boldsymbol{X}\hat{\boldsymbol{\eta}}^+ + \boldsymbol{h})\big] = 0. \tag{15}$$

Differentiating with respect to $\boldsymbol{y}$ and replacing (9) gives that the rate of change of the MLE vector $\hat{\boldsymbol{\eta}}^+$ with respect to the count vector $\boldsymbol{y}$, considered as continuous, is

$$\frac{\partial \hat{\boldsymbol{\eta}}^+}{\partial \boldsymbol{y}'} = \big(\boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{W}. \tag{16}$$

Conditional on N, the covariance estimate of $\boldsymbol{y}$ is $\hat{\boldsymbol{V}}_N$. Thus the delta method covariance estimate of $\hat{\boldsymbol{\eta}}^+$ is $(\partial\hat{\boldsymbol{\eta}}^+/\partial\boldsymbol{y}')\hat{\boldsymbol{V}}_N(\partial\hat{\boldsymbol{\eta}}^+/\partial\boldsymbol{y}')'$, yielding (13).

The rate of change of $\boldsymbol{\theta}^+$ with respect to $\boldsymbol{\eta}^+$ is

$$\boldsymbol{D} = \frac{\partial\boldsymbol{\theta}^+}{\partial(\boldsymbol{\eta}^+)'} = \frac{\partial(\log p_0, \boldsymbol{\theta}(\boldsymbol{\eta})')'}{\partial(C, \boldsymbol{\eta}')} = \begin{pmatrix} 1 & \dot{\psi}(\boldsymbol{\eta})' \\ 0 & \dot{\boldsymbol{\theta}}(\boldsymbol{\eta})' \end{pmatrix}, \tag{17}$$

so the rate of change of $\hat{\boldsymbol{\theta}}^+$ with respect to $\hat{\boldsymbol{\eta}}^+$ at $\hat{\boldsymbol{\eta}}$ is $\hat{\boldsymbol{D}} = \boldsymbol{D}(\hat{\boldsymbol{\eta}})$. The delta method covariance estimate of $\hat{\boldsymbol{\theta}}^+$ is $(\partial\hat{\boldsymbol{\theta}}^+/\partial(\hat{\boldsymbol{\eta}}^+)')\ \widehat{\mathrm{cov}}(\hat{\boldsymbol{\eta}}^+)(\partial\hat{\boldsymbol{\theta}}^+/\partial(\hat{\boldsymbol{\eta}}^+)')'$, yielding (14). □

Proposition 2. *The delta method covariance estimate of the empirical null fits* (9) *and the empirical alternative component* (10) *are respectively* $\widehat{\mathrm{cov}}(\hat{\boldsymbol{y}}) = (\hat{\boldsymbol{V}}\boldsymbol{D}_y)\hat{\boldsymbol{V}}_N(\hat{\boldsymbol{V}}\boldsymbol{D}_y)'$ *and* $\widehat{\mathrm{cov}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = (I - \hat{\boldsymbol{V}}\boldsymbol{D}_y)\hat{\boldsymbol{V}}_N(I - \hat{\boldsymbol{V}}\boldsymbol{D}_y)'$, *where* $\boldsymbol{D}_y = \partial(\log\hat{\boldsymbol{y}})/\partial\boldsymbol{y}'$ *is given by*

$$\boldsymbol{D}_y = \boldsymbol{X}\big(\boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{W}. \tag{18}$$

Proof. The rate of change (18) of the vector $\log\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\eta}}^+$ with respect to $\boldsymbol{y}$ follows directly by (16). By the chain rule, $\partial\hat{\boldsymbol{y}}/\partial\boldsymbol{y}' = [\partial\hat{\boldsymbol{y}}/\partial(\log\boldsymbol{y})'][\partial(\log\boldsymbol{y})/\partial\boldsymbol{y}'] = \hat{\boldsymbol{V}}\boldsymbol{D}_y$, and similarly, $\partial(\boldsymbol{y} - \hat{\boldsymbol{y}})/\partial\boldsymbol{y}' = \boldsymbol{I} - \hat{\boldsymbol{V}}\boldsymbol{D}_y$. The result follows by the delta method. □

An empirical alternative to the delta method is is to use the bootstrap. Under the assumption that the test statistics are independent, resampling with replacement from $\{T_1, \ldots, T_N\}$ gives sets $\{T_1^*, \ldots, T_N^*\}$ which lead to parameter estimates $(\hat{\boldsymbol{\theta}}^+)^*$. Doing

Hosted by The Berkeley Electronic Press

this repeatedly, the bootstrap covariance estimate of $\hat{\boldsymbol{\theta}}^+$ is the empirical covariance of $(\hat{\boldsymbol{\theta}}^+)^*$. It is shown in Section 6 that, in the data example considered here, both the delta method and the bootstrap give very similar results.

3 FDR estimates and covariance

Mode matching is particularly convenient for FDR estimation, as FDR estimates follow immediately from the Poisson regression fits (9). Let $F(t) = p_0 F_0(t) + (1 - p_0)F_A(t)$ be the cumulative version of (1). Recall that the local FDR and (positive) right-tail FDR are given respectively by (Efron et al. 2001; Efron 2004)

$$\mathrm{fdr}(t) = \frac{p_0 f_0(t)}{f(t)}, \qquad \mathrm{Fdr}_R(t) = \frac{p_0\big(1 - F_0(t)\big)}{1 - F(t)} = \frac{\int_t^\infty \mathrm{fdr}(u) f(u)\, du}{\int_t^\infty f(u)\, du}. \tag{19}$$

Using (3) and (4), the local FDR at the bin centers t_k is estimated by

$$\widehat{\mathrm{fdr}}(t_k) = \frac{\hat{p}_0 \hat{f}_0(t_k)}{\hat{f}(t_k)} = \frac{\hat{\lambda}_k/(N\Delta)}{y_k/(N\Delta)} = \frac{\hat{y}_k}{y_k}, \tag{20}$$

defined whenever $y_k > 0$, or in vector form as

$$\log \widehat{\mathbf{fdr}} = \log \hat{\boldsymbol{y}} - \log \boldsymbol{y}, \tag{21}$$

where $\hat{\boldsymbol{y}}$ is given by (9). In contrast to Efron (2007b), we estimate Fdr_R at the bin centers t_k based on the right side of (19) by

$$\widehat{\mathrm{Fdr}}_R(t_k) = \frac{\frac{1}{2}\widehat{\mathrm{fdr}}(t_k)\hat{f}(t_k) + \sum_{j=k+1}^K \widehat{\mathrm{fdr}}(t_j)\hat{f}(t_j)}{\frac{1}{2}\hat{f}(t_k) + \sum_{j=k+1}^K \hat{f}(t_j)} = \frac{\frac{1}{2}\hat{y}_k + \sum_{j=k+1}^K \hat{y}_j}{\frac{1}{2} y_k + \sum_{j=k+1}^K y_j} \tag{22}$$

where (3) and (20) were used. This can be written in vector form as

$$\log \widehat{\mathbf{Fdr}}_R = \log(\boldsymbol{S}\hat{\boldsymbol{y}}) - \log(\boldsymbol{S}\boldsymbol{y}) \tag{23}$$

where $\boldsymbol{S}$ is an upper triangular matrix with entries $1/2$ on the diagonal and 1 above the diagonal. The estimate (23) is easy to analyze theoretically in terms of bias (see

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

Section 4). For the left tail FDR, definition (19) is changed to $\mathrm{Fdr}_L(t) = p_0F_0(t)/F(t)$ and is estimated similarly by

$$\log \widehat{\mathbf{Fdr}}_L = \log(\boldsymbol{S}'\hat{\boldsymbol{y}}) - \log(\boldsymbol{S}'\boldsymbol{y}) \tag{24}$$

where $\boldsymbol{S}'$ is the transpose of $\boldsymbol{S}$, a lower triangular matrix with entries 1/2 on the diagonal and 1 below the diagonal.

Proposition 3.

a) The delta method covariance estimate of the local FDR (21) *is* $\widehat{\mathrm{cov}}(\log \widehat{\mathbf{fdr}}) = \boldsymbol{A}\hat{\boldsymbol{V}}_N\boldsymbol{A}'$, *where* $\boldsymbol{A} = \partial(\log \widehat{\mathbf{fdr}})/\partial \boldsymbol{y}' = \boldsymbol{D}_y - \boldsymbol{V}^{-1}$, $\boldsymbol{V} = \mathrm{diag}(\boldsymbol{y})$ *and* $\boldsymbol{D}_y$ *is given by* (18).

b) The delta method covariance estimate of the right tail FDR (23) *is* $\widehat{\mathrm{cov}}(\log \widehat{\mathbf{Fdr}}_R) = \boldsymbol{B}\hat{\boldsymbol{V}}_N\boldsymbol{B}'$, *where* $\boldsymbol{B} = \partial(\log \widehat{\mathbf{Fdr}}_R)/\partial \boldsymbol{y}' = \hat{\boldsymbol{U}}^{-1}\boldsymbol{S}\hat{\boldsymbol{V}}\boldsymbol{D}_y - \boldsymbol{U}^{-1}$ *and* $\boldsymbol{U} = \mathrm{diag}(\boldsymbol{S}\boldsymbol{y})$, $\hat{\boldsymbol{U}} = \mathrm{diag}(\boldsymbol{S}\hat{\boldsymbol{y}})$. *The formula for the left tail FDR* (24) *has the same form with* $\boldsymbol{S}$ *replaced by* $\boldsymbol{S}'$.

Proof.

a) Follows immediately by the delta method and the definition of $\boldsymbol{A}$.

b) To evaluate the rate of change of the vector $\log(\boldsymbol{S}\hat{\boldsymbol{y}})$ with respect to $\boldsymbol{y}$, compute

$$\frac{\partial(\log(\boldsymbol{S}\hat{\boldsymbol{y}}))_k}{\partial y_l} = \frac{\partial}{\partial y_l}\log\left(\frac{1}{2}\hat{y}_k + \sum_{j=k+1}^{K}\hat{y}_j\right) = \frac{\frac{1}{2}\frac{\partial \hat{y}_k}{\partial y_l} + \sum_{j=k+1}^{K}\frac{\partial \hat{y}_j}{\partial y_l}}{\frac{1}{2}\hat{y}_k + \sum_{j=k+1}^{K}\hat{y}_j}$$

Thus

$$\frac{\partial(\log(\boldsymbol{S}\hat{\boldsymbol{y}}))}{\partial \boldsymbol{y}'} = \hat{\boldsymbol{U}}^{-1}\boldsymbol{S}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}'} = \hat{\boldsymbol{U}}^{-1}\boldsymbol{S}\cdot\hat{\boldsymbol{V}}\frac{\partial(\log \hat{\boldsymbol{y}})}{\partial \boldsymbol{y}'} = \hat{\boldsymbol{U}}^{-1}\boldsymbol{S}\hat{\boldsymbol{V}}\boldsymbol{D}_y$$

where we have used the fact that $\partial(\log \hat{y}_k)/\partial y_l = (1/\hat{y}_k)\partial \hat{y}_k/\partial y_l$. The result now follows by the delta method and the definition of $\boldsymbol{B}$. □

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

4 Tuning parameters and bias

Mode matching is controlled by two tuning parameters, the bin width Δ and the fitting interval S_0. Each is connected to a different source of bias: the bin width Δ controls the bias incurred by using a first order approximation in (2); the fitting interval S_0 controls the bias incurred by the inclusion of the alternative component $(1-p_0)f_A$ in the fit of the empirical null. In what follows, we refer to the following two simulation scenarios of model (1):

$$T_i \overset{\text{ind}}{\sim} \begin{cases} f_0 = N(0.2, 1.2^2), & \text{probability } p_0 \\ f_A = N(3, 1.2^2), & \text{probability } 1-p_0 \end{cases} \tag{25}$$

$$T_i \overset{\text{ind}}{\sim} \begin{cases} f_0 = 0.8\chi^2(3), & \text{probability } p_0 \\ f_A = \text{noncentral } \chi^2(3, \delta = 3), & \text{probability } 1-p_0 \end{cases} \tag{26}$$

where $\delta = 3$ denotes the noncentrality parameter. The fitting interval is set to $S_0 = [0.2 - t_0, 0.2 + t_0]$ in the normal case and $S_0 = [0, t_0]$ in the χ^2 case, so that in both cases S_0 is tuned by the single number t_0.

4.1 The bin width

The first and smallest source of bias is the use of a first order approximation in (2). Under the zero assumption, we can approximate the error by the next expansion term

$$\pi_k - f_0(t_k)\Delta \approx \frac{f_0''(t_k)}{24}\Delta^3, \qquad t_k \in S_0. \tag{27}$$

As in nonparametric density estimation, bias is reduced by thinning the bins. However, the size of the bias depends also on the curvature of the empirical null. If f_0 is $N(\mu, \sigma^2)$, the largest curvature occurs at the mode μ, where $f_0''(\mu) \approx -0.4/\sigma^3$. Thus the error (27) is bounded by $0.017\Delta^3/\sigma^3$. This is about 1.3×10^{-4} if $\Delta = 0.2\sigma$. If f_0 is $a\chi^2(\nu)$, the curvature depends strongly on ν. For $\nu < 6$ except $\nu = 2$, the curvature

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

is unbounded at $t = 0$, but it decreases rapidly as t increases away from 0. For $\nu \geq 2$, more important is the curvature at the mode, where most of the probability mass lays. This curvature decreases rapidly as ν increases away from 2. For example, for $\nu = 3$, the curvature at the mode $\nu - 2$ is $f_0''(\nu - 2) \approx -0.12/a^3$ so the error (27) is bounded by $0.005\Delta^3/a^3$. This is about 4×10^{-5} if $\Delta = 0.2a$.

The effect of Δ is illustrated in Figure 1. The plotted empirical null estimates are averages over 100 simulated instances of models (25) and (26) with $p_0 = 1, N = 10000$, and fixed $t_0 = 1$ in the normal case and $t_0 = 4$ in the χ^2 case. Both the bias and the variance, although small, increase with Δ within the plotted range. In contrast to nonparametric density estimation, the variance is remarkably insensitive to Δ. A large Δ implies large counts within each bin, reducing variance. But for fixed S_0, a small Δ implies a large number of bins $K = |S_0|/\Delta$ and thus a large number of design points for the Poisson regression, which also reduces variance. Roughly, the variance of $\hat{\pi}_k$ is inversely proportional to $N\Delta$, so a large N implies that Δ can be afforded to be very small, reducing bias at the same time. On the other hand, a large number of bins K is computationally expensive as it implies inverting large matrices in the fitting of the empirical null. Based on Figure 1 and computational considerations, $\Delta = 0.1$ seems a reasonable choice for the normal and $\chi^2(\nu)$ with $\nu \geq 2$. If $\nu < 2$ the curvature near $t = 0$ demands much smaller values of Δ to avoid substantial bias. For large ν, the increase in the effective support of the density may require increasing Δ in order to reduce computations.

4.2 The fitting interval

The largest source of bias in the estimation of the null density is the inclusion of the alternative component $(1 - p_0)f_A$ within S_0. The asymptotic bias as a result of the alternative density is quantified in the following proposition.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Proposition 4. *Define the vectors* $\boldsymbol{f}_0 = f_0(\boldsymbol{t})$ *and* $\boldsymbol{f}_A = f_A(\boldsymbol{t})$. *For large* N, *the respective biases in the estimation of the canonical parameters* $\boldsymbol{\eta}^+$ *and the natural parameters* $\boldsymbol{\theta}^+$ *are given approximately by*

$$\hat{\boldsymbol{\eta}}^+_\infty - \boldsymbol{\eta}^+ \approx (1-p_0)\big(\boldsymbol{X}'\boldsymbol{W}\mathrm{diag}(\boldsymbol{f}_0)\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{W}\big(\boldsymbol{f}_A - \boldsymbol{f}_0\big) - (\log(p_0), \boldsymbol{0}')' \qquad (28)$$

and $\hat{\boldsymbol{\theta}}^+_\infty - \boldsymbol{\theta}^+ \approx \boldsymbol{D}(\hat{\boldsymbol{\eta}}^+_\infty - \boldsymbol{\eta}^+)$.

Proof. Dividing the score equation (15) by $\gamma\Delta$ and applying the law of large numbers as $\gamma \to \infty$ gives that $\hat{\boldsymbol{\eta}}^+$ converges to the solution $\hat{\boldsymbol{\eta}}^+_\infty$ of the equation

$$\boldsymbol{X}'\boldsymbol{W}\big[p_0\boldsymbol{f}_0 + (1-p_0)\boldsymbol{f}_A - \mathrm{diag}(g_0(\boldsymbol{t}))\exp(\boldsymbol{X}\hat{\boldsymbol{\eta}}^+_\infty)\big] = 0. \qquad (29)$$

In particular, if $p_0 = 1$, we have that $\hat{\boldsymbol{\eta}}^+$ is asymptoticaly unbiased, i.e. $\hat{\boldsymbol{\eta}}^+_\infty = (0, \boldsymbol{\eta}')'$. The idea is to find a first order expansion of $\hat{\boldsymbol{\eta}}^+$ near $p_0 = 1$. Differentiating (29) with respect to p_0, we obtain that, at $p_0 = 1$, $d\hat{\boldsymbol{\eta}}^+_\infty/dp_0 = \big(\boldsymbol{X}'\boldsymbol{W}\mathrm{diag}(\boldsymbol{f}_0)\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{W}\big(\boldsymbol{f}_0 - \boldsymbol{f}_A\big)$. The bias in the estimation of $\boldsymbol{\eta}^+$ is approximately

$$\hat{\boldsymbol{\eta}}^+_\infty - \boldsymbol{\eta}^+ = \hat{\boldsymbol{\eta}}^+_\infty - (0, \boldsymbol{\eta}')' - (\log(p_0), \boldsymbol{0}')' \approx \left.\frac{d\hat{\boldsymbol{\eta}}^+_\infty}{dp_0}\right|_{p0=1}(p_0 - 1) - (\log(p_0), \boldsymbol{0}')'$$

yielding (28). Similarly, the bias in the estimation of $\boldsymbol{\theta}^+$ is a first order expansion of $\boldsymbol{\theta}^+$ with respect to $\boldsymbol{\eta}^+$ near $p_0 = 1$. □

Roughly, the asymptotic bias (28) is proportional to the likelihood ratio between f_A and f_0 within S_0, but modified by the Poisson regression. For fixed Δ, the fitting interval controls the number of columns of $\boldsymbol{X}$. Notice that setting $\boldsymbol{f}_A = \boldsymbol{f}_0$ results in a bias of $(\log(p_0), \boldsymbol{0}')'$, reflecting the fact that $\mathrm{E}(\log(\hat{p}_0)) = 1$ in this case. The accuracy of the approximation (28) is shown in Table 1.

Figure 2 shows the empirical null parameter estimates averaged over 100 instances of the simulations (25) and (26) with $p_0 = 0.9$ and $N = 10000$. The simulations were repeated for varying t_0 and fixed $\Delta = 0.1$. Increasing t_0 increases the bias due to

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

the inclusion of the alternative component. On the other hand, increasing t_0 also increases the number of design design points for the Poisson regression, reducing variance. The bias is worse in the χ^2 simulation because the null and the alternative densities overlap more than in the normal simulation, even though they have a similar separation in their mean. All parameters except the d.f. ν of the χ^2 tend to be biased upwards. This implies that the empirical null is conservative, predicting a smaller contribution of the alternative density in the mixture than there is.

Guidelines for choosing t_0 are suggested by the MSE plots. Notice that the optimal t_0 is not the same for all the parameters. In the normal simulation, the optimal t_0 is in the range $1.4 \sim 1.9$, corresponding to about $1.2 \sim 1.6$ standard deviations of the true null $N(0.2, 1.2^2)$. In the χ^2 simulation, the optimal t_0 is in the range $3.2 \sim 6$, corresponding to the $74 \sim 94$ percentiles of the true null $0.8\chi^2(3)$.

4.3 Bias in FDR estimation

One issue overlooked by Efron (2007b) is that the local FDR estimate (20) is biased by definition. The bias is given by the following proposition.

Proposition 5.

a) If p_0 and f_0 are known then $\mathrm{E}\big[\widehat{\mathrm{fdr}}_k \big| y_k > 0\big] = \mathrm{fdr}_k\ \zeta(\lambda_k)$, *where*

$$\zeta(\lambda) = \frac{\lambda}{e^\lambda - 1} \int_0^\lambda \frac{e^u - 1}{u}\, du. \tag{30}$$

b) If p_0 and f_0 are estimated by mode matching then, for large N and $t_k \notin S_0$, $\mathrm{E}\big[\widehat{\mathrm{fdr}}_k \big| y_k > 0\big] \approx \mathrm{fdr}_k\ b_k\ \zeta(\lambda_k)$, *where b_k is the k-th entry of the asymptotic bias vector $\boldsymbol{b}_\infty = \exp\big(\boldsymbol{X}(\hat{\boldsymbol{\eta}}^+_\infty - \boldsymbol{\eta}^+)\big)$ with the inner parenthesis given by* (28).

Proof.

a) For known p_0 and f_0, (20) says $\widehat{\mathrm{fdr}}_k = \lambda_{k,0}/y_k$, where $y_k \sim Po(\lambda_k)$, $\lambda_k = \gamma \Delta f(t_k)$

COBRA BEPRESS REPOSITORY Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

and $\lambda_{k,0} = \gamma \Delta p_0 f_0(t_k)$. Thus

$$\mathrm{E}\big[\widehat{\mathrm{fdr}}_k \big| y_k > 0\big] = \mathrm{E}\left(\frac{\lambda_{k,0}}{y_k} \middle| y_k > 0 \right) = \frac{\lambda_{k,0}}{\lambda_k} \mathrm{E}\left(\frac{\lambda_k}{y_k} \middle| y_k > 0 \right) = \mathrm{fdr}_k \cdot \zeta(\lambda_k)$$

where $\zeta(\lambda)$ is defined for a generic $y \sim Po(\lambda)$ as $\zeta(\lambda) = \mathrm{E}(\lambda/y \big| y > 0)$. By direct evaluation,

$$\zeta(\lambda) = \frac{1}{1-e^{-\lambda}} \sum_{j=1}^{\infty} \frac{\lambda}{j} \, \frac{e^{-\lambda}\lambda^j}{j!} = \frac{\lambda}{e^{\lambda}-1} \int_0^{\lambda} \sum_{j=1}^{\infty} \frac{u^{j-1}}{j!} \, du = \frac{\lambda}{e^{\lambda}-1} \int_0^{\lambda} \frac{du}{u} \sum_{j=1}^{\infty} \frac{u^j}{j!}$$

which is equal to (30).

b) When the empirical null is used, the local FDR estimate (20) is $\widehat{\mathrm{fdr}}_k = \hat{y}_k / y_k$, where $\hat{y}_k = N \Delta \hat{p}_0 \hat{f}_0(t_k)$. Notice that when evaluated at $t_k \notin S_0$, the numerator $\hat{y}_k$ is independent of the denominator y_k. Thus

$$\mathrm{E}\big[\widehat{\mathrm{fdr}}_k \big| y_k > 0\big] = \mathrm{E}\left(\frac{\hat{y}_k}{y_k} \middle| y_k > 0 \right) = \frac{\lambda_{k,0}}{\lambda_k} \frac{\mathrm{E}(\hat{y}_k)}{\lambda_{k,0}} \mathrm{E}\left(\frac{\lambda_k}{y_k} \middle| y_k > 0 \right) \approx \mathrm{fdr}_k \; b_k \; \zeta(\lambda_k)$$

where the vector $\boldsymbol{b}_{\infty}$ is obtained as follows. By (8) and (9), $\boldsymbol{\lambda}_0 = \exp(\boldsymbol{X}\boldsymbol{\eta}^+ + \boldsymbol{h})$ and $\mathrm{E}(\hat{\boldsymbol{y}}) = \mathrm{E}[\exp(\boldsymbol{X}\hat{\boldsymbol{\eta}}^+ + \boldsymbol{h})] \approx \exp(\boldsymbol{X}\hat{\boldsymbol{\eta}}^+_{\infty} + \boldsymbol{h})$ for large N. Dividing entry by entry gives $\boldsymbol{b}_{\infty} = \exp\big(\boldsymbol{X}(\hat{\boldsymbol{\eta}}^+_{\infty} - \boldsymbol{\eta}^+)\big)$. □

Figure 3a shows a plot of the function $\zeta(\lambda)$, closely related to the so-called exponential integral (Abramowitz and Stegun 1972, Ch.5). Since λ_k increases with γ (i.e. N), most bins fall on the right end of the plot, making the local fdr estimate in most bins slightly conservatively biased above the correct value. However λ_k is aways small in the far tails of the null density f_0, making the FDR estimate biased down. This can be misleading when the true FDR is close to 1. Figure 3b shows the local FDR estimates for 100 instances of the normal simulation (25) with known $p_0 = 0.9$ and $N = 10000$. Because the alternative density sits on the right side of the plot, the true local FDR at the left tail is 1. The local FDR estimate, however, follows the graph of $\zeta(\lambda)$, as the bin counts get smaller towards the left, with the variance proportional to that graph. Any particular realization of the FDR curve here may

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

give the impression that there is something to discover at the left tail, while in reality there is not. The zoom-in in panel c shows that the phenomenon still occurs in the right tail, as the average FDR estimate is first biased up with higher variance and then dips below the truth as the bin counts get very small. This is not noticeable in Panel b because, by Propostion 5, the bias is proportional to the true FDR, which is low in this region. When the empirical null is used, the additional bias, captured by the factor b_k in Proposition 5b is visible in Panel d. The variance is also higher, as expected. Notice that the bias is up, so the FDR estimates are conservative.

The tail FDR also suffers from a similar bias phenomenon, being sensitive to small cumulative bin counts in the far tails. Following a similar argument as in Proposition 5a, when p_0 and f_0 are known, (22) says $\widehat{\text{Fdr}}_k = (\boldsymbol{S}\boldsymbol{\lambda}_0)_k/(\boldsymbol{S}\boldsymbol{y})_k$, where the cumulative denominator $y_k/2 + \sum_{l=k+1}^{K} y_l$ behaves similarly to a Poisson random variable with mean $(\boldsymbol{S}\boldsymbol{\lambda})_k = \lambda_k/2 + \sum_{l=k+1}^{K} \lambda_l$. Therefore $\text{E}[\widehat{\text{Fdr}}_k \big| y_k > 0] = \text{Fdr}_k \cdot \text{E}[(\boldsymbol{S}\boldsymbol{\lambda})_k/(\boldsymbol{S}\boldsymbol{y})_k \big| (\boldsymbol{S}\boldsymbol{y})_k > 0]$, where the conditional expectation behaves approximately like $\zeta((\boldsymbol{S}\boldsymbol{\lambda})_k)$. A simulation using $p_0 = 1$ (not shown) gives FDR curves that are very similar to those on the left end of panels b and d. In Figure 3e (zoom-in not shown), the bias is visible but small because the FDR itself is low in the right tail. Panel f shows the increase in bias and variance when the empirical null is used.

The bias phenomenon does not contradict the the results of Storey et al. (2004), which claim asymptotic unbiasedness of the tail FDR estimator. Consider a fixed bin k. As N increases, the expected bin count λ_k increases and the operating point in Figure 3a moves the right, making the FDR estimate asymptotically unbiased. The $\zeta(\lambda)$ phenomenon appeals to practical cases where N is large but finite, so that the bin counts at the tails are still small.

The $\zeta(\lambda)$ phenomenon stems in essence from the fact that the FDR denominator y_k may be equal to zero in some bins at the far tails. Efron (2004, 2007b) avoids

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

this problem by smoothing the histogram with a spline fit. This helps because the compunding of data at each bin again pushes the operating point in the $\zeta(\lambda)$ graph (Figure 3a) to the right. However, smoothing introduces a bias of its own. Simulations using smoothing show that the resulting estimates at the far tails, especially when $p_0 = 1$, are not reliable, as they are very sensitive to the choice of knots or smoothing bandwidth.

5 The effect of correlation

In linear regression problems with mildly correlated errors, the least squares solution ceases to maximize the likelihood but may still give a reasonable fit. The main effect of the correlation is then to inflate the variance of the least square estimates. Similarly for mode matching, if the bin counts are mildly correlated, the most important change is in the variance estimates. Specifically, in Propositions 1, 2 and 3, the multinomial covariance estimate $\widehat{\mathrm{cov}}(\boldsymbol{y}) = \hat{\boldsymbol{V}}_N = \mathrm{diag}(\hat{\boldsymbol{y}}) - \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}'/N$ gets replaced by the correlated multinomial covariance estimate $\hat{\boldsymbol{V}}_N^*$ given by the following proposition.

Proposition 6. *Let f_0 be one of the Lancaster distributions, i.e. the exponential families normal, gamma, Poisson or negative binomial (Koudou 1998). Assume that under the complete null every pair of test statistics (T_i, T_j) has a bivariate density $f_0(t_i, t_j; \rho_{ij})$ with marginals $f_0(t)$ and $\mathrm{corr}(T_i, T_j) = \rho_{ij}$. Let $E(\rho^n)$, $n = 1, 2, \ldots$ denote the empirical moments of the $N(N-1)$ correlations ρ_{ij}, $i < j$. Then the estimate of $\mathrm{cov}(\boldsymbol{y})$ is*

$$\hat{\boldsymbol{V}}_N^* = \hat{\boldsymbol{V}}_N + (1 - 1/N)\ \mathrm{diag}(\hat{\boldsymbol{y}})\ \boldsymbol{\delta}\ \mathrm{diag}(\hat{\boldsymbol{y}}) \tag{31}$$

where $\boldsymbol{\delta}$ is a $K \times K$ matrix with entries

$$\delta_{kl} = \sum_{n=1}^{\infty} \frac{E(\rho^n)}{n!} L_n(t_k) L_n(t_l) \tag{32}$$

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

and $L_n(t)$ are the Lancaster orthogonal polynomials with respect to f_0: Hermite if f_0 is normal, generalized Laguerre if f_0 is gamma, Charlier if f_0 is Poisson, and normalized Meixner if f_0 is negative binomial.

Proof. The form of expression (31) follows from Efron (2007a, Lemma 1). A similar argument as in Efron (2007a, Lemma 2) gives that

$$\delta_{kl} \approx \int_{-1}^{1} R_{kl}(\rho)\, dG(\rho), \qquad R_{kl}(\rho) = \frac{f_0(t_k, t_l; \rho)}{f_0(t_k) f_0(t_l)} \tag{33}$$

where $G(\rho)$ denotes the empirical distribution of the $N(N-1)$ pairwise correlations ρ_{ij}, $i < j$. If f_0 is a Lancaster distribution, then $R_{kl}(\rho)$ admits the expansion $R_{kl}(\rho) = \sum_{n=1}^{\infty} \rho^n L_n(t_j) L_n(t_k)/n!$, where the polynomials $L_n(t_j)$ are orthognoal with respect to f_0 (Koudou 1998). Plugging back into (33) gives (32). □

Proposition 6 generalizes the result of Efron (2007a, Theorem 1). When $f_0(t)$ is $N(\mu, \sigma^2)$ and $f_0(t_i, t_j; \rho)$ is the corresponding bivariate normal with correlation ρ then

$$R_{kl}(\rho) = \sum_{n=1}^{\infty} \frac{\rho^n}{n!} H_n\left(\frac{t_j - \mu}{\sigma}\right) H_n\left(\frac{t_l - \mu}{\sigma}\right)$$

where $H_n(t)$ are the Hermite polynomials. In particular, truncating the series at $n = 2$ and setting $\mu = 0$, $\sigma = 1$, gives precisely the expansion in Efron (2007a, Lemma 3), with $H_2(t) = t^2 - 1$ being the "wing function" described there. When $f_0(t)$ is the $a\chi^2(\nu)$ density (11) then

$$R_{kl}(\rho) = \sum_{n=1}^{\infty} \frac{\rho^n}{n!} \frac{\Gamma(\nu/2)}{\Gamma(\nu/2 + n)} L_n\left(\frac{t_j}{2a}\right) L_n\left(\frac{t_k}{2a}\right)$$

where $L_n(t)$ are the generalized Laguerre polynomials of degree $\nu/2 - 1$: $L_0(t) = 1$, $L_1(t) = -t + \nu/2$, $L_2(t) = t^2 - 2(\nu/2 + 1)t + (\nu/2)(\nu/2 + 1)$, and so on.

Apart from Proposition 6, another way to take correlation into account is to include an overdispersion parameter in the Poisson regression. The overdispersion parameter ϕ is estimated by the quasi-likelihood MLE $\hat{\phi} = (1/K) \sum_{k=1}^{K} (y_k - \hat{\lambda}_k)/\hat{\lambda}_k$. The fit of the Poisson regression is the same as before but the covariance estimates are inflated by a factor ϕ.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

6 SNP data example

Large scale χ^2 testing occurs frequently in the analysis of SNP data. The following example is a family-based study of genome-wide association between genetic variants and obesity based on the Framingham Heart Study (FHS) (Herbert et al. 2006). To obtain genetic markers, 1400 probands from the family-plates were genotyped on an Affymetrix 100K SNP-chip containing 116,204 SNPs. The selected phenotype of interest was body-mass index (BMI) at exams 1, 2, 3 and 4. Each SNP was tested for association with all 4 BMI measurements using the multivariate FBAT-GEE statistic (Lange et al. 2003). Excluding SNPs for which the number of informative families was less than 20, a total of $N = 95,810$ test statistics were generated. The theoretical null for this analysis is $\chi^2(4)$.

Figure 4 shows a histogram of the test statistics using bins of width $\Delta = 0.1$ starting from zero. The mismatch between the histogram and the theoretical null can be seen in the zoom-in inlet plot. The empirical null, which fits the data better, was obtained using χ^2 mode matching (Appendix A.2). The fitting interval was defined as $S_0 = [0, 20]$, wide enough to use most of the data without including the far tail region $t > 20$ where the discoveries are likely to be made (see description of Figure 5 below). The estimated natural parameters $\boldsymbol{\theta}^+$ are listed in Table 2. The associated standard errors (SE) were computed both as the square root of the diagonal of (14) and using the bootstrap. The delta-method SE estimates are only slightly smaller than the bootstrap SEs.

The CIs for a and ν do not include the theoretical values 1 and 4, indicating a significant departure from the theoretical null in both scaling and degrees of freedom. The CI for $\log(p_0)$ includes 0. This does not prove that there are no significant SNPs, but it shows that the study may not have enough power to discover them. The lower bound of the CI for $\log(p_0)$ suggests that the fraction of non-null SNPs may be as

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

high as $1-\hat{p}_0 = 3.18\times10^{-5} \approx -\log(\hat{p}_0)$, which is about 3 SNPs out of $N = 95,810$. If instead of fitting the full empirical null, p_0 is estimated alone believing the theoretical null, the result is $\log(\hat{p}_0) = 1.24\times10^{-4}$ with standard CI $[1.96\times10^{-5}, 2.28\times10^{-4}]$. The theoretical null does not admit an estimate of p_0 that is less than 1, again indicating that the theoretical null is unsuitable for this data.

The FDR analysis is summarized in Figure 5. Here we focus on the local FDR, which in this case is more powerful than the tail FDR (and has been proven to be more powerful in general (Sun and Cai 2007)). The local FDR estimates are compared to the expected local FDR under the complete null. In agreement with Proposition 5, the expected local FDR estimate goes down at the far tails where the number of counts is small. Here the expected local FDR was estimated replacing λ_k for $\hat{\lambda}_k$ in Proposition 5 and setting $\text{fdr}_k = 1$ and $b_k = 1$ as specified by the complete null. Three bins stand out with local FDR estimates significantly below the dashed line. The results for these bins are reported in Table 3. While the observed local FDR values are relatively low, they turn out to be not as low when their bias is taken into account. The adjusted local FDR values in Table 3 are not precise at correcting the bias, but they hint that about 50% of the 8 SNPs contained in these 3 bins might be associated with obesity. This is consistent with the estimate above that the non-null distribution may contain about 3 SNPs.

The methods of Section 5 were not implemented in this analysis because pairwise correlations are hard to estimate for these data. Given the widespread locations of the SNPs on the genome, the test statistics are not expected to be highly correlated (Herbert et al. 2006). This is consistent with the estimated overdispersion $\hat{\phi} = 1.090$ (bootsrap SE = 0.0655) whose standard CI contains the value $\phi = 1$ corresponding to independent statistics.

COBRA
BIOSTATISTICS
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

7 Summary and Discussion

In this article, we have extended the central matching method for estimating the null density in large-scale multiple testing to a mode matching method, applicable when the theoretical null belongs to any exponential family or a related distribution such as t or F. The empirical null estimate is accompanied by an estimate of p_0, the proportion of true null tests in the data. Further, the empirical null estimates can be used directly to estimate local and tail FDR curves for FDR inference. Delta method covariance estimates and bias formulas have been derived. We have seen that FDR estimates are biased down at the far tails and should be taken cautiously whenever the corresponding observed bin counts are small. The effect of correlation has been taken into account by a modification of the covariance estimates that generalizes Efron's "wing function".

Efron (2005, 2007b) discusses several reasons why the empirical null may not match exactly the theoretical null in observational studies. It should be emphasized that mode matching does not necessarily increase power with respect to the theoretical null (Efron (2004) provides counterexamples). Instead, the empirical null answers a question of model validity.

In the normal case, another empirical null method called MLE fitting (Efron 2007b) has been reported to give similar empirical null estimates with slightly lower variance. Mode matching is easier to analyze and is appealing because of its application to exponential families. It is also easier to implement in practice because of available software and computational efficiency. On the other hand, we forsee that, just like mode matching, MLE fitting could be extended to other distributions beyond the normal too.

At least two aspects of mode matching may benefit from further study beyond this paper. One aspect is the possibility of choosing data-dependent limits for the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

fitting interval S_0. Fixed limits based on the theoretical null may be inappropriate precisely because the empirical null is expected to be displaced or scaled with respect the theoretical null. In an analysis of χ^2-scores, Schwartzman et al. (2008) used as upper limit the 90th percentile of the test statistic distribution. In the simulations scenarios of Section 4.2, the 80th percentile may be more appropriate. In the SNP data example above, a bootstrap analysis showed that setting the upper limit to the 99.95th percentile of the test statistic distribution rather than the 99.95th percentile of the theoretical $\chi^2(4)$ density (the value 20 used previously), results in empirical null variance estimates that are about 50% higher than those obtained when the limit was fixed. This suggests that the cost of a data-dependent limit might not be too high. Unfortunately, as shown in Section 4 above, the choice of the limit depends very much on the alternative distribution, which is unknown.

Another aspect is the possibility of using the two-step approach of Efron (2007b) of estimating the mixture density nonparametrically before fitting the empirical null by mode matching. As noted above, the most crucial issue is the bias in the tails of the density. Future exploration may yield an answer to what is the best way to estimate the mixture density for mode matching. For example, different bin widths Δ could be used inside and outside the fitting interval S_0 since they serve different purposes. Inside S_0 one could optimize Δ for empirical null estimation, while outside S_0 one could optimize Δ for FDR estimation.

Matlab functions implementing the methods described in this paper are available at `http://biowww.dfci.harvard.edu/~armin/software.html`.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

A Appendix: Special cases

A.1 Normal family

The empirical null $N(\mu, \sigma^2)$ with $\boldsymbol{\theta} = (\mu, \sigma^2)$ has exponential family form

$$x(t) = (t, t^2)' \qquad \psi(\boldsymbol{\eta}) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$$
$$\boldsymbol{\eta} = (\eta_1, \eta_2)' = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)' \qquad g_0(t) = \frac{1}{\sqrt{2\pi}}$$

The Poisson regression (8) using t and t^2 as predictors and $\log(N\Delta/\sqrt{2\pi})$ as offset gives estimates $\hat{\eta}_1$, $\hat{\eta}_2$ and $\hat{C}$. From these we obtain

$$\hat{\mu} = -\frac{\hat{\eta}_1}{2\hat{\eta}_2}, \qquad \hat{\sigma}^2 = -\frac{1}{2\hat{\eta}_2}, \qquad \log \hat{p}_0 = \hat{C} + \psi(\hat{\boldsymbol{\eta}}).$$

The derivative (17) required for computing the covariance (14) of $\hat{\boldsymbol{\theta}}^+ = (\log \hat{p}_0, \hat{\mu}, \hat{\sigma}^2)'$ is

$$\hat{\boldsymbol{D}} = \frac{\partial \hat{\boldsymbol{\theta}}^+}{\partial (\hat{\boldsymbol{\eta}}^+)'} = \begin{pmatrix} 1 & -\frac{\hat{\eta}_1}{2\hat{\eta}_2} & \frac{\hat{\eta}_1^2}{4\hat{\eta}_2^2} - \frac{1}{2\hat{\eta}_2} \\ 0 & -\frac{1}{2\hat{\eta}_2} & \frac{\hat{\eta}_1}{2\hat{\eta}_2^2} \\ 0 & 0 & \frac{1}{2\hat{\eta}_2^2} \end{pmatrix} = \begin{pmatrix} 1 & \hat{\mu} & \hat{\mu}^2 + \hat{\sigma}^2 \\ 0 & \hat{\sigma}^2 & 2\hat{\mu}\hat{\sigma}^2 \\ 0 & 0 & 2\hat{\sigma}^4 \end{pmatrix}.$$

The normal family $N(\mu, \sigma^2)$ lends itself to two exponential subfamilies.

A.1.1 Estimate $\theta = \mu$

The empirical null $N(\mu, \sigma_0^2)$ with fixed σ_0^2 has the exponential family form $x(t) = t$, $\eta = \mu/\sigma_0^2$, $\psi(\eta) = \sigma_0^2\eta^2/2$ and $g_0(t) = e^{-t^2/(2\sigma_0^2)}/\sqrt{2\pi\sigma_0^2}$. Poisson regression using t as predictor and $\log(N\Delta g_0(t))$ as offset gives estimates $\hat{\eta}$ and $\hat{C}$. From these we obtain

$$\hat{\mu} = \sigma_0^2\hat{\eta}, \qquad \log \hat{p}_0 = \hat{C} + \frac{\sigma_0^2\hat{\eta}^2}{2}, \qquad \hat{\boldsymbol{D}} = \begin{pmatrix} 1 & \hat{\mu} \\ 0 & \sigma_0^2 \end{pmatrix}.$$

A.1.2 Estimate $\theta = \sigma^2$

The empirical null $N(\mu_0, \sigma^2)$ with fixed μ_0 has the exponential family form $x(t) = (t - \mu_0)^2$, $\eta = -1/(2\sigma^2)$, $\psi(\eta) = (-1/2)\log(-2\eta)$ and $g_0(t) = 1/\sqrt{2\pi}$. Poisson

http://biostats.bepress.com/harvardbiostat/paper77

regression using $(t-\mu_0)^2$ as predictor and $\log(N\Delta/\sqrt{2\pi})$ as offset gives estimates $\hat{\eta}$ and $\hat{C}$. From these we obtain

$$\hat{\sigma}^2 = -\frac{1}{2\hat{\eta}}, \qquad \log \hat{p}_0 = \hat{C} - \frac{1}{2}\log(-2\hat{\eta}), \qquad \hat{\boldsymbol{D}} = \begin{pmatrix} 1 & \hat{\sigma}^2 \\ 0 & 2\hat{\sigma}^4 \end{pmatrix}.$$

A.2 Scaled χ^2 family (Gamma)

The empirical null $a\chi^2(\nu)$ (11) with $\boldsymbol{\theta} = (a, \nu)'$ has exponential family form

$$x(t) = (t, \log t)' \qquad \psi(\boldsymbol{\eta}) = \log\left(\frac{\Gamma(\eta_2+1)}{(-\eta_1)^{\eta_2+1}}\right)$$
$$\boldsymbol{\eta} = (\eta_1, \eta_2)' = \left(-\frac{1}{2a}, \frac{\nu}{2} - 1\right)' \qquad g_0(t) = 1$$

The Poisson regression (8) using t and $\log t$ as predictors gives estimates $\hat{\eta}_1$, $\hat{\eta}_2$ and $\hat{C}$. From these we obtain

$$\hat{a} = -\frac{1}{2\hat{\eta}_1}, \qquad \hat{\nu} = 2(\hat{\eta}_2 + 1), \qquad \log \hat{p}_0 = \hat{C} + \psi(\hat{\boldsymbol{\eta}}).$$

The derivative (17) required for computing the covariance (14) of $\hat{\boldsymbol{\theta}}^+ = (\log \hat{p}_0, \hat{a}, \hat{\nu})'$ is

$$\hat{\boldsymbol{D}} = \begin{pmatrix} 1 & -\frac{\hat{\eta}_2+1}{\hat{\eta}_1} & \Psi(\hat{\eta}_2+1) - \log(-\hat{\eta}_1) \\ 0 & \frac{1}{2\hat{\eta}_1^2} & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & \hat{a}\hat{\nu} & \Psi(\hat{\nu}/2) + \log(2\hat{a}) \\ 0 & 2\hat{a}^2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

where $\Psi(z) = (d/dz)\log\Gamma(z)$ is the Digamma function. The scaled χ^2 family lends itself to two exponential subfamilies.

A.2.1 Estimate $\theta = a$

The empirical null $a\chi^2(\nu_0)$ with fixed ν_0 has exponential family form $x(t) = t$, $\eta = -1/(2a)$, $\psi(\eta) = -(\nu_0/2)\log(-\eta)$ and $g_0(t) = t^{\nu_0/2-1}/\Gamma(\nu_0/2)$. Poisson regression

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

using t as a predictor and $\log(N\Delta g_0(t))$ as offset gives estimates $\hat{\eta}$ and $\hat{C}$. From these we obtain

$$\hat{a} = -\frac{1}{2\hat{\eta}}, \qquad \log \hat{p_0} = \hat{C} - \frac{\nu_0}{2}\log(-\hat{\eta}), \qquad \hat{\boldsymbol{D}} = \begin{pmatrix} 1 & \hat{a}\nu_0 \\ 0 & 2\hat{a}^2 \end{pmatrix}.$$

A.2.2 Estimate $\theta = \nu$

The empirical null $a_0\chi^2(\nu)$ with fixed a_0 has exponential family form $x(t) = \log t$, $\eta = \nu/2 - 1$, $\psi(\eta) = \log\Gamma(\eta+1) + (\eta+1)\log(2a_0)$ and $g_0(t) = e^{-t/(2a_0)}$. Poisson regression using $\log t$ as a predictor and $\log(N\Delta g_0(t))$ as offset gives estimates $\hat{\eta}$ and $\hat{C}$. From these we obtain

$$\hat{\nu} = 2(\hat{\eta}+1), \qquad \log \hat{p_0} = \hat{C} + \psi(\hat{\eta}), \qquad \hat{\boldsymbol{D}} = \begin{pmatrix} 1 & \Psi(\hat{\nu}/2) + \log(2a_0) \\ 0 & 2 \end{pmatrix}.$$

References

Abramowitz, M. and Stegun, I. A. (eds.). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover, 9th edition (1972).

Benjamini, Y. and Hochberg, Y. "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing." *J R Statist Soc B*, 57(1):289–300 (1995).

Efron, B. "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis." *J Amer Statist Assoc*, 99(465):96–104 (2004).

—. "Bayesians, frequentists and scientists." *J Amer Statist Assoc*, 100(469):1–5 (2005).

—. "Correlation and Large-Scale Simultaneous Hypothesis Testing." *J Amer Statist Assoc*, 102(477):93–103 (2007a).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

—. "Size, power and false discovery rates." *Ann Statist*, 35(4):1351–1377 (2007b).

Efron, B. and Tibshirani, R. "Using especially designed exponential families for density estimation." *Ann Statist*, 24(6):2431–2461 (1996).

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. "Empirical Bayes Analysis of a Microarray Experiment." *J Amer Statist Assoc*, 96(456):1151–1160 (2001).

Everitt, B. S. and Bullmore, E. T. "Mixture Model Mapping of Brain Activation in Functional Magnetic Resonance Images." *Human Brain Mapping*, 7:1–14 (1999).

Genovese, C. R. and Wasserman, L. "A stochastic process approach to false discovery control." *Ann Statist*, 32:1035–1061 (2004).

Ghahremani, D. and Taylor, J. E. "Empirical and Theoretical False Discovery Rate Analyses for fMRI Data." Poster, Organization for Human Brain Mapping (2005).

Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T., Wichmann, H.-E., Meitinger, T., Hunter, D., Hu, F. B., Colditz, G., Hinney, A., Hebebrand, J., Koberwitz, K., Zhu, X., Cooper, R., Ardlie, K., Lyon, H., Hirschhorn, J. N., Laird, N. M., Lenburg, M. E., Lange, C., and Christman, M. F. "A Common Genetic Variant Is Associated with Adult and Childhood Obesity." *Science*, 312:279–283 (2006).

Jin, J. and Cai, T. T. "Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons." *J Amer Statist Assoc*, 102(478):495–506 (2007).

Kong, S. W., Pu, W. T., and Park, P. J. "A multivariate approach for integrating genome-wide expression data and biological knowledge." *Bioinformatics*, 22(19):2373–2380 (2006).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Koudou, A. E. "Lancaster bivariate probability distributions with Poisson, negative binomial and gamma margins." *Test*, 7(1):95–110 (1998).

Lange, C., Silverman, E. K., Xu, X., Weiss, S. T., and Laird, N. M. "A multivariate family-based association test using generalized estimating equations: FBAT-GEE." *Biostatistics*, 4(2):195–206 (2003).

Lee, J., Shahram, M., Schwartzman, A., and Pauly, J. M. "A complex data analysis in high-resolution SSFP fMRI." *Magn Reson Med*, 57:905–917 (2007).

Schwartzman, A., Dougherty, R. F., and Taylor, J. E. "Cross-subject comparison of principal diffusion direction maps." *Magn Reson Med*, 53:1423–1431 (2005).

—. "False Discovery Rate Analysis of Brain Diffusion Direction Maps." *Ann Appl Statist*, to appear (2008). Currently available at http://www.imstat.org/aoas/next_issue.html.

Storey, J. D. "The positive false discovery rate: a Bayesian interpretation and the q-value." *Ann Statist*, 31(6):2013–2035 (2003).

Storey, J. D., Taylor, J. E., and Siegmund, D. "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach." *J R Statist Soc B*, 66(1):187–205 (2004).

Sun, W. and Cai, T. T. "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control." *J Amer Statist Assoc*, 102(479):901–912 (2007).

Van Steen, K., McQueen, M. B., Herbert, A., Raby, B., Lyon, H., DeMeo, D. L., Murphy, A., Su, J., Datta, S., Rosenow, C., Christman, M., Silverman, E. K., Laird, N. M., Weiss, S. T., and Lange, C. "Genomic screening and replication using the same data set in family-based association testing." *Nature Genetics*, 37(7):683–691 (2005).

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

Table 1: Asymptotic bias in the estimation of $\boldsymbol{\theta}^+$. $\hat{\boldsymbol{\theta}}^+_\infty = \boldsymbol{\theta}(\hat{\boldsymbol{\eta}}^+_\infty)$ was computed from the solution to the limiting score equation (29).

Null	$\boldsymbol{\theta}^+$	$\hat{\boldsymbol{\theta}}^+_\infty$	$\hat{\boldsymbol{\theta}}^+_\infty - \boldsymbol{\theta}^+$	Formula (28)
Normal	$\log(p_0) =$ -0.1054	-0.0801	0.0253	0.0289
	$\mu = 0.2$	0.2265	0.0265	0.0250
	$\sigma^2 = 1.44$	1.4907	0.0507	0.0480
χ^2	$\log(p_0) =$ -0.1054	-0.0331	0.0723	0.0749
	$a = 0.8$	0.8449	0.0449	0.0457
	$\nu = 3$	2.9789	-0.0211	-0.0210

Table 2: Empirical null estimates for the SNP data example. Column 3 indicates the teoretical null values for reference. Columns 5 and 6 are the delta-method and bootstrap standard errors (SE). Column 7 contains standard 95% confidence intervals based on the delta-method SE.

$\boldsymbol{\theta}^+$	Theory	$\hat{\boldsymbol{\theta}}^+$	SE (14)	Bootstrap SE	95% CI
$\log(p_0)$	0	7.49×10^{-5}	5.45×10^{-5}	5.89×10^{-5}	$[-3.18 \times 10^{-5}, 1.82 \times 10^{-4}]$
a	1	0.9509	0.0046	0.0048	[0.9419, 0.9600]
ν	4	4.2675	0.0183	0.0199	[4.2316, 4.3034]

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 3: Bins with local FDR significantly below the expected under the complete null for the SNP data example. Column 3 contains standard 95% confidence intervals based on the delta-method SE. Column 4 is the observed local FDR from column 2 divided by the expected local FDR for the complete null at that bin. Column 4 is the total number of SNPs in the bin.

t_k	$\widehat{\text{fdr}}_k$	95% CI	Adjusted $\widehat{\text{fdr}}_k$	# SNPs
21.65	0.2831	[0.1546, 0.5184]	0.4169	3
21.85	0.2575	[0.1445, 0.4587]	0.4082	3
24.35	0.1173	[0.0726, 0.1896]	0.5310	2

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

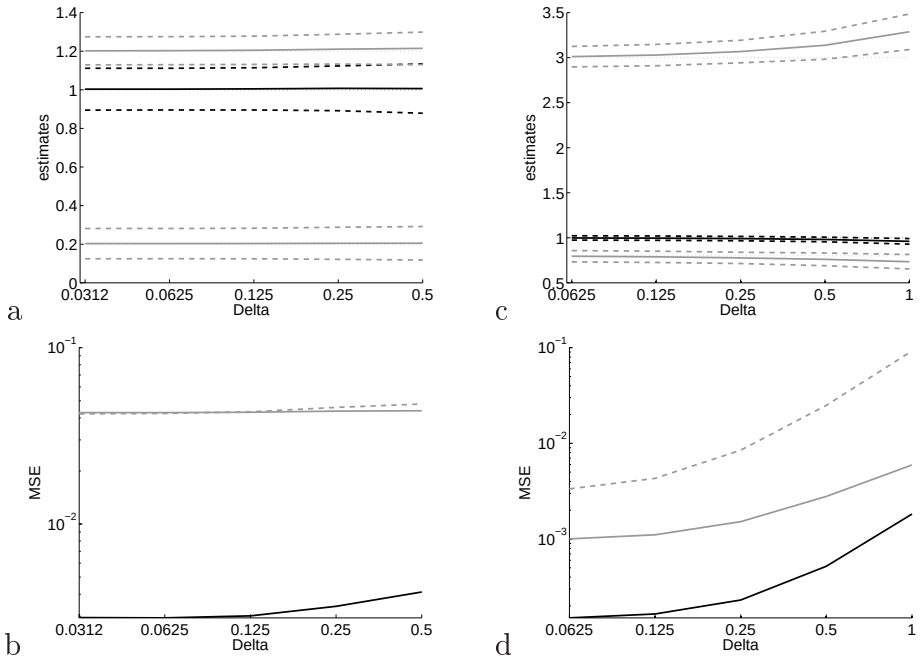

Figure 1: Effect of the bin width Δ. Left panels: Normal simulation. (a) Estimates of p_0 (black solid), μ (lower gray solid) and σ (upper gray solid). The thick dashed lines indicate simulated ± 2 standard errors, very close to the values predicted by the formula (14). The thin dashed lines indicate the true values $p_0 = 1$, $\mu_0 = 0.2$, $\sigma_0 = 1.2$. (b) Simulated MSE for p_0 (black), μ (solid gray) and σ (dashed gray). Right panels: χ^2 simulation. (c) Estimates of p_0 (black solid), a (lower gray solid) and ν (upper gray solid). The thick dashed lines indicate simulated ± 2 standard errors, very close to the values predicted by the formula (14). The thin dashed lines indicate the true values $p_0 = 1$, $a_0 = 0.8$, $\nu_0 = 3$. (d) Simulated MSE for p_0 (black), a (solid gray) and ν (dashed gray).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

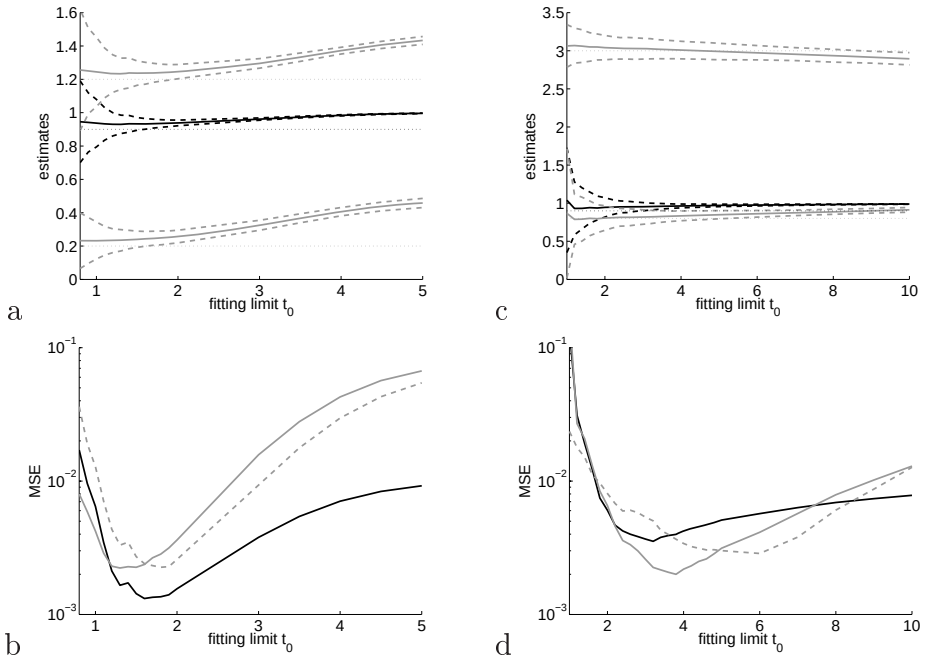

Figure 2: Effect of the fitting interval width t_0. Left panels: Normal simulation. (a) Estimates of p_0 (black solid), μ (lower gray solid) and σ (upper gray solid). The thick dashed lines indicate simulated ± 2 standard errors, very close to the values predicted by the formula (14). The thin dashed lines indicate the true values $p_0 = 0.9$, $\mu_0 = 0.2$, $\sigma_0 = 1.2$. (b) Simulated MSE for p_0 (black), μ (solid gray) and σ (dashed gray). Right panels: χ^2 simulation. (c) Estimates of p_0 (black solid), a (lower gray solid) and ν (upper gray solid). The thick dashed lines indicate simulated ± 2 standard errors, very close to the values predicted by the formula (14). The thin dashed lines indicate the true values $p_0 = 0.9$, $a_0 = 0.8$, $\nu_0 = 3$. Notice the vertical scale is different from panel (a). (d) Simulated MSE for p_0 (black), a (solid gray) and ν (dashed gray).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper77

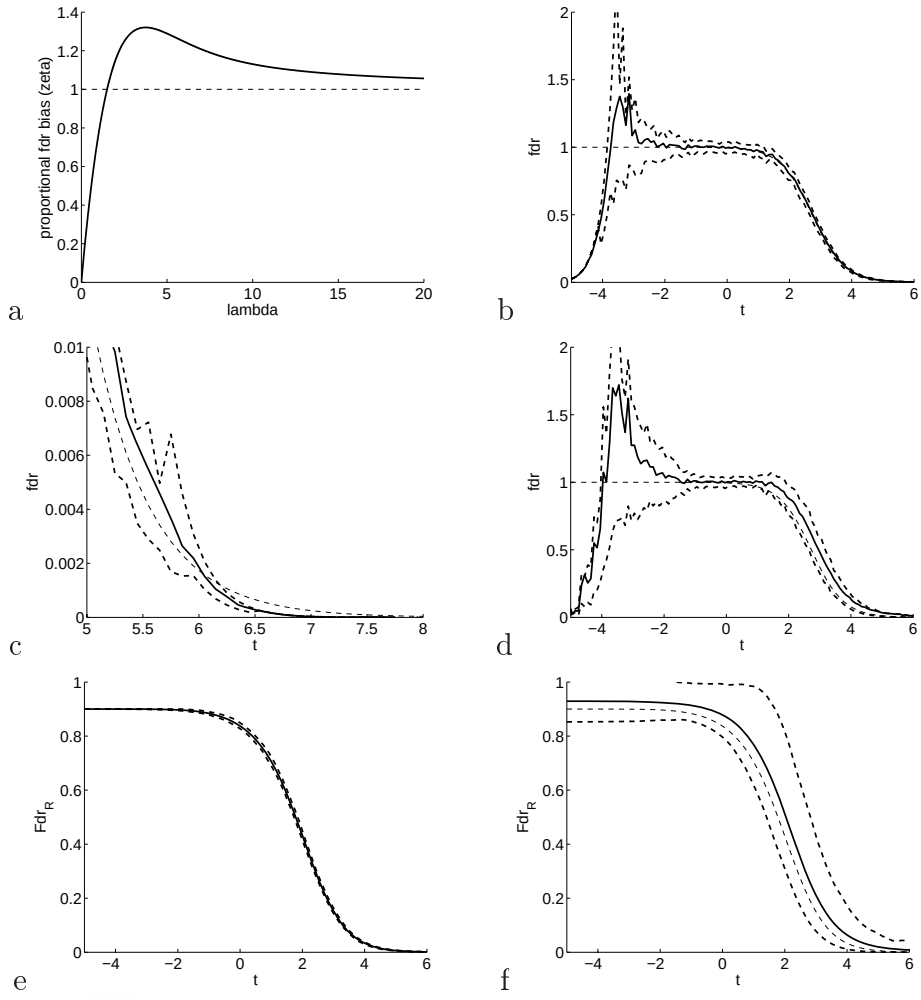

Figure 3: FDR bias in the normal simulation. (a) The proportional FDR bias function $\zeta(\lambda)$. In plots (b)-(f), the black solid line is the simulated average estimate, the thick dashed lines are 5 and 95 percentiles of the simulation, and the thin dashed line is the true value. (b) Local FDR using theoretical null. (c) Local FDR using theoretical null (zoom in). (d) Local FDR using empirical null. (e) Right tail FDR using theoretical null. (f) Right tail FDR using empirical null.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

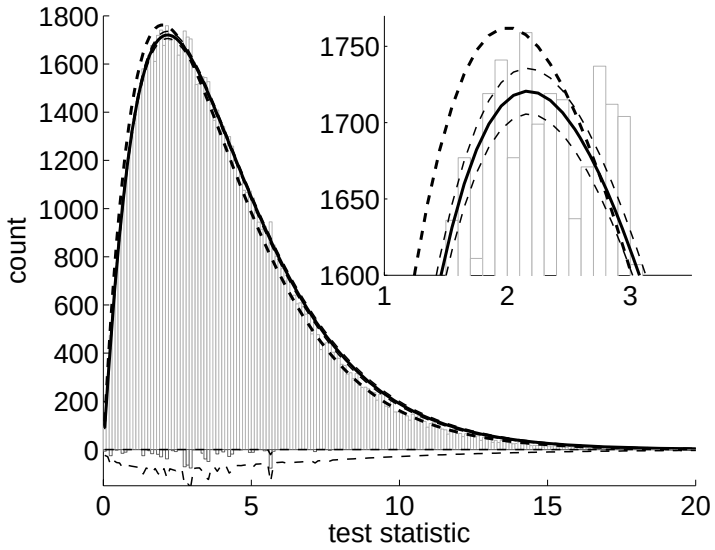

Figure 4: Histogram of test statistics in the SNP example. Superimposed densities are the theoretical null $\chi^2(4)$ (dashed) and the empirical null (solid), each with pointwise standard 95% CIs. The histogram of the estimated alternative component and corresponding upper standard CI are shown in inverted scale.

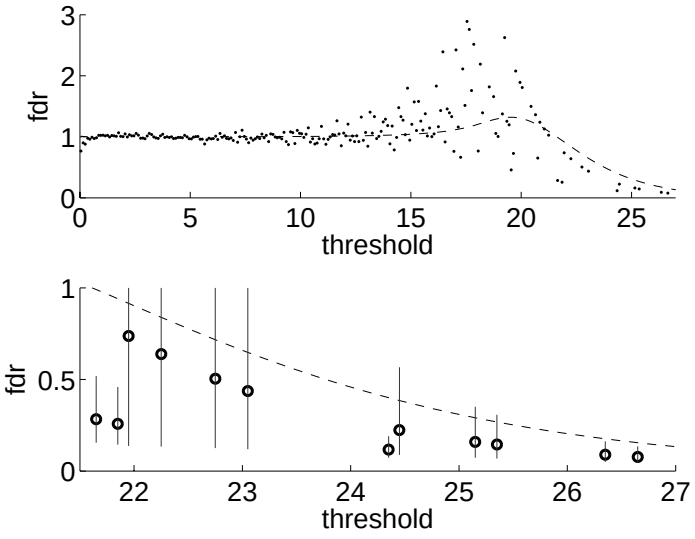

Figure 5: Local FDR estimates in the SNP example using the empirical null. Top panel: full range. Bottom panel: zoom-in including standard 95% CIs. In both panels the dashed line is the expected local FDR under the complete null.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive