Harvard University
Harvard University Biostatistics Working Paper Series

Year 2009 *Paper* 106

The Effect of Correlation in False Discovery Rate Estimation

Armin Schwartzman* Xihong Lin†

*Harvard School of Public Health and Dana Farber Cancer Institute, armin@jimmy.harvard.edu

†Harvard University, xlin@hsph.harvard.edu

The Effect of Correlation in False Discovery Rate Estimation

Armin Schwartzman and Xihong Lin
Department of Biostatistics, Harvard School of Public Health

July 6, 2009

Abstract

Current false discovery rate (FDR) methods mostly ignore the correlation structure in the data. The objective of this paper is to quantify the effect of correlation in FDR analysis. Specifically, we derive practical approximations for the mean, variance, distribution, and quantiles of the FDR estimator for arbitrarily correlated data. This is achieved using a negative binomial model for the number of false discoveries, where the parameters are found empirically from the data. We show that correlation increases the bias and variance substantially with respect to the independent case for practical FDR levels, and that in some cases, such as an exchangeable correlation structure, the FDR estimator fails to be consistent as the number of tests gets large.

1 Introduction

Large-scale multiple testing is a common statistical problem in the analysis of high-dimensional data, particularly in genomics (Dudoit et al., 2003; Efron, 2004; Roeder et al., 2006), proteomics (Tibshirani et al., 2005) and medical imaging (Genovese et al., 2002; Worsley et al., 2004; Schwartzman et al., 2009). An increasingly popular global measure of error in these applications is the false discovery rate (FDR) (Benjamini and Hochberg, 1995), defined as the expected proportion of false discoveries among the total number of discoveries. The majority of research on FDR assumes conveniently that the test statistics corresponding to each hypothesis are independent (Genovese and Wasserman, 2004; Storey et al., 2004; Efron, 2007b; Sun and Cai, 2007), and little is known about how to take the correlation into account. As a result, FDR methods are often used on data ignoring the correlation between the test statistics.

However, high-dimensional data are often highly correlated, e.g. gene expression levels in microarray experiments (Qiu et al., 2005; Owen, 2005; Klebanov et al., 2006). As a typical example, a look ahead at Figure 5b shows the distribution of sample pairwise correlations between genes obtained from the diabetes study of Mootha et al. (2003). The raw pairwise correlations range from -0.9 to 0.96. Other applications exhibit a correlation structure that can be modeled as a spatial random field, as in brain imaging (Genovese et al., 2002; Worsley et al., 2004), or as a time series, as in proteomic mass spectrometry (Harezlak et al., 2008). As a result of correlation, some FDR controlling methods that assume independence have been shown to fail (Qiu et al., 2005), while the variance of the number of false discoveries has been shown to be greatly inflated (Owen, 2005). While there

exist procedures that control FDR under arbitrary dependence (Yekutieli and Benjamini, 1999; Benjamini and Yekutieli, 2001), they have substantially less power than procedures that assume independence (Farcomeni, 2008) and the latter are often preferred. Because of the current widespread use of FDR, it is important to understand the effect of correlation in FDR analysis as it is typically used in practice, both for correct inference using current methods and as a guide for developing new FDR methods for correlated data.

The goal of this paper is to quantify the effect of correlation in FDR analysis. As a benchmark, we use the FDR estimator of Genovese and Wasserman (2004) and Storey et al. (2004). This estimator is appealing because it provides estimates of FDR at all thresholds simultaneously. Furthermore, thresholding of this estimator is equivalent to the original FDR algorithm (Benjamini and Hochberg, 1995) under independence, and under specific forms of dependence such as positive regression dependence (Benjamini and Yekutieli, 2001) and weak dependence such as dependence in finite blocks (Storey et al., 2004). However, the generality of the estimator makes it conservative under correlation and it can perform poorly in genomic data (Qiu and Yakovlev, 2006). We show that correlation increases both the bias and variance of the estimator substantially compared to the independent case, but less so for small FDR levels. From a theoretical point of view, we show that in some cases such as an exchangeable correlation structure, the FDR estimator fails to be consistent as the number of tests gets large.

Other related approaches that incorporate correlation in FDR analysis include the use of an empirical null (Efron, 2007a; Schwartzman, 2008), and procedures adapted to clusters of highly correlated genes (Dahl and Newton, 2007; Tibshirani and Wasserman, 2006) and clusters of locally correlated image pixels (Pacifico et al., 2004; Heller et al., 2007). Owen (2005) quantified the variance of the number of discoveries given an arbitrary correlation structure, but did not provide results about the FDR. His analysis was also restricted to the complete null hypothesis and to a particular test statistic, the correlation coefficient between the gene expression and a covariate.

Our contributions are as follows. First, we provide approximations for the mean, variance, distribution, and quantiles of the FDR estimator given an arbitrary correlation structure. This is achieved by modeling the number of discoveries with a negative binomial (NB) distribution whose parameters are estimated from the data based on the empirical distribution of the pairwise correlations. Our results are derived for common test statistics whose marginal distribution is either normal or χ^2. Second, we identify a necessary condition for consistency of the FDR estimator as the number of tests increases and show that it is violated in situations such as exchangeable correlation.

The structure of the paper is as follows. We first study the mean-variance structure of the number of discoveries and present the asymptotic results. Next we show how to quantify the overdispersion in the number of discoveries for normal and χ^2 statistics based on the empirical distribution of the pairwise correlations. Finally, based on the above mean-variance structure, we propose the NB model for the number of discoveries and derive the distributional properties of the FDR estimator based on this model. Results are evaluated by simulations and illustrated with a microarray data example.

2 Theory

2.1 The false discovery rate estimator

Let $H_1, \ldots, H_m$ be m null hypotheses with associated test statistics $T_1, \ldots, T_m$. The test statistics are assumed to have marginal distributions

$$T_j \sim \begin{cases} F_0, & H_j \text{ is true} \\ F_j, & H_j \text{ is false} \end{cases} \tag{1}$$

where F_0 is a common distribution under the null hypothesis and $F_1, F_2, \ldots$ are specific alternative distributions for each test. The test statistics may be dependent with $\text{corr}(T_i, T_j) = \rho_{ij}$. In particular, if the test statistics are z-scores, then this is the same as the correlation between the original observations.

The fraction $p_0 = m_0/m$ of tests where the null is true is called the null proportion. The complete null model is the one where $p_0 = 1$ and $T_j \sim F_0$ for all j. Without loss of generality, we focus on one-sided tests, where for each hypothesis H_j and a threshold u, the decision rule is $D_j(u) = 1(T_j > u)$. Two-sided tests may be incorporated, for example, by defining $T_j = \tilde{T}_j^2$ or $T_j = |\tilde{T}_j|$, where $\tilde{T}_j$ is a two-sided test statistic.

Let $R_m(u) = \sum_{j=1}^m D_j(u)$ and $V_m(u) = \sum_{j=1}^m D_j(u)1(H_j \text{ is true})$ be the number of rejected null hypotheses or discoveries, and the number of false positives, respectively. The FDR is the expected proportion of false positives among the tests where the null hypothesis is rejected, i.e.

$$\text{FDR} = \text{E}\left[\frac{V_m(u)}{R_m(u) \vee 1}\right] \tag{2}$$

where the expectations are computed under the true model (1) (Benjamini and Hochberg, 1995). When m is large, the FDR is empirically estimated by (Genovese and Wasserman, 2004; Storey et al., 2004)

$$\widehat{\text{FDR}}_m(u) = \frac{\hat{p}_0 m \alpha(u)}{R_m(u) \vee 1} \tag{3}$$

where $\hat{p}_0$ is an estimate of the null proportion p_0, and $\alpha(u)$ is the marginal type-I-error level $\alpha(u) = \text{E}[D_j(u)] = P[T_j > u]$, computed under the assumption that H_j is true. A heuristic argument for this estimator is that the expectation of the false discovery proportion numerator, $\text{E}[V_m(u)]$, is equal to $p_0 m \alpha(u)$ under the true model (1). There are several ways to estimate the null proportion p_0 (Storey et al., 2004; Efron, 2007b; Jin and Cai, 2007). In applications p_0 is often close to 1 and setting $\hat{p}_0 = 1$ biases the estimate only slightly and in a conservative fashion (Efron, 2004). In this article we study the effect that correlation has on the FDR estimator (3) via the number of discoveries $R_m(u)$.

2.2 The number of discoveries

As a stepping stone towards studying the FDR estimator (3), we first study the number of rejected null hypotheses or discoveries $R_m(u)$. Under the complete null model, and if the tests are independent, $R_m(u)$ is binomial with number of trials m and success probability

$\alpha(u)$. In general, under model (1) and allowing dependence, we have that

$$\begin{aligned} \mathrm{E}[R_m(u)] &= \sum_{i=1}^{m} \mathrm{E}[D_i(u)] = \sum_{i=1}^{m} \beta_i(u) \\ \mathrm{var}[R_m(u)] &= \sum_{i=1}^{m}\sum_{j=1}^{m} \mathrm{cov}[D_i(u), D_j(u)] = \sum_{i=1}^{m}\sum_{j=1}^{m} \Psi_{ij}(u) \end{aligned} \tag{4}$$

where

$$\begin{aligned} \beta_i(u) &= P_i(T_i > u) \\ \Psi_{ij}(u) &= P(T_i > u, T_j > u) - P(T_i > u)P(T_j > u) \end{aligned} \tag{5}$$

The quantity $\beta_i(u)$ is the per-test power. For those tests where the null hypothesis is true, this is equal to the marginal type-I-error level $\alpha(u)$. Summing over the diagonals in (4) reveals the mean-variance structure

$$\begin{aligned} \mathrm{E}[R_m(u)] &= m\overline{\beta}(u) \\ \mathrm{var}[R_m(u)] &= m\left[\overline{\beta}(u) - \overline{\beta^2}(u)\right] + m(m-1)\overline{\Psi}_m(u), \end{aligned} \tag{6}$$

where

$$\overline{\beta}(u) = \frac{1}{m}\sum_{i=1}^{m}\beta_i(u), \qquad \overline{\beta^2}(u) = \frac{1}{m}\sum_{i=1}^{m}\beta_i^2(u), \qquad \overline{\Psi}_m(u) = \frac{2}{m(m-1)}\sum_{i<j}\Psi_{ij}(u) \tag{7}$$

The quantities $\overline{\beta}(u)$ are $\overline{\beta^2}(u)$ are empirical moments of the power, while $\overline{\Psi}_m(u)$ is the average covariance of the decisions $D_i(u)$ and $D_j(u)$ for $i \neq j$, a function of the pairwise correlations $\{\rho_{ij} = \mathrm{corr}(T_i, T_j)\}$.

We observe that the dependence between the test statistics does not affect the mean of $R_m(u)$ but does affect its variance. It does so by adding the overdispersion term $m(m-1)\overline{\Psi}_m(u)$ to the independent-case variance $m[\overline{\beta}(u) - \overline{\beta^2}(u)]$. The special case of (6) under the complete null is an unconditional version of the conditional variance in expression (8) of Owen (2005). Similar expressions have appeared before in estimation problems with correlated binary data (Crowder, 1985; Prentice, 1986).

Another observation from (5) is that $\Psi_{ij}(u)$ vanishes asymptotically as $u \to \infty$, so the effect of the correlation becomes negligible for very high thresholds. We will see later in Section 2.4 that the rate of decay is in fact quadratic exponential times a polynomial.

Example (Exchangeable correlation)**.** Suppose the test statistics T_i have an exchangeable correlation structure so that $\rho_{ij} = \rho > 0$ is a constant for all $i \neq j$. Under the complete null, such test statistics may be generated as $T_i = \sqrt{\rho}Z + \sqrt{1-\rho}\varepsilon_i$, where $Z, \varepsilon_1, \ldots, \varepsilon_m$ are i.i.d. with mean zero. In this case, for any fixed threshold u, $\Psi_{ij}(u) = \Psi(u) > 0$ is a constant for all $i \neq j$ and $\overline{\Psi}_m(u) = \Psi(u)$. The mean and variance of $R_m(u)$ can be computed explicitly using (6). For example, if the tests are one-sided and the tests statistics are marginally $N(0,1)$ (obtained when $Z, \varepsilon_1, \ldots, \varepsilon_m$ are i.i.d. $N(0,1)$) then the mean and variance of $R_m(u)$ are given by

$$\begin{aligned} \mathrm{E}[R_m(u)] &= m\Phi(u) \\ \mathrm{var}[R_m(u)] &= m[\Phi(u) - \Phi^2(u)] + m(m-1)[\Phi_2(u,u;\rho) - \Phi^2(u)], \end{aligned} \tag{8}$$

where $\Phi(u)$ is the standard normal survival function and $\Phi_2(u, u; \rho)$ is the bivariate standard normal survival function with marginals $N(0, 1)$ and correlation ρ. As expected, the variance increases with ρ.

2.3 Asymptotic inconsistency of the FDR estimator

The overdispersion of the number of discoveries $R_m(u)$ has implications for the behavior of the FDR estimator (3). We consider the asymptotic case $m \to \infty$ in this section and treat the finite m case in Sections 2.4 and 2.5.

Asymptotically as $m \to \infty$, a sufficient condition for consistency of the estimator (3) is that the test statistics are independent or weakly dependent, e.g. dependent in finite blocks (Storey et al., 2004). On the other hand, taking m in (3) to the denominator, we see that a necessary condition for consistency is that the fraction of discoveries $R_m(u)/m$ has asymptotically vanishing variance. By (6), the variance of $R_m(u)/m$ is

$$\operatorname{var}\left[\frac{R_m(u)}{m}\right] = \frac{\overline{\beta}(u) - \overline{\beta^2}(u)}{m} + \left(1 - \frac{1}{m}\right)\overline{\Psi}_m(u) \tag{9}$$

and is asymptotically zero if and only if the overdispersion $\overline{\Psi}_m$ is asymptotically zero, provided that $\overline{\beta}$ and $\overline{\beta^2}$ grow slower than linearly with m.

Example (Exchangeable correlation, continued)**.** Suppose the test statistics T_i have an exchangeable correlation structure with pairwise correlation $\rho > 0$. Then, for any fixed threshold u, $\Psi_{ij}(u) = \Psi(u) = \Phi_2(u, u; \rho) - \Phi^2(u) > 0$ is the same for all i, j. By (9), $\operatorname{var}[R_m(u)/m] \to \Psi(u) > 0$ and the FDR estimator (3) is inconsistent. The case $\rho < 0$ is asymptotically moot because positive definiteness of the covariance of the T_i's requires $\rho > -1/(m-1)$, so ρ cannot be negative in the limit $m \to \infty$.

The following result characterizes covariance structures that may be called asymptotically exchangeable.

Theorem 1. *Assume* $[\overline{\beta}(u) - \overline{\beta^2}(u)]/m \to 0$ *as* $m \to \infty$. *Fix an ordering of the test statistics* $T_1, \ldots, T_m$ *and assume the autocovariance sequence* $\Psi_{i,i+k}(u) = \Psi_{i+k,i}(u)$ *in* (9) *has a limit* $\Psi_\infty(u) \geq 0$ *as* $k \to \infty$ *for every* i. *Then, as* $m \to \infty$, $\operatorname{var}[R_m(u)/m] \to \Psi_\infty(u) \geq 0$.

Example (Stationary ergodic covariance)**.** Suppose the index i represents time or position along a genome sequence, and suppose the test statistic sequence T_i is a stationary ergodic process, e.g. M-dependent or ARMA, with Toeplitz correlation $\rho_{ij} = \rho_{i-j}$. Then $\rho_{i,i+k} = \rho_k \to 0$ for all i as $k \to \infty$, and so $\Psi_\infty(u) = 0$, pointwise for all u. By Theorem 1, the variance (9) converges to zero as $m \to \infty$.

Example (Finite blocks)**.** Suppose the test statistics T_i are dependent in finite blocks, so that they are correlated within blocks but independent between blocks. Suppose the largest block has size K. If K increases with m such that $K/m \to 0$ as $m \to \infty$ then $\Psi_\infty(u) = 0$ for all u. On the other hand, if $K/m \to \gamma > 0$ then $\Psi_\infty(u) > 0$ and the FDR estimator (3) is inconsistent.

Example (Strong mixing)**.** Suppose the test statistics T_i have a strong mixing or α-mixing dependence; that is, the supremum over k of $|P(A \cap B) - P(A)P(B)|$, where A is in the σ-field generated by $T_1, \ldots, T_k$ and B is in the σ-field generate by $T_{k+1}, T_{k+2}, \ldots$, tends to 0 as $m \to \infty$ (Zhou and Liang, 2000). By taking $A = \{T_i > u\}$ and $B = \{T_j > u\}$ in (5), we have that $\Psi_\infty = 0$ and therefore, by Theorem 1, the variance (9) converges to zero as $m \to \infty$.

2.4 Quantifying overdispersion for finite m

We are interested in estimating the distributional properties of the FDR estimator. This requires estimating the overdispersion $\overline{\Psi}_m(u)$ in (6) and (9). This quantity is easy to write in terms of the pairwise correlations between the decision rules, as in (4), but not necessarily as a function of the pairwise test statistic correlations ρ_{ij}. In this section we provide expressions for $\overline{\Psi}_m(u)$ for finite m assuming a specific bivariate probability model for every pair of test statistics, but without the need to assume a higher-order correlation structure. We consider commonly used z and χ^2 tests.

Suppose first that every pair of test statistics (T_i, T_j) has the bivariate normal density with marginals $N(\mu_i, 1)$ and $N(\mu_j, 1)$, and $\text{corr}(T_i, T_j) = \rho_{ij}$. Denote by $\phi(t)$ and $\phi_2(t_i, t_j; \rho_{ij})$ the univariate and bivariate standard normal densities. Mehler's formula (Patel and Read, 1996; Kotz et al., 2000) states that the joint density $f_{ij}(t_i, t_j) = \phi_2(t_i - \mu_i, t_j - \mu_j; \rho_{ij})$ of (T_i, T_j) can be written as

$$f_{ij}(t_i, t_j) = \phi(t_i - \mu_i)\phi(t_j - \mu_j) \sum_{k=0}^{\infty} \frac{\rho_{ij}^k}{k!} \mathcal{H}_k(t_i - \mu_i)\mathcal{H}_k(t_j - \mu_j), \tag{10}$$

where $\mathcal{H}_k(t)$ are the Hermite polynomials: $\mathcal{H}_0(t) = 1$, $\mathcal{H}_1(t) = t$, $\mathcal{H}_2(t) = t^2 - 1$, and so on.

Theorem 2. *Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$. Under the bivariate normal model* (10)*, the overdispersion factor in* (7) *is given by*

$$\overline{\Psi}_m(u; \boldsymbol{\mu}) = \sum_{k=1}^{\infty} \frac{\rho_k^*(\boldsymbol{\mu})}{k!} \tag{11}$$

where

$$\rho_k^*(\boldsymbol{\mu}) = \frac{2}{m(m-1)} \sum_{i<j} \rho_{ij}^k \phi(u - \mu_i)\phi(u - \mu_j)\mathcal{H}_{k-1}(u - \mu_i)\mathcal{H}_{k-1}(u - \mu_j). \tag{12}$$

Under the complete null, (11) *reduces to*

$$\overline{\Psi}_m(u) = \phi^2(u) \sum_{k=1}^{\infty} \frac{\overline{\rho^k}}{k!} \mathcal{H}_{k-1}^2(u), \tag{13}$$

where $\overline{\rho^k}$, $k = 1, 2, \ldots$ denote the empirical moments of the $m(m-1)/2$ correlations ρ_{ij}, $i < j$.

Another common null distribution in multiple testing problems is the χ^2 distribution. Let $f_\nu(t)$ and $F_\nu(t)$ be the density and distribution functions of the $\chi^2(\nu)$ distribution with ν d.f. Under the complete null, the pair of test statistics (T_i, T_j) admits a Lancaster bivariate model where both T_i and T_j have the same marginal density f_ν, their correlation is ρ_{ij}, and their joint density is given by

$$f_\nu(t_i, t_j; \rho_{ij}) = f_\nu(t_i) f_\nu(t_i) \sum_{k=0}^{\infty} \frac{\rho_{ij}^k}{k!} \frac{\Gamma(\nu/2)}{\Gamma(\nu/2+k)} \mathcal{L}_k^{(\nu/2-1)}\left(\frac{t_i}{2}\right) \mathcal{L}_k^{(\nu/2-1)}\left(\frac{t_j}{2}\right) \tag{14}$$

where $\mathcal{L}_k^{(\nu/2-1)}(t)$ are the generalized Laguerre polynomials of degree $\nu/2-1$:

$$\begin{aligned}
\mathcal{L}_0^{(\nu/2-1)}(t) &= 1 \\
\mathcal{L}_1^{(\nu/2-1)}(t) &= -t + \nu/2 \\
\mathcal{L}_2^{(\nu/2-1)}(t) &= t^2 - 2(\nu/2+1)t + (\nu/2)(\nu/2+1)
\end{aligned}$$

and so on (Koudou, 1998). Notice that, in contrast to the normal, the χ^2 distribution is not a location family with respect to the non-centrality parameter, so the Lancaster expansion does not hold in the non-central case.

Theorem 3. *Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and assume the complete null. Under the χ^2 normal model* (14)*, the overdispersion factor* (7) *is given by*

$$\overline{\Psi}_m(u) = f_{\nu+2}^2(u) \sum_{k=1}^{\infty} \frac{\overline{\rho^k}}{k!} \frac{\nu^2 \Gamma(\nu/2)}{k^2 \Gamma(\nu/2+k)} \left[L_{k-1}^{(\nu/2)}\left(\frac{u}{2}\right) \right]^2. \tag{15}$$

where $\overline{\rho^k}$, $k = 1, 2, \ldots$ denote the empirical moments of the $m(m-1)/2$ correlations ρ_{ij}, $i < j$.

Theorems 2 and 3 show that the overdispersion is a function of modified empirical moments of the pairwise correlations ρ_{ij} and depends on m only through these moments. This provides an efficient way to evaluate $\overline{\Psi}_m(u)$ for a given set of pairwise correlations obtained from data, as the expansions (13) and (15) may be approximated evaluating only the first few low order terms. This is contrast to direct computation of (7), which requires $m(m-1)$ evaluations of the bivariate probabilities in (5), one for each value of ρ_{ij}. As shown later in the simulations, only the first few terms in the sums (13) and (15) are needed.

Theorems 2 and 3 also provide insight into the nature of the overdispersion. First, as a function of u, both expressions (13) and (15) decay as a quadratic exponential times a polynomial, implying that the effect of correlation quickly becomes negligible for high thresholds. Second, under the complete null, (13) and (15) indicate that the overdispersion, and thus also the variance (9), vanish asymptotically as $m \to \infty$ for all u if and only if $\overline{\rho^k} \to 0$ as $m \to \infty$ for all $k = 1, 2, \ldots$. The usefulness of this result is illustrated in the following example.

Example (Factor analysis)**.** Suppose we have observations $\boldsymbol{y} = (y_1, \ldots, y_m)'$ with factor analysis covariance $\mathrm{cov}(\boldsymbol{y}) = \sigma^2 I + a\delta\delta'$, where I is the $m \times m$ identity matrix, $a > 0$ and

δ is a unit m-vector. Suppose for each i, $y_i = \mu_i + \varepsilon_i$ and we want to test $H_i : \mu_i = 0$ with the z-statistic $T_i = y_i(\sigma^2 + a\delta_i^2)^{-1/2}$. Then

$$\overline{\rho^k} = \frac{2}{m(m-1)} \sum_{i<j} \rho_{ij}^k = \frac{2}{m(m-1)} \sum_{i<j} \left(\frac{a\delta_i \delta_j}{\sqrt{\sigma^2 + a\delta_i^2}\sqrt{\sigma^2 + a\delta_j^2}} \right)^k$$

and $R_m(u)/m$ has asymptotically vanishing variance if and only if $\overline{\rho^k} \to 0$ for all k. Suppose $\delta_i = 1/\sqrt{m}$ (to ensure unit length). Then one may obtain vanishing limits for example if a is constant, but not if a increases linearly with m.

2.5 Distributional properties of the FDR estimator: the negative binomial model

We now apply the above results about the number of discoveries $R_m(u)$ for finite m towards quantifying the distributional properties of the FDR estimator (3). In view of the mean-variance structure of $R_m(u)$, our strategy is to model $R_m(u)$ as a negative binomial (NB) variable and derive the properties of the FDR estimator based on this model.

Specifically, seen as a function of m, the mean-variance structure (6) resembles that of the beta-binomial distribution, which has been used to model overdispersed binomial data (Prentice, 1986; McGullagh and Nelder, 1989). Since the beta-binomial distribution is difficult to work with, we propose instead to model $R_m(u)$ using an NB distribution. The rationale is as follows. If a random variable R is binomial with number of trials n and success probability p, then when n is large and p is small, the distribution of R can be approximated by a Poisson distribution with mean parameter np. Similarly, if p has a beta distribution with mean μ, then when n is large and μ is small, the distribution of np can be approximated by a gamma distribution with mean $n\mu$. Therefore the beta-binomial mixture model can be approximated by a gamma-Poisson mixture model, which is equivalent to a NB distribution (Hilbe, 2007).

It is convenient to parametrize the NB distribution with parameters $\lambda \geq 0$ and $\omega \geq 0$, such that the mean and variance of $R \sim NB(\lambda, \omega)$ are

$$\mathrm{E}(R) = \lambda, \qquad \mathrm{var}(R) = \lambda + \omega\lambda^2. \tag{16}$$

Here λ has the interpretation of a mean parameter while ω controls the overdispersion with respect to the Poisson distribution. When $\omega = 0$, the NB distribution becomes Poisson with mean λ, and when $\lambda = 0$, it becomes a point mass at $R = 0$.

We describe how to estimate λ and ω from data in the next section. Given λ and ω, the moments, distribution, and quantiles of the FDR estimator can be obtained directly from the NB model. These are given by the following theorem. For simplicity of notation we omit the indices m and u.

Theorem 4. *Suppose R is distributed as*

$$R \sim \begin{cases} NB(\lambda, \omega), & \text{if } \lambda \geq 0,\ \omega > 0 \\ Poisson(\lambda), & \text{if } \lambda \geq 0,\ \omega = 0 \end{cases}$$

with moments (16), *cdf* $F(k) = P[R \le k]$, *and quantiles* $F^{-1}(q) = \inf\{x : P[R \le x] \ge q\}$. *Assume* p_0 *is known. Then*
(i) Probability of discovery:

$$\gamma(\lambda,\omega) \triangleq P[R>0] = \begin{cases} 1-(1+\omega\lambda)^{-1/\omega}, & \omega > 0 \\ 1-\exp(-\lambda), & \omega = 0 \end{cases} \tag{17}$$

(ii) Mean:

$$\mathrm{E}\left[\widehat{\mathrm{FDR}}\right] = p_0 m\alpha\left[1-\gamma(\lambda,\omega)+\gamma(\lambda,\omega)\zeta(\lambda,\omega)\right] \tag{18}$$

where

$$\zeta(\lambda,\omega) = \mathrm{E}\left[\frac{1}{R}\middle| R>0\right] = \int_0^\lambda \frac{\gamma^{-1}(\lambda,\omega)-1}{\gamma^{-1}(x,\omega)-1}\cdot\frac{dx}{x(1+\omega x)} \tag{19}$$

(iii) Variance:

$$\mathrm{var}\left[\widehat{\mathrm{FDR}}\right] = (p_0 m\alpha)^2\left[1-\gamma(\lambda,\omega)+\gamma(\lambda,\omega)\zeta_2(\lambda,\omega)\right] - \mathrm{E}^2\left[\widehat{\mathrm{FDR}}\right] \tag{20}$$

where

$$\zeta_2(\lambda,\omega) = \mathrm{E}\left[\frac{1}{R^2}\middle| R>0\right] = \int_0^\lambda \frac{\gamma^{-1}(\lambda,\omega)-1}{\gamma^{-1}(x,\omega)-1}\cdot\frac{\zeta(x,\omega)\,dx}{x(1+\omega x)} \tag{21}$$

(iv) Distribution:

$$P\left[\widehat{\mathrm{FDR}} \le x\right] = \begin{cases} 0, & x<0 \\ 1-F(k), & a_{k+1} \le x < a_k, \quad k=1,2,\ldots \\ 1, & a_1 \le x \end{cases} \tag{22}$$

where $a_k = p_0 m\alpha/k$.
(v) Quantiles:

$$\inf\left\{x : P\left[\widehat{\mathrm{FDR}} \le x\right] \ge q\right\} = \begin{cases} a_{k+1}, k = F^{-1}(1-q), & q \le 1-F(1) \\ a_1, & q > 1-F(1) \end{cases} \tag{23}$$

where $a_k = p_0 m\alpha/k$.

In terms of notation, the functions $\zeta(\lambda,\omega)$ and $\zeta_2(\lambda,\omega)$ above were defined so that both are bounded, tend to 0 as $\lambda \to 0$, and tend to 1 as $\lambda \to \infty$. In particular, (19) with $\omega = 0$ is the same as Equation (27) in Schwartzman (2008) divided by λ.

2.6 Estimation of distributional properties of the FDR estimator

Next we show how the parameters λ and ω can be estimated from the data. We set the parameters λ and ω of the NB model for $R_m(u)$ by the method of moments. First, based on (6) but respecting the form (16), we propose the estimates

$$\widehat{\mathrm{E}}[R_m(u)] = m\alpha(u), \qquad \widehat{\mathrm{var}}[R_m(u)] = m\alpha(u) + m^2\widehat{\overline{\Psi}}_m(u;0)$$

Here the estimate for the mean is simply the mean under the complete null. The form of the variance ensures that $\widehat{\mathrm{var}}[R_m(u)] \geq \widehat{\mathrm{E}}[R_m(u)]$ as required by the NB model. The overdispersion $\overline{\Psi}_m(u;0)$ is also estimated conservatively under the complete null. In the normal case, (13) is estimated by

$$\widehat{\overline{\Psi}}_m(u;0) = \phi^2(u) \sum_{k=1}^{K} \frac{\overline{\hat{\rho}^k}}{k!} \mathcal{H}_{k-1}^2(u),$$

where $\overline{\hat{\rho}^k}$, $k = 1, 2, \ldots, K$ denote the empirical moments of the $m(m-1)/2$ empirical pairwise correlations $\hat{\rho}_{ij}$, $i < j$, after correction for sampling variability. In practice, $K = 3$ suffices.

Matching the estimated moments of $R_m(u)$ to those of the NB model (16) leads to the parameter estimates

$$\hat{\lambda}_m(u) = m\alpha(u), \qquad \hat{\omega}_m(u) = \frac{\widehat{\overline{\Psi}}_m(u;0)}{\alpha^2(u)} \tag{24}$$

Finally, the moments and quantiles of $\widehat{\mathrm{FDR}}$ are estimated using the formulas in Theorem 4 by plugging in the parameter estimates (24).

3 Numerical studies and data examples

3.1 Numerical studies

The goal of the following simulations is to illustrate the effect of correlation on the FDR estimator and to assess the accuracy of the NB model in quantifying that effect. In the following simulations, $N = 10000$ datasets were drawn at random from the model $Y_j = \mu_j + Z_j$, $j = 1, \ldots, m$, where μ_j is the signal and Z_j has marginal distribution $N(0,1)$. The tests were set up as $H_0 : \mu_j = 0$ vs. $H_j : \mu_j > 0$. The test statistics were taken as the z-scores $T_j = Y_j$, the signal-to-noise ratio being controled by the strength of the signal. The true FDR was computed according to (2) where the expectation was replaced by an average over the N datasets. Similarly, the true moments and distribution of $R_m(u)$ and $\widehat{\mathrm{FDR}}_m(u)$ (3) were obtained from the N simulated datasets.

Example (Exchangeable correlation, complete null)**.** To obtain an exchangeable correlation structure, the Z_j were generated as $Z_j = \sqrt{\rho}Z + \sqrt{1-\rho}\varepsilon_j$ with $Z, \varepsilon_1, \ldots, \varepsilon_m$ i.i.d. $N(0,1)$. The signal was set to $\mu_j = 0$, $j = 1, \ldots, m$.

Figure 1 shows the population mean and standard deviation of the discovery rate $R_m(u)/m$ for $m = 100$ under the complete null. Dependency does not affect the mean (panel (a)) but affects the variance substantially (panel (b)). The estimated NB mean and standard deviation of $R_m(u)/m$ coincide with the true mean and standard deviation by design because the moments were matched, except that only 3 terms were used in the polynomial expansion (13) for $\overline{\Psi}_m(u)$. This shows the accuracy of this expansion. While three terms seems enough, the largest contribution by far is provided by the first term.

Figure 2 shows the effect of dependency on the FDR estimator for $m = 100$ under the complete null. First, note that the FDR estimator is always biased up, even greater than

1 for intermediate thresholds, where it is likely for the denominator to be smaller than the numerator. Dependence causes the expectation of the FDR estimator to go further up while it causes the true FDR to go down, further increasing the bias. The true FDR in this case is equal to the FWER and can be easily seen to be

$$\mathrm{FWER}_m(u) = 1 - \int_{-\infty}^{\infty} \left[1 - \Phi\left(\frac{u - \sqrt{\rho}z}{\sqrt{1-\rho}}\right)^m\right] \phi(z)\, dz = \begin{cases} 1 - [1 - \Phi(u)]^m, & \rho = 0 \\ \Phi(u), & \rho = 1 \end{cases}$$

Dependence also causes the variability of the estimator to go up dramatically, as indicated by the 5th and 95th percentiles in panel (b). Here, the ragged lines are a consequence of the discrete nature of the distribution. On both panels (a) and (b), the expectation and percentiles of $\widehat{\mathrm{FDR}}$ under dependency are surprisingly well approximated by the NB model for most thresholds.

The bias and standard error of the FDR estimator are easier to assess when plotted against the true FDR in each case, as shown in panels (c) and (d). Here we see that at practical FDR levels such as 0.2, the bias of the FDR estimator under independence is about 6% while under correlation is 60%, about 10 times larger. Similarly, the standard error under independence is about 8% while under correlation is 25%, about 3 times larger. Again, the NB model is able to capture this effect quite accurately.

Other simulations for increasing m from $m = 10$ to $m = 10000$ indicate that, when plotted against the true FDR as in panels (c) and (d), the bias and standard error of the FDR estimator increase only slightly as a function of m both under independence and correlation. This is because increasing m increases the expectation and variance of the FDR estimator for any fixed threshold, but the threshold required for controling FDR at a given level also increases accordingly.

Example (Exchangeable correlation, non-complete null)**.** Keeping the same exchangeable correlation structure as before with $m = 100$, the signal was set to $\mu_j = 2$, $j = 1, \ldots, 5$, providing a null fraction $p_0 = 0.95$. Figure 3 shows the population mean and standard deviation of the discovery rate $R_m(u)/m$. Taking into account the dependency approximates the variance better than if independence is assumed. Since the NB parameters were chosen assuming the complete null, the estimate for the mean is slightly smaller, and the estimate for the variance is slighlty smaller for high thresholds.

Figure 4 shows the effect of dependency on the FDR estimator. Panel (a) shows that the bias persists as in the complete null case. Panel (b) shows that correlation increases the variability of $\widehat{\mathrm{FDR}}$, visible in this panel mostly in terms of the 5th percentile. The NB quantile estimates assuming dependency capture the variability better than if the dependency were ignored. When plotted against the true FDR, we see in panels (c) and (d) that the NB model with complete null parameters slightly underestimates both the bias and standard error of $\widehat{\mathrm{FDR}}$ for small FDR levels.

3.2 Data Example

As a specific data example we use the genetic microarray dataset from Mootha et al. (2003). Briefly, $m = 10983$ expression levels were measured among diabetes patients. For the purposes of this article, standard two-sample t-statistics were computed at each gene between

the two groups labeled DM2 (Type 2 diabetes mellitus, $n_1 = 17$) and NGT (normal glucose tolerance, $n_2 = 17$). The t-statistics, having 17+17-2=32 degrees of freedom, were converted to the normal scale by a one-to-one quantile transformation (Efron, 2004, 2007a). Figure 5(a) shows a histogram of the $m = 10983$ test statistics. The histogram follows the $N(0, 1)$ density pretty well with mean 0.059 and standard deviation 0.977, making an empirical null unnecessary.

Figure 5(b) shows a histogram of 499500 sample pairwise correlations computed from 1000 randomly sampled genes out of $m = 10983$. The pairwise correlations were computed between the gene expression levels across all 34 subjects after subtracting the means of both groups separately. These are approximately the same as the correlations between the z-scores gven the moderate sample size of 34. The first three empirical moments of the pairwise correlations, obtained from a random sample of 2000 genes, were $\overline{\rho} = 0.0044$, $\overline{\rho^2} = 0.0846$ and $\overline{\rho^3} = 0.0020$. To correct for sampling variability, as recommended by Owen (2005) and Efron (2007a), we applied Fisher's transformation to the sample correlations, shrunk them towards zero in that domain using empirical Bayes, and transformed them back with the inverse of Fisher's transformation. This process resulted in the superimposed black histogram, with empirical moments $\overline{\rho} = 0.0029$, $\overline{\rho^2} = 0.0586$ and $\overline{\rho^3} = 0.0012$.

Figure 5(c) shows the FDR estimator as a function of the threshold u. Superimposed are the 5th and 95th percentiles of the distribution of $\widehat{\text{FDR}}$, estimated by the NB model under the complete null, both assuming indpendence and asumming the correlation structure in the data. When the dependence is taken into account, the bands are realistically wider and indicate the variability of the FDR estimate. For example, at $u = 4$, the estimated FDR is 0.17, but the bands indicate that with 90% probability it could have been as low as 0.12 or as high as 0.35. For reference, superimposed are the 5th and 95th percentiles of the empirical distribution of $\widehat{\text{FDR}}$ obtained by permuting the subject labels between the two groups, while keeping genes belonging to the same subject together. We see that the permutation estimates closely resemble those of the NB model under dependence.

4 Discussion

In this article we have developed explicit expressions for estimating the mean, variance, distribution, and quantiles of the FDR estimator based on a NB model, where the mean and variance parameters are chosen to match those of the number of discoveries under the complete null. These expressions allow and arbitrary correlation structure, requiring only the first few empirical moments of the distribution of pairwise correlations in the normal and χ^2 cases. The NB model has been found to work well for practical FDR levels and for both small and large number of tests, giving an accurate representation of the effect of correlation in FDR analysis. Moreover, we have shown that accumulated correlation can cause the FDR estimator to be inconsistent, for instance if the correlation structure is exchangeable, has a block form with blocks that increase in size, or has a factor analysis form with loadings increasing linearly with m.

The distribution of the FDR estimator is discrete and highly skewed, so rather than standard errors, as in Efron (2007a) and Schwartzman (2008), we chose to indicate the variability of the FDR estimator by its quantiles. The 5th percentile indicates how decep-

tively low the FDR estimator can be even when there is no signal. The 95th percentile is almost always equal to $m\alpha(u)$, the Bonferroni-adjusted level, showing that the FDR estimator can be sometimes as conservative as the Bonferroni method. We emphasize that the band between the 5th and 95th percentiles describe the behavior under the complete null and is not a 90% confidence interval for the true FDR, but it provides a reference for variability of the estimator under dependence when the true signal is not far from the complete null. It is interesting that in our data example, correlation played an important role but did not cause a departure from the null distribution, as may have been predicted by Efron (2007a).

The method used in this paper requires in principle few assumptions about the data. The NB parameters can be computed via (6) and (24) for any pairwise distribution of the test statistics. The main motivation for assuming the normal and χ^2 models was to reduce computations, as the Mehler and Lancaster expansions of Section 2.4 allowed reducing the correlation structure to the first few empirical moments of the pairwise correlations. A similar reduction was used by Owen (2005) and Efron (2007a). As shown in the data example, t-statistics can be handled easily by a quantile transformation to z-scores (Efron, 2007a). Similarly, F statistics can be transformed to χ^2 scores (Schwartzman, 2008).

An issue that requires further work for the NB model is the estimation of $\lambda_m(u)$ and $\omega_m(u)$ not under the complete null. Since $\lambda_m(u)$ is the expected number of discoveries, one could estimate it by $R_m(u)$, but this estimate is too noisy. In terms of $\omega_m(u)$, one could estimate the required overdispersion term using Theorem 2, but the provided expression depends on the true signal, which is unknown. Here, estimating the signal as null performs reasonably well when the signal is weak and/or p_0 is close to 1, but better approximations may be found using a highly regularized estimator of the signal vector $\boldsymbol{\mu}$. From the form of (12), it is expected that more weight would be given to pairwise correlations were the signal is large.

About the FDR estimator itself, we have learned that it is inherently biased and highly variable. The positive bias ensures FDR control, as this keeps the true FDR below the estimated one on average. But a smaller bias implies that the control is less conservative. The exchangeable correlation structure used above is a special case of positive regression dependence and therefore FDR control is guaranteed (Benjamini and Yekutieli, 2001). Its conservativeness is reflected in the large positive bias of the FDR estimator, a result of overdispersion in the number of discoveries. It is possible that a correlation structure that produced underdispersion instead would not guarantee FDR control. Because the FDR estimator was derived from the point of view of FDR control, it is also possible that better estimators might be found by approaching the problem directly as an estimation problem. This is a topic left for future work.

Appendix: Proofs

Proof of Theorem 1.
(i) Let $\mathbf{\Psi}(u)$ denote the covariance matrix of the vector $(D_1, \ldots, D_m)'$ with entries Ψ_{ij}. Define a mapping of $\mathbf{\Psi}_m$ to the unit square $[0,1)^2$ by the function

$$g_m(x,y) = \sum_{i=1}^{m}\sum_{j=1}^{m} \Psi_{ij}\ 1\left(\frac{i-1}{m} \le x < \frac{i}{m}\ \cap\ \frac{j-1}{m} \le y < \frac{j}{m}\right)$$

Then the variance (4) is equal to the integral $\int_0^1\int_0^1 g_m(x,y)\,dx\,dy$. The limit assumptions on Ψ_{ij} imply that as $m \to \infty$, $g_m(x,y) \to \Psi_\infty$ pointwise for every (x,y). Therefore, by the bounded convergence theorem, the integral converges to $\int_0^1\int_0^1 \Psi_\infty\,dx\,dy = \Psi_\infty$.
(ii) Follows immediately from Theorem 2 and the fact that $|\overline{\rho^k}|$ is bounded by 1, so the infinite sum on k always converges for every m. □

Proof of Theorem 2.
Let $\Phi(\cdot)$ and $\Phi_2(\cdot,\cdot;\rho)$ denote respectively the standard normal survival function and the standard bivariate normal survival function. Integrating (10) over the quadrant $[t_i-\mu_i,\infty]\times[t_j-\mu_j,\infty]$ gives

$$\begin{aligned}\Phi_2(t_i-\mu_i, t_j-\mu_j;\rho_{ij}) =& \Phi(t_i-\mu_i)\Phi(t_j-\mu_j)\\ &+ \phi(t_i-\mu_i)\phi(t_j-\mu_j)\sum_{k=1}^{\infty}\frac{\rho_{ij}^k}{k!}\mathcal{H}_{k-1}(t_i-\mu_i)\mathcal{H}_{k-1}(t_j-\mu_j)\end{aligned}$$

where we have used the property that the integral of $\phi(t)\mathcal{H}_k(t)$ over $[u,\infty)$ is $-\phi(u)\mathcal{H}_{k-1}(u)$ for $k \ge 1$. Replacing in (5) and then (7) gives that the overdispersion term $\overline{\Psi}_m(u)$ is (13). □

Proof of Theorem 3.
Let $F_\nu(u,v;\rho)$ denote the bivariate χ^2 distribution function with ν d.f. corresponding to the density (14). Integrating (14) over the quadrant $[u,\infty]\times[v,\infty]$ gives

$$F_\nu(u,v;\rho_{ij}) = F_\nu(u)F_\nu(v) + f_{\nu+2}(u)f_{\nu+2}(v)\sum_{k=1}^{\infty}\frac{\rho_{ij}^k}{k!}\frac{\nu^2\Gamma(\nu/2)}{k^2\Gamma(\nu/2+k)}\mathcal{L}_{k-1}^{(\nu/2)}\left(\frac{u}{2}\right)\mathcal{L}_{k-1}^{(\nu/2)}\left(\frac{v}{2}\right),$$

where we have used the property that the integral of $f_\nu(t)\mathcal{L}_k^{(\nu/2-1)}(t/2)$ over $[u,\infty)$ is $(\nu/k)f_{\nu+2}(u)\mathcal{L}_{k-1}^{(\nu/2)}(u/2)$ for $k \ge 1$. Replacing in (5) and then (7) gives that the overdispersion term $\overline{\Psi}_m(u)$ is (15). □

Proof of Theorem 4. Let $R \sim NB(r,p)$ denote the common parametrization of the NB distribution with probability mass function

$$P[R=k] = \frac{\Gamma(k+r)}{k!\Gamma(r)}p^r(1-p)^k, \qquad k=0,1,2,\ldots \tag{25}$$

where $0<p<1$ and $r \ge 0$. This parametrization is related to ours by

$$\lambda = r\frac{1-p}{p} \ge 0, \quad \omega = \frac{1}{r} > 0 \qquad \Longleftrightarrow \qquad r = \frac{1}{\omega}, \quad p = \frac{1}{1+\omega\lambda} \tag{26}$$

It is easy to show that the case $\omega = 0$, $R \sim Poisson(\lambda)$, is obtained as the continuous limit of the above NB distribuition as $\omega \to 0$ such that λ remains constant. The same is true for the moments of R, which are continuous functions of ω.
(i) From (25), $P[R > 0] = 1 - p^r = 1 - (1 + \omega\lambda)^{-1/\omega}$, which becomes $1 - \exp(-\lambda)$ in the limit as $\omega \to 0$.
(ii) For the mean, we have that

$$\mathrm{E}[\widehat{\mathrm{FDR}}] = \mathrm{E}\left[\frac{p_0 m\alpha}{R \vee 1}\right] = p_0 m\alpha\left\{P[R = 0] + \mathrm{E}\left[\frac{1}{R}\middle| R > 0\right] P[R > 0]\right\}$$

where $P[R > 0] = \gamma(\lambda, \omega)$ and $P[R = 0] = 1 - \gamma(\lambda, \omega)$ by definition. All that remains to get (18) is the conditional expectation, which for $\omega > 0$ is equal to

$$\begin{aligned}\mathrm{E}\left[\frac{1}{R}\middle| R > 0\right] &= \frac{1}{1 - p^r}\sum_{k=1}^{\infty}\frac{1}{k}\frac{\Gamma(k + r)}{k!\Gamma(r)}p^r(1 - p)^k = \frac{p^r}{1 - p^r}\sum_{k=1}^{\infty}\frac{\Gamma(k + r)}{k!\Gamma(r)}\int_p^1 (1 - t)^{k-1}\, dt \\ &= \frac{p^r}{1 - p^r}\int_p^1 \frac{dt}{(1 - t)t^r}\sum_{k=1}^{\infty}\frac{\Gamma(k + r)}{k!\Gamma(r)}t^r(1 - t)^k = \frac{p^r}{1 - p^r}\int_p^1 \frac{1 - t^r}{(1 - t)t^r}\, dt.\end{aligned}$$

Replacing (26) and making the change of variable $t = 1/(1+\omega x)$ gives (19). The case $\omega = 0$ is obtained by similar calculations for the Poisson distribution or by taking the limit of (19) as $\omega \to 0$.
(iii) For the second moment, we have

$$\mathrm{E}[\widehat{\mathrm{FDR}}^2] = \mathrm{E}\left[\left(\frac{p_0 m\alpha}{R \vee 1}\right)^2\right] = (p_0 m\alpha)^2\left\{P[R = 0] + \mathrm{E}\left[\frac{1}{R^2}\middle| R > 0\right] P[R > 0]\right\}$$

Again, to get (20), we only need the conditional expectation, which for $\omega > 0$ is equal to

$$\begin{aligned}\mathrm{E}\left[\frac{1}{R^2}\middle| R > 0\right] &= \frac{1}{1 - p^r}\sum_{k=1}^{\infty}\frac{1}{k^2}\frac{\Gamma(k + r)}{k!\Gamma(r)}p^r(1 - p)^k = \frac{p^r}{1 - p^r}\sum_{k=1}^{\infty}\frac{\Gamma(k + r)}{k!\Gamma(r)}\int_p^1 \frac{(1 - t)^{k-1}}{k}\, dt \\ &= \frac{p^r}{1 - p^r}\int_p^1 \frac{dt}{(1 - t)t^r}\sum_{k=1}^{\infty}\frac{1}{k}\frac{\Gamma(k + r)}{k!\Gamma(r)}t^r(1 - t)^k.\end{aligned}$$

Following the argument in part (ii) of the proof, the last sum above is equal to $(1 - t^r)\mathrm{E}[(1/S)|S > 0]$, where $S \sim NB(r, t)$. Then the change of variable $t = 1/(1 + \omega x)$ and replacing (26) gives (21).
(iv) For the distribution, $\widehat{\mathrm{FDR}}$ takes values on the discrete set $a_1, a_2, \ldots$ where $a_k = p_0 m\alpha/k$. If $a_{k+1} \le x < a_k$, $k = 1, 2, \ldots$, then

$$P[\widehat{\mathrm{FDR}} \le x] = P\left[\frac{p_0 m\alpha}{R \vee 1} \le \frac{p_0 m\alpha}{k + 1}\right] = P[R \ge k + 1] = 1 - P[R \le k] = 1 - F(k)$$

If $x \ge a_1 = p_0 m\alpha$, then x is greater than the largest value that $\widehat{\mathrm{FDR}}$ can take, so clearly $P[\widehat{\mathrm{FDR}} \le x] = 1$.
(v) If $q \le 1 - F(1)$, $k = F^{-1}(1 - q)$, then

$$P\left[\widehat{\mathrm{FDR}} \le a_{k+1}\right] = P\left[\widehat{\mathrm{FDR}} \le \frac{p_0 m\alpha}{k + 1}\right] = P[R \ge k + 1] = 1 - F(k) \ge q$$

because $F(k) \leq 1 - q$ by definition of k. But for all $\varepsilon > 0$,

$$P\left[\widehat{\text{FDR}} \leq a_{k+1} - \varepsilon\right] = P\left[\widehat{\text{FDR}} \leq \frac{p_0 m \alpha}{k+2}\right] = P[R \geq k+2] = 1 - F(k+1) < q$$

If $q > 1 - F(1)$,

$$P\left[\widehat{\text{FDR}} \leq a_1\right] = P\left[\widehat{\text{FDR}} \leq p_0 m \alpha\right] = P[R \vee 1 \geq 1] = 1 \geq q$$

but for all $\varepsilon > 0$,

$$P\left[\widehat{\text{FDR}} \leq a_1 - \varepsilon\right] = P\left[\widehat{\text{FDR}} \leq \frac{p_0 m \alpha}{2}\right] = P[R \geq 2] = 1 - F(1) < q$$

□

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann Statist*, 29:1165–1188, 2001.

Martin Crowder. Gaussian estimation for correlated binomial data. *J R Statist Soc B*, 47 (2):229–237, 1985.

David B. Dahl and Michael A. Newton. Multiple hypothesis testing by clustering treatment effects. *J Amer Statist Assoc*, 102(478):517–526, 2007.

Sandrine Dudoit, Juliet P. Shaffer, and Jennifer C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

Bradley Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Amer Statist Assoc*, 99(465):96–104, 2004.

Bradley Efron. Correlation and large-scale simultaneous hypothesis testing. *J Amer Statist Assoc*, 102(477):93–103, 2007a.

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

Alessio Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17(4): 347–388, 2008.

Christopher R. Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *Ann Statist*, 32:1035–1061, 2004.

Christopher R. Genovese, Nicole A. Lazar, and Thomas E. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15: 870–878, 2002.

Jaroslaw Harezlak, Michael C. Wu, Mike Wang, Armin Schwartzman, David C. Christiani, and Xihong Lin. Biomarker discovery for arsenic exposure using functional data. analysis and feature learning of mass spectrometry proteomic data. *Journal of Proteome Research*, 7(1):217–224, 2008.

Ruth Heller, Damian Stanley, Daniel Yekutieli, Nava Rubin, and Yoav Benjamini. Cluster-based analysis of fMRI data. *Neuroimage*, 33:599–608, 2007.

Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, Cambrdige, UK, 2007.

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

Lev Klebanov, Craig Jordan, and Andrei Yakovlev. A new type of stochastic dependence revealed in gene expression data. *Statist. Appl. in Gen. and Mol. Bio.*, 5(1):7, 2006.

S. Kotz, N. Balakrishnan, and N. L. Johnson. Bivariate and trivariate normal distributions. In *Continuous Multivariate Distributions, Vol. 1: Models and Applications*, pages 251–348. John Wiley & Sons, New York, 2000.

Angelo Efoévi Koudou. Lancaster bivariate probability distributions with poisson, negative binomial and gamma margins. *Test*, 7(1):95–110, 1998.

P. McGullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition, 1989.

V. K. Mootha, C. M. Lindgren, F. K. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, , M. Ridderstraale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tomayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1 -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.

Art B. Owen. Variance of the number of false discoveries. *J R Statist Soc B*, 67:411–426, 2005.

M. Perone Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman. False discovery control for random fields. *J Amer Statist Assoc*, 99(468):1002–1014, 2004.

Jagdish K. Patel and Campbell B. Read. *Handbook of the Normal Distribution*. Marcel Dekker Inc., New York, 2nd edition, 1996.

Ross L. Prentice. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J Amer Statist Assoc*, 81(394):321–327, 1986.

Xing Qiu and Andrei Yakovlev. Some comments on instability of false discovery rate estimation. *J Bioinform Comput Biol*, 4(5):1057–1068, 2006.

Xing Qiu, Lev Klebanov, and Andrei Yakovlev. Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Statist. Appl. in Gen. and Mol. Bio.*, 4(1):34, 2005.

Kathryn Roeder, Silvi-Alin Bacanu, Larry Wasserman, and B. Devlin. Using linkage genome scans to improve power of association in genome scans. *American Journal of Human Genetics*, 78:243–252, 2006.

Armin Schwartzman. Empirical null and false discovery rate inference for exponential families. *Ann Appl Statist*, 2(4):1332–1359, 2008.

Armin Schwartzman, Robert F. Dougherty, Jongho Lee, Dara Ghahremani, and Jonathan E. Taylor. Empirical null and false discovery rate analysis in neuroimaging. *Neuroimage*, 44 (1):71–82, 2009.

John D. Storey, Jonathan E. Taylor, and David Siegmund. 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.

Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J Amer Statist Assoc*, 102(479):901–912, 2007.

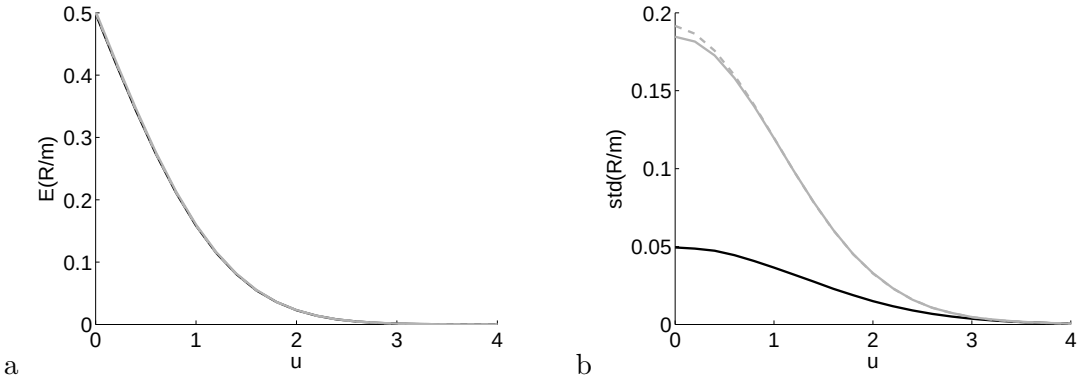
Robert Tibshirani and Larry Wasserman. Correlation-sharing for detection of differential gene expression. *http://arxiv.org/abs/math/0608061v1*, 2006.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *J R Statist Soc B*, 67(1):91–108, 2005.

Keith J. Worsley, Jonathan E. Taylor, F. Tomaiuolo, and J. Lerch. Unified univariate and multivariate random field theory. *Neuroimage*, 23:S189–195, 2004.

D. Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Inference*, 82:171–196, 1999.

Yong Zhou and Hua Liang. Asymptotic normality for l_1 norm kernel estimator of conditional median under α-mixing dependence. *J Multivariate Analysis*, 73:136–154, 2000.



a b

Figure 1: Effect of correlation on the rate of discovery $R_m(u)/m$ under the complete null for $m = 100$: (a) Expectation. (b) Standard deviation. Plotted in both panels are: the true value assuming independence (black solid), the true value assuming an exchangeable correlation structure with $\rho = 0.2$ (gray solid), and the value calculated using the polynomial expansion (13) (gray dashed).

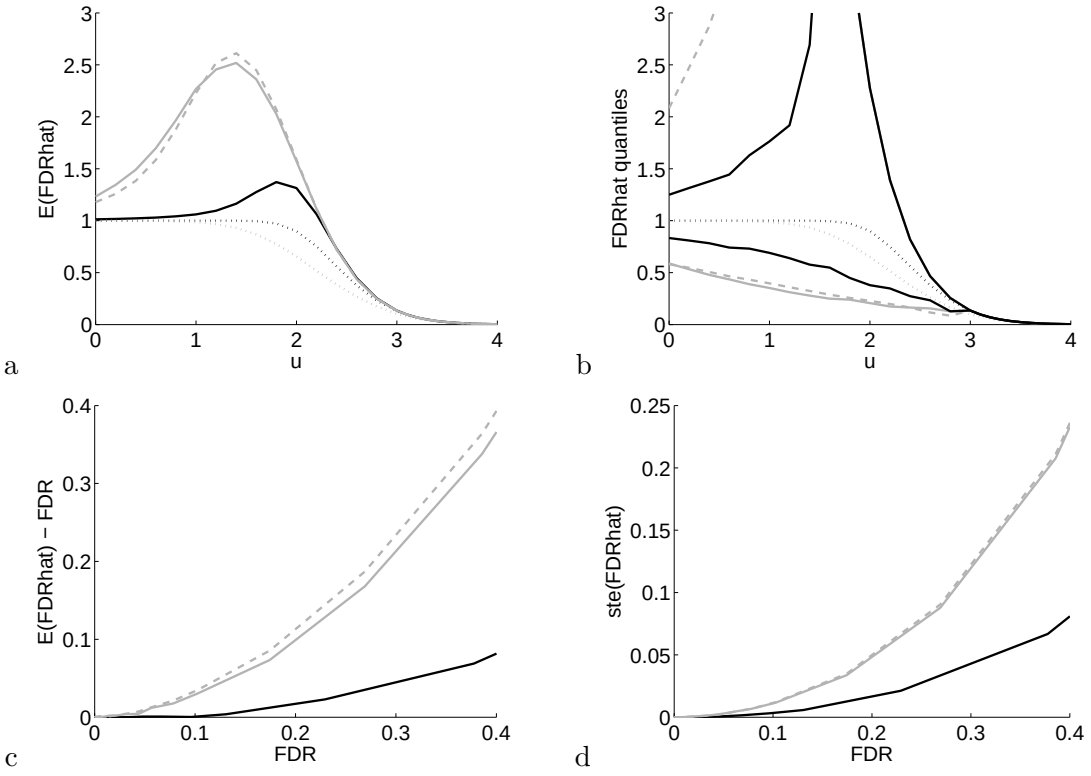


Figure 2: Effect of correlation under the complete null for $m = 100$. (a) Expectation of $\widehat{\text{FDR}}$ as a function of threshold. (b) Percentiles 5 and 95 of $\widehat{\text{FDR}}$ as a function of threshold. All the 95th percentile lines overlap at the right edge. (c) Bias of $\widehat{\text{FDR}}$ as a function of the true FDR. (d) Standard error of $\widehat{\text{FDR}}$ as a function of the true FDR. Plotted in all panels are: the true value for the estimator assuming independence (black solid), the true value for the estimator assuming an exchangeable correlation structure with $\rho = 0.2$ (gray solid), and the value estimated by the NB model (gray dashed). Also, in panels (a) and (b): true FDR assuming independence (black dotted), true FDR assuming exhangeable correlation (gray dotted).

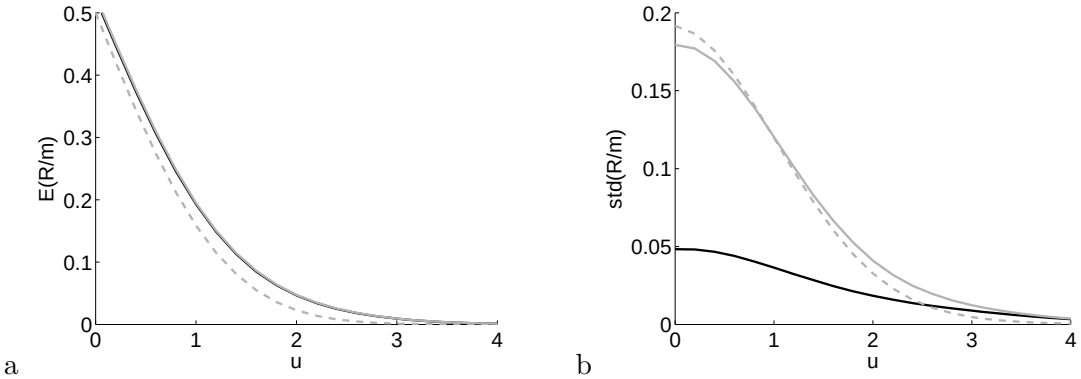


Figure 3: Effect of correlation on the rate of discovery $R_m(u)/m$ for $m = 100$ and 95% signal with $\mu = 2$. (a) Expectation. (b) Standard deviation. Plotted in both panels are: the true value assuming independence (black solid), the true value assuming an exchangeable correlation structure with $\rho = 0.2$ (gray solid), and the value calculated using the polynomial expansion (13) (gray dashed).

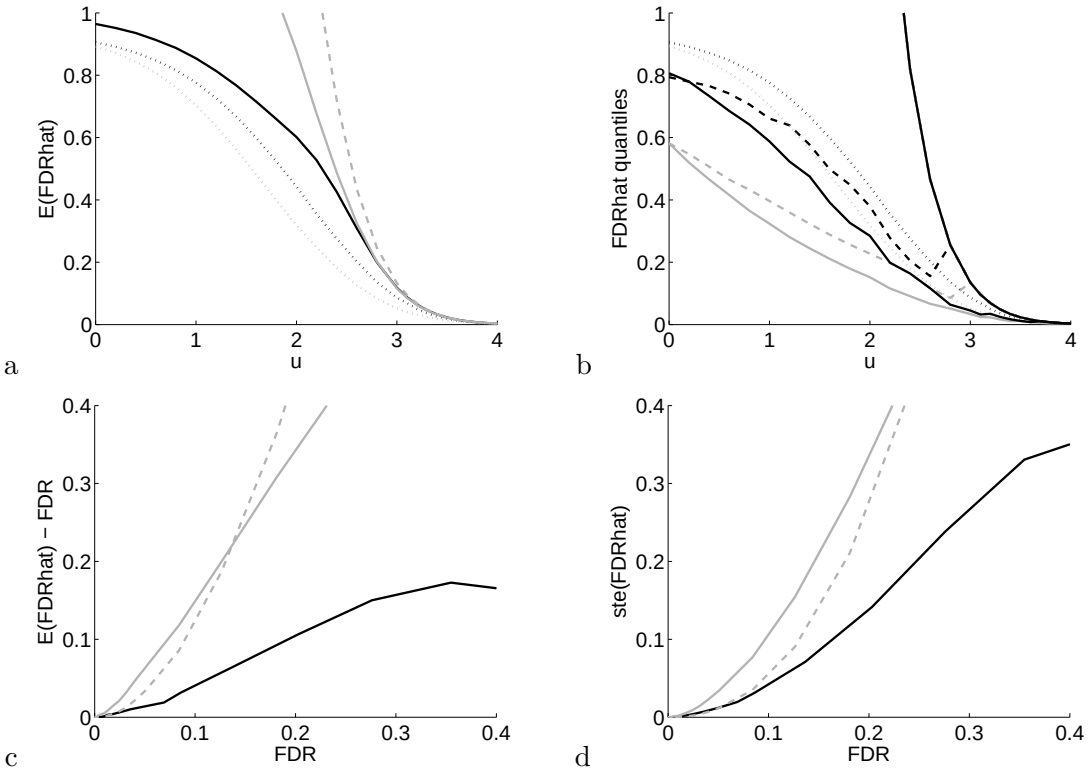


Figure 4: Effect of correlation for $m = 100$ and 95% signal with $\mu = 2$. (a) Expectation of $\widehat{\text{FDR}}$ as a function of threshold. (b) Percentiles 5 and 95 of $\widehat{\text{FDR}}$ as a function of threshold. All the 95th percentile lines overlap at the right edge. (c) Bias of $\widehat{\text{FDR}}$ as a function of the true FDR. (d) Standard error of $\widehat{\text{FDR}}$ as a function of the true FDR. Plotted in all panels are: the true value for the estimator assuming independence (black solid), the true value for the estimator assuming an exchangeable correlation structure with $\rho = 0.2$ (gray solid), and the value estimated by the NB model (gray dashed). Also, in panels (a) and (b): true FDR assuming independence (black dotted), true FDR assuming exhangeable correlation (gray dotted).

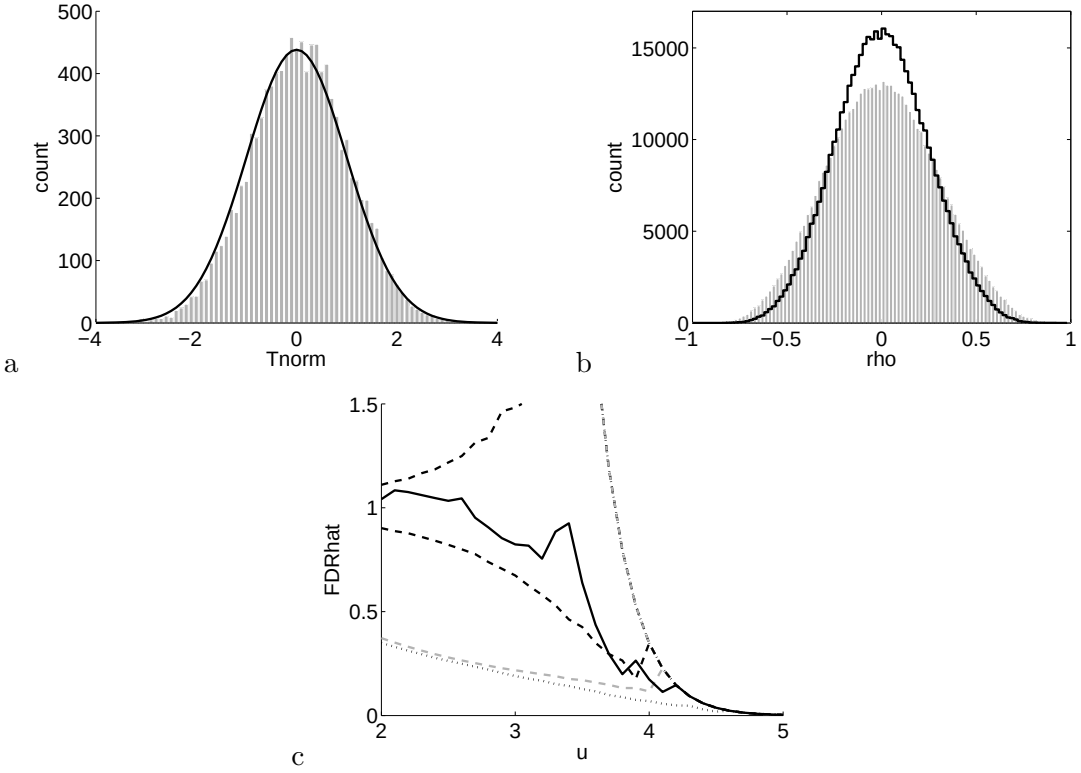


Figure 5: The diabetes microarray data. (a) Histogram of the $m = 10983$ test statistics converted to normal scale. Superimposed is the $N(0, 1)$ density. (b) Histogram of 499500 pairwise sample correlations from 1000 genes randomly sampled out of $m = 10983$. Superimposed in black: histogram corrected for random sampling. (c) FDR curves: $\widehat{\text{FDR}}$ (black solid); percentiles 5 and 95 estimated by the NB model assuming independence (black dashed); percentiles 5 and 95 estimated by the NB model assuming the pairwise correlations from the data (gray dashed); percentiles 5 and 95 estimated by permutations (black dotted). All the 95th percentile lines overlap at the right edge.