Statistica Sinica **22** (2012), 1689-1716
doi:http://dx.doi.org/10.5705/ss.2010.255

ESTIMATING THE PROPORTION OF TRUE NULL HYPOTHESES UNDER DEPENDENCE

Irina Ostrovnaya and Dan L. Nicolae

Memorial Sloan-Kettering Cancer Center and The University of Chicago

Abstract: Multiple testing procedures, such as the False Discovery Rate control, often rely on estimating the proportion of true null hypotheses. This proportion is directly related to the minimum of the density of the p-value distribution. We propose a new estimator for the minimum of a density that is based on constrained multinomial likelihood functions. The proposed method involves partitioning the support of the density into several intervals, and estimating multinomial probabilities that are a function of the density. The motivation for this new approach comes from multiple testing settings where the test statistics are dependent, since this framework can be extended to the case of the dependent observations by using weighted univariate likelihoods. The optimal weights are obtained using the theory of estimating equations, and depend on the discretized pairwise joint distributions of the observations. We discuss how optimal weights can be estimated when the test statistics have multivariate normal distribution, and their correlation matrix is available or estimated. We evaluate the performance of the proposed estimator in simulations that mimic the testing for differential expression using microarray data.

Key words and phrases: Constrained multinomial likelihood, correlated tests, FDR, minimum of a density, multiple testing under dependence, proportion of null hypotheses.

1. Introduction

The proportion of true null hypotheses, denoted throughout the paper as π_0, has been shown to be an important quantity in many multiple comparison procedures. When the number of tests is large, this proportion is the basis of calculation of almost any total error measure. For example, the false discovery rate (FDR), defined as the expected proportion of falsely rejected hypotheses at a given threshold of significance, is a function of π_0 (Storey (2002)). The various modifications of the Benjamini-Hochberg sequential procedure use π_0 for identification of the rejection threshold that will guarantee the total error rate below α (see Benjamini and Hochberg (1995); Genovese and Wasserman (2001)). Also, π_0 can be a quantity of interest itself in a number of areas, for example, astrophysics (Miller et al. (2001)).

Multiple comparison procedures are often used in genetic association studies and microarray experiments, where the test statistics might be correlated. For example, expression microarrays are used for identification of signaling pathways or tissue classification. Since genes in the same genetic pathways will generate correlated expression values, the test statistics are not independent. In genetic association studies, for example (Hoffjan et al. (2004, 2005)), multiple associations between related phenotypes, genetic markers, and environmental covariates are tested, resulting in dependence between the test statistics. As a consequence of dependence the effective null distribution may appear considerably wider or more narrow (Efron (2007)). That is, correlations may produce more or fewer extreme test statistics than expected, even though the null marginal distributions still hold. Also, since many error measures involve expectations of sums of random variables, the variance of the error estimates may significantly increase. The goal of this paper is to construct an estimator of the proportion of true null hypotheses, π_0, that takes into account the dependence.

The commonly used estimator of π_0 (Storey (2002)) for the independent data is

$$\hat{\pi}_0^S(\lambda) = \frac{\#\{\text{p-values} > \lambda\}}{(1-\lambda)N}, \tag{1.1}$$

where λ is a tuning parameter selected by a bootstrap procedure and N is a total number of tests. In Storey and Tibshirani (2003) a cubic spline is fitted to (1.1), and the estimate at $\lambda = 1$ is used. Other methods proposed in this context include Poisson regression based estimation of the density of the test statistics (Efron (2004)), parametric models for the distribution of the test statistics under the alternative hypothesis (e.g., Pounds and Morris (2003); Allison et al. (2002); Markitsis and Lai (2010)), and methods combining parametric models and splines (Ruppert, Nettleton, and Hwang (2007)). Langaas, Lindqvist, and Ferkingstad (2005) review several available methods and compare them to their own nonparametric MLE method with convexity and monotonicity constraints.

Literature on estimating π_0 for dependent data is much more scarce. Some authors argue that the methods created for the independent data will work for dependent data as well, as long as dependence is weak or moderate (Langaas, Lindqvist, and Ferkingstad (2005); Benjamini and Yekutieli (2001); Farcomeni (2007)). Sun and Cai (2009) offer an FDR control procedure with the built-in estimate of π_0 for a special case when the data come from a two-state hidden Markov model. Many multiple comparison procedures developed for dependent data either use estimates of π_0 under independence (Hunt, Cheng and Pounds (2009)), or conservatively assume $\pi_0 = 1$ (Efron (2007); Chen, Tong, and Zhao (2008)).

Two articles approach estimation of π_0 under dependence more directly. Singular value decomposition of the distribution of the test statistics is used by Pawitan, Calza and Ploner (2006) to characterize the dependence structure, which is estimated using permutations. The expected false discovery proportion is calculated using a latent FDR model, and π_0 is estimated as one of the parameters of the model. Lu and Perkins (2007) use permutations to adjust for variability in π_0 brought by dependence. They resample cases and controls for each gene separately and calculate p-values and π_0 using methods created for independent data. The 3rd quartile of the resampled π_0 distribution serves as a final estimate of π_0. We compare these two methods, as well as Langaas' method created for independent data, to the proposed estimator in simulations in Section 4.

The main idea behind estimation of π_0 is that the p-value density can be modeled as a mixture of two densities with mixing proportion π_0. The component corresponding to the null hypothesis is the uniform distribution with the density equal to 1. Let $f_A(x)$ be the unknown density of the p-values generated under the alternative hypotheses. Then the p-value density is

$$f(x) = \pi_0 + (1 - \pi_0) f_A(x), \text{ for any } x \in [0, 1].$$

The problem can be equivalently formulated in terms of the distribution of the test statistics, but the p-value density is attractive because it usually has a similar shape regardless of the test used. The p-values coming from the alternative hypotheses are expected to be small, thus, it is reasonable to assume that $f(x)$ is a decreasing function. Additionally, in most applications the density is convex. Thus, the minimum of the density is achieved at 1 and $f(1) = \pi_0$ if $f_A(1) = 0$ (the only case when π_0 is identifiable). This reduces the task to estimating the minimum of non-increasing convex density. We propose to use a constrained multinomial likelihood on data obtained by partitioning the support of the density. Then we extend the model to the dependent case - we use a weighted marginal likelihood with optimal weights defined by the correlation structure of the test statistics.

In Section 2 we present a framework for estimating π_0 based on multinomial likelihood, which is modified for dependent data in Section 3. The performance of the method is evaluated using simulations in Section 4, and finally in Sections 5 and 6 we present an application of our method to a gene expression dataset and discussion.

2. Proposed Method for Independent Test Statistics

The goal of this section is to introduce the framework for estimating the minimum of a monotone density with bounded support assuming independence. We describe it for any sample of iid random variables $Z_1, \ldots, Z_N$ having density

$f(x)$ and cdf $F(x)$. Since we are motivated by estimating π_0, we assume the density is non-increasing and convex, the support is $[0,1]$, and the value of the minimum is of interest. However, the method can be easily modified for any monotone density on a bounded support.

2.1. The multinomial likelihood

We introduce a new function, g, determined solely by $f(x)$ or $F(x)$, which is used for the multinomial modeling of the data:

$$g(s) = \frac{1 - F(s)}{1 - s} = \frac{\int_s^1 f(x)dx}{1 - s}, \qquad 0 \le s \le 1. \tag{2.1}$$

Figure 1 shows a simple example of the shape of g. For this plot, $f(x)$ is a mixture of a Uniform(0,1) density and a density of p-values based on one-sided normal test with $\sigma^2 = 1$ and difference in means equal to 2. Note that $g(0) = 1$ and g carries many properties of f: g is non-increasing and convex when f is non-increasing and convex (see the first theorem in the Appendix), and both g and f converge to the same value at the point 1,

$$g(1-) = \lim_{s \to 1} g(s) = \lim_{s \to 1} \frac{-f(s)}{-1} = f(1-).$$

Thus, the estimation of π_0 can be accomplished by finding the minimum of the function g instead of f.

In order to construct a multinomial model using g, let k be a positive integer and consider the partition of the unit interval into k subintervals

$$0 = t_0 < t_1 < t_2 < \cdots < t_k = 1.$$

Although the final goal is to estimate $g(1)$, instead we estimate the whole function g on the grid defined by the partition, so $g_i = g(t_i)$, $i = 1, \ldots, k-1$, are the parameters of interest. Let $g_0 := g(t_0) = 1$. For $i = 1, \ldots, k$ we define

$$\theta_i = \int_{t_{i-1}}^{t_i} f(s)ds = (1 - t_{i-1})g_{i-1} - (1 - t_i)g_i,$$

and it is easy to see that $\sum_{i=1}^k \theta_i = 1$. Each θ_i is the probability that an observation falls into the ith interval. There is a one-to-one correspondence between $(g_1, \ldots, g_{k-1})$ and $(\theta_1, \ldots, \theta_k)$, and the inverse is given by

$$g_i = \frac{\sum_{j=i+1}^k \theta_j}{(1 - t_i)}, \quad i = 1, \ldots, (k-1).$$

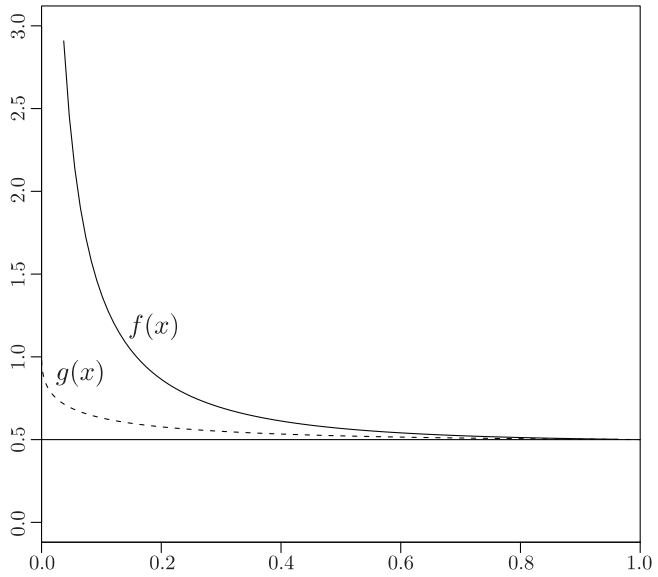

Figure 1. Example of $f(x)$ and $g(x)$. The function f is a mixture of U(0,1) density and density of p-values based on the one-sided normal test with $\sigma^2 = 1$ and difference in means equal to 2. The mixing proportion is equal to 0.5.

Let X_i be the number of observations in the interval $(t_{i-1}, t_i]$. Then $X = (X_1, \ldots, X_k)$ follows a multinomial distribution with $N = \sum_{i=1}^{k} X_i$, and probabilities $\theta = (\theta_1, \ldots, \theta_k)$. The multinomial log-likelihood function,

$$l(g; X) \sim \sum X_i \log[(1 - t_{i-1})g_{i-1} - (1 - t_i)g_i], \tag{2.2}$$

is maximized subject to the shape constraints on the function g. The discretization of the support allows the construction of a parametric likelihood without imposing a parametric model on the distribution of the data.

The value $g_{k-1} = g(t_{k-1})$ is the closest to $g(1)$, and thus our estimate of the minimum of the density is

$$\hat{\pi}_0 = \hat{g}_{k-1}. \tag{2.3}$$

In this formulation of the problem, it is not possible to estimate $g(1)$ directly because there are no data points to the right of 1. Since we assume that the density is non-increasing, $g_{k-1} \geq \pi_0$ and the estimator is possibly biased. In the multiple testing context, most of the large p-values correspond to the null hypotheses, so the density is often nearly flat around 1. It gets flatter as π_0 increases, so the bias in most applications is small, especially when t_{k-1} is chosen to be close to 1. Note that a simple estimator of $g(1)$ can be obtained using linear

interpolation of g_{k-2} and g_{k-1}, but this does not dramatically change the results shown in this paper.

2.2. Constrained optimization of the likelihood

We impose a monotonicity and convexity constraint on the function f and, therefore, g. The constraints can be written as:

$$1 \geq g_1 \geq \cdots \geq g_{k-1} \geq 0, \tag{2.4}$$

$$\begin{aligned} \beta_1 &:= \frac{1-g_1}{t_1} - \frac{g_1-g_2}{t_2-t_1} \geq 0, \\ \text{for } i = 2, \ldots, k-2, \qquad \beta_i &:= \frac{g_{i-1}-g_i}{t_i-t_{i-1}} - \frac{g_i-g_{i+1}}{t_{i+1}-t_i} \geq 0, \\ \beta_{k-1} &:= \frac{g_{k-2}-g_{k-1}}{t_{k-1}-t_{k-2}} \geq 0. \end{aligned} \tag{2.5}$$

Again there is a one to one correspondence between parameters β_i and g_i:

$$g_1 = 1 - t_1 \sum_{j=1}^{k-1} \beta_j, g_i = 1 - [\sum_{j=1}^{i-1} \beta_j t_j + (\beta_i + \beta_{i+1} + \ldots + \beta_{k-1}) t_i].$$

The log-likelihood (2.2) is maximized as a function of β_i's with respect to the constraints (2.4) and (2.5). The MLE is computed numerically using the quasi-Newton method with box constraints built in the procedure *optim* in R.

The proposed estimator is based on the constrained likelihood maximization, but many of the properties of the unconstrained likelihood can be used to prove the asymptotic results for a fixed partition, including consistency and asymptotic normality (See the Appendix for results and proofs).

The partition of the $[0, 1]$ interval plays the role of a tuning parameter in this method. The bigger is the number of intervals, k, the smoother is the estimated curve of g, but since the number of observations in each interval gets smaller, the variance of each $\hat{g}_i$ increases. There is a similar bias-variance trade-off in the choice of the last interval - the closer is t_{k-1} to 1, the smaller is the bias, but the variance increases.

2.3. Comparison to the other available methods and simulations

There are connections between the proposed method and other methods available for estimating π_0. Many other procedures use the fact that the problem of estimating π_0 is related to estimating the non-increasing p-value density, and some also include the convexity assumption (e.g., Langaas, Lindqvist, and Ferkingstad (2005)). Several methods also partition the $[0, 1]$ interval and model the

distribution of the number of p-values per interval (e.g., Efron (2004)), and some methods are related to the function $g(x) = (1 - F(x))/(1 - x)$ (Wu, Guan and Zhao (2006)). Note also that the estimator (1.1) is $\hat{g}(\lambda)$, where g is estimated directly from the empirical cdf of F.

We ran simulations to compare the performance of the proposed estimator with four partitions: $k = 6$ $(0.2, 0.4, 0.6, 0.8, 0.95, 1)$, $k = 11$ $(0.1, 0.2, 0.3, \ldots, 0.9, 0.95, 1)$, $k = 15$, and $k = 30$ (equally spaced intervals ending with 0.95,1). Note that the last interval is the same for all partitions. We compared these to three other methods. The first method is the smoothing spline based on the function (1.1) (Storey and Tibshirani (2003)). The second method is described in Efron (2004) and is based on partitioning the $[0, 1]$ interval and fitting the Poisson distribution to the number of p-values in the intervals. The third method is described in Langaas, Lindqvist, and Ferkingstad (2005) and is considered to be state-of-the-art (Ruppert, Nettleton, and Hwang (2007)).

Two types of distributions are considered. Both are densities of one-sided p-values from the normal test, where the null test statistics are drawn from the standard normal distribution. For the first scenario the alternative distribution is normal with mean $\mu_A = 2$ and variance 1. For the second scenario the variance is still 1, but μ_A is drawn from $N(2, 0.75^2)$ and constrained from below by 1. This scenario is more realistic since different hypotheses have different signal intensities. There were N=1,000 observations in both situations, and the mean squared errors(MSE) were calculated based on 1,000 replicates. The simulation was run for $\pi_0 = 0.1, 0.5$ and 0.9.

Table 1 contains MSE, bias, and variance of the estimators, with the true value of the parameter taken to be $f(1) = \pi_0$. Table 1 shows that the performance of the proposed method improves as the number of the intervals increases, but the change in MSE is very small. This is one of the advantages of the proposed estimator - its performance is not affected greatly by the choice of the partition as long as the number of tests is much larger than the number of intervals. In part that is because we are only interested in the estimate of one point of the curve, so the bias-variance trade-off in the choice of partition does not play a big role.

Note that when μ_A is fixed the proposed estimator had negative bias for $\pi_0 = 0.5, 0.9$, but not for the $\pi_0 = 0.1$. We believe that the reason is that for higher π_0 the density is nearly flat near 1, and the decreasing and convex assumptions are pushing the estimates down. The scenario with $\pi_0 = 0.1$, while unrealistic in practice, was included here to demonstrate that the bias depends on the shape of the curve.

When μ_A's are variable and can be as low as 1, it is harder to estimate π_0, and the estimate has larger bias and variance. Scenarios with smaller π_0 are

Table 1. Estimators of π_0 assuming independence. Comparison of the proposed method with four partitions ($k = 6, 11, 15, 30$), and three other methods available in the literature (Storey and Tibshirani (2003), Efron (2004), Langaas, Lindqvist, and Ferkingstad (2005)).

		Proposed method						
		$k = 6$	$k = 11$	$k = 15$	$k = 30$	Storey -spline	Efron	Langaas
π_0	Normal test statistics - $\mu_A = 2$							
0.1	MSE	0.0006	0.0007	0.0006	0.0006	0.0011	0.0236	0.0003
	Var	0.0006	0.0007	0.0006	0.0005	0.0005	0.0009	0.0003
	Bias	0.0003	0.0005	0.0021	0.0052	-0.0242	-0.1507	0.0064
0.5	MSE	0.0014	0.0012	0.0012	0.0010	0.0028	0.0050	0.0012
	Var	0.0014	0.0012	0.0012	0.0010	0.0027	0.0008	0.0012
	Bias	-0.0058	-0.0032	-0.0024	-0.0007	-0.0125	-0.0646	-0.0037
0.9	MSE	0.0016	0.0012	0.0012	0.0011	0.0043	0.0049	0.0012
	Var	0.0013	0.0011	0.0010	0.0009	0.0043	0.0017	0.0010
	Bias	-0.0158	-0.0139	-0.0134	-0.0134	-0.0044	-0.0560	-0.0122
	Normal test statistics - $\mu_A = \max(1, N(2, 0.75^2))$							
0.1	MSE	0.0013	0.0013	0.0013	0.0016	0.0008	0.0042	0.0013
	Var	0.0009	0.0008	0.0008	0.0008	0.0007	0.0009	0.0005
	Bias	0.0189	0.0225	0.0242	0.0287	-0.0054	-0.0578	0.0285
0.5	MSE	0.0018	0.0015	0.0015	0.0015	0.0027	0.0020	0.0015
	Var	0.0017	0.0014	0.0013	0.0013	0.0027	0.0009	0.0013
	Bias	0.0086	0.0123	0.0133	0.0152	-0.0030	-0.0324	0.0115
0.9	MSE	0.0015	0.0012	0.0011	0.0010	0.0045	0.0047	0.0011
	Var	0.0014	0.0011	0.0010	0.0009	0.0045	0.0020	0.0011
	Bias	-0.0112	-0.0091	-0.0084	-0.0081	-0.0019	-0.0513	-0.0071

more affected by it. The bias was positive for $\pi_0 = 0.5$ since the density is not as flat near 1 as for fixed μ_A, but otherwise the results were similar. For most scenarios our estimator outperformed Efron's and Storey's estimators, but has a similar performance when compared to Langaas' method. This is not surprising since both of these methods rely on nonparametric estimators of decreasing and convex densities.

3. Extension to the Dependent Data

The main strength of our procedure is that it is likelihood-based and can be extended to the dependent case. The basic idea is that the correlated observations carry less information and should be used with less confidence, or down-weighted in comparison with the independent observations. As before, we formulate the method for general dependent identically distributed random

variables $Z_1, \ldots, Z_N$ from a density with support $[0, 1]$. Denote the jth interval of the partition as I_j. Then the number of observations in the jth interval is $X_j = \sum_{i=1}^{N} 1_{Z_i \in I_j}$. The multinomial log-likelihood (2.2) can be written as the sum of the individual contributions of observations:

$$l(\theta, X) \sim \sum_{j=1}^{k} X_j \log(\theta_j) = \sum_{i=1}^{N} \sum_{j=1}^{k} 1_{Z_i \in I_j} \log(\theta_j). \tag{3.1}$$

If the observations are dependent, this function is no longer the true log-likelihood. In general, it is not possible to specify the joint distribution of all the observations, especially if N is large; approximations to the likelihood can be used instead. We propose to use the marginal likelihoods that are weighted according to how strongly the observation is correlated with the rest of the sample. The weighted log-likelihood becomes

$$l_w(\theta, X) \sim \sum_{i=1}^{N} [w_i \sum_{j=1}^{k} 1_{Z_i \in I_j} \log(\theta_j)], \tag{3.2}$$

where w_i are the weights.

The function (3.2) is a special case of the composite likelihood introduced by Lindsay (1988). This idea is motivated by the following fact. It has been suggested (Cox and Reid (2004)) that when full likelihood is not available one can use all possible conditional distributions of one component given another. In the case of the multivariate normal distribution, using all possible conditional distributions is equivalent to using a likelihood function obtained by a weighted average of the marginal univariate distributions, where the weights are a function of the correlation matrix. The advantages of this approach are that the weighted log-likelihood is still a simple function, and we do not need to specify any pairwise or higher order joint distributions. Asymptotic properties of composite likelihoods based on marginal likelihoods are investigated in Cox and Reid (2004).

3.1. Derivation of optimal weights

Our goal is to find weights for (3.2) that correspond to 'optimal' estimating equations in the class of scores of weighted marginal likelihoods. For notational convenience, denote the number of free parameters by $m = k - 1$. If we had no constraints, we would differentiate the weighted log likelihood (3.2) with respect to θ_j, j=1, $\ldots$, m, and solve the system of equations

$$\sum_{i=1}^{N} w_i e_{ij} = 0, \text{ where } e_{ij} = \frac{1_{Z_i \in I_j}}{\theta_j} - \frac{1_{Z_i \in I_k}}{\theta_k}, \quad j = 1, \ldots, m. \tag{3.3}$$

Each e_{ij} $(i = 1, \ldots, N,\ j = 1, \ldots, m)$ represents a derivative of the log-likelihood of the ith observation with respect to θ_j, and $E(e_{ij}) = 0$. Functions (3.3) are examples of estimating equations, functions of the data and parameters that have expectation 0, and they share a lot of properties with the derivatives of the true likelihood.

Let $e = \{e_{11}, e_{12}, \ldots, e_{1m}, e_{21}, e_{22}, \ldots, e_{2m}, \ldots, e_{N1}, e_{N2}, \ldots, e_{Nm}\}^T$ be the mN-dimensional vector of e_{ij}'s. Consider an $mN \times m$ weight matrix α and an m-dimensional vector $Q = \alpha^T e$ of unbiased estimating equations. Define the class $\mathcal{Q}$ of all estimating equations in which $\alpha^T = (W_1 W_2 \cdots W_N)$ consists of the diagonal $m \times m$ block matrices W_i with the value w_i on the diagonal, and $\sum_{i=1}^{N} w_i = 1$. Thus, only one weight w_i per observation is allowed. Denote the class of all the matrices α by $\mathcal{A}_d$.

Following Heyde (1997), the analog of Fisher's information matrix for estimating equations is

$$\mathcal{E}(Q) = (E\dot{Q})^T E(QQ^T)^{-1}(E\dot{Q}), \tag{3.4}$$

where $\dot{Q}$ denotes the gradient of Q. Estimating equation Q^* is called Loewner-optimal in a class $\mathcal{Q}$ if, for any other estimating equation $Q \in \mathcal{Q}$, $\mathcal{E}(Q^*) - \mathcal{E}(Q)$ is non-negative definite (nnd). If Q^* is Loewner-optimal, it is closest to the true score function in terms of minimizing the dispersion distance.

Loewner-optimal weights α^* have to satisfy $\mathcal{E}(\alpha^{*T} e) - \mathcal{E}(\alpha^T e)$ is nnd for all $\alpha \in \mathcal{A}_d$. Using algebraic derivations and properties of nnd matrices, it can be shown that this is equivalent to

$$\alpha^T S \alpha - \alpha^{*T} S \alpha^* \text{ is nnd}, \tag{3.5}$$

where S is an $mN \times mN$ covariance matrix of e. Thus, maximizing the information matrix is equivalent to minimizing the covariance matrix.

We can think about the covariance matrix S as consisting of $N \times N$ blocks S_{ij}, where each S_{ij} is an $m \times m$ matrix describing the association between ith and jth observations. For a given pair of observations (i, j) and a pair of intervals (a, b), let

$$\gamma_{ij,ab} = \frac{P(Z_i \in I_a, Z_j \in I_b)}{P(Z_i \in I_a) P(Z_j \in I_b)} = \frac{P(Z_i \in I_a, Z_j \in I_b)}{\theta_a \theta_b}. \tag{3.6}$$

The off-diagonal blocks S_{ij} are symmetric square matrices and contain the elements $(a, b = 1, \ldots, m)$

$$S_{ij,ab} = E(e_{ia} e_{jb}) = \gamma_{ij,ab} - \gamma_{ij,ak} - \gamma_{ij,kb} + \gamma_{ij,kk}.$$

The diagonal matrices S_{ii} are all equal and depend only on θ_a, $a = 1, \ldots, m$. If ith and jth observations are independent, $\gamma_{ij,ab} = 1$, $S_{ij,ab} = 0$, and S_{ij} is a zero matrix.

There is no guarantee that an optimal α^* satisfying condition (3.5) exists. However, according to the Theorem 2.2 from Heyde (1997), if a Loewner-optimal estimating equation exists, it coincides with the tr-optimal equation. The estimating equation is tr-optimal if the trace of its information matrix is the largest compared to the other estimating equations in the class. Note that it is much easier to minimize the trace, since it is a scalar function, while we would have to minimize in the space of matrices to verify Loewner optimality.

Thus, instead of solving (3.5) we choose to minimize the function $\mathrm{tr}(\alpha^T S \alpha)$. Indeed, let the vector of weights be $w = (w_1, \ldots, w_N)$. Let R be $N \times N$ matrix composed of traces of block matrices of S, $R_{ij} = tr(S_{ij})$. Since S is nnd as a covariance matrix, R is also nnd, see Theorem A.3 in the Appendix. Then

$$tr(\alpha^T S \alpha) = w^T R w. \tag{3.7}$$

In order to obtain the optimal weights we minimize the quadratic form $w^T R w$ subject to the constraint that w_i's sum up to 1. By differentiating the Lagrangian we find to the following solution:

$$w_i^* = \frac{\sum_{j=1}^N R_{ij}^{-1}}{\sum_{j,l=1}^N R_{jl}^{-1}}. \tag{3.8}$$

For each observation we have found the optimal weight that guarantees that the trace of the corresponding covariance matrix of estimating equations is the smallest. The weights (3.8) may or may not be positive. In our simulations where all the correlations are positive, (3.8) often provides positive weights; however, we do allow negative weights in our calculations. We see negative weights as compensation for overly high weights of some other correlated statistics. Notice that the weights depend on the partition t. Negative weights appear when the diagonal elements R_{ii} are not substantially higher than the off-diagonal elements R_{ij}, all of them dependent on t. We have noticed in simulations (not shown) that as number of intervals k grows, negative weights become rare and eventually disappear. Since the performance of the π_0 estimate does not change significantly with higher k, see Section 4, we prefer to use smaller k for computational efficiency even when weights are negative.

3.2. Some properties of the optimal weights

Consider an extreme case. If $Z_i \equiv Z_j$ and the two statistics i and j are perfectly dependent, then $S_{ij} = S_{ii}$ and $R_{ij} = R_{ii}$. Since Z_i and Z_j have equal

correlation with all other observations, $R_{il} = R_{jl}$ for any $l \neq i, j$. As a result, R is singular. However, if the Moore-Penrose generalized inverse of matrix R is used, the optimal weights are one half the weight of an independent observation. Similarly, for the block of size n of perfectly dependent statistics, the optimal weights are $1/n$ of the weight of an independent observation.

If $N_0 < N$ observations are dependent according to some correlation matrix Σ_0 and independent of all other observations, then, regardless of N and the dependence structure of the other observations, the submatrix of R and, therefore, R^{-1}, corresponding to these N_0 observations, is the same. This implies that the weights of these N_0 dependent observations, divided by the weight of an independent observation, depend only on Σ_0 and not on N or the rest of the covariance matrix.

Notice that the criteria for optimality of weights does not involve any constraints. The weights are designed to represent the dependence between observations rather than any knowledge about their distribution. Assumptions about the density are used only in the maximization of the weighted likelihood.

3.3. Optimal weights in practice

Here we consider how the optimal weights can be calculated when the full distribution is known, or estimated from the data. Suppose that the test statistics T_i, $i = 1, \ldots, N$, have a multivariate normal distribution $MVN(\mu, \Sigma)$, where $\mu_i = \mu_0$ if the test statistic comes from the null hypothesis, $\mu_i = \mu_A$ otherwise, and diagonal values of Σ are equal to σ^2. The proportion of the true null hypotheses is π_0. Let Φ_μ be the cdf of the corresponding $N(\mu, \sigma^2)$ distribution.

The one-sided p-value is $p_i = 1 - \Phi_{\mu_0}(T_i)$. Let $t'_a = \Phi^{-1}_{\mu_0}(1 - t_a)$. The multinomial probabilities θ_a, $a = 1, \ldots, k$, are

$$\theta_a = P(p_i \in I_a) = \pi_0(t_a - t_{a-1}) + (1 - \pi_0)(\Phi_{\mu_A}(t'_{a-1}) - \Phi_{\mu_A}(t'_a)). \tag{3.9}$$

The joint probability of the p-values i and j falling into a pair of intervals can be calculated likewise using the mixture model. Let $\Phi_{(\mu_0,\mu_A),\Sigma_{ij}}$ be bivariate normal distribution, where Σ_{ij} is a 2×2 submatrix of Σ corresponding to the ith and jth statistics. Knowing π_0 and assuming null and alternative hypotheses are independent, we obtain

$$\begin{aligned} P(p_i \in I_a, p_j \in I_b) = {} & \pi_0\{\Phi_{(\mu_0,\mu_0),\Sigma_{ij}}(t'_{a-1}, t'_{b-1}) - \Phi_{(\mu_0,\mu_0),\Sigma_{ij}}(t'_a, t'_b)\} \\ & + (1-\pi_0)\{\Phi_{(\mu_A,\mu_A),\Sigma_{ij}}(t'_{a-1}, t'_{b-1}) - \Phi_{(\mu_A,\mu_A),\Sigma_{ij}}(t'_a, t'_b)\}. \end{aligned} \tag{3.10}$$

The optimal weights (3.8) can be easily calculated as functions of the marginal probabilities (3.9) and joint probabilities (3.10). Alternatively, other bivariate

distributions can be used in place of $\Phi_{(\mu_0,\mu_A),\Sigma_{ij}}$. The formulas for the two-sided test can be derived similarly.

Consider microarray data for which the parameters of the distribution of the test statistics are not known, and suppose there are n_1 cases and n_2 controls. Let $X_{\cdot i} = X_{1i}, \ldots, X_{Ni}$ be the overall intensities corresponding to N genes observed on the microarray from the ith person of the cases group. Similarly, $Y_{\cdot i} = Y_{1i}, \ldots, Y_{Ni}$ are the expression levels obtained from the ith person of the control group. Vectors $X_{\cdot i}$'s are independent for any person i, and so are the $Y_{\cdot i}$'s. The controls are independent of the cases. Suppose the vector $X_{\cdot i}$ of ith individual's expression levels is multivariate normal distribution with mean μ_x, variance σ^2, and correlation matrix Σ, with the $Y_{\cdot i}$'s distributed likewise with mean vector μ_y.

Let $\bar{X}_{j\cdot}$ and $\bar{Y}_{j\cdot}$ be the mean intensity levels for gene j in the cases and controls groups respectively. If the variance is known, the z test statistics $T_j = (\bar{X}_{j\cdot} - \bar{Y}_{j\cdot})/\sqrt{\sigma^2/n_1 + \sigma^2/n_2}$ are multivariate normal with the same correlation matrix Σ. If the variance is not known and the t-statistics are calculated, their distribution would be a multivariate t-distribution (Krishnan (1972)). In our simulations (not shown) we concluded that for a sufficient number of degrees of freedom and small to moderate effect sizes, the distribution of the t-statistics can be approximated by the multivariate normal distribution with the same means and correlation matrix Σ as for z-statistics. These facts suggest a route to estimating weights in practice.

- Obtain the correlation matrix of the test statistics as the sample correlation matrix between vectors of gene expression intensities.
- Obtain initial estimates of π_0 and the means of the test statistics under null and alternative distributions. Alternatively, for simplicity, assume $\pi_0 = 1$.
- Plug the above estimates into (3.9) and (3.10), then calculate the matrix R and the weights (3.8).
- Maximize the weighted log-likelihood (3.2) using the estimated weights to obtain the final estimates of θ's, and then π_0.

4. Simulation Studies under Dependence

4.1. Characteristics of weights

We used normally distributed test statistics to investigate how the weights change when dependence block size or correlation change. Consider the situation where $N = 100$, half of the observations are independent and the other half have pairwise correlations ρ. Suppose also that $\mu_0 = 0$, $\mu_A = 2$, and $\sigma = 1$. Let $\pi_0 = 0.5$. Since weights sum up to 1, we compare the ratios of the weights for

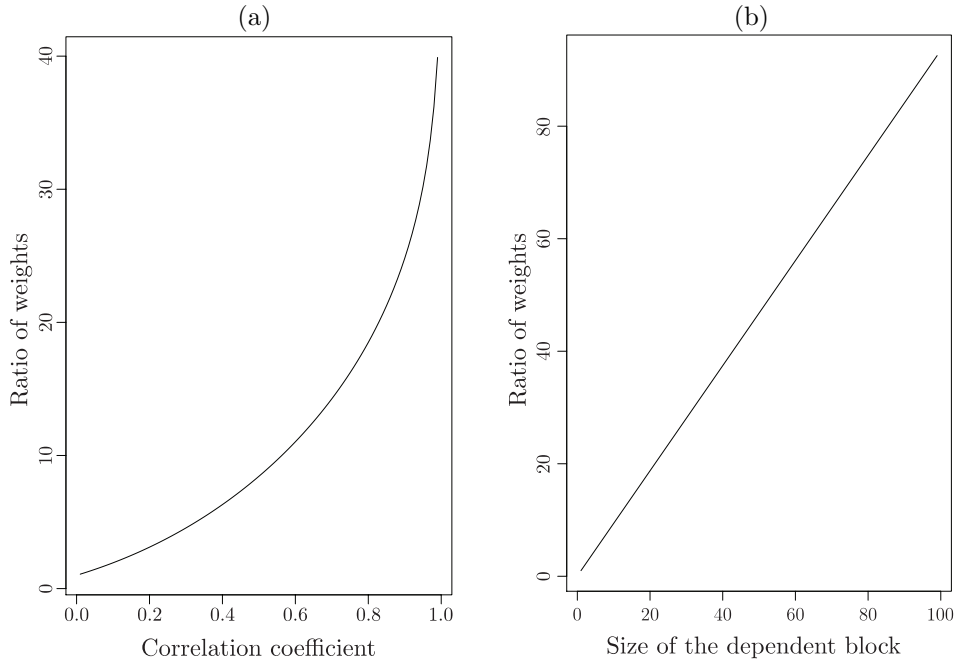

Figure 2. Ratio between the weights of independent and dependent observations against a) correlation coefficient in the block 50x50; b) size of the dependent block with correlation 0.999. For both simulations $N = 100$ and $\pi_0 = 1$.

independent observations over weights for dependent observations. Figure 2a demonstrates how the ratio of weights changes as ρ increases. As expected, for higher values of ρ the ratio is higher, so that the more dependent p-values are down-weighted more. For small ρ, the ratio is close to 1. Figure 2b shows that if $\pi_0 = 1$, $\rho = 0.999$ and the size of the dependent block changes, then the ratio of weights is almost linearly inversely proportional to the size of the dependent block. Next we present several simulations that show how the optimal weights affect the performance of the estimator of π_0.

4.2. Performance of the weighted estimator of π_0

For this simulation we generated multivariate normal "gene expression" datasets as described in Section 3.3. We assumed N=1,000, $n_1 = n_2 = 20$, $\sigma = 1$, μ_x=0, and $\mu_y = 2/\sqrt{n_1/2}$. Thus, the mean of the t-statistics under the alternative distribution is about 2. Half of the genes were dependent in blocks of 50, where all off-diagonal values in a block were ρ, and blocks were independent of each other. The other half of the genes were independent. Half of all the null genes were dependent, as were half of all non-null, differentially expressed

genes; null and non-null genes were independent. We repeated the simulation for $\rho = 0.3, 0.5, 0.9$, and $\pi_0 = 0.5, 0.9$, and each combination of these two parameters is referred to as a scenario. For each dataset we calculated the t-statistics and the corresponding p-values which were used to estimate π_0.

For each particular set of parameters we calculated the true optimal weights using a partition with k=11 intervals ($k = 6$ was not used since it had slighly worse performance in the simulations with independent data, $k = 15$ was very similar to $k = 11$). These weights were applied to all "datasets" within a scenario. We also estimated π_0 assuming independence, i.e., using equal weights. For each dataset we also estimated the optimal weights as outlined in Section 3.3, using the sample correlation matrix of genes and either π_0 and μ_A estimated under independence, or $\pi_0 = 1$. To be precise, means of each gene were subtracted within cases and controls separately, and then the centered expression levels were combined and correlation matrix of genes was computed. To reduce the time needed to calculate the weights we rounded the correlation matrix to 2 digits. The initial estimate of π_0 was the estimate with our proposed method, under independence. The absolute value of the mean of the test statistics under the alternative distribution, μ_A, was estimated from the mixture model as $(\sum_{i=1}^{N} |T_i|/N - 0.794\pi_0)/(1-\pi_0)$, where T_i is the test statistic for gene i, π_0 is an initial estimate and 0.794 is a mean of the absolute value of a standard normal random variable. We also decided to estimate the weights assuming $\pi_0 = 1$ since we are concerned that the noise in initial estimates of π_0 and μ_A might make the weights inefficient. We took a bivariate normal distribution in (3.9, 3.10).

To check whether the derived optimal weights were indeed optimal, we used weights suggested in Chen, Tong, and Zhao (2008). They considered indicator variables, ν_i, for whether the ith hypothesis is falsely rejected. The weight for the test statistic i is inversely proportional to the sum of correlation coefficients between ν_i and ν_j, $j \neq i$, $j = 1, \ldots N$. We calculated such weights under the known multivariate model and the predefined threshold for significance, 0.01. These weights were plugged into our estimator.

Comparisons were made with four methods: the estimator of π_0 built into the ELF algorithm (Pawitan, Calza, and Ploner (2006)); the median and the 3rd quartile of the SampG estimator from Lu and Perkins (2007); Langaas' method (Langaas, Lindqvist, and Ferkingstad (2005)) which was shown to be robust to the violation of the independence assumption.

Datasets were repeated 100 times for each scenario, and estimates of π_0 were summarized by the mean squared error (MSE), variance, and bias. Results for the one-sided tests are shown in Table 2. In all scenarios MSE was much higher for the estimator assuming independence compared to the estimator with the optimal weights. The difference ranged from 3-fold for small correlation 0.3

to 10-fold for $\rho = 0.9$. Both variance and bias increased as the strength of the correlation increased, but less for the weighted estimator compared to the estimator under independence. As in simulations with independent data, bias was negative. The MSE for estimated weights was only slightly larger or similar to MSE for the estimator with the true weights, and weights estimated under the full or the null model seemed to have similar performance.

The ELF estimator had similar performance for small correlation, 0.3, but much higher variance for higher values of ρ. The median of SampG estimators had better performances than the third quartile, but had larger MSE than the weighted estimator, especially for higher ρ or lower π_0. Langaas' method is equivalent to our method that assumes independence, but had much higher variance than the proposed weighted estimator. Note that the weights from Chen, Tong, and Zhao (2008) result in roughly 2-fold higher variance of the estimator compared to the optimal weights.

In simulations not shown we also assessed the MSE of the weighted estimator when the weights were estimated using numerical optimization with positivity constraints. In all scenarios such constrained weights resulted in slightly worse performance of the estimator.

The proposed estimator appears to have better performance than its competitors, and estimating weights from the data does not lead to substantial loss of precision.

The results for the two-sided tests are shown in the Supplementary Table 3. The simulations were performed in the same way except that half of the non-null test statistics had negative means (they were not correlated with the non-null test statistics with positive means). MSE and bias were calculated around the true value of the density at 1. Note that for the two-sided normal tests $f(1) > \pi_0$ due to the identifiability issue. The weighted estimator does not provide gain in precision for small correlation $\rho = 0.3$, but the reduction in MSE was roughly 1.5-fold for $\rho = 0.5$ and 4.5-fold for $\rho = 0.9$ compared to method assuming independence, or Langaas' method. We believe that the dependence between the null p-values is reduced under the two-sided test compared to the one-sided test, hence the improvement in performance is not as dramatic. The rest of the conclusions still hold.

In Supplementary Tables 4 and 5 we have shown results of simulations with the one- and two-sided tests and μ_A drawn at random from $N(2, 0.75^2)$ (constrained from below by 1). These results are similar to the results with fixed μ_A.

Table 2. Performance of the estimators of π_0 under dependence. One-sided test, $\mu_A = 2$.

	$k = 11$					Other methods			
	Fixed weights			Estimated weights					
	Equal	Full	eQTL	Full	Null	ELF	SampG3Q	SampGM	Langaas
	π_0=0.5, ρ=0.3								
MSE	0.00567	0.00234	0.00442	0.00352	0.00351	0.0060	0.02633	0.01564	0.00558
Var	0.00560	0.00229	0.00438	0.00349	0.00347	0.0035	0.00116	0.00103	0.00552
Bias	-0.00804	-0.00687	-0.00655	-0.00579	-0.00595	0.0497	0.15863	0.12086	-0.00739
	π_0=0.5, ρ=0.5								
MSE	0.01023	0.00210	0.00550	0.00290	0.00283	0.0089	0.02984	0.01843	0.01001
Var	0.00928	0.00209	0.00520	0.00286	0.00279	0.0073	0.00268	0.00251	0.00906
Bias	-0.03091	-0.00251	-0.01753	-0.00618	-0.00590	0.0401	0.16479	0.12621	-0.03076
	π_0=0.5, ρ=0.9								
MSE	0.01693	0.00184	0.00339	0.00172	0.00180	0.0140	0.03020	0.01935	0.01605
Var	0.01569	0.00182	0.00330	0.00171	0.00179	0.0122	0.00344	0.00329	0.01492
Bias	-0.03527	-0.00382	-0.00950	-0.00262	-0.00292	0.0424	0.16357	0.12673	-0.03353
	π_0=0.9, ρ=0.3								
MSE	0.00600	0.00228	0.00429	0.00280	0.00279	0.0024	0.00514	0.00147	0.00602
Var	0.00498	0.00185	0.00357	0.00227	0.00227	0.0024	0.00079	0.00095	0.00506
Bias	-0.03201	-0.02096	-0.02695	-0.02306	-0.02293	-0.0042	0.06595	0.02285	-0.03101
	π_0=0.9, ρ=0.5								
MSE	0.01261	0.00206	0.00608	0.00244	0.00252	0.0031	0.00565	0.00281	0.01163
Var	0.01073	0.00186	0.00536	0.00211	0.00219	0.0030	0.00146	0.00204	0.00989
Bias	-0.04346	-0.01385	-0.02683	-0.01812	-0.01811	0.0098	0.06474	0.02777	-0.04172
	π_0=0.9, ρ=0.9								
MSE	0.02596	0.00217	0.00413	0.00213	0.00218	0.0099	0.00576	0.00464	0.02513
Var	0.01941	0.00164	0.00318	0.00167	0.00170	0.0097	0.00285	0.00422	0.01897
Bias	-0.08097	-0.02305	-0.03097	-0.02155	-0.02196	-0.0111	0.05395	0.02041	-0.07846

5. Application to a Lymphoma Dataset

We have analyzed the lymphoma dataset published in Rosenwald et al. (2002), data also explored by Pawitan, Calza, and Ploner (2006). The 102 survivors and 138 non-survivors of diffuse large B-cell lymphoma were compared using a custom-made microarray consisting of 7,399 markers. Censoring was ignored and a two-sided t-test was used to compare the two groups. The p-value distribution appears to be non-increasing and convex as shown by the histogram in Figure 3a. The histogram of pairwise correlation coefficients of the marker expression values is shown in panel b. The median of this distribution is 0.05, and about 25% of all correlations are greater in absolute value than 0.2, thus there is significant amount of dependence; if all markers were independent, having 240 observations per marker we would expect only 0.2% of correlation coefficients outside of [-0.2,0.2].

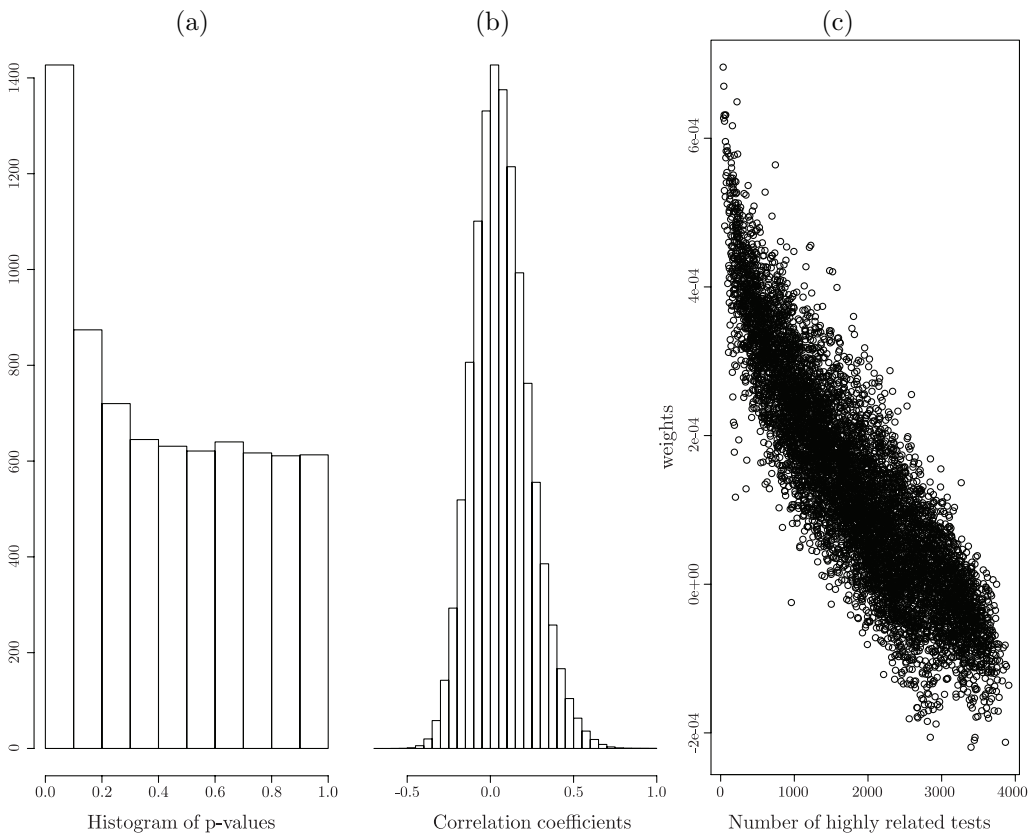

Figure 3. Lymphoma data: a) histogram of p-values; b) histogram of the estimated correlation coefficients; c) weight versus the number of related tests (with $|\hat{\rho}| \geq 0.2$)

ELF (Pawitan, Calza, and Ploner (2006)) produces an estimate of π_0 of 0.92. Langaas' method and our proposed estimator under independence and $k = 11$ give the estimate 0.83. Estimated weights assuming a multivariate normal distribution under both full and null models provide weighted estimates equal to 84% and 85%, respectively. The difference between these estimates is small, but with N=7,399 it corresponds potentially to at least 70 more discoveries. To illustrate the relationship between the dependence structure and the weights, we first calculated "dependence scores". For each particular marker we looked at its pairwise correlations with all other markers, and then counted the number of correlations that exceeded 0.2 in absolute value. A plot of the weights (under the full model) against these "dependence scores" is shown in panel c of Figure 3. A strong negative association is present: if a marker is correlated with many other markers, it is downweighted.

Of course we do not know which estimate is closest to the truth. Therefore, we performed another simulation study based on the lymphoma data. Denote the estimated correlation matrix of the marker expression values as Σ. We created a vector of means μ equal to the test statistics in the dataset, except that the values between -2 and 2 were replaced with 0 and declared null. As a result the 'true' π_0 was 88%. We estimated the optimal weights using the correlation matrix Σ and the means μ. Then we generated N=7,399 test statistics from the multivariate normal distribution with means μ and correlation matrix Σ, transformed them into two-sided p-values using the standard normal as a null distribution, and applied our proposed estimators with equal and with optimal weights. The simulation was repeated 100 times. When π_0 was estimated assuming independence, MSE, variance, and bias were 0.0137, 0.01 and -0.05 respectively. For the weighted estimator MSE, variance and bias were 0.0038, 0.0036 and -0.013; variance and bias were reduced about 3-fold.

It was not computationally feasible to perform a more comprehensive simulation generating the expression levels rather than the test statistics and, therefore, we did not compare our method to other methods. Of course, if the estimated weights were used instead of the "true" weights, the performance would have dropped but, based on the simulations from the previous section, we believe that the reduction in accuracy would have been small. Nevertheless, these results suggest that even in more realistic scenarios, where correlation matrix is not block diagonal and the test statistics under the alternative distribution have different means, the proposed weighted estimator can offer substantial decrease in variance and bias.

6. Discussion

We have developed a method for estimating the proportion of true null hypotheses based on constrained multinomial likelihood. It is free of distributional assumptions other than the non-increasing and convex shape of the density of the p-values. Partition of the support interval serves as a tuning parameter and has to be specified apriori. One of the main advantages of our method is that it does not depend crucially on the choice of the partition.

When the tests are dependent we propose to use weighted marginal likelihoods within the same framework. Each test statistic receives a weight that is roughly inversely proportional to the extent of its correlation with all other test statistics. Optimal weights are derived to guarantee that the trace of the corresponding covariance matrix of the estimating equations is the smallest. When the parametric bivariate distributions of the test statistics are known, we can easily calculate the weights using closed-form formulas. We suggested a simple

procedure in which the weights can be estimated from the data if the test statistics have multivariate normal distribution, that can also be used when the test statistics come from the t-test. Simulations show that the weighted estimator offers a considerable improvement in bias and variance compared to the estimator under independence. When the weights are estimated from the data the drop in performance seems negligible.

The computation of the optimal weights requires estimation of the correlation matrix of the test statistics and, optionally, preliminary estimates of π_0 and μ_A. Our simulations show that rounding the correlation matrix to two digits has a very small effect on the weights but provides substantial savings in the computational time: for example, it takes about 10 seconds to estimate the weights for $N = 1,000$ and 8 minutes for $N = 7,399$ on a computer with 2.33GHz CPU. The greatest computational challenge in estimating weights is storing and inverting an NxN matrix. As the number of tests increases to more than 10,000, estimating weights might not be feasible unless some simplifications are made. For such cases the test statistics should be split into smaller blocks that are independent of each other based, for example, on genomic locations or known pathway information. Then the computation of the weights can be achieved for each block separately, with the restriction that the weights sum up to the size of the block over the total number of tests. The R code for estimating the weights and π_0 is available from the authors.

Based on the results of the simulations, weights that are calculated using the multivariate normal distribution work well even for t-statistics. We believe it is possible to estimate the weights under distributions other than normal, although we have not investigated this issue further. It would be necessary to specify the parametric form of the bivariate distribution of the test statistics that would depend on a small number of parameters that can be estimated from the data. Some form of the multivariate t distribution in Kotz and Nadarajah (2004) can be a useful approximation, or in the case of χ^2 tests the multivaraite χ^2 distribution can be applied (Kotz, Johnson, and Balakrishnan (2000)).

In our derivation of the optimal weights for multivariate normal test statistics we made some assumptions: the null and alternative test statistics are independent, all the non-null test statistics have the same mean, correlation between the two test statistics is the same whether they come from the null or alternative distributions. Of course, in reality these assumptions are violated and the data structure can be much more complex. Our simulations with means and correlation matrix from the lymphoma dataset showed good performance of the estimator even when some of the assumptions did not hold, e.g., some null and non-null statistics were correlated. Note that the model for bivariate distributions (3.10) can be easily modified to accommodate more complex data structures.

In this paper we have not touched upon how the estimate of π_0 under dependence should be used in the multiple comparison procedures, or how it affects their power and other properties. There is a host of error metrics in the literature developed under dependence, including conditional FDR (Friguet, Kloareg, and Causeur (2009)), upper bound of false discovery proportion (Chen, Tong, and Zhao (2008)), empirical FDR (Hunt, Cheng, and Pounds (2009)), and others. These available methods operate under different settings and assumptions, e.g., testing of linear contrasts, regression analysis or the beta-binomial distribution. Many methods either use values of π_0 obtained by methods developed for independent data, or assume $\pi_0 = 1$. We have shown that the proposed estimator of π_0, while requiring minimum assumptions on the data structure, has much smaller variance and bias if the dependence is taken into account; we expect it will improve the power of the multiple testing procedures in a variety of models. However, quantifying its effect on various metrics and under various settings is beyond the scope of this paper, although it is the area of our current research.

Acknowledgements

This research was supported in part by the National Science Foundation (DMS-0072510) and the National Institutes of Health (HL084715).

Appendix

A.1. Theorems and proofs

Theorem A.1. Assume that f is a non-increasing continuously differentiable function, and let g defined as in (2.1). Then

(i) *g is a non-increasing function.*

(ii) $0 \le g(x) \le f(x)$*, for any* $x \in [0, 1]$.

(iii) *If f is convex, then g is also convex.*

Proof of Theorem A.1. The following lemma is needed for the proof of Theorem A.1.

Lemma A.1. Let f be a non-increasing, continuously differentiable, convex function. Then for any a and b in the support of f,

$$\int_a^b f(y)dy \ge f\left(\frac{a+b}{2}\right)(b-a).$$

Proof. Convexity of f implies that for any a_1, a_2, $\delta \in [0, (b-a)/2]$

$$f\left(\frac{a+b}{2}\right) \le \frac{1}{2}\left(f\left(\frac{a+b}{2}+\delta\right)+f\left(\frac{a+b}{2}-\delta\right)\right)$$

Table 3. Supplementary table. Performance of the estimators of π_0 under dependence. Two-sided test, $\mu_A = 2$.

	$k = 11$					Other methods			
	Fixed weights			Estimated weights					
	Equal	Full	eQTL	Full	Null	ELF	SampG3Q	SampGM	Langaas
	π_0=0.5, ρ=0.3								
MSE	0.00236	0.00216	0.00228	0.00225	0.00226	0.0041	0.01109	0.00497	0.00219
Var	0.00233	0.00211	0.00224	0.00222	0.00222	0.0022	0.00105	0.00107	0.00215
Bias	-0.00569	-0.00683	-0.00622	-0.00564	-0.00627	0.0431	0.10021	0.06246	-0.00615
	π_0=0.5, ρ=0.5								
MSE	0.00364	0.00260	0.00278	0.00306	0.00282	0.0041	0.01069	0.00491	0.00366
Var	0.00360	0.00252	0.00271	0.00296	0.00269	0.0033	0.00205	0.00195	0.00363
Bias	-0.00652	-0.00932	-0.00865	-0.00986	-0.01139	0.0297	0.09292	0.0544	-0.00523
	π_0=0.5, ρ=0.9								
MSE	0.01275	0.00378	0.00421	0.00402	0.00353	0.0107	0.01516	0.00906	0.0129
Var	0.01223	0.00366	0.00409	0.00385	0.00339	0.0068	0.00574	0.00553	0.01237
Bias	-0.02271	-0.01063	-0.01100	-0.01305	-0.01190	0.0619	0.09704	0.05943	-0.02313
	π_0=0.9, ρ=0.3								
MSE	0.00240	0.00202	0.00223	0.00220	0.00221	0.0024	0.00390	0.00109	0.00252
Var	0.00217	0.00183	0.00202	0.00194	0.00194	0.0022	0.00083	0.00093	0.00233
Bias	-0.01520	-0.01359	-0.01446	-0.01605	-0.01627	-0.0108	0.05539	0.01257	-0.01349
	π_0=0.9, ρ=0.5								
MSE	0.00439	0.00286	0.00326	0.00352	0.00331	0.0053	0.00399	0.00228	0.00428
Var	0.00392	0.00253	0.00290	0.00307	0.00288	0.0049	0.00172	0.00220	0.00388
Bias	-0.02167	-0.01836	-0.01898	-0.02121	-0.02076	-0.0217	0.04769	0.00909	-0.02011
	π_0=0.9, ρ=0.9								
MSE	0.01691	0.00348	0.00468	0.00435	0.00369	0.0157	0.00563	0.00573	0.01620
Var	0.01420	0.00273	0.00378	0.00322	0.00274	0.0149	0.00421	0.00569	0.01352
Bias	-0.05205	-0.02738	-0.02999	-0.03372	-0.03085	-0.0295	0.03777	0.00604	-0.05181

$$\text{and } f\left(\frac{a+b}{2}-\delta\right)-f\left(\frac{a+b}{2}\right)\geq f\left(\frac{a+b}{2}\right)-f\left(\frac{a+b}{2}+\delta\right).$$

Denote $h_1(\delta) = f\left([(a+b)/2]-\delta\right) - f((a+b)/2)$ and $h_2(\delta) = f\left((a+b)/2\right) - f\left([(a+b)/2]+\delta\right)$. Notice that $h_1(\delta) \geq h_2(\delta) \geq 0$ for any δ, therefore

$$\int_0^{\frac{b-a}{2}} h_1(\delta)d\delta \geq \int_0^{\frac{b-a}{2}} h_2(\delta)d\delta. \tag{A.1}$$

Calculating the integrals, we get

$$\int_a^{\frac{b+a}{2}} f(y)dy - f\left(\frac{a+b}{2}\right)\frac{b-a}{2} \geq f\left(\frac{a+b}{2}\right)\frac{b-a}{2} - \int_{\frac{b+a}{2}}^{b} f(y)dy \Rightarrow$$
$$\int_a^b f(y)dy \geq f\left(\frac{a+b}{2}\right)(b-a).$$

Table 4. Supplementary table. Performance of the estimators of π_0 under dependence. One-sided test, μ_A is random.

	$k = 11$					Other methods			
	Fixed weights			Estimated weights					
	Equal	Full	eQTL	Full	Null	ELF	SampG3Q	SampGM	Langaas
	π_0=0.5, ρ=0.3								
MSE	0.00707	0.00268	0.00305	0.00379	0.00377	0.01665	0.04178	0.02737	0.00676
Var	0.00702	0.00266	0.00304	0.00379	0.00376	0.00259	0.00108	0.00102	0.00671
Bias	-0.00760	0.00342	0.00220	0.00247	0.00263	0.11860	0.20174	0.16232	-0.00722
	π_0=0.5, ρ=0.5								
MSE	0.00971	0.00289	0.00304	0.00396	0.00395	0.01403	0.03718	0.02395	0.00917
Var	0.00960	0.00267	0.00284	0.00380	0.00379	0.00317	0.00193	0.00180	0.00907
Bias	-0.01026	0.01496	0.01419	0.01261	0.01254	0.10423	0.18776	0.14882	-0.00981
	π_0=0.5, ρ=0.9								
MSE	0.02259	0.00325	0.00321	0.00367	0.00379	0.01839	0.04375	0.02978	0.02118
Var	0.02071	0.00323	0.00319	0.00362	0.00374	0.00731	0.00479	0.00476	0.01965
Bias	-0.04333	0.00410	0.00416	0.00744	0.00664	0.10528	0.19739	0.15817	-0.03921
	π_0=0.9, ρ=0.3								
MSE	0.00472	0.00234	0.00241	0.00289	0.00296	0.00239	0.00555	0.00177	0.00475
Var	0.00454	0.00218	0.00228	0.00278	0.00284	0.00231	0.00072	0.00092	0.00455
Bias	-0.01349	-0.01255	-0.01137	-0.01062	-0.01084	0.00900	0.06950	0.02914	-0.01408
	π_0=0.9, ρ=0.5								
MSE	0.01062	0.00215	0.00226	0.00296	0.00292	0.00425	0.00574	0.00279	0.00991
Var	0.00916	0.00203	0.00214	0.00285	0.00280	0.00424	0.00156	0.00208	0.00852
Bias	-0.03825	-0.01083	-0.01083	-0.01038	-0.01068	0.00189	0.06464	0.02680	-0.03725
	π_0=0.9, ρ=0.9								
MSE	0.02093	0.00233	0.00234	0.00243	0.00243	0.01157	0.00645	0.00559	0.01856
Var	0.01613	0.00208	0.00209	0.00217	0.00217	0.01156	0.00303	0.00470	0.01409
Bias	-0.06931	-0.01581	-0.01590	-0.01620	-0.01635	-0.00094	0.05849	0.02987	-0.06688

Proof of Theorem A.1. (i) Put $s_1 < s_2 < 1$. Then g is non-increasing if and only if

$$\frac{\int_{s_1}^{s_2} f(s)ds}{s_2 - s_1} - \frac{\int_{s_2}^{1} f(s)ds}{1 - s_2} \geq 0.$$

Since f is decreasing,

$$\int_{s_1}^{s_2} f(s)ds \geq (s_2 - s_1)f(s_2) \text{ and } \int_{s_2}^{1} f(s)ds \leq (1 - s_2)f(s_2).$$

It follows that

$$\frac{\int_{s_1}^{s_2} f(s)ds}{s_2 - s_1} - \frac{\int_{s_2}^{1} f(s)ds}{1 - s_2} \geq f(s_2) - f(s_2) = 0.$$

(ii) Since f is decreasing $\int_s^1 f(s)ds \leq (1-s)f(s)$. It follows that $g(x) \leq f(x)$, for any x in (0,1).

Table 5. Supplementary table. Performance of the estimators of π_0 under dependence. Two-sided test, μ_A is random.

	$k = 11$					Other methods			
	Fixed weights			Estimated weights					
	Equal	Full	eQTL	Full	Null	ELF	SampG3Q	SampGM	Langaas
	π_0=0.5, ρ=0.3								
MSE	0.00387	0.00351	0.00324	0.00371	0.00361	0.00883	0.01883	0.00998	0.00381
Var	0.00212	0.00183	0.00175	0.00201	0.00194	0.00177	0.00097	0.00097	0.00207
Bias	0.04185	0.04107	0.03868	0.04125	0.04080	0.08399	0.13366	0.09496	0.04169
	π_0=0.5, ρ=0.5								
MSE	0.00419	0.00318	0.00239	0.00355	0.00320	0.00864	0.01617	0.00847	0.00435
Var	0.00365	0.00273	0.00221	0.00310	0.00279	0.00257	0.00200	0.00193	0.00381
Bias	0.02336	0.02110	0.01348	0.02124	0.02036	0.07791	0.11902	0.08085	0.02339
	π_0=0.5, ρ=0.9								
MSE	0.01160	0.00426	0.00338	0.00412	0.00379	0.01293	0.02196	0.01315	0.01184
Var	0.01135	0.00323	0.00222	0.00328	0.00282	0.00550	0.00461	0.00444	0.01162
Bias	0.01580	0.03216	0.03414	0.02891	0.03113	0.08622	0.13169	0.09334	0.01471
	π_0=0.9, ρ=0.3								
MSE	0.00210	0.00204	0.00201	0.00201	0.00199	0.00202	0.00388	0.00115	0.00213
Var	0.00194	0.00189	0.00187	0.00183	0.00181	0.00202	0.00071	0.00091	0.00201
Bias	-0.01269	-0.01205	-0.01188	-0.01365	-0.01363	-0.00160	0.05636	0.01542	-0.01089
	π_0=0.9, ρ=0.5								
MSE	0.00450	0.00342	0.00276	0.00385	0.00390	0.00416	0.00413	0.00221	0.00435
Var	0.00416	0.00312	0.00237	0.00344	0.00346	0.00401	0.00149	0.00203	0.00405
Bias	-0.01838	-0.01752	-0.01980	-0.02024	-0.02106	-0.01195	0.05140	0.01355	-0.01721
	π_0=0.9, ρ=0.9								
MSE	0.01273	0.0027	0.00170	0.00338	0.00297	0.01235	0.00508	0.00492	0.01204
Var	0.01151	0.0026	0.00165	0.00313	0.00280	0.01232	0.00304	0.00464	0.01084
Bias	-0.03485	-0.0101	-0.00712	-0.01573	-0.01292	-0.00480	0.04523	0.01671	-0.03464

(iii) The first and second derivatives of g are

$$g'(x) = \frac{g(x) - f(x)}{(1-x)} \text{ and } g''(x) = \frac{2g'(x) - f'(x)}{1-x}.$$

Fix arbitrary x. By the first mean value theorem, there exists a in $[x, 1]$, such that

$$\int_x^1 f(y)dy = (1-x)f(a),$$

and therefore $g'(x) = (f(a) - f(x))/(1-x)$. Also there exists $b \in [x, a]$, such that $\int_x^a b'(y)dy = (a-x)f'(b)$. It follows that

$$g'(x) = \frac{(a-x)f'(b)}{1-x} \text{ and } g''(x) = \frac{2\frac{(a-x)f'(b)}{1-x} - f'(x)}{1-x}.$$

Since $f''(x) \geq 0$, $f'(x)$ is non-decreasing and $0 \geq f'(b) \geq f'(x)$ (remember $0 \leq x \leq b \leq a \leq 1$)

$$g''(x) \geq -\frac{f'(x)}{1-x}\frac{1+x-2a}{1-x}.$$

Now it remains to show that $a \leq (x+1)/2$. Since f is non-increasing, this occurs if and only if $f(a) \geq f\left((1+x)/2\right)$, or equivalently

$$(1-x)f\left(\frac{1+x}{2}\right) \leq \int_x^1 f(y)dy = (1-x)f(a).$$

This inequality is a special case of Lemma A.1, thus $a \leq (x+1)/2$ and g is convex.

Theorem A.2. Let $l_\theta(\theta, X) \sim \sum_{i=1}^k X_i \log(\theta_i)$. Let S^C be the constrained set of multinomial probabilities

$$S^C = \{\theta = (\theta_1, \ldots \theta_k) : \quad \theta_i \geq 0, \quad \sum_{i=1}^k \theta_i = 1, \quad \psi(\theta) \leq 0\},$$

where $\psi(\theta)$ is the vector of linear constraints, and let the unconstrained set be

$$S^U = \{\theta = (\theta_1, \ldots, \theta_k) : \quad \theta_i \geq 0, \quad \sum_{i=1}^k \theta_i = 1\}.$$

If $\hat{\theta}^C$ is the constrained MLE, $\operatorname{argmax}_{\{S^C\}} l_\theta(\theta, X)$, and $\hat{\theta}^U$ is the MLE over the unconstrained set S^U, then if θ is an interior point of S^C, $\hat{\theta}^C$ is a consistent estimator of θ and $\hat{\theta}^C$ has the same asymptotic distribution as the unconstrained estimator $\hat{\theta}^U$.

For simplicity, Theorem A.2 is formulated in terms of parameters θ_i's. It implies that, as N goes to infinity, $\hat{\theta}^C \sim N(\theta, \Sigma)$, where

$$\Sigma_{ii} = \frac{\theta_i(1-\theta_i)}{N} \qquad \text{and} \qquad \Sigma_{ij} = -\frac{\theta_i\theta_j}{N}.$$

Thus $\hat{\pi}_0 = \hat{g}_{k-1}$ has an asymptotically normal distribution with mean g_{k-1} and variance $g_{k-1}[1-g_{k-1}(1-t_{k-1})]/N(1-t_{k-1})$. This distribution does not take into account the constraint and cannot be used for a fixed sample size. Consistency of $\hat{g}$ implies the consistency of $\hat{\pi}_0$ if the length of the last interval goes to 0 as N goes to infinity.

Proof of Theorem A.2. We can write the log-likelihood function in terms of parameters θ: $l(\theta, \mathbf{x}) \sim \sum_{i=1}^k X_i \log(\theta_i)$.

First, notice that $\theta^{\mathbf{U}} \in S^C \Leftrightarrow \theta^{\mathbf{U}} = \theta^{\mathbf{C}}$. To see that, observe that $S^C \subset S^U \Rightarrow \max_{S^U} l(\theta, \mathbf{X}) \geq \max_{S^C} l(\theta, \mathbf{X})$ and $\theta^{\mathbf{U}} \in S^C \Rightarrow \max_{S^U} l(\theta, \mathbf{X}) = l(\theta^{\mathbf{U}}, \mathbf{X}) \leq \max_{S^C} l(\theta, \mathbf{X})$. Therefore $M := \max_{S^U} l(\theta, \mathbf{X}) = \max_{S^C} l(\theta, \mathbf{X})$.

There is no point $\theta' \in S^U$ with $l(\theta', \mathbf{X}) = M$ except $\theta^{\mathbf{U}}$, since the unconstrained MLE is unique, thus there is no such point in $S^C \subset S^U$ as well. The unconstrained and constrained maximum M must be attained in the same point and $\theta^{\mathbf{U}} = \theta^{\mathbf{C}} \Rightarrow \theta^{\mathbf{U}} \in S^C$.

We show that $P(\theta^{\mathbf{U}} \in S^C) \to 1$ as $n \to \infty$, true value $\theta \in S^C$. The MLE $\theta^{\mathbf{U}}$ is consistent and thus, for any $\epsilon > 0$, $P(|\theta^{\mathbf{U}} - \theta| \leq \epsilon) \to 1$ as $n \to \infty$. Since θ is an interior point of S^C, there exists $\epsilon_0 > 0$, such that $|\theta^{\mathbf{U}} - \theta| \leq \epsilon_0$ and this can only happen if $\theta^{\mathbf{U}} \in S^C$. That is, if $\theta^{\mathbf{U}}$ is close enough to θ, it must be in S^C also. So for ϵ_0 $P(|\theta^{\mathbf{U}} - \theta| \leq \epsilon_0) = P(\theta^{\mathbf{U}} \in S^C) \to 1$ as $n \to \infty$.

Now $P(\theta^{\mathbf{U}} \in S^C) = P(\theta^{\mathbf{U}} = \theta^{\mathbf{C}}) \to 1$ as $n \to \infty$, and thus $\theta^{\mathbf{U}} - \theta^{\mathbf{C}}$ converges to 0 in probability. It follows that $\theta^{\mathbf{C}}$ converges to θ in probability.

Theorem A.3. The matrix $R = \{tr(S_{ij})\}$ is nnd.

Proof. We need to show that $x^T R x \geq 0$ for any N-dimensional vector x. Notice that

$$R_{ij} = tr(S_{ij}) = \sum_{a=1}^{m} Cov(e_{ia}, e_{ja}).$$

Therefore

$$\begin{aligned} x^T R x &= \sum_{a=1}^{m} \sum_{i,j=1}^{N} x_i x_j Cov(e_{ia}, e_{ja}) \\ &= \sum_{a=1}^{m} \sum_{i,j=1}^{N} Cov(x_i e_{ia}, x_j e_{ja}) = \sum_{a=1}^{m} Var\left(\sum_{i=1}^{N} x_i e_{ia}\right) \geq 0, \end{aligned}$$

and R is nnd.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J., Lee, C. K., Prolla, T., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal.* **39**, 1-20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165-1188.

Chen, L., Tong, T. and Zhao, H. (2008). Considering dependence among genes and markers for false discovery control in eqtl mapping. *Bioinformatics* **24**, 2015-2022.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.

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

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102**, 93-103.

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scand. J. Statist.* **34**, 275-297.

Friguet, C., Kloareg, M. and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104**, 1406-1415.

Genovese, C. and Wasserman, L. (2001). Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* **64**, 499-517.

Heyde, C. (1997). *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer-Verlag, New York.

Hoffjan, S., Nicolae, D., Ostrovnaya, I., Roberg, K., Evans, M., Mirel, D. B., Steiner, L., Walker, K., Shult, P., Gangnon, R. E., Gern, J. E., Martinez, F. D., Lemanske, R. F. and Ober, C. (2005). Gene-environment interaction effects on the development of immune responses in the 1st year of life. *Amer. J. Human Genetics* **76**, 696-704.

Hoffjan, S., Ostrovnaja, I., Nicolae, D., Newman, D., Nicolae, R., Gangnon, R., Steiner, L., Walker, K., Reynolds, R., Greene, D., Mirel, D., Gern, J., Lemanske, R. J. and Ober, C. (2004). Genetic variation in immunoregulatory pathways and atopic phenotypes in infancy. *J. Allergy Clin. Immunol.* **113**, 511-518.

Hunt, D., Cheng, C. and Pounds, S. (2009). The beta-binomial distribution for estimating the number of false rejections in microarray gene expression studies. *Comput. Statist. Data Anal.* **53**, 1688-1700.

Kotz, S., Johnson, N. and Balakrishnan, N. (2000). *Continuous Multivariate Distributions: Models and Applications*. Wiley-Interscience.

Kotz, S. and Nadarajah, S. (2004). *Multivariate T-Distributions and Their Applications*. Cambridge University Press.

Krishnan, M. (1972). Series representations of a bivariate singly noncentral t-distribution. *J. Amer. Statist. Assoc.* **67**, 228-231.

Langaas, M., Lindqvist, B. and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. Roy. Statist. Soc. Ser. B* **67**, 555-572.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Math.* **80**, 221-239.

Lu, X. and Perkins, D. (2007). Re-sampling strategy to improve the estimation of number of null hypotheses in fdr control under strong correlation structures. *BMC bioinformatics* **8**, 157.

Markitsis, A. and Lai, Y. (2010). A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics* **26**, 640-646.

Miller, C. J., Genovese, C., Nichol, R. C., Wasserman, L., Connolly, A., Reichart, D., Hopkins, A., Schneider, J. and Moore, A. (2001). Controlling the false discovery rate in astrophysical data analysis. *Astronomical J.* **122**, 3492-3505.

Pawitan, Y., Calza, S. and Ploner, A. (2006). Estimation of false discovery proportion under general dependence. *Bioinformatics* **22**, 3025-3031.

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**(10), 1236-1242.

Rosenwald et al, A. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England J. Medicine* **346**, 1937-1947.

Ruppert, D., Nettleton, D. and Hwang, G. (2007). Exploring the information in p-values for the analysis and planning of multiple-test experiments. *Biometrics* **63**, 483-495.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B* **64**, 479-498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Nat. Acad. Sci.* **100**, 9440-9445.

Sun, W. and Cai, T. (2009). Large-scale multiple testing under dependency. *J. Roy. Statist. Soc. Ser. B* **71**, 393-424.

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 307 East 63rd Street, New York NY 10021, USA.

E-mail: ostrovni@mskcc.org

Departments of Medicine and Statistics, The University of Chicago, 5734 S. University Ave., Chicago IL 60637, USA.

E-mail: nicolae@galton.uchicago.edu

(Received November 2010; accepted August 2011)