More Powerful Control of the False Discovery Rate Under Dependence

Article *in* Statistical Methods and Applications · May 2006

DOI: 10.1007/s10260-006-0002-z · Source: RePEc

1 author:

Alessio Farcomeni
Sapienza University of Rome

Statistical Methods & Applications (2006) 15: 43–73
DOI 10.1007/s10260-006-0002-z

ORIGINAL ARTICLE

Alessio Farcomeni

More powerful control of the false discovery rate under dependence

Accepted: 10 November 2005 / Published online: 7 February 2006

Abstract In a breakthrough paper, Benjamini and Hochberg (J Roy Stat Soc Ser B 57:289–300, 1995) proposed a new error measure for multiple testing, the FDR; and developed a distribution-free procedure to control it under independence among the test statistics. In this paper we argue by extensive simulation and theoretical considerations that the assumption of independence is not needed. Along the lines of (Ann Stat 32:1035–1061, 2004b), we moreover provide a more powerful method, that exploits an estimator of the number of false nulls among the tests. We propose a whole family of iterative estimators that prove robust under dependence and independence between the test statistics. These estimators can be used to improve also classical multiple testing procedures, and in general to estimate the weight of a known component in a mixture distribution. Innovations are illustrated by simulations.

Keywords Dependence · False discovery rate · Multiple testing

1 Introduction

Many modern applications require the testing of hundreds, or even thousands of hypotheses. Examples include DNA Microarrays (Bovenhuis and Spelman, 2000; Mosig et al., 2001; Reiner et al., 2003), brain imaging (Worsley et al., 1996; Ellis et al., 2000; Merriam et al., 2003), etc. Traditional methods that involve control of the familywise error rate (FWER), i.e., the probability of making one or more false positives, can be unduly conservative when the number of tests m is as large as in those applications. For this reason, in a breakthrough paper, Benjamini and

A. Farcomeni
Dipartimento di Statistica, Probabilitá e Statistiche Applicate,
Piazzale Aldo Moro 5, 00185 Rome Italy
E-mail: alessio.farcomeni@uniroma1.it

Table 1 Categorization of the outcome

	H_0 not rejected	H_0 rejected	Total
H_0 True	$N_{0\|0}$	$N_{1\|0}$	M_0
H_0 False	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$m - R$	R	m

Hochberg (1995) proposed a new error measure for multiple testing, the false discovery rate (FDR), or expected proportion of false positives over the number of rejections. They proposed also a distribution-free procedure to control it under independence of the test statistics. They proved that control of the FDR is more liberal than control of FWER, with good power also when the number of tests is very high. The problem of controlling the FDR when the test statistics are dependent has been open since the introduction of the error measure: it is well known that in many applications the test statistics are dependent. For instance, expression levels of genes are dependent by their own intrinsic nature, and the nature of the experiment. Among other possible applications of dependent multiple tests, refer to Ip (2001) for testing for local dependency in item response data, or to Yekutieli and Benjamini (1999) for testing for significant correlations in correlation maps. The problem has been partially tackled by Benjamini and Yekutieli (2001), who elegantly proved conditions under which the standard technique works under dependence.

The contributions of this paper can be divided into two parts: in the first part, that includes sections 2 and 3, we argue by simulations and theoretical considerations that the assumption of independence is not needed for the classical procedures to work, provided that it is used an estimator for the number of true false hypotheses that is robust with respect to dependence. The simulations imply that the conditions on the dependence used in Benjamini and Yekutieli (2001) can be sharpened; and that the classical estimators of the proportion of false nulls break down under dependence. Suitable estimators are proposed in section 4, where a whole family of iterative estimators are proposed. The discussion is of general interest for the problem of estimating the weight of a component in a mixture model (see Swanepoel (1999) and Example 2). We moreover point out that in the literature, usually, uncertainty brought about by estimation of the proportion of false nulls is not taken into account, and propose a way to do this. Next section reviews the necessary background on multiple testing.

1.1 Background

Consider a multiple testing situation in which m tests are being performed. Suppose M_0 of the m hypotheses are true, and M_1 are false. Table 1 shows a categorization of the outcome of the tests. R is the number of rejections. Note that $N_{0|1}$ and $N_{1|0}$ give us the exact (unknown) number of errors committed.

Traditional methods for multiple testing, like the well known Bonferroni correction, attempted control on the FWER, i.e., $P(N_{1|0} > 0)$. This error measure proves too strict if m is big, resulting in an unsatisfactory low number of rejections and high number of false negatives $N_{0|1}$.

For this reason, Benjamini and Hochberg (1995) defined a new error measure, the FDR. Control of FDR leads to much higher power than control of FWER.

Define the false discovery proportion (FDP), proportion of incorrect rejections, to be $\frac{N_{1|0}}{R+1_{\{R=0\}}}$; and the false non-discovery proportion (FNP), proportion of incorrect non-rejections, to be $\frac{N_{0|1}}{m-R+1_{\{R=m\}}}$. The FDR and false non-discovery rate (FNR) are usually defined to be the expected values of this two quantities; and the procedures attempt a control on the FDR. FNR is used as a measure of power of the multiple testing procedure. Refer to Genovese and Wasserman (2002) for a detailed discussion. The intuitive justification of FDR is related to classification theory: if one thinks about the problem of hypothesis testing as unsupervised classification, then the FDR is the expected classification error for the "zero" labels.

Moreover, control of the FDR is also a weak control on the FWER (Benjamini and Hochberg 1995), and it actually is FWER control when $R = 1$.

The procedures for control of the FDR are as follows: suppose we are testing using p-values, and that we reject the tests for which the p-value is not bigger than a cut-off point T, i.e. if $I_i = 1_{\{p_i<T\}}$ is equal to 1, $1 \leq i \leq m$. Here, we make the assumption that all the prospective rejection regions will be fixed equal.

Define $\widehat{G}(t) = \frac{1}{m}\sum I_i$, the empirical distribution of the p-values, and denote by a the proportion M_1/m, i.e., the proportion of false nulls among the tests.

The BH method, as described in Simes (1986) and then by Benjamini and Hochberg (1995), fixes the threshold T as $\sup\{t : \widehat{G}(t) = \frac{t}{\alpha}\}$, where α is the desired upper bound for the FDR. The plug-in method, as described in Genovese and Wasserman (2002), fixes T as $\sup\{t : \widehat{G}(t) = \frac{(1-\widehat{a})t}{\alpha}\}$, where $\widehat{a}$ is a suitable estimator for a. The most common estimator is Storey's estimator, proposed in Storey (2002), defined as: $\widehat{a} = \widehat{G}(t_0) - t_0/1 - t_0$ for some $t_0 \in (0, 1)$.

The BH method yields an FDR lower than $(1-a)\alpha$, while the plug-in method is asymptotically less conservative. Plug-in yields FDR $\leq \alpha$ as m goes to infinity; thus being more powerful. Note that these methods are distribution free, in the sense that they not depend on the distribution of the p-values when the null is false.

Benjamini and Hochberg (1995) note that controlling the FDR yields more powerful tests than the ones controlling the FWER; and that control on the FDR is a liberal form of control on the FWER. Genovese and Wasserman (2002) give an intuitive interpretation of the FDR and the controlling method, and introduce the FNR. Efron et al. (2001) developed an innovative empirical Bayes approach to multiple testing, with interesting connections with the FDR. Storey (2002, 2003) introduced the positive FDR, which proves even more powerful than the original BH method, and provided interesting extensions of the methods. Genovese and Wasserman (2004b) introduce estimators for a and $F(\cdot)$, suggest ways to build confidence thresholds for the FDP and prove asymptotic results, in particular on the limiting distributions of the quantities of interest.

Benjamini and Yekutieli (2001) prove that the BH procedure can never control the FDR at level higher than $\alpha\sum_{i=1}^{m} 1/i$, under arbitrary dependence. Note that this is possibly conservative. They also prove that, under conditions of positive regression dependency on subset S_0 (PRDS), the BH procedure is still valid, controlling the FDR at level α. Let $X = \{X_1, \ldots, X_i, \ldots, X_n\}$ be a vector of random variables, and $S_0 \subseteq \{1, \ldots, n\}$. The condition of PRDS introduced in Benjamini and Yekutieli (2001) is as follows: For any increasing set D and for each $i \in S_0$, let $\Pr(X \in D|X_i = x)$ be non decreasing in x. Recall that a set is said to be increasing if for any $x \in D$ and $y \geq x$, $y \in D$. This is a relaxed version of positive regression dependency (see Esary et al. 1967).

Distributions satisfying this property include multivariate normal distributions with positive correlations, all uni-dimensional latent variable distributions, and few other cases. Sarkar (2002) also extends the results of Benjamini and Yekutieli (2001) by generalizing their results to a whole class of step-up/step-down procedures to control the FDR. In this paper, we show examples and simulations that suggest that, in practice, the procedure might work also when their conditions are not respected.

Storey and Tibshirani (2001) show how to estimate the pFDR, positive false discovery rate of Storey (2003), under general dependence between the test statistics and apply the methodology to estimate the FDR under dependence for pre-fixed rejection region.

Yekutieli and Benjamini (1999), and Pollard and van der Laan (2003) propose resampling based procedures to control the FDR when the test statistics are correlated.

Storey et al. (2004) propose a unified estimation approach for the FDR, showing methods for control of the FDR either fixing the threshold or the rejection region, or asymptotically over all rejection regions simultaneously. Moreover, they present several theorems that all require almost sure pointwise convergence of the empirical distributions of the null p-values and alternative p-values.

In the next two sections we argue, with extensive simulations and theoretical considerations, that the BH technique is robust with respect to dependence. On the other hand, usual estimators of the quantity a are not robust and break down under strong dependence. This leads the plug-in technique to break down under strong (positive) dependence. In section 4, to solve this problem, we propose a family of iterative estimators of a and show that, though being conservative, plug-in method with any of those estimators controls the FDR at the desired level and still dominates the BH procedure in terms of power, under different assumptions on the dependence.

2 Mean and variance of the FDP

We show, in this section, that the expected value of the FDP is unchanged by dependence between the test statistics; while its variance will be. Suppose that the test statistics are not independent, so that the p values will not be independent. Let $I^m = (I_1, \ldots, I_m)$, where $I_j = 1_{\{p_j \leq t\}}$. $F(\cdot)$ will denote the distribution of the p-values under the alternative hypothesis; a is the true proportion of false nulls, and $G(t) = (1-a)t + aF(t)$ will be the marginal distribution of the p-values. H_i will be the indicator function of the i^{th} null hypothesis to be false.

Note that the FDP as a stochastic process in t will be[1]

$$\Gamma(t) = \frac{\sum(1 - H_i)I_i}{\sum I_i + \prod(1 - I_i)}.$$

[1] While the FNP as a stochastic process in t will be $\frac{\sum H_i(1-I_i)}{\sum(1-I_i)+\prod I_i}$.

Table 2 Correlation ranges, 10 × 10 grid, exponential model

$1/\tau$	0.02	0.05	0.1	0.3	0.5	0.6	0.7	0.8	1	1.2
min	0.77	0.53	0.28	0.02	0.00	0.00	0.00	0.00	0.00	0.00
max	0.98	0.95	0.90	0.74	0.61	0.54	0.50	0.45	0.37	0.30

Table 3 Correlation ranges (40 × 40 grid, exponential model)

τ	50	20	10	3.33	2	1.66	1.43	1.25	1	0.833
min	0.33	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
max	0.98	0.95	0.90	0.74	0.61	0.54	0.50	0.45	0.37	0.30

Table 4 Correlation ranges (100 × 100 grid, exponential model)

τ	50	20	10	3.33	2	1.66	1.43	1.25	1	0.833
min	0.06	0.00	0.00	.00	0.00	0.00	0.00	0.00	0.00	0.00
max	0.98	0.95	.90	.74	0.61	.54	.50	.45	.37	.30

Let moreover $Q(t) = (1-a)t/G(t)$.

Genovese and Wasserman (2002) prove that $E\Gamma(t) = Q(t)(1-(1-G(t))^m)$. As long as I_i is independent of H_j given I_j, it's easy to see that[2] $E(\Gamma(t)|I^m) = Q(t)1_{\{some\ p_i \leq t\}}$. Taking expectations, we have $E(\Gamma(t)) = Q(t)(1-(1-G(t))^m)$, which is the same expression as above. For the variance of the FDP process, we derive an expression in next lemma.

Lemma 1 *The variance of the FDP process $V(\Gamma(t))$ is equal under dependence to*

$$Q(t)^2(1-(1-G(t))^m)(1-G(t))^m$$
$$+E\left[\frac{\sum_{i\neq j} I_i I_j [\Pr(H_i=0, H_j=0|I^m) - Q(t)^2]}{(\sum I_i + \prod(1-I_i))^2}\right].$$

Proof To compute the variance, we will just apply the well known formula:

$$V(\Gamma(t)) = V(E(\Gamma(t)|I^m)) + E(V(\Gamma(t)|I^m)).$$

Note that $\Pr(H_i = 0|I_i = 1) = Q(t)$.

Then, independently of the correlation structure,

$$E(\Gamma(t)|I^m) = Q(t)1_{\{\text{some } p_i \leq t\}}$$

and[3]

$$V(Q(t)1_{\{\text{some } p_i \leq t\}}) = Q^2(t)(1-(1-G(t))^m)(1-G(t))^m.$$

[2] If the p-values are dependent, it is not reasonable to assume p_i is independent of H_j, $i \neq j$, tout court. It is reasonable, by the way, to assume conditional independence between p_i and H_j, $i \neq j$.

[3] Note that $I_i \sim Bernoulli < G(t) >$.

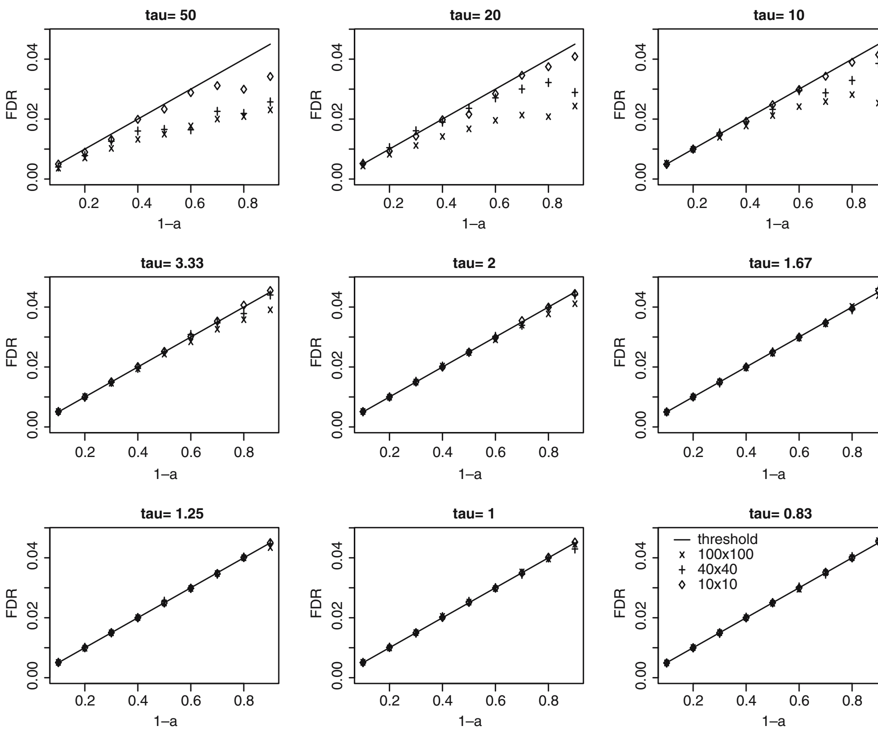

Fig. 1 False discovery rate (FDR) for BH method, normal random variables with $\mathrm{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}\, d((i,j),(i',j'))}$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5)

On the other hand,

$$V(\Gamma(t)|I^m) = E(\Gamma^2(t)|I^m) - Q^2(t)1_{\{\text{some } p_i \leq t\}},$$

and

$$\begin{aligned} E(\Gamma^2(t)|I^m) &= E\left(\frac{\sum_{ij} I_i I_j (1-H_i)(1-H_j)}{\left(\sum I_i + \prod(1-I_i)\right)^2}|I^m\right) \\ &= \frac{\sum_i I_i \Pr(H_i = 0|I_i)^2 + \sum_{i\neq j} I_i I_j \Pr(H_i = 0, H_j = 0|I^m)}{\left(\sum I_i + \prod(1-I_i)\right)^2} \\ &= Q(t)^2 1_{\{\text{some } p_i < t\}} + \frac{\sum_{i\neq j} I_i I_j [\Pr(H_i = 0, H_j = 0|I^m) - Q(t)^2]}{\left(\sum I_i + \prod(1-I_i)\right)^2}. \end{aligned}$$

In the end,

$$\begin{aligned} V(\Gamma(t)) = {} & Q(t)^2(1-(1-G(t))^m)(1-G(t))^m \\ & + E\left[\frac{\sum_{i\neq j} I_i I_j [\Pr(H_i = 0, H_j = 0|I^m) - Q(t)^2]}{\left(\sum I_i + \prod(1-I_i)\right)^2}\right]. \end{aligned}$$

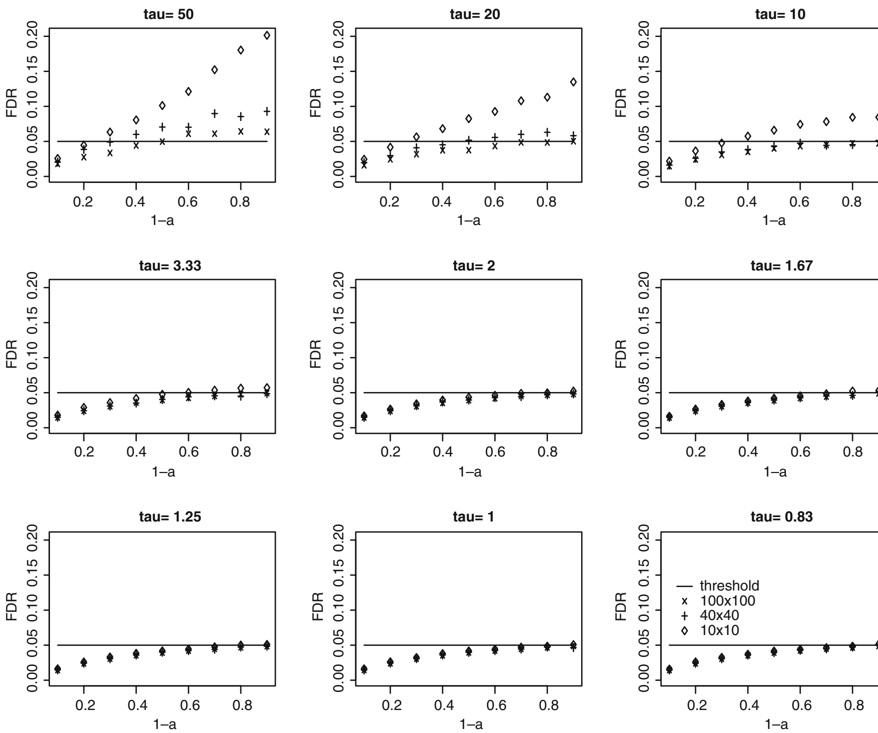

Fig. 2 FDR for plug-in method, normal random variables with $\mathrm{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}\, d((i,j),(i',j'))}$, 1,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Storey's estimator used for a

It is apparent that, if all the p values are independent; then the second term in the expression of $V(\Gamma(t))$ as derived in Lemma 1 will be zero. This reasoning suggests that in general it may not be sensible to control the FDR under strong dependence: since the variance of the FDP may be increased by dependence, control of the quantiles of the distribution of the FDP is in general more desirable. A similar argument is made in Bickel (2004), and such procedures are proposed in van der Laan et al. (2003) and Genovese and Wasserman (2004a).

3 The simulations

3.1 Gaussian data

We applied the BH method and the plug-in method proposed by Genovese and Wasserman (2004b) to correlated spatial data; to illustrate the effects of dependence on the outcomes of standard iid methods for controlling the FDR.

The case of stationary and isotropic spatial data is particularly interesting; since data are usually correlated when close to each other, but as distance increases the correlations fade to 0. A good reference on this setting is Smith (2001).

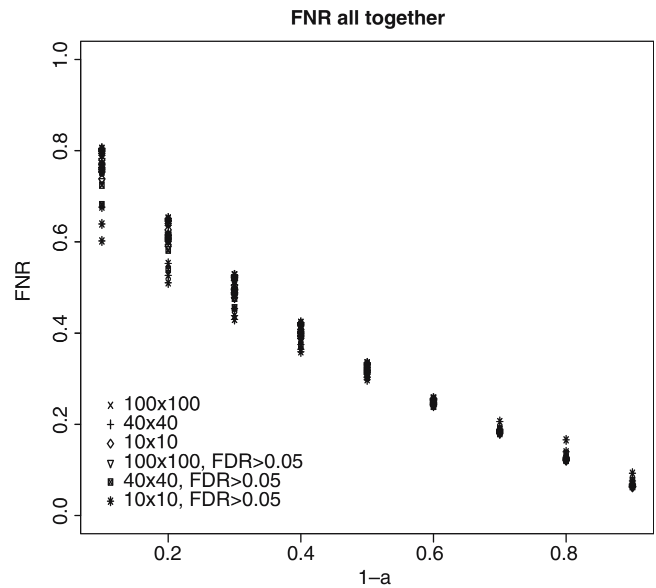

Fig. 3 False non-discovery rate (FNR) for both plug-in and BH method, normal random variables with $\mathrm{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}\, d((i,j),(i',j'))}$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Storey's estimator used for a

We will simulate data scattered on a regular quadratic grid of n by n pixels, forming a $m = n^2$ multivariate normal. We want to test the mean on each pixel, to discover if it is different than zero. So we will need to do m tests. Note that this can be a very common setting, for instance, in neuroimaging; where some variables in m spots of the brain are measured to see if there is neuron activity (Worsley et al., 1996).

We will randomly assign m_1 pixels to a non-zero mean (uniform in (0, 5)); and the variance/covariance matrix will remain the same throughout the iterations. For each set of parameters, we will do 1,000 iterations.

Our purpose is to compare the outcomes of the BH and plug-in procedures in the correlated case with the outcomes in the independent case.

3.1.1 The covariance structures

We will use three covariance structures. The first is the independent one, in which the covariance between two different pixels is 0, while the variance will be taken to be 1. The second is a simplified version of the usual exponential covariance structure: the covariance between two different pixels will always be non-negative, and determined by $e^{-\frac{1}{\tau}\, d(x,y)}$, where x and y are the coordinates on the plane of the two pixels, $d(\cdot,\cdot)$ is the euclidean distance function and $1/\tau$ is just a tuning parameter. The higher τ, the more slowly decaying the correlation.

The third covariance structure will allow for both positive and negative covariances, and a suitable kernel will be given by the "damped cosine" function:

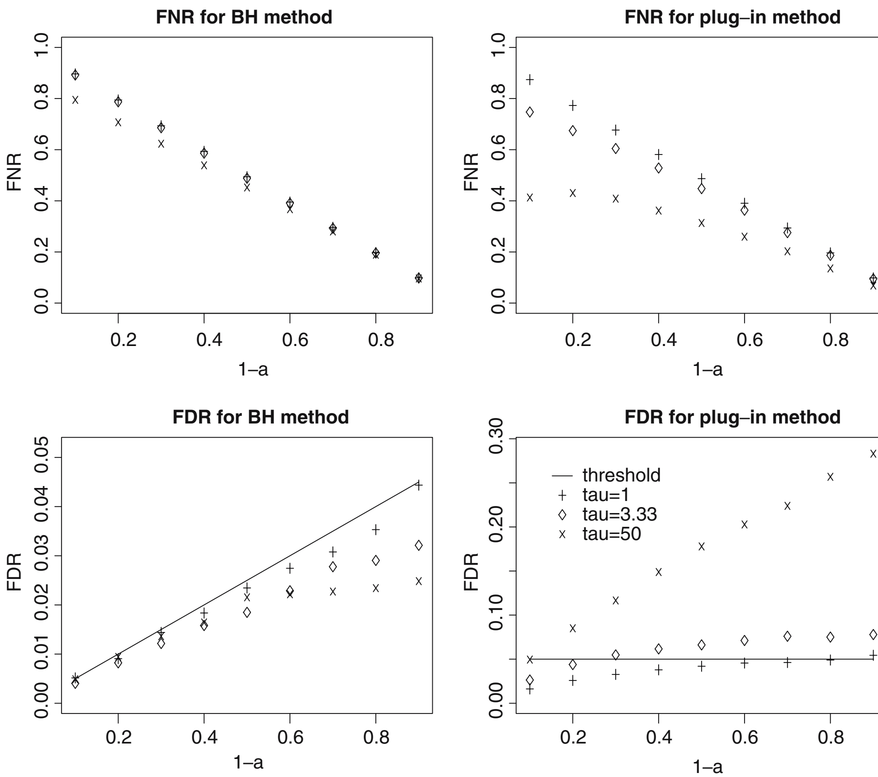

Fig. 4 FNR and FDR for both plug-in and BH method, $m = 100$ normal random variables with $\mathrm{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}\, d((i,j),(i',j'))}$, 10,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean equal to 1. Storey's estimator used for a

$$e^{-\frac{1}{\tau}\, d(x,y)} \cos\left(\frac{1}{\tau}\, d(x, y)\right).$$

Note that this covariance function does *not* meet the condition of Benjamini and Yekutieli (2001).

Abrahmsen (1997) states that the lower bound for the correlation value of normal is -0.4 in two dimensions; and otherwise the variance–covariance matrix will not be positive definite. This structure will allow us to have correlations as low as -0.39; thus being almost as extreme as possible.

The following analyses are done for $n = 10 \times 10$, 40×40 and 100×100 grids.

3.1.2 Results, positive correlations

Tables 2, 3 and 4 show the range of the covariances on each grid for the tuning parameters chosen, in the positive case[4].

[4] All the simulations were programmed in C language.

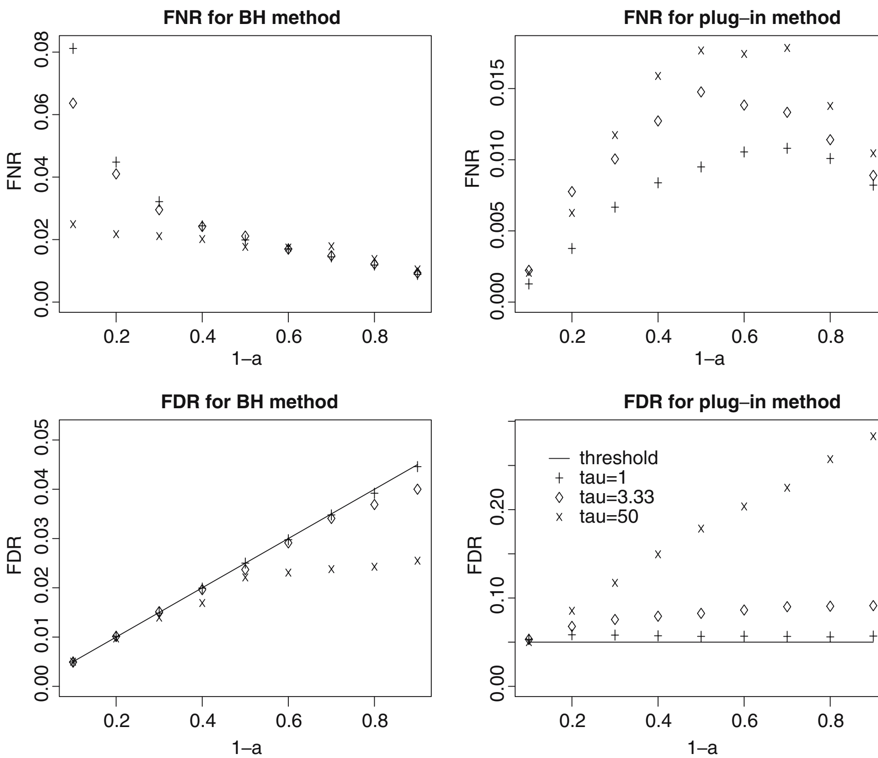

Fig. 5 FNR and FDR for both plug-in and BH method, $m = 100$ normal random variables with $\text{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau} d((i,j),(i',j'))}$, 10,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean equal to 4. Storey's estimator used for a

Table 5 Correlation ranges, negative model

τ	0.2	0.33	0.73	0.77	0.83	0.98	14.29	16.67	20
min	0.00	−0.05	−0.10	−0.16	−0.24	−0.36	−0.39	−0.34	−0.22
max	0.00	0.01	0.14	0.14	0.11	0.13	0.91	0.92	0.94

Figure 1 shows the average FDR controlled using the BH method. Figure 2 shows the average FDR controlled using the plug-in method.

It is apparent that, as long as the relationship between the variables fades to independence fast enough, the methods are still working; and plug-in is sensibly less conservative than BH procedure. When the correlation becomes too strong, BH is valid but becomes even more conservative, while the plug-in violates the threshold condition and gets bigger than .05. We will show later that this problem is caused by the estimator of a.

The behavior of FNR is particularly interesting, and leaves way for further research. Figure 3 shows that the FNR was practically the same in all correlation settings, and the number of tests (m) did not bring about changes in the behavior of the FNR. The lower points come from the plug-in method, while the upper ones

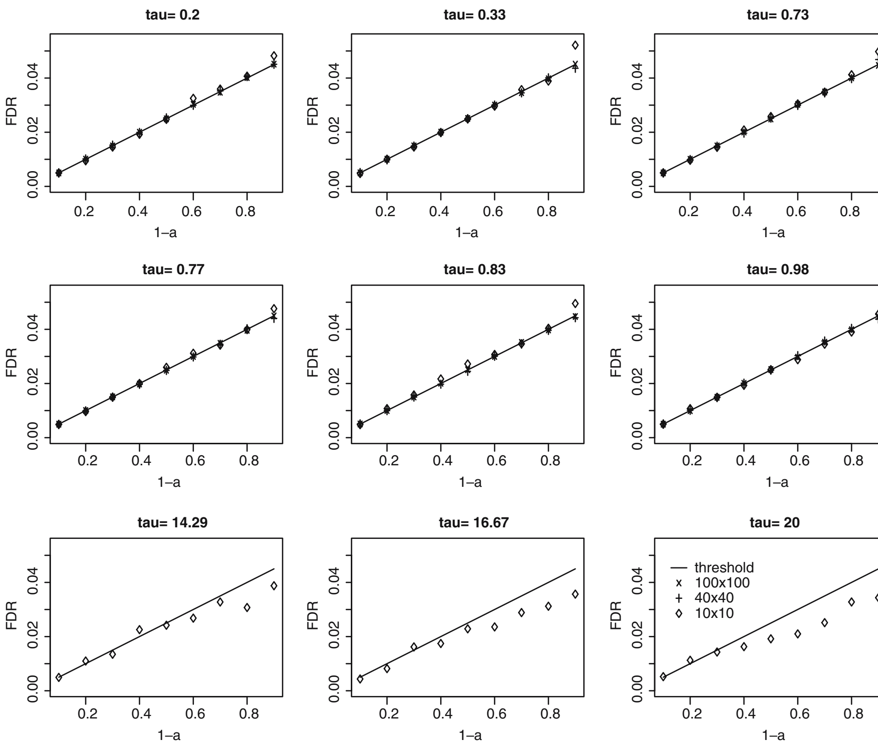

Fig. 6 FDR for BH method, normal random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}\, d((i,j),(i',j'))}$ $\cos(\frac{1}{\tau}\, d((i, j), (i', j')))$, 1000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5)

come from the BH method. The only significant differences can be seen in the cases in which the FDR is not controlled by the methods (and the grid is small), which were plotted with a different symbol. This is not surprising: the violation of the threshold comes from an higher number of rejections, which leads also to an artificially lower FNR. This is more clearly visible in Figures 4 and 5 described below. Other simulations with different covariance structures are not shown for reasons of space. We should point out, by the way, that the FDR itself is also surprisingly robust with respect to dependence, and it is probably partly the scale of the graphs that leads us to judge the FNR as even more robust.

Figures 4 and 5 show the effect of the distribution of the p-values under the alternative. When the non-zero means are taken all equal to 1, strong positive dependence leads to a lower FNR in the cases discussed above. When the non-zero means are taken all equal to 4, a clear pattern can not be seen but still it can be said that strong positive dependence may lead to an higher FNR in case the plug-in method is used, and that the magnitude of the error measure is much lower.

Our main setting, in which the non-zero means are sampled from a uniform on (0, 5), is a sort of "average out" of the effects of taking the non-zero means all equal to b, for $b \in (0, 5)$; with the effects for the higher values of the means masked out because of the much smaller magnitude of the error measure.

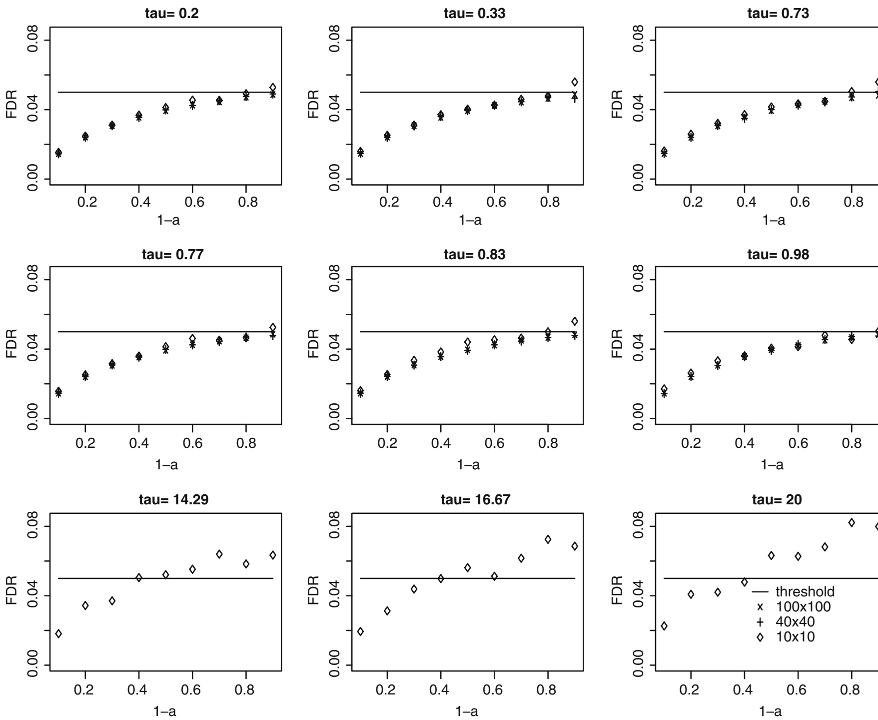

Fig. 7 FDR for plug-in method, normal random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}d((i,j),(i',j'))}$ $\cos(\frac{1}{\tau}d((i,j),(i',j')))$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on $(0, 5)$. Storey's estimator used for a

3.1.3 Results, negative correlations

Table 5 shows the range of the covariances for the negative correlation structure, for all the grids. During the simulation of the bigger grids, certain cases were dropped because the variance covariance matrix lost the positive definiteness property, due to approximation error.

Figure 6 shows the average FDR controlled using the BH method. Figure 7 shows the average FDR using the plug-in method. In both figures and all the following the "threshold" is the nominal level at which the FDR is to be controlled, which is always $(1-a)\alpha$ if the BH method is used, and α if the plug-in method is used. In all cases α was taken equal to 0.05.

Figure 8 shows that the FNR was practically the same in all correlation settings. The line is an interpolation line.

The negative case yields results at least as good as the positive one. Note that there were problems when the correlation became too high (bigger than 0.9), and nothing wrong was observed in the cases in which the correlation was too low.

In order to better distinguish the effect of positive and negative correlations in this case, a simulation with a different covariance structure was done. The correlation between two variables was taken equal to $-1/m$ (recall that, in case of

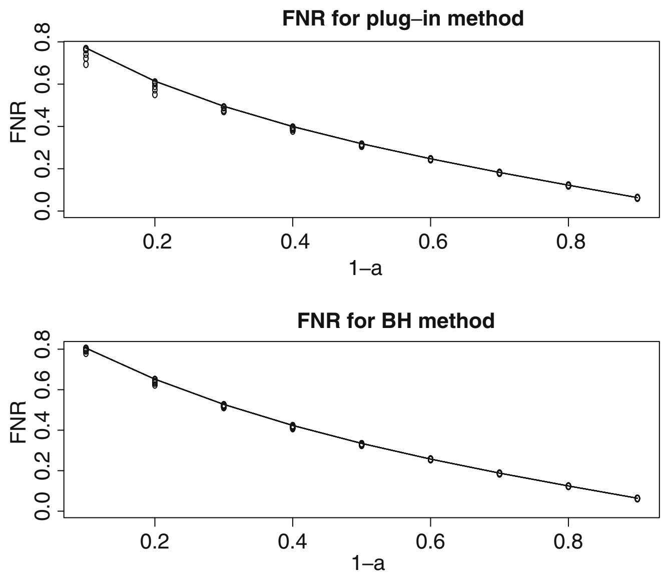

Fig. 8 FNR for both plug-in and BH method, normal random variables with $cor(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}\ d((i,j),(i',j'))} \cos(\frac{1}{\tau}\ d((i,j),(i',j')))$, 1000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Storey's estimator used for a

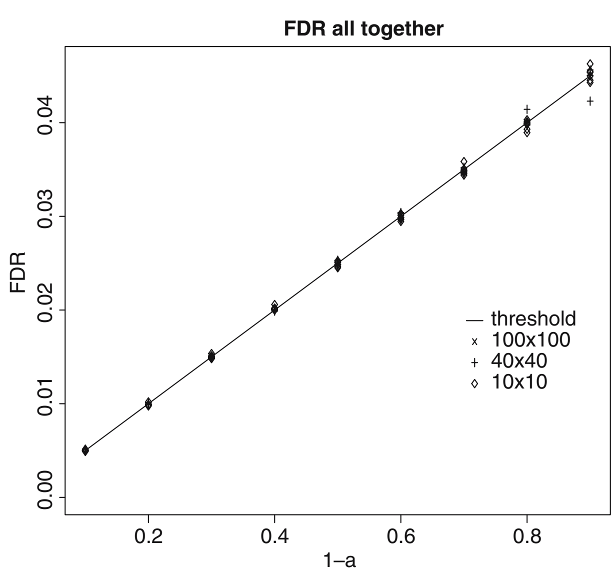

Fig. 9 FDR for plug-in method, normal random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'})$ equal to $0, -1/m$, $-2/m$ or $-1/2$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Storey's estimator used for a

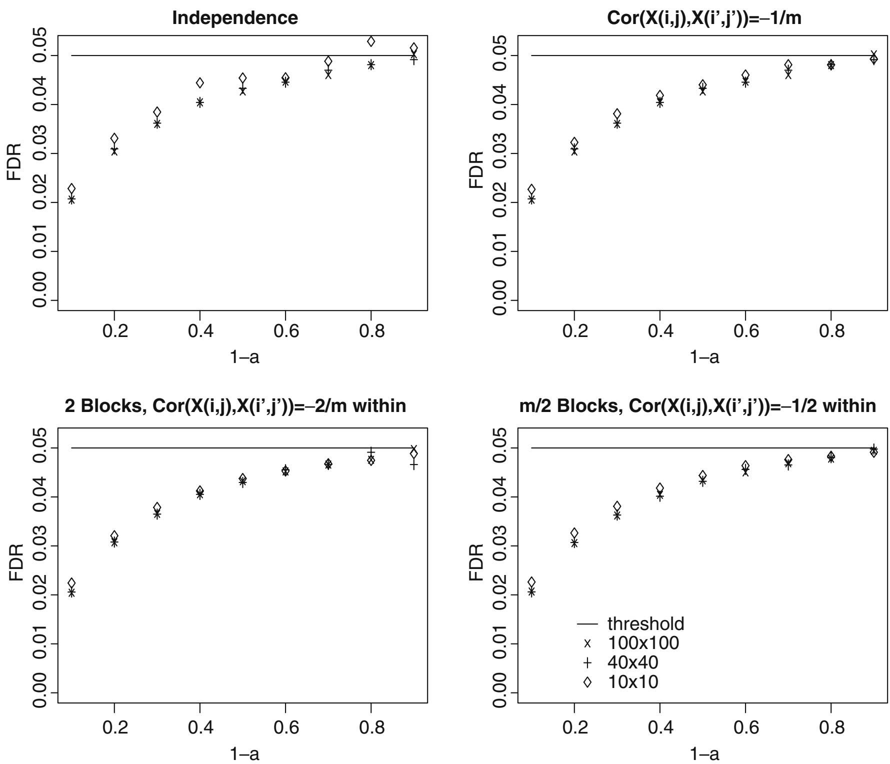

Fig. 10 FDR for plug-in method, normal random variables with $\text{cor}(X_{i,j}, X_{i',j'})$ equal to 0, $-1/m$, $-2/m$ or $-1/2$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on $(0, 5)$

constant correlation, the minimum must be bigger than $-1/(m-1)$). Then, block independence was assumed, and two different configurations were used: first, the variables were divided in two blocks, with correlation between pairs of variables within the blocks equal to $-2/m$; then the variables were divided in $m/2$ blocks, with correlation between pairs of variables within the blocks equal to $-1/2$. Finally, independent random variables were used.

Figure 9 shows the average FDR controlled using the BH method. Figure 10 shows the average FDR using the plug-in method. It can be seen that the average FDR is always very close to the threshold; and it cannot be distinguished among the correlation structures, and the three values of m used.

Figures 11 and 12 show that the FNR presents the same behaviour as the previous cases, with no significant changes brought about by m or by the correlation structures.

3.2 Pearson Type VII data

The same random fields were simulated with Pearson Type VII random variables. The degrees of freedom for each t variable were chosen to be 3, because these

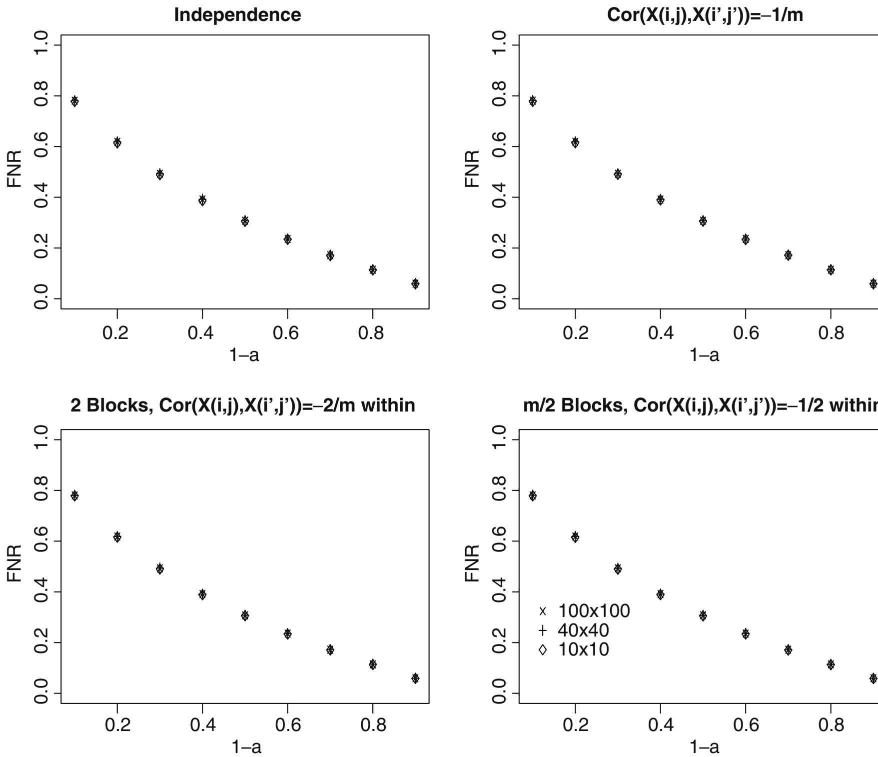

Fig. 11 FNR for BH method, normal random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'})$ equal to 0, $-1/m$, $-2/m$ or $-1/2$, 1,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Storey's estimator used for a

random variables were the farthest from normality, but with finite mean and variance. Using three degrees of freedom makes also straightforward the comparison with the normal case, since the correlation matrix, using the same parameters, will be unchanged. Refer to Johnson (1987) for the methods used to do this multivariate simulation.

Figure 13 shows the average FDR controlled using the BH method. The stronger correlation structures made the procedure be conservative also in this case. There is also more instability in the observed FDRs (and even a couple of FDRs just over the threshold), most likely due to higher simulation variability. Figure 14 shows the average FDR controlled using the plug-in method. The break down for high correlation is slightly worse than the normal case. Note that previously $\tau = 3.33$ did not show any violation of the threshold.

Once again, in order to better distinguish the effect of positive and negative correlations, the correlation between two variables was taken equal to $-1/m$; and then block dependence was assumed, and two different configurations used: first, the variables were divided in two blocks, with correlation between pairs of variables within the blocks equal to $-2/m$; then the variables were divided in $m/2$ blocks,

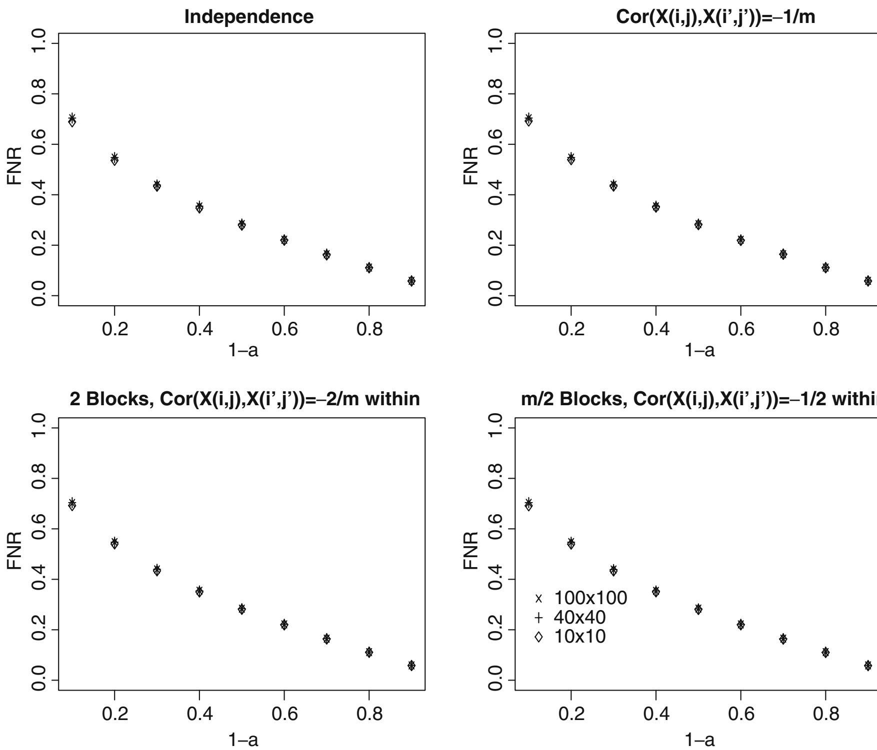

Fig. 12 FNR for plug-in method, normal random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'})$ equal to 0, $-1/m$, $-2/m$ or $-1/2$, 1,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Storey's estimator used for a

with correlation between pairs of variables within the blocks equal to $-1/2$. Finally, independent random variables were used.

Figure 15 shows the average FDR controlled using the BH method. Figure 16 shows the average FDR using the plug-in method.

Figures 17 and 18 show that the FNR presents the same behaviour as the previous cases.

In summary, it seems like strong positive correlations can lead the BH method to be more conservative than it already is; but it will not lead it to violate the threshold for the FDR. The FNR is not affected by dependence in the data, unless the FDR goes out of control. Plug-in method leads to violation of the threshold when dependence becomes too strong. Simulations in the next section show that this is due to the estimators used for a.

3.3 Estimating a with dependent data

A key thing in the plug-in method is the choice of the estimator for a. It is obvious that, as long as $0 \leq \widehat{a} \leq a$, the power is increased with respect to the BH procedure; while the FDR is still below the desired threshold. So, if anything, a statistic

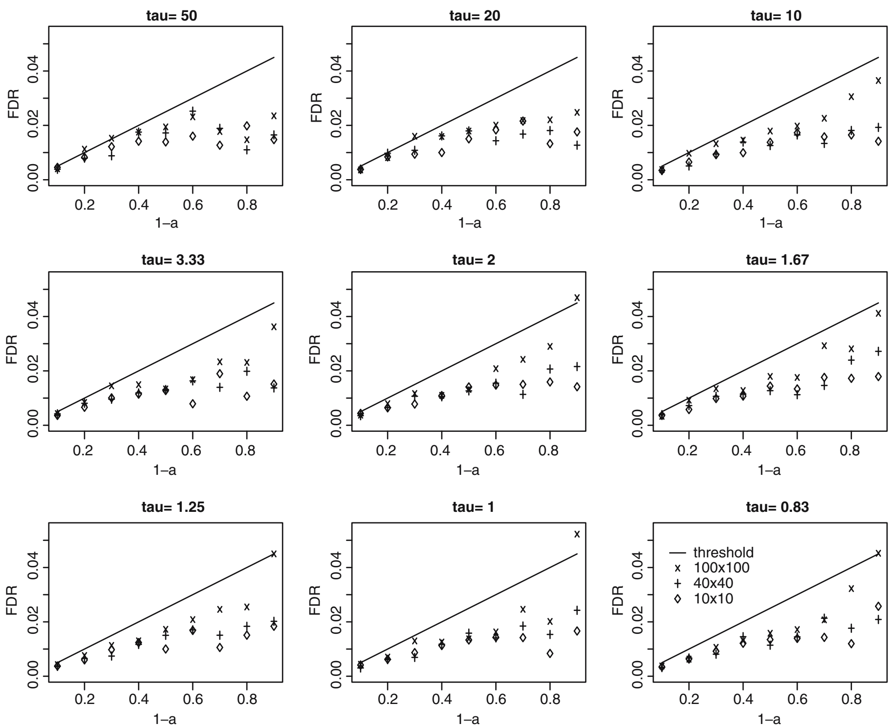

Fig. 13 FDR for BH method, Pearson Type VII random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau} d((i,j),(i',j'))}$ and three degrees of freedom, 1,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5)

that underestimates a is desirable because it improves on BH while being at least conservative. We see now that common estimators of a are not in general conservative under dependence, and propose conservative estimators (under arbitrary dependence) in the next section.

3.3.1 Oracle simulations

In an attempt to understand why the plug-in method failed when the correlation between close variables was very high, we implemented a simulation in which the proportion of true nulls, $1 - a$, was considered known and used in the procedure, i.e., we took $\widehat{a} = a$. Figure 19 shows the results: the estimator used for a was the thing that broke down when using strong correlations, while now the plug-in method works just as it should. The case of strong correlations, when parameter is close to 0, brings about just an increase in the variance (less stability).[5]

It is easy to see how Storey's estimator breaks down by strongly overestimating a. Figure 20 shows $\frac{\widehat{a}}{a}$, where $\widehat{a}$ is Storey's estimator, in the usual 10×10 Gaussian

[5] Note that in real cases it is reasonable to assume sparseness, i.e., $1 - a > .5$ or even much more. Then, as we can see, it is reasonable not to expect problems even from Storey's estimator.

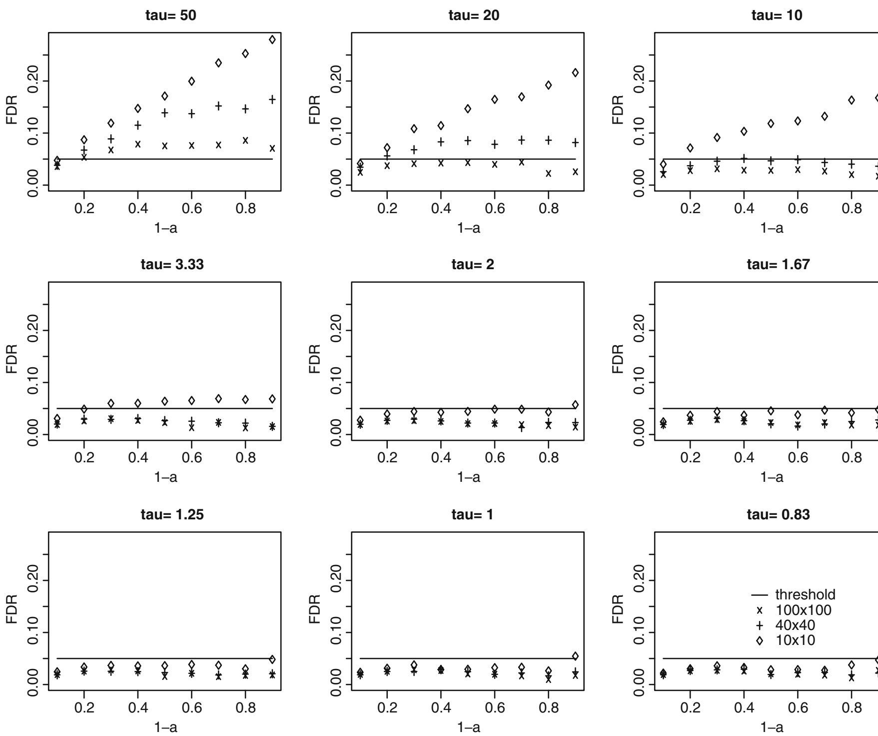

Fig. 14 FDR for plug-in method, Pearson Type VII random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}\, d((i,j),(i',j'))}$ and three degrees of freedom, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on $(0, 5)$. Storey's estimator used for a

random fields. We observed values as much as 3.5 times the real a. Also the other classical estimators, like the one in Swanepoel (1999) or the one in Woodroofe and Sun (1999), are seen to break down under dependence.[6]

4 Estimating the number of false null hypotheses

In this section, we will briefly give some results and explicit proposals for estimating the number of false null hypotheses. Such estimates, as seen, can greatly improve the power of many multiple testing procedures (MTPs), for instance in passing from the BH to the plug-in method. Many other procedures can benefit from a good estimator of M_1, as in many cases m is used as a rough approximation of M_0. For instance it is well known that, if M_0 were known, rejection of $p_j < \alpha/M_0$ would yield (exact[7]) control the FWER at level α. Here we propose the estimate $\widehat{M}_0 = m - \widehat{M}_1$.

[6] Simulations of the other estimators not shown for reasons of space.
[7] Of course, one may not have either strong or weak control in this way. See Hochberg and Tamhane (1987) for a discussion.

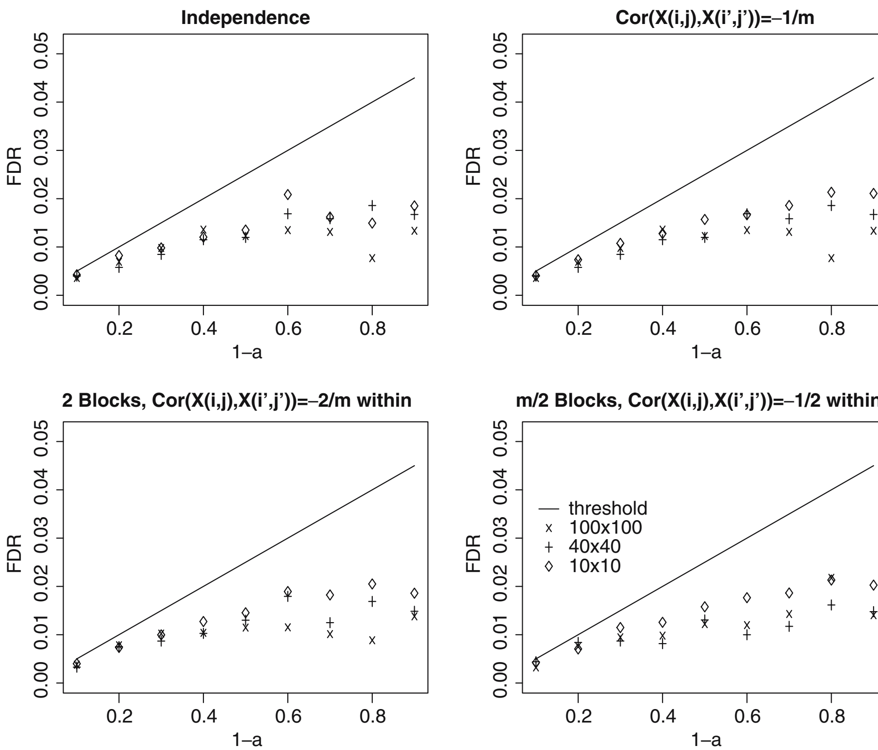

Fig. 15 FDR for plug-in method, Pearson Type VII random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'})$ equal to 0, $-1/m$, $-2/m$ or $-1/2$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on $(0, 5)$. Storey's estimator used for a

The usual estimators are seen to break down under dependence (as shown in section 3.3). We will propose here a class of estimators robust with respect to dependence.

Since we need to estimate M_1, the number of false nulls, we thought the most natural estimator was the number of rejected hypotheses. If we reject R hypotheses, then it is natural to think that $R = M_1$, i.e., that the actual number of false nulls is the number of rejected hypotheses. Then, a can be estimated as $\widehat{a} = R/m$. We suggest, then, to choose a multiple testing procedure, use it to estimate a; and then repeat the same MTP (or another one) to do the actual multiple test, using the estimator for a determined before.

We will do now precise considerations in order to give probabilistic statements on such iterative estimators, and refine them depending on the MTP chosen to determine the number of rejections R.

First, recall that a suitable estimator for M_1 is a conservative one. That is, we want $\widehat{m}_1 \leq M_1$, but as close as possible to the upper bound. This will increase the power without violating the condition on the Type I error rate. This is equivalent to looking for a confidence interval for M_1, which will be in the form $[\widehat{m}_1, m]$.

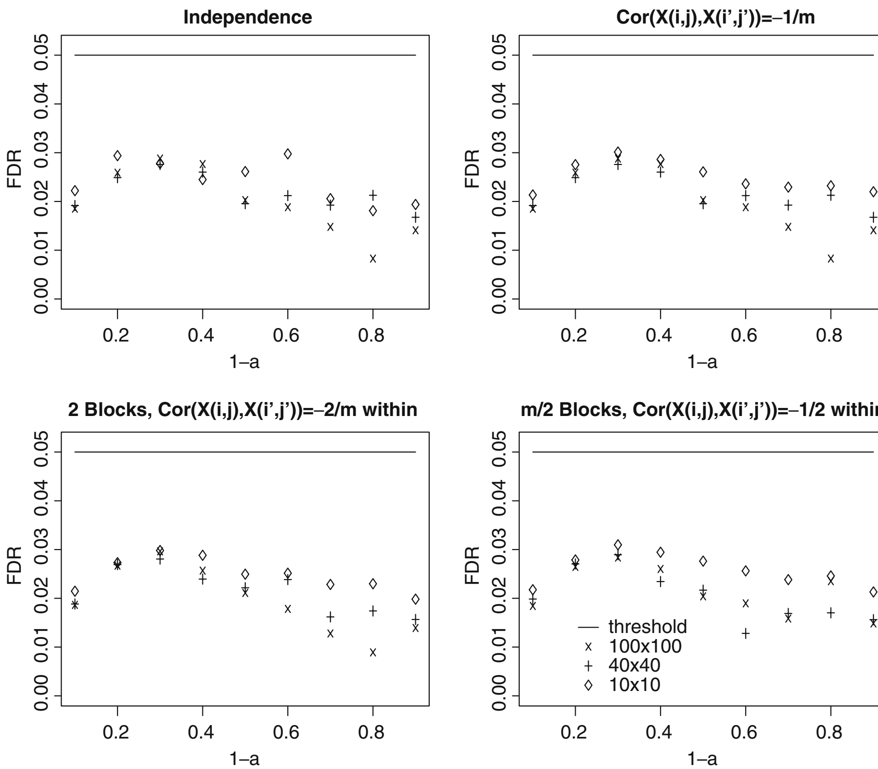

Fig. 16 FDR for plug-in method, Pearson Type VII random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'})$ equal to 0, $-1/m$, $-2/m$ or $-1/2$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5)

The basic idea is as follows: first of all, note that in the notation of Table 1,

$$M_1 = R - N_{1|0} + N_{0|1}.$$

We usually do not know much directly on M_1, while the random variables on the right hand side are dealt with in MTPs.

We can conservatively approximate $N_{0|1}$ to 0 and require that with high probability $\widehat{m}_1 \leq R - N_{1|0}$, on which we have bounds. A more complex path is to include $N_{0|1}$ in the considerations.

Note that whenever $N_{0|1} \geq N_{1|0}$, R is a good conservative estimator of M_1. All depends on the controlled error measure, α, m and $F(\cdot)$. In general, anyway, experience and simulations suggest that this is often true, especially for big m.

In what follows, we will always take into account the uncertainty brought about by the estimation of M_1 when controlling the desired error measure. This will be done by controlling the error measure at a certain level $\alpha_2 \leq \alpha$, which will be exactly determined. It is common in literature *not* to incorporate this uncertainty. Obviously, in what follows, this corresponds to using $\alpha_2 = \alpha$. We will make some comparisons between these two choices at the end of the section.

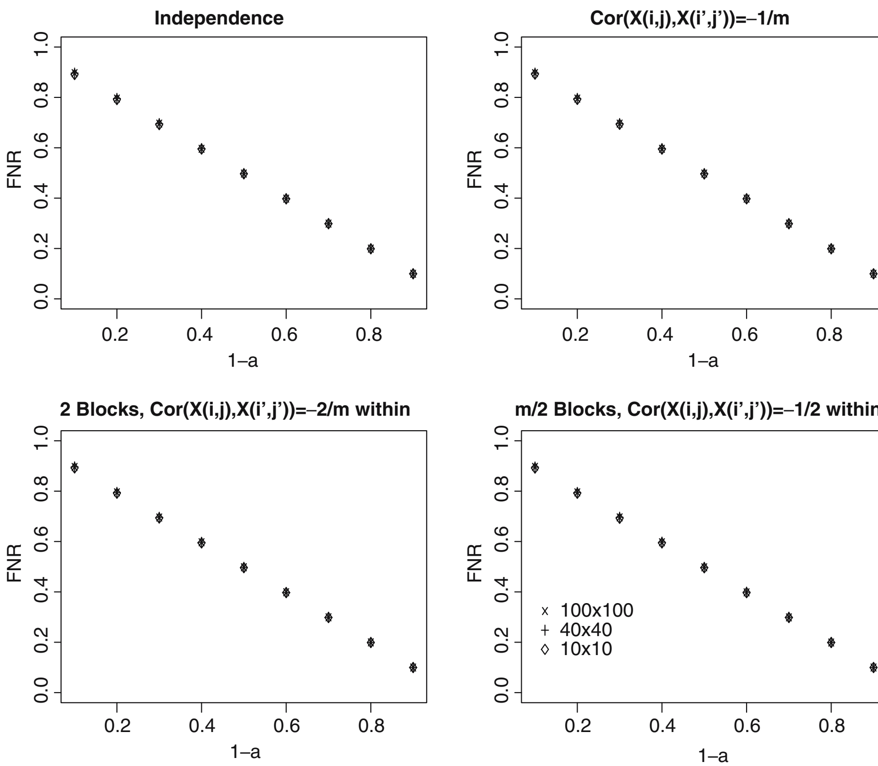

Fig. 17 FNR for BH method, Pearson Type VII random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'})$ equal to 0, $-1/m$, $-2/m$ or $-1/2$, 1,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on $(0, 5)$. Storey's estimator used for a

4.1 Two-step procedures

We define a k-step procedure as a procedure that estimates M_1 through $k - 1$ MTP steps, and then controls a pre-specified error rate using the last estimate found. One can either fix k, or iterate till the estimator does not change in two subsequent iterations. In this case, k is unknown and random, but finite almost surely: there are only $m + 1$ possible values for $\widehat{a}$; and the random variable given by the difference between the current and the previous estimate is discrete and puts a non null probability mass at 0, which is our stopping rule. Such "random k" iterative estimators possess an interesting internal coherence property: the final number of rejected hypotheses is used as an estimator of the number of false nulls. This is noted also in Benjamini, Krieger, and Yekutieli (2004), who independently derived a similar estimator with a modified BH procedure at the estimation steps, and proved that when $k = 2$ their estimator is conservative almost surely.

A particular case is given by two step procedures, i.e., by the choice $k = 2$, which we will consider now. In section 4.2 we will then generalize to arbitrary k. In our calculations, we will always condition on R, since we will know it from the previous step.

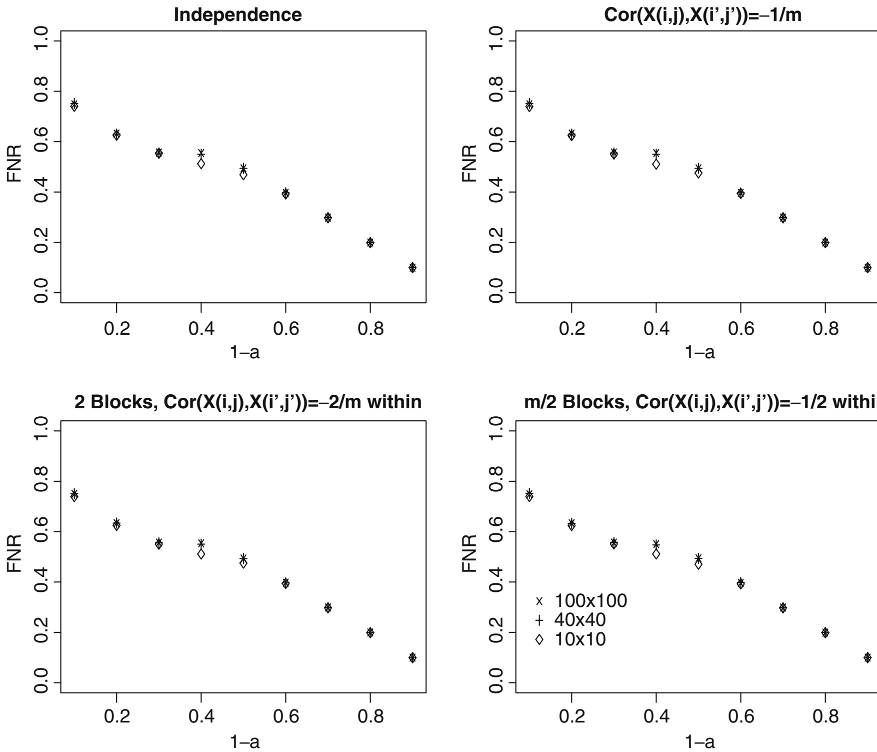

Fig. 18 FNR for plug-in method, Pearson Type VII random variables with $\mathrm{cor}(X_{i,j}, X_{i',j'})$ equal to 0, $-1/m$, $-2/m$ or $-1/2$, 1,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Storey's estimator used for a

4.1.1 Two-step procedures based on FWER control

Many FWER controlling procedures work under arbitrary dependence (like Bonferroni). Let R_{Bonf} denote the number of rejected hypotheses controlling the FWER at level α_1.

It is easily seen that, under arbitrary dependence, $\widehat{m}_1 = R_{\mathrm{Bonf}}$ is conservative with high probability: $\Pr(R_{\mathrm{Bonf}} < R_{\mathrm{Bonf}} - N_{1|0} + N_{0|1}) \geq \Pr(N_{1|0} = 0) > 1 - \alpha_1$. Note that the second inequality in practice is always strict, and that typically $\Pr(N_{0|1} > 0) >> 0$, so that the bound is far from being sharp.

Two-step control of FWER under arbitrary dependence At the second step, one can reject the hypotheses corresponding to all the p-values smaller than $\alpha_2/(m - R_{\mathrm{Bonf}})$ and exactly control FWER at level $\alpha = 1 - (1 - \alpha_1)(1 - \alpha_2)$ in this way, under arbitrary dependence, as stated in next proposition.

Proposition 1 *Two-step Bonferroni just described is such that* $\Pr(N_{0|1} > 0) \leq 1 - (1 - \alpha_1)(1 - \alpha_2)$ *under arbitrary dependence.*

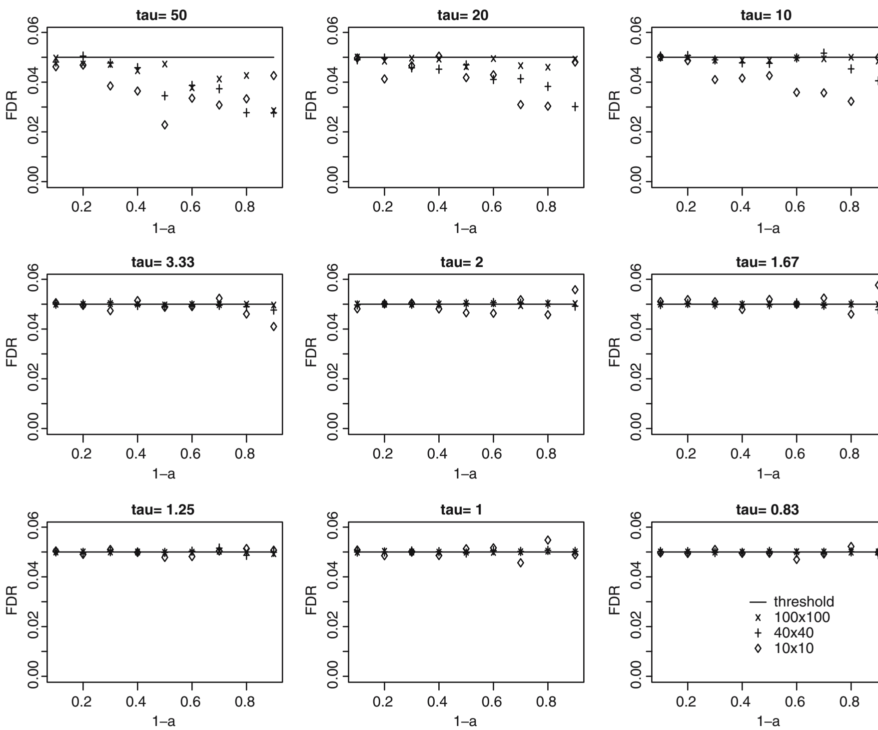

Fig. 19 FDR for plug-in method, normal random variables with $\mathrm{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau} d((i,j),(i',j'))}$ $\cos(\frac{1}{\tau}\, d((i, j), (i', j')))$, 1,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). The proportion a is known and *not* estimated

Proof

$$\begin{aligned}
\Pr(N_{0|1} = 0) &= \Pr(N_{0|1} = 0 | R_{\text{Bonf}} < M_1) \Pr(R_{\text{Bonf}} < M_1) \\
&\quad + \Pr(N_{0|1} = 0 | R_{\text{Bonf}} > M_1) \Pr(R_{\text{Bonf}} > M_1) \\
&\geq (1 - \alpha_2)(1 - \alpha_1) + \Pr(N_{0|1} = 0 | R_{\text{Bonf}} > M_1) \Pr(R_{\text{Bonf}} > M_1) \\
&\geq (1 - \alpha_2)(1 - \alpha_1)
\end{aligned}$$

Two-step control of FDR under PRDS assumption Now we can improve on the results of Benjamini and Yekutieli (2001). If at the second step a plug-in procedure at level α_2 is done, FDR is controlled at level $\alpha = \alpha_2(1 + (R_{\text{Bonf}}/(m - R_{\text{Bonf}}))\alpha_1)$ for any m under PRDS assumptions (as defined in Benjamini and Yekutieli 2001), as stated in next proposition.

Proposition 2 *Two-step plug-in with Bonferroni at first step is such that* $\text{FDR} \leq \alpha_2(1 + (R_{\text{Bonf}}/(m - R_{\text{Bonf}}))\alpha_1)$ *under* PRDS.

Table 6 Observed FDR for multi-step control, using Bonferroni at estimation steps

	$\alpha_1 = 0$ (BH)	$\alpha_1 = 0.05$, no correction	$\alpha_1 = 0.05$, corrected
$M_1 = 0.1 * m$			
$m = 10$	0.0418	0.0471	0.0468
$m = 30$	0.0448	0.0467	0.0465
$m = 100$	0.0408	0.0424	0.0423
$m = 500$	0.0445	0.0459	0.0459
$m = 1000$	0.0449	0.0465	0.0464
$m = 5000$	0.0451	0.0465	0.0464
$M_1 = 0.5 * m$			
$m = 10$	0.0266	0.0354	0.0347
$m = 30$	0.0241	0.0315	0.0309
$m = 100$	0.0254	0.0312	0.0309
$m = 500$	0.0249	0.0288	0.0286
$m = 1000$	0.0251	0.0286	0.0284
$m = 5000$	0.0248	0.0289	0.0288
$M_1 = 0.9 * m$			
$m = 10$	0.0053	0.0112	0.0104
$m = 30$	0.0047	0.0084	0.0078
$m = 100$	0.0050	0.0074	0.0073
$m = 500$	0.0050	0.0067	0.0063
$m = 1000$	0.0048	0.0064	0.0064
$m = 5000$	0.0048	0.0061	0.0060

Table 7 Comparison of estimators of M_1

m	M_1	$E[\widehat{m}_1]$ Bonferroni	$E[\widehat{m}_1]$ multistep Bonferroni	$E[\widehat{m}_1]$ BH
5	1	0.581	0.589	0.477
10	1	0.526	0.529	0.430
30	3	1.309	1.316	1.165
100	10	3.48	3.506	3.036
200	20	6.184	6.211	6.311
500	50	13.47	13.53	18.29
1000	100	23.58	23.67	37.68
5000	500	86.95	87.25	218.172
5	3	1.621	1.694	1.364
10	5	2.438	2.531	2.034
30	15	6.272	6.495	5.918
100	50	17.317	17.855	23.903
200	100	31.010	31.866	50.400
500	250	66.261	67.941	129.236
1000	500	118.257	120.957	264.429
5000	2500	483.219	486.191	1629.731
5	4	2.133	2.306	1.802
10	9	4.394	4.772	3.696
30	27	11.187	12.007	12.471
100	95	32.930	35.160	52.000
200	190	58.760	62.333	106.034
500	475	126.606	133.376	272.398
1000	950	222.750	233.520	568.560
5000	4750	826.700	857.420	2903.570

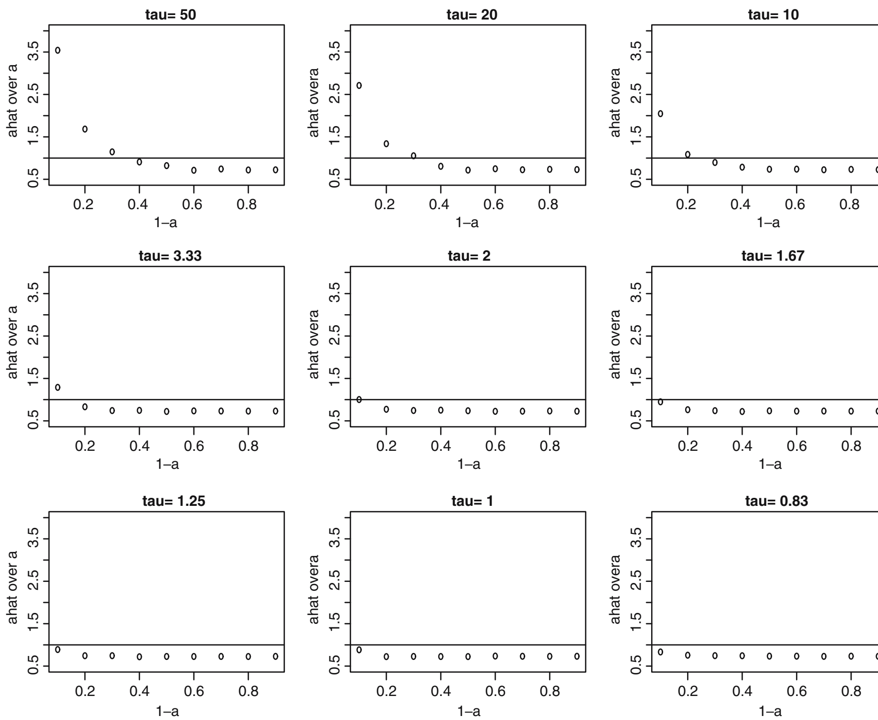

Fig. 20 Storey's Estimator divided by actual value of a, normal random variables with $\mathrm{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}\, d((i,j),(i',j'))} \cos(\frac{1}{\tau}\, d((i,j),(i',j')))$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5)

Proof

$$\begin{aligned} E[\mathrm{FDP}] &= E[\mathrm{FDP}|R_{\mathrm{Bonf}} < M_1]\,\Pr(R_{\mathrm{Bonf}} < M_1) \\ &\quad + E[FDP|R_{\mathrm{Bonf}} > M_1]\,\Pr(R_{\mathrm{Bonf}} > M_1) \\ &\leq \alpha_2 \Pr(R_{\mathrm{Bonf}} < M_1) + \frac{(m - M_1)\alpha_2}{m - R_{\mathrm{Bonf}}}\alpha_1 \\ &\leq \alpha_2 + \frac{m}{m - R_{\mathrm{Bonf}}}\alpha_2\alpha_1. \end{aligned}$$

Note that, that we can explicitly use R_{Bonf} in the bound since it is known. So, taking any $\alpha_1 \in (0, 1)$ and $\alpha_2 = \alpha/(1 + (m/(m - R_{\mathrm{Bonf}})\alpha_1))$ controls the FDR at the desired level α for any finite m under PRDS assumptions. This is a slight generalization of the results of Benjamini and Yekutieli (2001): they proved the result only for $\alpha_1 = 0$. Moreover, it is now possible to achieve better power using a sort of plug-in procedure under dependence. Note moreover that taking no correction and using $\alpha_2 = \alpha$ is sensible since in many cases FDR will be controlled with high probability at the desired level (see below). It is so because $m/(m - R_{\mathrm{Bonf}}))\alpha_1$ is always very close to zero, and ignoring it is just a weak counter balance of the conservative procedure.

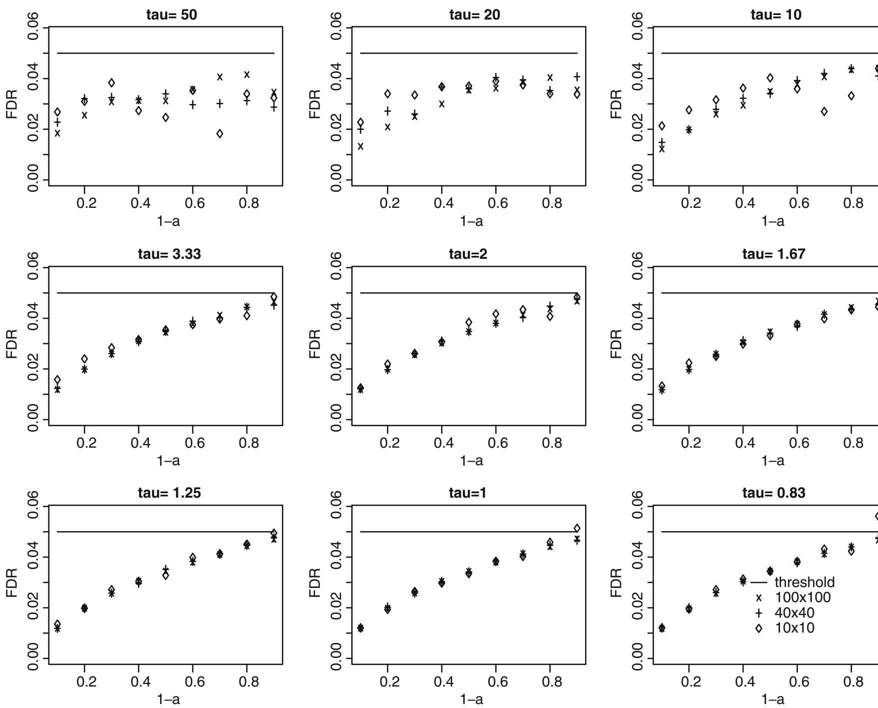

Fig. 21 FDR for plug-in method, normal random variables with $\text{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}d((i,j),(i',j'))}$, 1,000 iterations, a random proportion of $1 - a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Iterative estimator used for a

Table 8 compares the choice of $\alpha_1 = 0$ (BH procedure) with the choice of $\alpha_1 = \alpha$. Simulation shows that choosing $\alpha_1 = \alpha$ yields better power. Recall that the higher the FDR (still below α), the better the procedure in terms of power. This improvement is more and more evident as M_1 increases, for smaller m. Note that a very small increase in the FDR, especially for big m, can result in a much higher number of rejections, and hence in a much higher power.

4.1.2 Two-step procedures based on FDR control

Suppose BH procedure is used at the first step. It is easy to extend our approach to estimation of M_1 by taking an estimator that is good "on average". It is straightforward to see, in fact, that $\lfloor R(1 - \alpha_1) \rfloor$ is on average smaller than M_1 if the FDR is controlled at level α_1 at the first step.

Alternatively, by an easy application of Markov inequality, it is immediately seen that any FDR controlling procedure is such that $\Pr(\text{FDP} > c) < \alpha_1/c$, so that $R_{\text{BH}}(1-c)$ is conservative with probability α_1/c. If a correction for this uncertainty is used (like the ones proposed in the previous sections), the chosen error measure will be controlled at the desired level even if α_1/c is not close to zero.

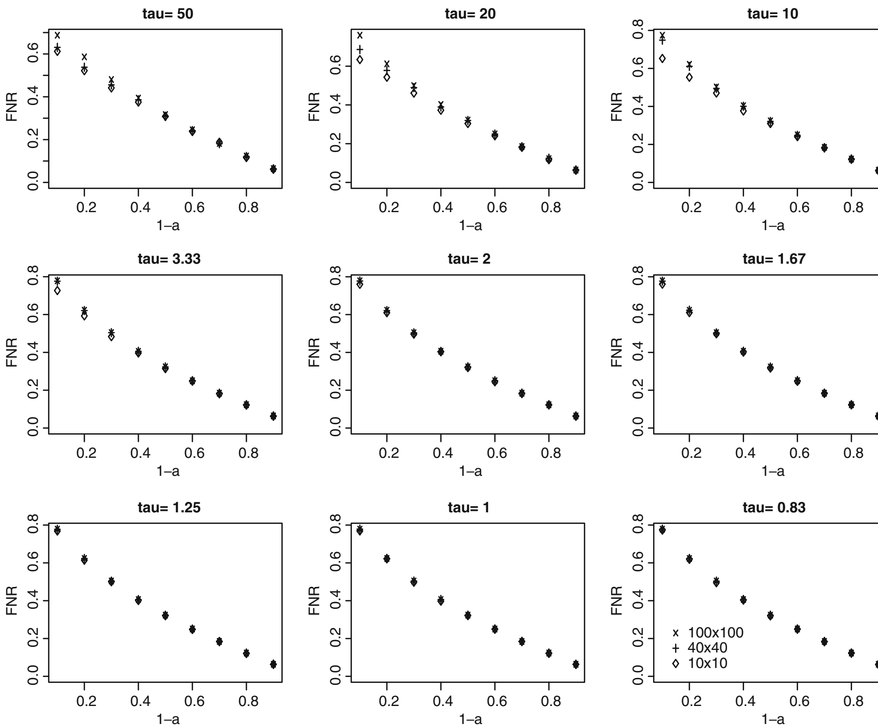

Fig. 22 FNR for plug-in method, normal random variables with $\mathrm{cov}(X_{i,j}, X_{i',j'}) = e^{-\frac{1}{\tau}d((i,j),(i',j'))}$, 1,000 iterations, a random proportion of $1-a$ random variables have a zero mean, the remaining have a non-zero mean sampled from uniform on (0, 5). Iterative estimator used for a

4.2 Multi-step procedures

A generalization of the iterative estimator proposed in the previous chapter is as follows:

1. Pick any multiple testing procedure to estimate M_1.
2. Update the estimator of M_1 by repeating Step 1 with the previous estimate $\widehat{m}_1$.
3. Iterate $(k-1)$-times or till Steps 1 and 2 give the same estimator.
4. Control the desired error measure making use of the most recent value of $\widehat{m}_1$.

It is intuitive that multi-step procedures are less conservative than two-step procedures, and that iterating till the estimator does not change in two subsequent steps is as least conservative as possible.

In practice, the change in the estimate $\widehat{m}_1$ will be smaller and smaller as the number of iterations increase. An appreciation of the improvement in passing from two-step to multi-step estimation of M_1 is given in a comparison of the first two columns of Table 7.

To fix the ideas, we describe the algorithm for a particular choice of Steps 1 and 2:

1. Let $R_B := 0$.

Table 8 Observed FDR for two-step control, using Bonferroni at first step

	$\alpha_1 = 0$ (BH)	$\alpha_1 = 0.05$, no correction	$\alpha_1 = 0.05$, corrected
$M_1 = 0.1 * m$			
$m = 10$	0.0418	0.0471	0.0462
$m = 30$	0.0448	0.0467	0.0448
$m = 100$	0.0408	0.0424	0.0429
$m = 500$	0.0445	0.0459	0.0459
$m = 1000$	0.0449	0.0465	0.0460
$m = 5000$	0.0451	0.0465	0.0457
$M_1 = 0.5 * m$			
$m = 10$	0.0266	0.0349	0.0329
$m = 30$	0.0241	0.0311	0.0301
$m = 100$	0.0254	0.0310	0.0300
$m = 500$	0.0249	0.0287	0.0284
$m = 1000$	0.0251	0.0285	0.0270
$m = 5000$	0.0248	0.0287	0.0274
$M_1 = 0.9 * m$			
$m = 10$	0.0053	0.0099	0.0083
$m = 30$	0.0047	0.0081	0.0081
$m = 100$	0.0050	0.0072	0.0071
$m = 500$	0.0050	0.0067	0.0062
$m = 1000$	0.0048	0.0064	0.0063
$m = 5000$	0.0048	0.0061	0.0060

Table 9 Average estimators for a, for different methods and true a, $g(\cdot)$ following a Beta density with parameters ρ and 1

m	ρ	a	Two-Step Bonf.	Two-Step BH	k-Step Bonf.	k-Step BH
30	0.1	0.1	0.0549	0.0555	0.0626	0.0634
30	0.3	0.1	0.0155	0.0161	0.0188	0.0191
100	0.1	0.1	0.0465	0.0468	0.0586	0.0592
100	0.3	0.1	0.0109	0.0109	0.0148	0.0149
30	0.1	0.3	0.1572	0.1607	0.1962	0.2026
30	0.3	0.3	0.0471	0.0490	0.0671	0.0701
100	0.1	0.3	0.1399	0.1422	0.1975	0.2051
100	0.3	0.3	0.0313	0.0317	0.0571	0.0588

2. Let $R_B := |\{j : p_j < \alpha_1/(m - R_B)\}|$, where $|\cdot|$ denotes the cardinality of a set.
3. Iterate till Steps 1 and 2 give the same estimator.
4. Let $\widehat{m}_1$ be the number of rejected hypotheses at the previous step. Do a plug-in method to control the FDR, taking $\widehat{a} = \widehat{m}_1/m$.

Table 6 compares the BH procedure ($\alpha_1 = 0$) with the multi step just described, with and without correction (note that, as said, no correction is done and $\alpha_2 = \alpha$).

4.3 Applications

We now provide two examples of application of our iterative estimators. In the first example, we show that our iterative estimators can be used to apply the plug-in method under dependence. In the second we use our iterative estimators to anti-conservatively estimate the weight of a known component of a mixture, and describe an application to cosmology.

Example 1 (Iterative plug-in method) Let us consider this iterative estimator: do a plug-in procedure starting with $\widehat{a} = 0$ (BH procedure) and then with $\widehat{a}$ equal to the number of rejections at the previous step, till two subsequent steps don't yield the same estimator $\widehat{a}$. Then, apply again the plug-in procedure and reject all the hypotheses corresponding to p-values smaller than the threshold determined in the last step.

For the usual Gaussian simulations with positive correlation structure, Figure 21 shows the results of the plug-in procedure with such iterative estimator.

The average number of steps was always between 2 and 7. Note that this procedure manages to control the FDR at level 5% when the correlation is very strong (robustness), while it behaves just like the old one-step procedure when the correlation is weak (it just seems to be a little bit more conservative). If we compare Figure 21 with Figures 2, 7, 14, we can see that the plug-in method with iterative estimator succeeds in controlling the FDR in the cases in which the use of other estimators lead to violation of the threshold $\alpha = 0.05$.

Figure 22 shows the FNR obtained using the plug-in procedure with the iterative estimator. In this case also the FNR seems very stable with respect to m and τ.

Example 2 (Estimating the weight of a known component of a mixture distribution) Suppose we observe data from the mixture $f(x) = (1 - a)g_1(x) + ag(x)$, where $g_1(\cdot)$ is a known density, and $g(\cdot)$ is an unknown density. It is not necessary that $f(\cdot)$ be a two-component mixture distribution, since $g_1(\cdot)$ can be a mixture itself. We are interested in estimating the weight a. In applications in cosmology, data from $f(\cdot)$ can be high energy photon arrival times. Each arrival time is either noise (hence, a uniform arrival time) or signal, a pulsed radiation. Arrival time of the pulsed radiation is distributed according to an unknown density $g(\cdot)$, which is not of interest. The parameter a describes the *strength* of the pulsed signal, and is of interest to cosmologists. See Swanepoel (1999) and references therein for a complete discussion of the problem. A possibility is to estimate a using an iterative estimator like the ones proposed in this section. This results in a conservative statement about the strength of the pulse (i.e., it will not be over estimated with high probability). In Table 9, we show the average estimator for a, for different methods of estimation and different sample size m of the arrival times, $B = 1000$ iterations; and $g(\cdot)$ always taken to be a Beta density with parameters ρ and 1 ($g_1(\cdot)$, as said, is a uniform density). We see that the estimators are always conservative, and close to the real a for lower values of ρ. This happens because lower values of ρ correspond to stronger signal, which makes easier the elimination of the noise component.

Acknowledgements The author wishes to thank an anonymous referee for suggestions that lead to improvements of the paper. Acknowledgments go moreover to Marco Alfò for careful reading of a first draft of the manuscript. This paper is part of the author's dissertation (Farcomeni 2004).

References

Abrahmsen P (1997) A review of gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center, Norway

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

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188

Benjamini Y, Krieger AM, Yekutieli D (2004) Two staged linear step up FDR controlling procedures. Technical report Department of statistics, Tel Aviv University, Tel Aviv

Bickel DR (2004) On "strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates": does a large number of tests obviate confidence intervals of the FDR? Technical report, Office of Biostatistics and Bioinformatics, Medical College of Georgia, Georgia

Bovenhuis H, Spelman RJ (2000) Selective genotyping to detect quantitative trait loci for multiple traits in outbred populations. J Dairy Sci 83(1):173–180

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160

Ellis SP, Underwood MD, Arango V (2000) Mixed models and multiple comparisons in analysis of human neurochemical maps. Psychiat Res-Neuroim 9:111–119

Esary JD, Proschan F, Walkup DW (1967) Association of random variables, with applications. Ann Math Stat 38:1466–1474

Farcomeni A (2004) Multiple testing procedures under dependence, with applications. Ph.D. Thesis, Università degli studi di Roma "La Sapienza"; Dipartimento di Statistica Probabilità e Statistiche Applicate.

Genovese CR, Wasserman L (2002) Operating characteristics and extensions of the FDR procedure. J Roy Stat Soc (Ser B) 64:499–518

Genovese CR, Wasserman L (2004a) Exceedance control of the false discovery proportion. Technical report, Department of statistics, Carnegie Mellon University, Pittsburgh

Genovese CR, Wasserman L (2004b) A stochastic process approach to false discovery control. Ann Stat 32:1035–1061

Hochberg Y, Tamhane AC (1987) Multiple comparisons procedures. Wiley, New York

Ip EH (2001) Testing for local dependency in dichotomous and polytomous item response models. Psychometrika 66(1):109–132

Johnson ME (1987) Multivariate statistical simulation. Wiley, New York

van der Laan MJ, Dudoit S, Pollard KS (2003) Multiple testing. Part III. Procedures for control of the generalized family-wise error rate and proportion of false positives. Technical Report 141, Division of Biostatistics, UC Berkeley

Merriam EP, Genovese CR, Colby CL (2003) Spatial updating in human parietal cortex. Neuron 39:361–373

Mosig MO, Lipkin E, Khutoreskaya G, Tchourzyna E, Soller M, Fridmann AA (2001) Whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics 157:1683–1698

Pollard KS, van der Laan MJ (2003) Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data. Technical Report 121, Division of Biostatistics, UC Berkeley

Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19:368–375

Sarkar SK (2002) Some results on false discovery rate in stepwise multiple testing procedures. Ann Stat 30:239–257

Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Smith R (2001) Environmental statistics. http://www.math.uio.no/~glad/envnotes.ps

Storey JD (2002) A direct approach to false discovery rates. J Roy Stat Soc(Ser B) 64:479–498

Storey JD (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat 31:2013–2035

Storey JD, Tibshirani R (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical Report 2001-28, Department of Statistics, Stanford University, Stanford

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. J Roy Stat Soc(Ser B) 66:187–205

Swanepoel JWH (1999) The limiting behavior of a modified maximal symmetric $2s-$spacing with applications. Ann Stat 27:24–35

Woodroofe M, Sun J (1999) Testing uniformity versus a monotone density. Ann Stat 27:338–360

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC (1996) A unified statistical approach for determining significant signals in images of cerebral activation. Hum Brain Mapp 4:58–73

Yekutieli D, Benjamini Y (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plann Inference 82:171–196