

False Discovery Rates for Spatial Signals

Yoav Benjamini*

Department of Statistics and Operations Research

Tel Aviv University, Tel Aviv 69978, Israel

ybenja@post.tau.ac.il

Ruth Heller

Department of Statistics and Operations Research

Tel Aviv University, Tel Aviv 69978, Israel

rheller@post.tau.ac.il

June 15, 2006

Abstract

The problem of multiple testing for the presence of signal in spatial data can involve a large number of locations. Traditionally, each location is tested separately for signal presence but then the findings are reported in terms of clusters of nearby locations. This is an indication that the units of interests for testing are clusters rather than individual locations. The investigator may know a-priori these more natural units or an approximation to them. We suggest testing these cluster units rather than individual locations, thus increasing the signal to noise ratio within the unit tested as well as reducing the number of hypotheses tests conducted. Since the signal may be absent from part of each cluster, we define a cluster as containing signal if the signal is present in at least one subset of the cluster. We introduce powerful adaptive procedures for controlling the false discovery rate (FDR) on clusters, i.e. the proportion of clusters rejected erroneously out of all clusters rejected, or the size weighted FDR on clusters, which captures the size of erroneously rejected clusters out of the total size of clusters rejected. Moreover, we prove that the adaptive weighted procedure controls the weighted FDR. Once the cluster discoveries have been made, we suggest 'cleaning' locations in which the signal is absent. For this purpose we develop a hierarchical testing procedure that controls the expected proportion of locations in which false rejections occur under the fixed alternative model and show that it controls the desired error rates under less idealistic assumptions by extensive simulations. We discuss an application to functional neuroimaging which motivated this research and demonstrate the advantages of the proposed methodology on an example.

Key words : Signal detection, FDR, multiple testing, hierarchical testing, weighted testing procedures, functional MRI.

1 Introduction

Consider a spatial process $\{X(s) : s \in D \subset \mathfrak{R}^d\}$, generated by $X(s) = \mu(s) + \epsilon(s)$, where $\mu(\cdot)$ is a “piecewise smooth” signal and $\epsilon(\cdot)$ is a Gaussian mean zero process. The process $X(\cdot)$ is observed in some locations (possibly a grid) $\{s_k : k = 1, \dots, n\}$ in D , and we wish to infer on the subset of D with positive mean signal (the restriction to positive is for ease of notation).

The common approach to the problem is to test at each location separately for the signal’s presence

$$H_{0s_k} : \mu(s_k) = 0 \text{ versus } H_{1s_k} : \mu(s_k) > 0$$

and adjust the level of the test to the multiplicity of locations using random field theory, resampling methods, or the false discovery rate (FDR).

We argue, for two main reasons, that the data should instead be aggregated into clusters of locations before testing: (i) in many applications the fundamental units of interest are contiguous aggregates of measured locations that describe the regions of interest. For example, in functional magnetic resonance imaging (fMRI), the signal is recorded over time for a series of brain slices yielding measured signal at volumetric pixels (called voxels). The spatial resolution is set by the capacity of the MRI machine used, but the fundamental units of interest are the contiguous brain regions that participate in cognitive tasks and therefore are activated together (see e.g. Penny and Friston (2003)) rather than the individual voxels. (ii) spatial clusters of measured locations can have increased signal to noise ratio (SNR), since

if there is a signal in one location there tends to be signal in neighboring locations as well. For example, even if the SNR of every monitoring site in a region is too low to detect a pollutant in water or in the air, the SNR of the pooled information may be high enough. Another example is the across country mortality surveillance systems. Currently each municipality is tested separately for a suspicious increase in mortality (see e.g. sartorius et al. (2006)). These systems can benefit from our approach by testing clusters of municipalities with similar death rates first, possibly weighing the importance of a cluster by its population, and then only test municipalities within detected clusters.

Let C_1, C_2, \dots, C_m be a partition of D into m components we call clusters, and let $D_0 = \{s \in D : \mu(s) \leq 0\}$ and $D_1 = \{s \in D : \mu(s) > 0\}$ denote the unknown sets on which the null hypothesis is true and false respectively. We say that a cluster C_i does not contain a signal, or is a null cluster, if the signal is absent in all subsets of C_i (i.e. $C_i \subset D_0$). This is our necessary definition for a cluster that comes from the null hypothesis, since it is the only definition that guarantees the p-values to be uniform and stochastically larger than the p-values of clusters coming from H_1 . The collection of m hypotheses tested are therefore, for all $i = 1, \dots, m$,

$$H_{0i} : \mu(s) = 0 \forall s \in C_i \text{ versus } H_{1i} : \mu(s) > 0 \text{ for at least one } s \in C_i \quad (1)$$

How to aggregate the data into clusters of locations is problem-specific and depends on the information at hand: in the fMRI example, Heller et al. (2005) use the data from a preparatory fMRI scan to approximate the appropriate brain units (sub-units) by aggregating neighboring voxels that are highly correlated; in the pollutant example, the clusters can be defined by considering the geographic positions of the monitoring sites; in the mortality surveillance system example, the clusters can be

defined by aggregating municipalities with similar past death rates. Two points are important regarding the construction of the partition. First, it should be based on information outside the data we set out to analyze. This guarantees that if the partitioning has a stochastic component it does not affect the interpretation or the validity of the results. In practice, such information is often available for spatial data, as in the fMRI and pollutant examples above. Second, the quality of the partition does affect the potential gain from using clusters of locations rather than individual locations. Intuitively, if the data is partitioned into the smallest possible number of homogeneous clusters (in the sense that all its locations contain signal or the signal is absent from all), the gain will be largest. From extensive simulations and analytical examination of simple cases, we show that we still gain from using clusters rather than individual locations if on average the percent signal in the cluster, multiplied by the square root of the average cluster size, is greater than the standard deviation.

Starting with a given partition of the data into clusters, testing clusters of locations rather than individual locations raises the question of which error criterion we would like to control. Benjamini and Hochberg (1995) introduced the false discovery rate (FDR), which is in our context the expected proportion of erroneously rejected clusters out of all clusters rejected, say FDR on clusters (FDR_c). Benjamini and Hochberg (1997) also introduced the weighted FDR, which may be especially appropriate here. For example, we may want to control a size weighted FDR on clusters ($WFDR_s$), where the weights are proportional to the size of the clusters, which means on the one hand that it is important to reject a large cluster that contains signal since it considerably increases the weight of the total discoveries, but on the other hand it also increases the weight of the errors if in fact it is an error. Exact definitions and their relationships are given in section 2. In section

3 we give procedures to control the FDR error rates on clusters. In particular, we prove that the adaptive two stage procedure of Benjamini et al. (2006) can be used using weights and still control the FDR, a property of interest by itself.

We have already identified the advantage of using clusters as building blocks for inference. However, the researcher may want to control the FDR on locations as well, possibly at a more lenient level. For this purpose we introduce a Cluster Testing and Trimming (CTT) procedure for spatial data, that combines both goals by first testing clusters and then trimming individual locations from the cluster discoveries. We show its analytic properties for a simple model in section 4 and examine more realistic models using simulations in section 5. While the theory regarding cluster testing and trimming is developed for testing cluster means, it is later generalized to other test statistics.

We demonstrate our analysis approach on an fMRI example in section 6. This approach is especially useful for detecting activated brain regions, since the data is very noisy and the number of hypothesis tests (in the original units) is very large. Moreover, the CTT procedure shows the researcher not only the regions of activity, but the locations within the regions where the activity is most significant.

The first attempt to give FDR control over clusters was made by Pacifico et al. (2004), but their approach is very different from ours in principle as well as in detail. Moreover, their suggested procedure relies on power of single element intensities and therefore does not make use of the power gain made possible by averaging over elements within clusters (we come back to this point in section 7). An estimation approach using FDR with spatial signals was given in Shen et al. (2002).

2 False Discovery Rates for Spatial Signals

We first propose useful error rates to control when drawing inference about the set of locations that contain signal. For a rejected set $A \subseteq D$, and λ a measure on the domain D , define $\lambda_1(A) \equiv \lambda(A \cap D_1)$, $\lambda_0(A) \equiv \lambda(A \cap D_0)$, and Q_λ be the unobservable random quotient $Q_\lambda = \lambda_0(\mathbf{A})/\lambda(\mathbf{A})I_{\{\lambda(\mathbf{A})>0\}}$ where $I_{\{\lambda(\mathbf{A})>0\}}$ defines the quotient as 0 when $\lambda(\mathbf{A}) = 0$. For example, with the 2-dimensional Lebesgue measure, Q_λ is the proportion of area where the signal is absent out of the area rejected, unless the rejected set has measure zero, in which case Q_λ is zero. A natural generalization of the FDR for spatial applications is given in the following definition.

Definition 2.1. *Using the measure λ for rejected sets $A \subseteq D$, the false discovery rate of a testing procedure is*

$$\text{FDR}_\lambda = \text{E}(Q_\lambda) = \text{E}\left[\frac{\lambda_0(\mathbf{A})}{\lambda(\mathbf{A})}I_{\{\lambda(\mathbf{A})>0\}}\right].$$

FDR_λ has been suggested by Pacifico et al. (2004) for a threshold procedure: $A(T) = \{s \in D : X(s) \geq T(X)\}$. Here, FDR_λ is the expected proportion of the area above a (random) threshold T from which the signal is absent. In practice, the signal is usually measured in a known, finite number of locations. When the locations are equally spaced and represent an equal area, the regular FDR on locations approximate the FDR in terms of area.

When the fundamental unit of interest is a cluster rather than a location, a useful error rate (in the spirit of Pacifico et al. (2004)) for a testing procedure that rejects clusters is the expected proportion of falsely discovered clusters out of all clusters discovered.

Definition 2.2. Let C_1, \dots, C_m be a partition of D into clusters, $R_i = 1$ if C_i was rejected and 0 otherwise, and $V_i = 1$ if $R_i = 1$ but $C_i \subset D_0$. The FDR on clusters is

$$\text{FDR}_c = E\left[\frac{\sum_{i=1}^m V_i}{\sum_{i=1}^m R_i} I_{\{\sum_{i=1}^m R_i > 0\}}\right]$$

When clusters are completely homogeneous, weighing each cluster by its size leads to an error rate that is identical to FDR_λ :

Definition 2.3. The size weighted FDR on clusters is

$$\text{WFDR}_s = E\left[\frac{\sum_{i=1}^m \lambda(C_i) V_i}{\sum_{i=1}^m \lambda(C_i) R_i} I_{\{\sum_{i=1}^m R_i > 0\}}\right]$$

The above two error rates are special cases of the weighted FDR (defined in Benjamini and Hochberg (1997) for general hypotheses)

$$\text{WFDR} = E\left[\frac{\sum_{i=1}^m w_i V_i}{\sum_{i=1}^m w_i R_i} I_{\{\sum_{i=1}^m R_i > 0\}}\right]$$

where w_i is the weight of cluster i and $\sum_{i=1}^m w_i = m$. Specifically, the weight of cluster i for FDR_c and WFDR_s is $w_i = 1$ and $w_i = m\lambda(C_i) / \sum_{i=1}^m \lambda(C_i)$ respectively.

The relation between FDR on clusters with that on the original testing units for a testing procedure that rejects clusters is

$$\text{FDR}_\lambda = E\left[\frac{\sum_{i=1}^m R_i \lambda_0(C_i)}{\sum_{i=1}^m R_i \lambda(C_i)} I_{\{\sum_{i=1}^m R_i > 0\}}\right] = \text{WFDR}_s + E\left(\frac{\sum_{i=1}^m S_i \lambda_0(C_i)}{\sum_{i=1}^m R_i \lambda(C_i)}\right)$$

where $S_i = 1$ if cluster C_i is rejected rightfully and 0 otherwise. Control of FDR_c or WFDR_s does not guarantee control of FDR_λ , but the simple relation with WFDR_s can be used to get an upper bound on FDR_λ .

Recall that our motivation for using clusters rather than locations is primarily

that the units of interest are clusters. These units may have higher SNR than locations. So the primary interest is to achieve control over WFDR on clusters, where the choice of weights is goal oriented. This is done in the following section 3. But once the cluster discoveries have been made, the investigator may be interested in 'cleaning' the rejected clusters from null locations, thus reducing the FDR_λ . This is done in section 4.

3 Controlling the Weighted FDR on Clusters

We can use a test statistic of our choice for testing the collection of m cluster hypotheses (1): any spatial summary measure of the cluster data is appropriate as long as a valid p-value p_i for testing H_{0i} in (1) can be computed.

The BH procedure in Benjamini and Hochberg (1995) guarantees $FDR_c \leq (m_0/m)q$, and its extension in Benjamini and Hochberg (1997) guarantees $WFDR \leq (\sum_{i=1}^{m_0} w_i/m)q$ for independent test statistics, where H_{01}, \dots, H_{0m_0} correspond to true null hypotheses. Benjamini and Yekutieli (2001) further prove that $FDR_c \leq q$ if the joint distribution of the test statistics is positive regression dependent (PRDS) on the subset of true nulls. For example, the PRDS property is satisfied if the correlation between cluster test statistics is non-negative and the testing hypotheses are one-sided. Kling (2005) proved that the bound on WFDR is valid under the same conditions.

Procedure 3.1. *The weighted BH procedure: Order the cluster p-values $p_{(1)} \leq \dots \leq p_{(j)} \leq \dots \leq p_{(m)}$. Let $w_{(j)}$ be the weight associated with $p_{(j)}$. Let $k = \max\{j : p_{(j)} \leq (\sum_{i=1}^j w_{(i)}/m)q\}$ and reject the clusters corresponding to the smallest k p-values.*

In line with many others (e.g. Storey et al. (2004), see Benjamini et al. (2006) for a review), Benjamini et al. (2006) suggest a two stage procedure that first estimates

the number of null hypotheses and then uses it to enhance the power when $w_i = 1$, $i = 1, \dots, m$. They prove that it controls the FDR at level q for independent test statistics. The generalization of their procedure, that is especially useful when $\sum_{i=1}^{m_0} w_i/m$ is believed to be small, is as follows

Procedure 3.2. *The weighted two stage procedure*

1. Use procedure 3.1 at level $q' = q/(q + 1)$.
2. Estimate the sum of weights of null clusters by $\hat{m}_0 = m - \sum_{i=1}^k w_{(i)}$.
3. Let $k_2 = \max\{j : p_{(j)} \leq (\sum_{i=1}^j w_{(i)}/\hat{m}_0)q'\}$. The clusters corresponding to the smallest k_2 p -values are rejected.

Theorem 3.1. *For independent test statistics, the weighted two stage procedure 3.2 controls the WFDR at level q .*

See appendix A.1 for a proof. Note that although the development of the weighted two stage procedure was motivated by the problem of testing clusters, it is more general and can be applied to any multiple testing problem with weights.

Benjamini et al. (2006) argue (by a simulation study) that the adaptive procedure for unit weights controls the FDR at level q under the PRDS assumption. The argument can be extended under weighting as well.

4 Hierarchical Testing Procedure

As explained before, even a researcher interested in clusters prefers to avoid errors in specific locations, so after applying an FDR procedure on clusters, we look within the rejected clusters and eliminate locations that contain no signal. This hierarchical testing approach is applicable to any application where the fundamental unit of

interest for testing is a cluster of basic elements, therefore we refer to our smallest unit of testing as an element rather than a location from now on.

The regular FDR on elements takes the following form:

$$\text{FDR}_e = \mathbb{E}\left[\frac{\sum_{i=1}^m \sum_{e=1}^{c_i} V_{ei}}{\sum_{i=1}^m \sum_{e=1}^{c_i} R_{ei}} \mathbb{I}_{\{\sum_{i=1}^m \sum_{e=1}^{c_i} R_{ei} > 0\}}\right]$$

where c_i is the number of elements in cluster i , $R_{ei} = 1$ if element e in cluster i is rejected and 0 otherwise, and $V_{ei} = 1$ if element e in cluster i is rejected erroneously and 0 otherwise.

We shall first confine ourselves to the **fixed alternative model** for data $\{X_{ei} \mid e = 1, \dots, c_i, i = 1, \dots, m\}$, where X_{ei} is the measurement on element e in cluster i . The model assumptions are: $X_{ei} \sim N(0, 1)$ or $X_{ei} \sim N(\mu, 1)$ for some fixed $\mu > 0$; the proportion of elements with $\mu > 0$ within a cluster, h , is the same for every cluster that contains non-null elements.

We suppose that h, μ and the number of null clusters m_0 are known. In practice, these parameters are unknown and moreover h can vary from one cluster to the next and μ can vary across elements. We will address these issues later.

For the fixed alternative model, the cluster average $\bar{X}_i = \sum_{e=1}^{c_i} X_{ei}/c_i$ is used to test for signal presence,

$$H_{0i} : E(\bar{X}_i) = 0, \quad H_{1i} : E(\bar{X}_i) = h\mu, \quad i = 1, \dots, m.$$

Next, each element included in a rejected cluster is tested for signal presence. The relevant p-value is no longer the marginal one, but conditional on it being in a rejected cluster. The following notations will be convenient in the derivation of the p-value: $\rho_{ei} = \text{corr}(X_{ei}, \bar{X}_i)$ (e.g. $\rho_{ei} = 1/\sqrt{c_i}$ if the measured signal between elements is independent), $\sigma_{\bar{X}_i}$ is the standard deviation of \bar{X}_i , and $\mu_i = h\mu/\sigma_{\bar{X}_i}$

(recall that $\sigma_{\bar{X}_i} = \sum_{e=1}^{c_i} \rho_{ei}/c_i$). If u_1 is the fixed threshold applied to the cluster p-values, let

$$p_{ei} = \begin{cases} \frac{\int_{x_{ei}}^{\infty} \left(\frac{m_0}{m} \right) \tilde{\Phi} \left(\frac{\tilde{\Phi}^{-1}(u_1) - \rho_{ei}u}{\sqrt{1-\rho_{ei}^2}} \right) + \left(1 - \frac{m_0}{m} \right) \tilde{\Phi} \left(\frac{\tilde{\Phi}^{-1}(u_1) - \rho_{ei}u - \mu_i}{\sqrt{1-\rho_{ei}^2}} \right) \phi(u) du}{\frac{m_0}{m} u_1 + \left(1 - \frac{m_0}{m} \right) \tilde{\Phi}(\tilde{\Phi}^{-1}(u_1) - \mu_i)} & \text{if } h < 1, \\ \frac{1}{u_1} \int_{x_{ei}}^{\infty} \tilde{\Phi} \left(\frac{\tilde{\Phi}^{-1}(u_1) - \rho_{ei}u}{\sqrt{1-\rho_{ei}^2}} \right) \phi(u) du & \text{if } h = 1 \end{cases} \quad (2)$$

where Φ , $\tilde{\Phi}$ and ϕ are respectively the cumulative distribution, the right tail probability, and the density of a standard normal distribution.

Lemma 4.1. *For the fixed alternative model p_{ei} given in (2) is a p-value for testing an element in a rejected cluster.*

Proof. We have to show that $p_{ei} \sim U(0, 1)$ if element e in cluster i contains no signal.

$$p_{ei} = P_0(X_{ei} \geq x_{ei} | \bar{X}_i / \sigma_{\bar{X}_i} \geq \tilde{\Phi}^{-1}(u_1)) = \frac{P_0(X_{ei} \geq x_{ei}, \bar{X}_i / \sigma_{\bar{X}_i} \geq \tilde{\Phi}^{-1}(u_1))}{P_0(\bar{X}_i / \sigma_{\bar{X}_i} \geq \tilde{\Phi}^{-1}(u_1))}$$

where the subscript zero indicates that the probabilities are calculated under the null hypothesis. The result follows directly from the fact that the joint distribution of an element with its cluster average is bivariate normal with unit variance, correlation ρ_{ei} , zero element mean and zero or μ_i mean for the standardized cluster average, depending on whether the cluster null hypothesis is true or false. \square

Moving from an ideal setting to a more realistic one we need to address the following points. First, since the parameters m_0 and $h\mu$ are generally unknown, they need to be estimated from the data. Note that we assume $h < 1$, since for $h = 1$ there is no reason to use the hierarchical procedure. Second, if the measured

signals in elements within a cluster are dependent, ρ_{ei} needs to be estimated from the data. For example, in section 6 we assume a constant covariance model within each cluster and estimate ρ_{ei} from the empirical variogram. Finally, we propose to use the adaptive FDR procedure of Benjamini et al. (2006) on the estimated p_{ei} 's. The gain in power over the BH FDR procedure may be high since the potential number of true element discoveries within discovered clusters is expected to be quite large. The following procedure summarizes the suggested analysis steps so far.

Procedure 4.1. *The Clusters Testing and Trimming (CTT) procedure:*

1. *Testing Stage: Apply procedure 3.1 at level q on the spatial average of every cluster. Let u_1 be the cut-off point of the largest p -value rejected (e.g. $u_1 = Rq/m$ for a BH procedure that rejects R clusters).*

2. *Trimming Stage:*

(a) *Let $\hat{m}_0 = \frac{(m - \sum_{i=1}^m r_i)}{1-q}$ and $\hat{h}\mu = \frac{\sum_{i=1}^m \sum_{e=1}^{c_i} x_{ei}}{\sum_{i=1}^m c_i}$.*

(b) *For every element in a rejected cluster, estimate (according to the assumed dependency structure) ρ_{ei} from the data to get $\hat{\rho}_{ei}$. Let $\hat{\sigma}_{\bar{X}_i} = \sum_{e=1}^{c_i} \hat{\rho}_{ei}/c_i$ and $\hat{\mu}_i = \hat{h}\mu/\hat{\sigma}_{\bar{X}_i}$.*

(c) *Use the estimated parameters above to estimate the p_{ei} 's.*

(d) *Sort the resulting $m_2 (= \sum_{i=1}^m r_i c_i)$ \hat{p}_{ei} 's: $\hat{p}_{(1)} \leq \dots \leq \hat{p}_{(m_2)}$. Let $q'_2 = q_2/(1 + q_2)$, $k_1 = \max\{j : \hat{p}_{(j)} \leq (j/m_2)q'_2\}$ and $\hat{m}_{02} = m_2 - k_1$, where q_2 is the desired level of FDR_e .*

(e) *Let $k_2 = \max\{j : \hat{p}_{(j)} \leq (j/\hat{m}_{02})q'_2\}$. Keep all elements with $\hat{p}_{ei} \leq (k_2/\hat{m}_{02})q'_2$ in the clusters, trimming the others.*

While the procedure is stated for a general correlation structure, we shall first

show its properties under independence. The properties under dependency will be discussed below.

Theorem 4.1. *Assume the data $\{X_{ei} \mid e = 1, \dots, c_i, i = 1, \dots, m\}$ are independent and satisfy the fixed alternative model assumptions. Consider the asymptotic case where $m \rightarrow \infty$ while m_0/m remains fixed. Moreover, assume that the weights are positive and i.i.d under the null as well as under the alternative (but the distribution need not be the same) and that they are independent of the data. Assume these same assumptions on the cluster sizes distribution. Finally assume that $m_0/m < 1$. Then procedure 4.1 controls the FDR_e at level q_2 .*

See appendix A.2 for a proof. Note that when $m_0/m = 1$ FDR_e is controlled even non-asymptotically. Note also that from corollary 3.1.7 in Farcomeni (2004) it is enough to assume m-dependence between the locations.

In the asymptotic case that the number of elements within each cluster grows to infinity the assumptions of the fixed alternative model can be relaxed and the CTT procedure can be simplified since then, for large enough cluster sizes, the p-value of each rejected element is approximately equal to the p-value of the element when ignoring the testing stage selection

Theorem 4.2. *Assume $\{X_{ei} \mid e = 1, \dots, c_i, i = 1, \dots, m\}$ are independent normally distributed random variables with unit variance and either 0 or $\mu_{ei} > 0$ expectation. Replace the computation of \hat{p}_{ei} in procedure 4.1 with $\hat{p}_{ei} = \tilde{\Phi}(x_{ei})$. Then the modified CTT procedure 4.1 controls the FDR_e at level q_2 in the asymptotic case that $c_i \rightarrow \infty$, for all $i = 1 \dots m$.*

Proof. As $c_i \rightarrow \infty$, $\rho_{ei} = 1/\sqrt{c_i} \rightarrow 0$ and therefore $p_{ei} \rightarrow \tilde{\Phi}(x_{ei})$. □

Note that although the assumption that $c_i \rightarrow \infty$ may seem unrealistic at first, this is the essence of the technological progress in various areas of research. In

fMRI, for example, the regions in the brain may remain constant while the technical resolution improves, leading to increasing c_i .

While the theory underlying the CTT procedure 4.1 is based on the signal average in the testing stage, and on the signal level in each element in the trimming stage, it can be generalized to any element-wise statistic of interest. Once the element-wise test statistic is chosen it defines the cluster test statistic as the standardized z-score average of its elements.

Procedure 4.2. *The CTT procedure for general element-wise test statistics:*

1. *For every element calculate its original p-value and let z_{ei} be the corresponding z-score for element e in cluster i .*
2. *Apply procedure 4.1 on the z-score data $\{z_{ei} | e = 1, \dots, c_i, i = 1, \dots, m\}$.*

The discoveries of the CTT procedure 4.1 are elements rather than clusters. If the control of FDR_e is of primary importance, then q_2 should be controlled at the desired level. It may though happen that the trimming stage will eliminate completely clusters that were discovered in the testing stage, in that no element remains within the previously discovered cluster. Moreover, the clusters retained are not necessarily those corresponding to the smallest p-values. Therefore, if the primary goal is to detect clusters with FDR_c control, and 'cleaning' the clusters or controlling the FDR_e is only a secondary goal, the testing level in the trimming stage can be more lenient. For example, within the detected clusters from the testing stage at level $q = 0.05$, the p-values of the individual locations from the trimming stage at level $q_2 = 0.25$ indicate where within each cluster the location can be more trusted.

5 A Simulation Study

We conducted a simulation study in order to (1) validate that the CTT procedure 4.1 controls the FDR_e at level q_2 under realistic settings, and (2) compare the performance of procedures 3.1 and 4.1 with that of an element-wise analysis on the entire data. Specifically, identify under which cluster configurations we loose from using the clusters rather than using the elements only.

5.1 Setting

For 1024 points, we chose 13 different cluster configurations: equal size clusters (of size 2, 4, 8, 16 and 32); uniform, \cap -shaped or \cup -shaped symmetric, right and left skewed distributions of sizes. For each cluster configuration the percent of active clusters considered was $p = 0, 0.01, 0.05, 0.10, 0.15$. The proportion of active elements within each active cluster was $h = 1, 0.75, 0.5, 0.25$. The proportion h was either fixed for a data set or varied among clusters with the above set average. The signal at each element was either zero or positive. Active elements either had identical signal intensity or variable signal intensity, and in the latter non-zero signals were drawn independently from a truncated normal distribution with fixed mean and variance (the coefficient of variation of the element means within a cluster ranged from 0.2 to 2). White noise was added to each element so the signal to noise ratio ranged from 0 to 5. The same noise was added for all signal and cluster configurations. We used 150 simulation repetitions. The simulations were performed in Matlab (version 6.5).

5.2 Data Analysis

The CTT procedure 4.1 was applied to each simulated data set. We restricted the analysis of the testing stage to the BH procedure with a 0.05 threshold. In the trimming stage q_2 was either 0.05 or 0.25. The estimated parameters in the trimming stage were obtained by using $q = 0.05$ or $q = 0.5$.

The FDR was estimated by averaging over the simulations the proportion of false units rejected among all rejected units. For FDR_c and FDR_e the units were clusters and elements respectively. Note that after the trimming stage a cluster is a discovery if at least one element within the cluster is rejected. If no clusters were rejected in the testing stage, the realized FDR_c and FDR_e were set to zero.

The power was estimated by averaging over the simulations the proportion of true discoveries out of the number of potential discoveries at the appropriate units: elements or clusters. If no discoveries were made in the testing stage, the power was set to zero. Four power measures were calculated: the power of clusters after the testing stage, the power of elements after the trimming stage with $q_2 = 0.05$ and with $q_2 = 0.25$, and the power of an element-wise analysis on the entire data.

5.3 Results

The procedures applied preserve their nominal thresholds under all simulation scenarios. Specifically, even for signal configurations with variable cluster size, variable signal intensity μ for non-null elements and a variable percent of active elements within an active cluster h , control over FDR_c , FDR_s and FDR_e is achieved. We show the results graphically for a small representative subset of signal configuration. In this subset, 15% of the clusters were active, the cluster sizes were equal and the data satisfied the fixed alternative model assumptions.

FDR Control Figure 1 shows that the FDR_c after the testing stage is around 0.05. After the trimming stage, the FDR_c is much lower than 0.05 for small μ . Note that although the CTT procedure with fixed q_2 does not guarantee that the FDR_c is preserved, it was preserved in all our simulations. The reason why it is low for small μ is that only very few elements were rejected, and they belonged mostly to true cluster rejections. The estimated FDR_c is most variable for small μ and large c (standard errors are at most 0.021, 0.010 and 0.013 after the testing stage and the trimming stage either with $q_2 = 0.05$ or with $q_2 = 0.25$ respectively). FDR_e is not preserved after the testing stage, yet for all signal and cluster configurations the FDR_e is below q_2 after the trimming stage. This holds even for fairly large deviations from the fixed alternative model (standard errors are at most 0.03 and 0.02 after the testing stage and trimming stage respectively).

Power Figure 2 shows the power improvements achieved by procedures 3.1 and 4.1 over a single element analysis of the entire data at the 0.05 level as a function of μ (with standard errors ≤ 0.02).

Note that the power advantage is largest when μ is not too small and not too large. In figure 2, only for $h = 0.25$ and $c = 8$, the single element analysis has more power. This is so since $\sqrt{ch} < 1$ in this case. Clearly, as \sqrt{ch} increase the advantage in power of our procedures over single element analysis is larger. This observation repeated itself in all signal configurations considered: when $\sqrt{ch} > 1$ (where c is the average cluster size), our procedures are more powerful than a single element analysis on the entire data at the 0.05 level, and the advantage is more pronounced for larger \sqrt{ch} .

Specifically, whenever $\sqrt{ch} > 1$, the CTT procedure is more powerful than a single element analysis on the entire data for controlling the FDR_c at the 0.05 level.

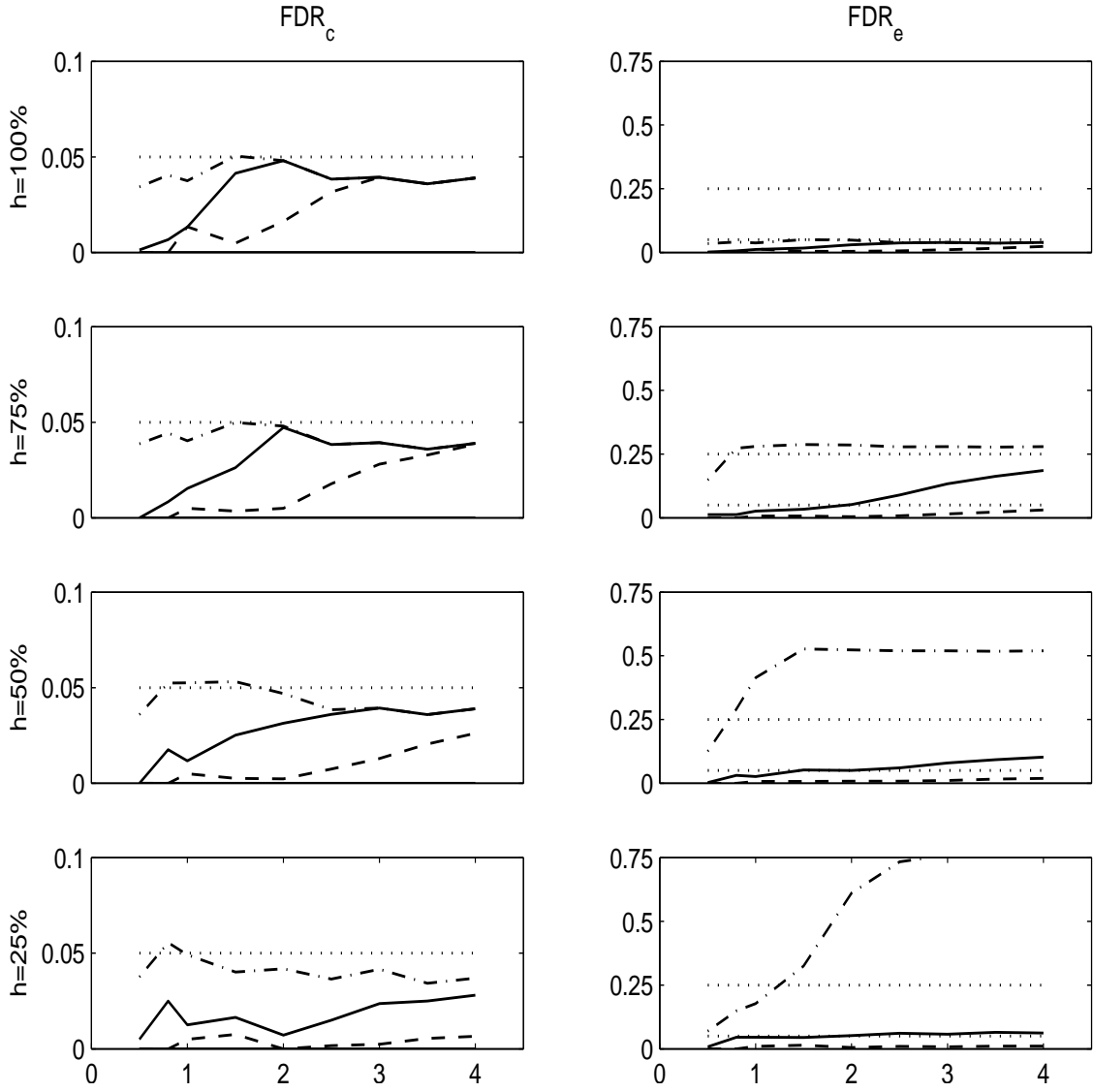


Figure 1: FDR_c (left) and FDR_e (right) as a function of μ for (1) cluster-wise analysis (dash-dot line); (2) CTT analysis with $q_2 = 0.05$ (dashed line); (3) CTT analysis with $q_2 = 0.25$ (solid line). The FDR_c after the testing stage is around 0.05. After the trimming stage, the FDR_c is much lower than 0.05 for small μ . The FDR_e is below q_2 after the trimming stage.

(the gain in power is of course even larger for $q_2 > 0.05$). This is so since the SNR is higher per cluster than per element, and moreover the number of hypotheses tested is smaller.

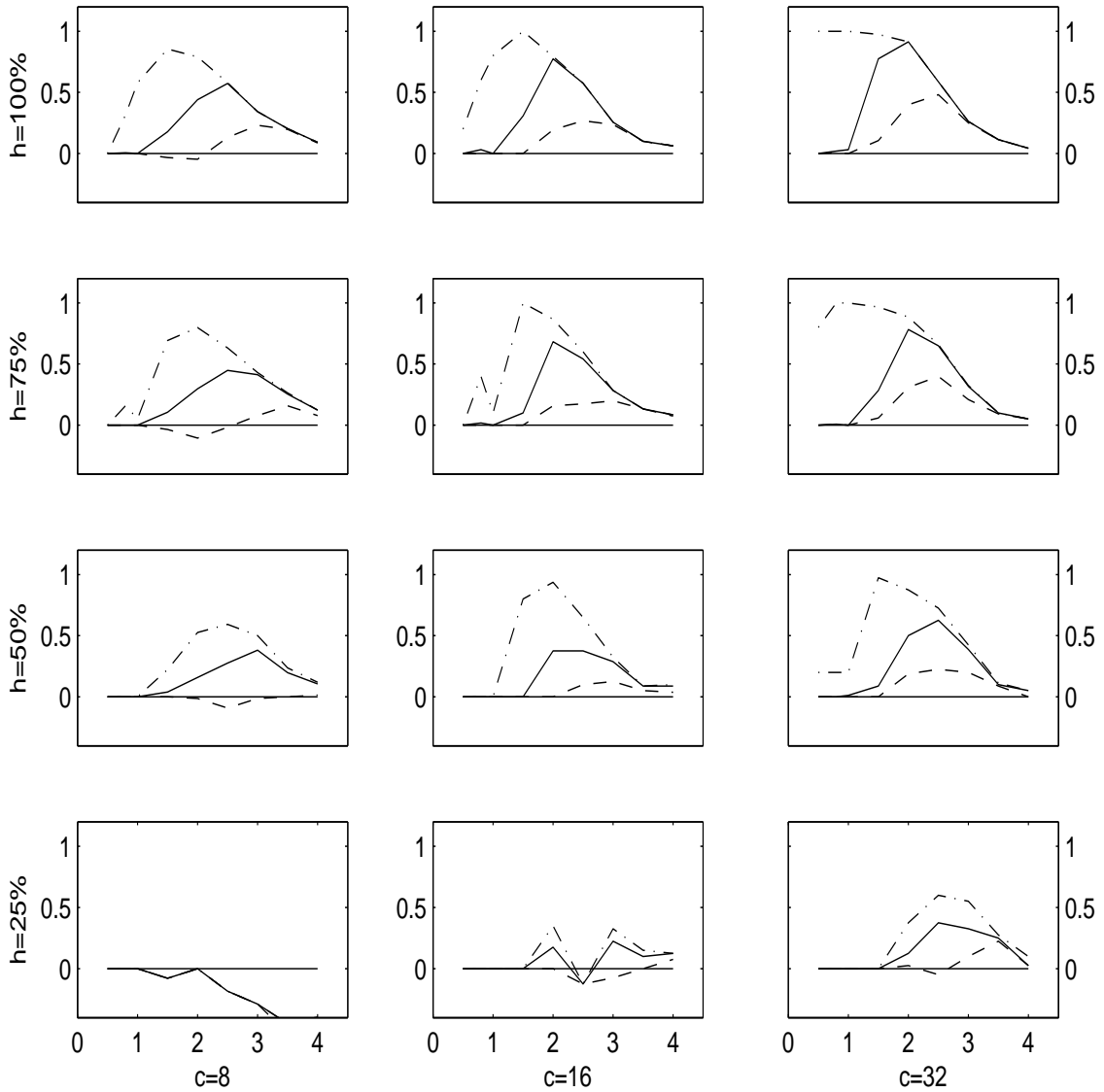


Figure 2: Power difference from element-wise analysis as a function of μ for (1) cluster-wise analysis (dash-dot line); (2) CTT analysis with $q_2 = 0.05$ (dashed line); (3) CTT analysis with $q_2 = 0.05$ (solid line). The power advantage is largest when μ is not too small and not too large. As \sqrt{ch} increases the advantage is larger. Disadvantage only when $\sqrt{ch} < 1$.

6 An fMRI Example

The functional magnetic resonance imaging (fMRI) signal is recorded over time for a series of brain slices, while the subject performs a sequence of behavioral tasks. The

fMRI researcher looks for brain regions that are correlated with the experimental paradigm in the form of clusters of voxels showing task related activity, whereas the unit of a 'voxel' is arbitrarily determined by the measurement technique and does not represent a primary neural entity. We used the correlation between neighboring voxels in order to identify clusters during a preparatory scan, see Heller et al. (2005) for details.

In this fMRI example, an observer viewed stimuli that contained perceptually completed (“illusory”) contours versus a control stimuli that shared local features but did not contain illusory contours in a block design experiment. The p-value of each voxel was calculated from a general linear model (GLM). Next, as suggested in procedure 4.2, the z-score of every voxel was calculated. The p-value of every cluster was based on the average z-score within the cluster. We estimated the isotropic correlation structure of the z-scores from the data. We applied the weighted FDR procedure (3.2) with both the FDR_c and $WFDR_s$ criteria at level 0.05, within a previously defined region of interest that contained 207 voxels grouped into 20 clusters (for a description of details of above analysis, see Heller et al. (2005)). Controlling the $WFDR_s$ we found 15 clusters and a total of 183 activated voxels, one more cluster (containing 11 voxels) than when controlling the FDR_c (both at level 0.05). Next, on the discovered clusters from the weighted procedure we calculated the p-values of the voxels in the rejected clusters. We applied the CTT procedure with $q_2 = 0.05$ and 0.25 and found 43 and 153 activated voxels respectively, coming from 14 and 15 different clusters respectively.

For comparison, we performed the standard single voxel analysis using an adaptive BH FDR procedure at level 0.05 and found 41 active voxels, coming from 14 different clusters. Using the current practice in fMRI of not reporting single non-contiguous voxel discoveries reduces the number of voxels to 36, coming from 13

different clusters. So even if the investigator wants control over FDR_e at the 0.05 level, the CTT procedure discovers few more voxels. However, if the investigator is willing to control the FDR_e at a higher level (say 0.25) as long as he already has control over the $WFDR_s$ at the 0.05 level, then many more voxels are discovered.

Figure 3 shows the activated voxels in slices 9-10 in the block design IC vs. control experiment computed with the three different procedures. Note that clusters extend across slices, which is why noncontiguous voxels within a slice can belong to the same cluster.

7 Discussion

We argued in favor of testing clusters of locations rather than individual locations in situations where (i) the investigator's main interest is in regions of activity rather than activity in individual locations; (ii) the SNR in individual locations is low but increases when pooling information from neighboring locations; and (iii) the number of locations tested is high. We suggested controlling the FDR on clusters, or the size weighted FDR which amounts to the expected proportion of area of rejected clusters which belongs to clusters rejected in error.

We generalized the adaptive FDR in Benjamini et al. (2006) to arbitrary weights. We showed the adaptive procedure controls the weighted FDR under independence of test statistics. When the test statistics are not independent, we know that the procedure 3.1 controls the WFDR if the PRDS property is satisfied. We believe that, as argued in Benjamini et al. (2006), the adaptive weighted procedure 3.2 controls the WFDR under dependency, but this remains to be proved.

The only previous attempt to control the FDR on clusters that is known to us is in Pacifico et al. (2004). They defined the FDR criterion on clusters as $E\Theta_\tau(T)$

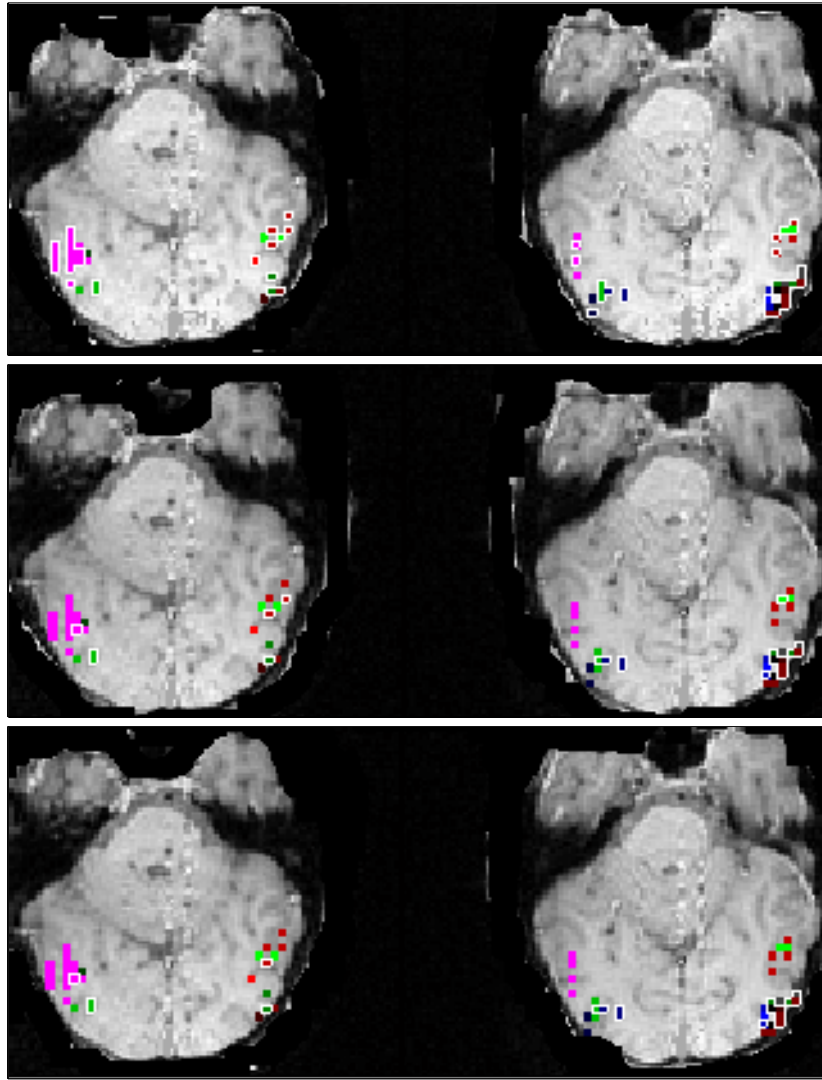


Figure 3: Activations within the LOC ROI in the block-design IC vs. Control experiment computed with three different procedures (sample slices 9-10; activated regions within the ROI indicated by white outlines; every cluster within the ROI is indicated in a different color; in these two slices all ROI clusters are detected as active). Top panel, CTT procedure with $q_1 = 0.05$ and $q_2 = 0.25$. Middle panel, CTT procedure with $q_1 = 0.05$ and $q_2 = 0.05$. Bottom panel, single voxel analysis with adaptive FDR at the 5% level. Note that many more clusters of activity are discovered when compared to the single voxel analysis in the bottom panel. If we insist on an FDR_e of 5% in the CTT procedure only few more discoveries are made over the single voxel analysis.

where $\Theta_\tau(T) = \#\{1 \leq k \leq m_T : \lambda_0(C_k)/\lambda(C_k) \geq \tau\}/m_T$, m_T is the number of clusters rejected and τ is the maximal proportion of null signal allowed within a true cluster discovery. Our definition of FDR on clusters looks at first glance similar to theirs, except that we take $\tau = 1$. However, the m_T clusters rejected in Pacifico et al. (2004) were found by considering a rejection threshold $T(X)$, and the rejection set $R_T = \{s \in S : X(s) \geq T(X)\}$. Then the decomposition of R_T into connected components C_1, \dots, C_{m_T} defines the set of clusters rejected. Thus their procedure is based on each element's individual test statistic, and no advantage is taken of the *SNR* increase that an aggregation of elements within a cluster can offer. In this paper we showed the advantage of such averaging, in terms of power, over single element analysis. The possible advantage of Pacifico et al. (2004) lies in the fact that the partition is obtained from the data used for testing, whereas our method relies on the assumption that the investigator can obtain a partition from other data or other information. While this is quite often feasible, it may be useful to develop methods that use the experimental data for partitioning as well as for testing.

The gain in power when testing clusters rather than individual locations depends on the quality of the partition. We showed that the power is higher for testing clusters if $h > 1/\sqrt{c}$, where h is the average percent of signal within clusters containing signal, and c is the average cluster size. Note that although a cluster is considered a true discovery if at least one element within it contains signal, i.e $ch > 1$, we require $\sqrt{ch} > 1$. So for example, if the average cluster size is 10 elements, then the cluster based analysis performs better if $h > 0.32$ (but $h > 0.1$ is not enough). A small simulation study that examined the damage in applying the cluster based analysis when in fact there was no cluster structure in the data showed that there is hardly any damage (though no advantage as well) when the fraction of elements containing signal in the data is at least $1/\sqrt{c}$.

For each cluster detected, we can only conclude that there is signal somewhere within the cluster. We developed hierarchical procedures, such as the Cluster Testing and Trimming procedure, that indicate where the signal is within the detected cluster. These procedures generalize existing theory about FDR controlling procedures, in the sense of Benjamini and Yekutieli (2002), to control the FDR at the desired level, even though the test statistics between the two levels of testing are dependent. The degree of confidence by which we report the discoveries depends on the FDR levels used. We can achieve the same degree of confidence as when testing individual locations only, and even gain power by pre-screening by cluster, but this may not be necessary. For example, the investigator may be satisfied in knowing that the detected clusters with no signal at all comprise no more than 5% of the clusters (on average), but that 25% of the detected locations within these clusters may be false. The willingness to allow a large FDR level for individual locations follows from the fact that the precision necessary is often that of a general region, rather than the exact spatial locations where the signal is present. We allow the investigator to choose the degrees of confidence that he feels are necessary for his application.

As a final note, the cluster p-value and the estimated conditional location p-value developed here may be useful when combined with the hierarchical testing scheme of Benjamini and Yekutieli (2002). In that approach, discoveries from the different levels of analysis (clusters and locations in our case) are combined into a single FDR measure. Introducing weights to their approach may open a new venue of its relevance even for the fMRI example we used.

Acknowledgement. We wish to thank Nava Rubin for useful discussion of the fMRI example that motivated this study and for supplying the fMRI data and Felix

Abramovich for valuable comments. This study was supported by a grant from the Adams Super Center for Brain Studies, Tel-Aviv University.

References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, 57 (1):289–300.
- Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24:407–418.
- Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika (To appear)*.
- Benjamini, Y. and Yekutieli, D. (2002). Hierarchical fdr testing of trees of hypotheses. *Technical Report RP-SOR-02-02*, URL <http://www.math.tau.ac.il/st/>.
- Benjamini, Y. and Yekutieli, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29 (4):1165–1188.
- Farcomeni, A. (2004). Multiple testing procedures under dependence, with applications. *Thesis*, URL <http://padis.uniroma1.it/getfile.py?recid=129>.
- Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J.R. Statist. Soc. B*, 64 (3):499–517.
- Heller, R., Stanley, D., Yekutieli, D., Rubin, N., and Benjamini, Y. (2005). Cluster based analysis of fmri data. *NeuroImage (Revisions under considerations)*.

- Kling, Y. (2005). Issues of multiple hypothesis testing in statistical process control. *Thesis, The Neiman Library of Exact Sciences & Engineering, Tel-Aviv University.*
- Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, 99 (468):1002–1014.
- Penny, W. and Friston, K. (2003). Mixtures of general linear models for functional neuroimaging. *IEEE Transactions on Medical Imaging*, 22:504–514.
- sartorius, B., Jacobsen, H., Torner, A., and Giesecke, J. (2006). Description of a new all cause mortality surveillance system in sweden as a warning system using threshold detection algorithms. *European Journal of Epidemiology*, 21:181 – 189.
- Shen, X., Huang, H.-C., and Cressie, N. (2002). Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association*, 97 (460):1122–1140.
- Storey, J., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66:187–205.
- Willink, R. (2004). Bounds on the bivariate normal distribution function. *Communications in Statistics: Theory and Methods*, 33(10):2281 – 2297.
- Yekutieli, D., Reiner, A., Elmer, G., Letwin, N., and Benjamini, Y. (to appear). Multiplicity issues related to complex research questions in microarrays analysis. *Statistica Neerlandica.*

A Appendix: Proofs of Theorems

A.1 Proof of theorem 3.1

$$\begin{aligned}
 WFDR &= E\left[\frac{\sum_{i=1}^{m_0} w_i V_i}{\sum_{i=1}^m w_i R_i} I_{\{\sum_{i=1}^m R_i > 0\}}\right] = \sum_{a>0} \frac{1}{a} E\left[\sum_{i=1}^{m_0} w_i V_i I_{\{\sum_{i=1}^m w_i R_i = a\}}\right] \\
 &= \sum_{i=1}^{m_0} w_i \sum_{a>0} \frac{1}{a} E[V_i I_{\{\sum_{i=1}^m w_i R_i = a\}}] = \sum_{i=1}^{m_0} w_i \sum_{a>0} \frac{1}{a} P(V_i = 1 \cap \sum_{i=1}^m w_i R_i = a) \quad (3)
 \end{aligned}$$

Note that the sum of weights a need not be an integer, but that its range is finite since we have a finite number of weights. Note also that under the PRDS assumption we can prove that the weighted FDR procedure is controlled at level $(\sum_{i=1}^{m_0} w_i/m)q$ by starting from (3) and following a similar proof to that in Benjamini and Yekutieli (2001).

From (3) it follows that the $WFDR$ can be expressed as

$$WFDR = \sum_{i=1}^{m_0} w_i \sum_a \frac{1}{a} P(H_{0i} \text{ is rejected and } \sum_{j=1, j \neq i}^m w_j R_j = a - w_i)$$

Let P_{0i} be the p-value associated with H_{0i} , $P^{(i)}$ a vector of p-values of the $m - 1$ hypotheses excluding H_{0i} .

Conditioning on $P^{(i)}$ we can express the $WFDR$ as

$$WFDR = \sum_{i=1}^{m_0} w_i E_{P^{(i)}} Q(P^{(i)}) \quad (4)$$

where $Q(P^{(i)}) = \sum_{a>0} P(H_{0i} \text{ is rejected and } \sum_{j=1, j \neq i}^m w_j R_j = a - w_i | P^{(i)})/a$.

For each value of $P^{(i)}$, let $r(P_{0i})$ be the sum of weighted rejections. Since the

p-values are independently distributed (specifically, P_{0i} is independent of $P^{(i)}$),

$$Q(P^{(i)}) = E_{P_{0i}}\left(\frac{1}{r(P_{0i})} I_{\{H_{0i} \text{ rejected}\}}\right) \quad (5)$$

Note that $\hat{m}_0 = \hat{m}_0(P_{0i}, P^{(i)})$ is non decreasing in P_{0i} for fixed $P^{(i)}$.

In the two-stage procedure, \hat{m}_0 can have two values. The smallest value, $m_{01}(i) = m - \sum_{l=1}^{j_1} w_{(l)} - w_i$, occurs when $P_{0i} \leq (\sum_{l=1}^{j_1} w_{(l)} + w_i)q'/m$, where $j_1 = \arg \max_{1 \leq k \leq m-1} \{P_{(k)}^{(i)} \leq (\sum_{l=1}^k w_{(l)} + w_i)q'/m\}$. The largest value occurs when H_{0i} is not rejected in the first stage, $m_{02}(i) = m - \sum_{k=1}^{j_2} w_{(k)}$, where $j_2 = \arg \max_{1 \leq k \leq m-1} \{P_{(k)}^{(i)} \leq (\sum_{l=1}^k w_{(l)})q'/m\}$.

Let the number of hypotheses rejected in the second stage along with H_{0i} be $r_h + 1$, where $r_h = \arg \max_{1 \leq k \leq m-1} \{P_{(k)}^{(i)} \leq (\sum_{l=1}^k w_{(l)} + w_i)q'/m_{0h}(i)\}$, $h = 1, 2$.

Using equation 5 we get

$$\begin{aligned} Q(P^{(i)}) &= \frac{1}{\sum_{l=1}^{r_1} w_{(l)} + w_i} \frac{\sum_{l=1}^{j_1} w_{(l)} + w_i}{m} q' \\ &+ \frac{1}{\sum_{l=1}^{r_2} w_{(l)} + w_i} \left(\frac{\sum_{l=1}^{r_2} w_{(l)} + w_i}{m_{02}(i)} q' - \frac{\sum_{l=1}^{j_1} w_{(l)} + w_i}{m} q' \right) \\ &\leq \frac{1}{\sum_{l=1}^{r_2} w_{(l)} + w_i} \left(\frac{\sum_{l=1}^{r_2} w_{(l)} + w_i}{m_{02}(i)} q' \right) = \frac{q'}{m_{02}} \end{aligned} \quad (6)$$

where the inequality follows since $r_2 \leq r_1$. Hence,

$$WFDR \leq q' \sum_{i=1}^{m_0} w_i E_{P^{(i)}} \frac{1}{m_{02}(i)}$$

A lower bound on $m_{02}(i)$ is $w_i + \sum_{j=1, j \neq i}^{m_0} w_j Y_j$, where $Y_j \sim B(1, 1 - q')$.

Lemma A.1.

$$\sum_{i=1}^{m_0} w_i E\left(\frac{1}{w_i + \sum_{j=1, j \neq i}^{m_0} w_j Y_j}\right) = \frac{1}{1 - q'} (1 - (q')^{m_0}),$$

where $Y_j \sim B(1, 1 - q')$, $j = 1 \dots m_0$ are independent Bernoulli random variables.

Proof. Let $S(k, i)$ and $S(k)$ denote all possible subsets of size k from $\{1, \dots, i - 1, i + 1, \dots, m_0\}$ and $\{1, \dots, m_0\}$ respectively.

$$\begin{aligned}
& \sum_{i=1}^{m_0} w_i E\left(\frac{1}{w_i + \sum_{j=1, j \neq i}^{m_0} w_j Y_j}\right) = \sum_{i=1}^{m_0} \sum_{k=0}^{m_0-1} \sum_{s \in S(k, i)} \frac{w_i}{w_i + \sum_{j \in s} w_j} (1 - q')^k (q')^{m_0-1-k} \\
&= \sum_{k=0}^{m_0-1} (1 - q')^k (q')^{m_0-1-k} \sum_{i=1}^{m_0} \sum_{s \in S(k, i)} \frac{w_i}{w_i + \sum_{j \in s} w_j} \\
&= \sum_{k=0}^{m_0-1} (1 - q')^k (q')^{m_0-1-k} \sum_{s \in S(k+1)} \sum_{j \in s} \frac{w_j}{\sum_{j \in s} w_j} \\
&= \sum_{k=0}^{m_0-1} (1 - q')^k (q')^{m_0-1-k} \binom{m_0}{k+1}
\end{aligned}$$

□

The result is immediate

$$\begin{aligned}
WFDR &\leq q' \sum_{i=1}^{m_0} w_i E_{P^{(i)}} \frac{1}{m_{02}(i)} \leq q' \sum_{i=1}^{m_0} w_i E\left(\frac{1}{w_i + \sum_{j=1, j \neq i}^{m_0} w_j Y_j}\right) \\
&\leq \frac{q'}{1 - q'} = q.
\end{aligned}$$

A.2 Proof of theorem 4.1

Proof. For notational convenience, let $A_0 = m_0/m$, $A_1 = 1 - A_0$, $F(u) = \tilde{\Phi}(\tilde{\Phi}^{-1}(u) - h\mu)$ and $F_i(u) = \tilde{\Phi}(\tilde{\Phi}^{-1}(u) - \mu_i)$. Let us first show, under the conditions of theorem 4.1, the asymptotic properties of our estimators:

1. $A_0^\infty = \lim_{m \rightarrow \infty} \hat{m}_0/m \geq A_0$ with probability 1:

$$\frac{\hat{m}_0}{m} = \frac{(m - R_1)}{m(1 - q)} \geq \frac{(m_0 - V_1)}{m(1 - q)} = \frac{1 - \frac{V_1}{m_0}}{1 - q} A_0$$

where V_1 and R_1 are the number of falsely rejected and rejected null hypotheses at the testing stage. Result follows since at the testing stage for a rejection its p-value has to be less than q so $\lim_{m \rightarrow \infty} V_1/m_0 < q$.

2. $h\mu^\infty = \lim_{m \rightarrow \infty} \hat{h}\mu \leq h\mu$ with probability 1:

$$\hat{h}\mu = \frac{\sum_{i=1}^m \sum_{e=1}^{c_i} x_{ei}}{\sum_{i=1}^m c_i} \xrightarrow{p} \frac{A_1 \nu_{c1}}{A_0 \nu_{c0} + A_1 \nu_{c1}} h\mu$$

where ν_{c0} and ν_{c1} are the mean cluster sizes under the null and under the alternative hypotheses respectively.

3. $u_1 \xrightarrow{p} u_1^\infty$: We shall first show that the weighted BH procedure corresponds to a (positive) threshold procedure asymptotically.

Lemma A.2. *Let u_1^∞ be the solution to $\tilde{F}(u) = \beta u$, where $\beta = (1/q - \frac{\mu_{w0}A_0}{\mu_{w0}A_0 + \mu_{w1}A_1}) / (\frac{\mu_{w1}A_1}{\mu_{w0}A_0 + \mu_{w1}A_1})$, μ_{w0} and μ_{w1} are the expected weights under the null and alternative hypotheses respectively, and $\tilde{F}(u) = \lim_{m \rightarrow \infty} \sum_{i=1}^{m_1} w_i F_i(u) / \sum_{i=1}^{m_1} w_i$. Let $R = \max\{j : p_{(j)} \leq (\sum_{i=1}^j w'_{(i)} / m)q\}$ where $w'_i = mw_i / \sum_{i=1}^m w_i$ is the normalized weight. Then (1) u_1^∞ is unique and positive and (2) $\sum_{i=1}^R w'_{(i)}q/m \xrightarrow{p} u_1^\infty$ under the assumptions of theorem 4.1.*

Proof. (1) follows by the concavity of $F_i(u)$ and the fact that $F'_i(0) > \beta$. The proof of (2) is a generalization of Theorem 1 in Genovese and Wasserman (2002) to arbitrary weights. Let $a_m = (1 - \epsilon_m)u_1^\infty m/q$ and $b_m = (1 + \epsilon_m)u_1^\infty m/q$, where $\lim_{m \rightarrow \infty} \epsilon_m = 0$ and $\lim_{m \rightarrow \infty} \sqrt{m}\epsilon_m = \infty$. For no-

tational convenience let $W'(k) = \sum_{i=1}^k w'_i$ and let I_0 and I_1 be the subset of indices corresponding to true and false null hypotheses. Note that

$$\begin{aligned} \{W'(R) < a_m\} &\subset \left\{ \sum_{i=1}^m w'_i I_{[p_i \leq a_m q/m]} < a_m \right\} \\ \left\{ \sum_{i=1}^R w'_i > b_m \right\} &= \bigcup_{\{k: W'(k) > b_m\}} \{p(k) \leq W'(k)q/m\} \\ &= \bigcup_{\{k: W'(k) > b_m\}} \left\{ \sum_{i=1}^m w'_i I_{[p_i \leq W'(k)q/m]} \geq W'(k) \right\} \end{aligned}$$

Let $\mu_w(t) = E(\sum_{i=1}^m w'_i I_{[p_i \leq tq/m]}) = \sum_{i \in I_0} w'_i tq/m + \sum_{i \in I_1} w'_i F_i(tq/m)$. The following properties can be easily established (1) $\lim_{m \rightarrow \infty} (\mu_w(a_m) - a_m) / \sqrt{m} = \infty$ (2) $\lim_{m \rightarrow \infty} (b_m - \mu_w(b_m)) / \sqrt{m} = \infty$ and (3) For m large enough $t - \mu_w(t)$ is an increasing function of t for $tq/m > u_1^\infty$. By Hoeffding's inequality,

$$\begin{aligned} \sum_{W'(k) > b_m} P\left(\sum_{i=1}^m w'_i I_{[p_i \leq W'(k)q/m]} \geq W'(k)\right) &\leq \sum_{W'(k) > b_m} \exp\left\{-\frac{2(W'(k) - \mu_w(W'(k)))^2}{m}\right\} \\ &\leq \sum_{W'(k) > b_m} \exp\left\{-\frac{2(b_m - \mu_w(b_m))^2}{m}\right\} \leq m \cdot \exp\left\{-\frac{2(b_m - \mu_w(b_m))^2}{m}\right\} \rightarrow 0 \end{aligned}$$

and similarly

$$P\left(\sum_{i=1}^m w'_i I_{[p_i \leq a_m q/m]} < a_m\right) \leq \exp\left\{-2\frac{(\mu_w(a_m) - a_m)^2}{m}\right\} \rightarrow 0$$

The result follows. For completeness, let us outline the proof of (2) (that of

(1) being very similar):

$$\begin{aligned}
b_m - \mu_w(b_m) &= \frac{(1 + \epsilon_m)u_1^\infty m}{q} \left(1 - \sum_{i=1}^{m_0} w'_i \frac{q}{m}\right) - \sum_{i=1}^{m_1} w'_i F_i((1 + \epsilon_m)u_1^\infty) \\
&= \frac{(1 + \epsilon_m)u_1^\infty m}{q} \left(1 - \sum_{i=1}^{m_0} w'_i \frac{q}{m}\right) - \sum_{i=1}^{m_1} w'_i \tilde{F}((1 + \epsilon_m)u_1^\infty) + O\left(\frac{1}{\sqrt{m}}\right) \\
&= \frac{(1 + \epsilon_m)u_1^\infty m}{q} \left(1 - \sum_{i=1}^{m_0} w'_i \frac{q}{m}\right) - \sum_{i=1}^{m_1} w'_i (\tilde{F}(u_1^\infty) + \tilde{F}'(u_1^\infty)\epsilon_m u_1^\infty + o(\epsilon_m)) + O(\sqrt{m}) \\
&= (1 + \epsilon_m)u_1^\infty \left(\frac{m}{q} - \sum_{i=1}^{m_0} w'_i\right) - \sum_{i=1}^{m_1} w'_i (\beta u_1^\infty + \tilde{F}'(u_1^\infty)\epsilon_m u_1^\infty + o(\epsilon_m)) + O(\sqrt{m}) \\
&= (1 + \epsilon_m)u_1^\infty \left(\frac{m}{q} - m \frac{\mu_{w0}A_0}{\mu_{w0}A_0 + \mu_{w1}A_1}\right) \\
&\quad - m \frac{\mu_{w1}A_1}{\mu_{w0}A_0 + \mu_{w1}A_1} (\beta u_1^\infty + \tilde{F}'(u_1^\infty)\epsilon_m u_1^\infty + o(\epsilon_m)) + O(\sqrt{m}) \\
&= (1 + \epsilon_m)u_1^\infty m \left(\frac{1}{q} - 1 \frac{\mu_{w0}A_0}{\mu_{w0}A_0 + \mu_{w1}A_1}\right) - m \left(\frac{1}{q} - 1 \frac{\mu_{w0}A_0}{\mu_{w0}A_0 + \mu_{w1}A_1}\right) u_1^\infty \\
&\quad - m \frac{\mu_{w1}A_1}{\mu_{w0}A_0 + \mu_{w1}A_1} (\tilde{F}'(u_1^\infty)\epsilon_m u_1^\infty + o(\epsilon_m)) + O(\sqrt{m}) \\
&= \epsilon_m u_1^\infty m \beta \frac{\mu_{w1}A_1}{\mu_{w0}A_0 + \mu_{w1}A_1} - m \frac{\mu_{w1}A_1}{\mu_{w0}A_0 + \mu_{w1}A_1} (\tilde{F}'(u_1^\infty)\epsilon_m u_1^\infty + o(\epsilon_m)) + O(\sqrt{m}) \\
&= m \frac{\mu_{w1}A_1}{\mu_{w0}A_0 + \mu_{w1}A_1} [\epsilon_m u_1^\infty (\beta - \tilde{F}'(u_1^\infty))] + o(\epsilon_m) + O(\sqrt{m})
\end{aligned}$$

By concavity of \tilde{F} and the fact that $\tilde{F}'(0) > \beta$ it follows that $(\beta - \tilde{F}'(u_1^\infty)) > 0$, so $(b_m - \mu_w(b_m))^2/m \rightarrow \infty$. The proof of (3) follows by showing that $d(t - \mu_w(t))/dt > 0$ for m large enough. First note that by the strong law of large numbers, for every $\epsilon_1 > 0$ and $\epsilon_2 > 0$ there exists an $m(\epsilon_1, \epsilon_2)$ such that for all $m \geq m(\epsilon_1, \epsilon_2)$

$$\begin{aligned}
\frac{\sum_{i=1}^{m_0} w'_i}{m} &\leq \epsilon_1 + \frac{A_0 \mu_{w0}}{A_0 \mu_{w0} + A_1 \mu_{w1}} \\
\frac{\sum_{i=1}^{m_1} w'_i F'_i(u_1^\infty)}{m} &\leq \epsilon_2 + \frac{A_1 \mu_{w1}}{A_0 \mu_{w0} + A_1 \mu_{w1}} \tilde{F}'(u_1^\infty)
\end{aligned}$$

Since $\tilde{F}(u_1^\infty) < \beta$ (by concavity of $\tilde{F}(u)$ and the fact that $\tilde{F}'(0) > \beta$) we can choose ϵ_1 and ϵ_2 so that $\epsilon_1 + \epsilon_2 \leq \frac{A_1\mu_{w1}}{A_0\mu_{w0} + A_1\mu_{w1}}(\beta - \tilde{F}(u^\infty))$. So for all $m \geq m(\epsilon_1, \epsilon_2)$ we have

$$\begin{aligned}
d(t - \mu_w(t))/dt &= 1 - \frac{\sum_{i=1}^{m_0} w'_i}{m} q - \frac{\sum_{i=1}^{m_1} w'_i F'_i(u^\infty)}{m} q \\
&\geq 1 - q(\epsilon_1 + \frac{A_0\mu_{w0}}{A_0\mu_{w0} + A_1\mu_{w1}} + \epsilon_2 + \frac{A_1\mu_{w1}}{A_0\mu_{w0} + A_1\mu_{w1}} \tilde{F}'(u^\infty)) \\
&\geq 1 - q(\frac{A_0\mu_{w0}}{A_0\mu_{w0} + A_1\mu_{w1}} + \frac{A_1\mu_{w1}}{A_0\mu_{w0} + A_1\mu_{w1}} \tilde{F}'(u^\infty) + \frac{A_1\mu_{w1}}{A_0\mu_{w0} + A_1\mu_{w1}}(\beta - \tilde{F}(u^\infty))) \\
&= 0
\end{aligned}$$

□

4. \hat{p}_i converges to a random variable that is stochastically larger than p_i :

Lemma A.3. *Let $X \sim N(0, 1)$, $T \sim N(\mu, 1)$ and $\text{corr}(X, T) = \rho \geq 0$. Then $f(\mu, \rho, x, t) = P(X \geq x | T \geq t)$ is a decreasing function of μ .*

Proof. The probability integral of the standard bivariate normal distribution with correlation ρ is $\Phi(h, k, \rho) = \int_{-\infty}^h \int_{-\infty}^k \frac{1}{2\pi\sqrt{1-\rho^2}} \exp[\frac{u^2 - 2\rho uv + v^2}{-2(1-\rho^2)}] dudv$. Simple calculus shows that

$$\begin{aligned}
f(\mu, \rho, x, t) &= \frac{\tilde{\Phi}(t - \mu) - \Phi(x) + \Phi(x, t - \mu, \rho)}{\tilde{\Phi}(t - \mu)} \\
\frac{\partial}{\partial \mu} f(\mu, \rho, x, t) &= \frac{\phi(t - \mu)[1 - \Phi(\frac{x - \rho(t - \mu)}{\sqrt{1 - \rho^2}})] - f(\mu, \rho, x, t)}{\tilde{\Phi}(t - \mu)}
\end{aligned}$$

A lower bound on $f(\mu, \rho, x, t)$ is $1 - \Phi(\frac{x - \rho(t - \mu)}{\sqrt{1 - \rho^2}})$ (see for example Willink (2004) for the derivation of this lower bound). Therefore $\frac{\partial}{\partial \mu} f(\mu, \rho, x, t) < 0$ and the result follows. □

The asymptotic p-value of an element in the trimming stage is

$$p_{ei} = p(u_1^\infty, A_0, \mu_i, \rho_{ei}) = \frac{A_0 f(0, \rho_{ei}, x_{ei}, \tilde{\Phi}^{-1}(u_1^\infty)) u_1^\infty + (1 - A_0) f(\mu_i, \rho_{ei}, x_{ei}, \tilde{\Phi}^{-1}(u_1^\infty)) \tilde{\Phi}(\tilde{\Phi}^{-1}(u_1^\infty) - \mu_i)}{A_0 u_1^\infty + (1 - A_0) \tilde{\Phi}(\tilde{\Phi}^{-1}(u_1^\infty) - \mu_i)}$$

where $\rho_{ei} = 1/\sqrt{c_i}$ for independent data.

The estimated p-value is $p(u_1, \hat{A}_0, \sigma_{\bar{X}_i} \hat{h}\mu, \rho_{ei})$. By Slutsky's theorem

$$\hat{p}_{ei}^\infty = \lim_{m \rightarrow \infty} p(u_1, \hat{A}_0, \sigma_{\bar{X}_i} \hat{h}\mu, \rho_{ei}) \stackrel{d}{=} p(u_1^\infty, A_0^\infty, \sigma_{\bar{X}_i} h\mu^\infty, \rho_{ei}).$$

By lemma A.3 and the fact that $A_0^\infty > A_0$ and $h\mu^\infty > h\mu$ we get that

$$\hat{p}_{ei}^\infty \geq p_{ei}(u^\infty, A_0, \mu_c, \rho_{ei})$$

Combining the above four results, since the BH procedure controls the FDR for independent test statistics as long as the p-values corresponding to null hypotheses are at least as large as the $U(0, 1)$ and since the set of p-values $\{\hat{p}_{ei}^\infty\}$ is independent if the set $\{p_{ei}(u^\infty, A_0, \mu_c, \rho_{ei})\}$ of asymptotic p-values is independent, we get that the FDR of a BH type procedure (e.g. adaptive weighted procedure) on $\{\hat{p}_{ei}^\infty\}$ is smaller than on $\{p_{ei}(u^\infty, A_0, \mu_c, \rho_{ei})\}$ because $P(\hat{p}_{ei}^\infty < a) \leq P(p_{ei}(u^\infty, A_0, \mu_c, \rho_{ei}) < a)$ for any a . Since the FDR on $\{p_{ei}(u^\infty, A_0, \mu_c, \rho_{ei})\}$ is less than q_2 , this upper bound is preserved. \square