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 28, 2007

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 somewhere within the cluster. We suggest controlling the false discovery rate (FDR) on clusters, i.e. the expected proportion of clusters rejected erroneously out of all clusters rejected, or its extension to general weights (WFDR). We introduce a powerful two-stage testing procedure and show that it controls the WFDR. 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 tests clusters first, then locations within rejected clusters. We show formally that this procedure controls the desired location error rate asymptotically, and conjecture that this is so also for realistic settings 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 signal $\{\mu(s) : s \in D \subset \Re^d\}$ that is deterministic and has spatial structure in the loose sense that the signal is present over regions rather than in singular points in the spatial domain. Let $D_0 = \{s \in D : \mu(s) \leq 0\}$ and $D_1 = \{s \in D : \mu(s) > 0\}$ denote the subsets of D with non-positive and positive signal respectively. We wish to infer on the subset D_1 of D with positive mean signal (the restriction to positive is for ease of notation) from the N realized processes $\{X_i(s) : i = 1, \ldots, N, s \in D \subset \Re^d\}$, where $X_i(s) = \mu(s) + \epsilon_i(s)$ and $\epsilon_i(\cdot)$ is a Gaussian mean zero process. While a large part of the spatial literature is devoted to the difficult case that $N = 1$, requiring extra assumptions on $\mu(\cdot)$ that capture the spatial structure of the signal, in recent important applications such as fMRI and satellite measurements we encounter the more general setting $N > 1$.

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

$$H_0(s) : \mu(s) = 0 \text{ versus } H_1(s) : \mu(s) > 0 \tag{1}$$

and adjust the level of the test to the multiplicity of locations to control the family-wise error (FWE, e.g. using random field theory or resampling methods) or the false discovery rate (FDR).

We argue, for two main reasons, that the data should be aggregated into clusters of locations before testing. First, 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. Second, spatial clusters of measured locations can have increased signal to noise ratio (SNR), since if the signal is present 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 $C_1, C_2, \ldots, C_m$ be a partition of D into m contiguous components we call clusters. We say that a cluster C_i does not contain 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 the only definition that guarantees that the p-values based on the spatially averaged signal are uniform or stochastically larger than the uniform under the null. The collection of m hypotheses tested are therefore

$$H_{0i} : \mu(s) = 0 \ \forall \ s \in C_i \text{ versus } H_{1i} : \mu(s) > 0 \text{ for at least one } s \in C_i \qquad (2)$$

for all $i = 1, \ldots, m$.

How to aggregate the data into spatial clusters is problem-specific and depends on the information at hand: in the fMRI example, Heller et al. (2006) use the data from a preparatory fMRI experiment to approximate the appropriate brain units (subunits) 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 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 can gain power by testing clusters rather than individual locations even if the partitioning is far from ideal. This is so, for example, for independent data if on average the proportion of locations with (same) signal in a cluster multiplied by the square root of the cluster size is greater than one.

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 ($\mathrm{FDR}_{\mathrm{CLUST}}$). Benjamini

and Hochberg (1997) also introduced the weighted FDR (WFDR) , which may be especially appropriate here. For example, we may want to control a size weighted FDR on clusters ($\text{WFDR}_{\text{CLUST}}$), 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 WFDR on clusters. In particular, we prove that the two stage procedure of Benjamini et al. (2006) can be applied using weights and still control the WFDR, 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 detect not only the clusters containing signal, but also the locations within the clusters where the signal is present. For this purpose we introduce a Cluster Testing and Trimming procedure for spatial data, that first tests clusters and then trims individual locations from the cluster discoveries to control the FDR on locations as well (possibly at a more lenient level). We show its asymptotic properties in section 4 and examine more realistic models using simulations in section 5.

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 tests (in the original units) is very large.

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 location intensities and therefore does not make use of the power gain made possible by averaging over locations within clusters (we come back to this point in section 7). An estimation

procedure of the spatial signal using the FDR 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 $\lambda_1(A) \equiv \lambda(A \cap D_1)$, $\lambda_0(A) \equiv \lambda(A \cap D_0)$, $Q_\lambda = \lambda_0(A)/\lambda(A) I_{\{\lambda(A)>0\}}$ where $I_{\{\lambda(A)>0\}}$ defines Q_λ as 0 when $\lambda(A) = 0$. 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*

$$\mathrm{FDR}_\lambda = \mathrm{E}(\mathrm{Q}_\lambda) = \mathrm{E}[\frac{\lambda_0(\mathrm{A})}{\lambda(\mathrm{A})} \mathrm{I}_{\{\lambda(\mathrm{A})>0\}}].$$

FDR_λ has been suggested by Pacifico et al. (2004) for a threshold procedure with data dependent threshold $T(X)$: $A(T) = \{s \in D : X(s) \geq T(X)\}$. 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 approximates the FDR in terms of area.

When the fundamental unit of interest is a cluster, 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 $C_1, \ldots, C_m$ be a partition of D into clusters, let I_0 be the subset*

of indices corresponding to null clusters, and let $R_i = 1$ *if cluster* i *is rejected and zero otherwise. The FDR on clusters is*

$$\mathrm{FDR_{CLUST}} = \mathrm{E}[\frac{\sum_{\mathrm{i}\in \mathrm{I}_0} \mathrm{R_i}}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{R_i}} \mathrm{I}_{\{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{R_i}>0\}}]$$

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*

$$\mathrm{WFDR_{CLUST}} = \mathrm{E}[\frac{\sum_{\mathrm{i}\in \mathrm{I}_0} \lambda(\mathrm{C_i})\mathrm{R_i}}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \lambda(\mathrm{C_i})\mathrm{R_i}} \mathrm{I}_{\{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{R_i}>0\}}]$$

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

$$\mathrm{WFDR} = \mathrm{E}[\frac{\sum_{\mathrm{i}\in \mathrm{I}_0} \mathrm{w_i}\mathrm{R_i}}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{w_i}\mathrm{R_i}} \mathrm{I}_{\{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{R_i}>0\}}]$$

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

Control of $\mathrm{FDR_{CLUST}}$ or $\mathrm{WFDR_{CLUST}}$ does not guarantee control of FDR_λ, but the following simple relation with $\mathrm{WFDR_{CLUST}}$ can be used to get an upper bound on FDR_λ for a testing procedure that rejects clusters:

$$\mathrm{FDR}_\lambda = \mathrm{E}[\frac{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{R_i}\lambda_0(\mathrm{C_i})}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{R_i}\lambda(\mathrm{C_i})} \mathrm{I}_{\{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{R_i}>0\}}] = \mathrm{WFDR_{CLUST}} + \mathrm{E}(\frac{\sum_{\mathrm{i}\in \mathrm{I}_1} \mathrm{R_i}\lambda_0(\mathrm{C_i})}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{R_i}\lambda(\mathrm{C_i})})$$

where I_1 is the subset of indices corresponding to non-null clusters.

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

units of interest are clusters and that 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

Consider any test for testing H_{0i} vs. H_{1i} in (2) and let P_i be the corresponding p-value and p_i its realized value. Note that any test statistic is appropriate as long as the p-value is valid, i.e. $P_i \overset{H_{0i}}{\sim} U(0,1)$ or $\underset{\text{st}}{\succ} U(0,1)$.

Benjamini and Hochberg (1995) introduced the FDR controlling linear step-up procedure, also called the BH procedure. Benjamini and Hochberg (1997) extend this procedure to general weights:

Procedure 3.1. *The weighted BH procedure: Order the cluster p-values* $p_{(1)} \leq \ldots \leq p_{(j)} \leq \ldots \leq p_{(m)}$. *Let* $w_{(j)}$ *be the weight associated with* $p_{(j)}$ *and* $\sum_{j=1}^{m} w_{(j)} = m$. *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.*

This procedure was shown in Benjamini and Hochberg (1997) to guarantee that for independent test statistics $\text{WFDR} \leq (\sum_{\text{i}\in\text{I}_0} \text{w}_\text{i}/\text{m})\text{q}$. Benjamini and Yekutieli (2001) further prove that $\text{FDR}_{\text{CLUST}} \leq \text{q}$ 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 cluster test statistics are Gaussian, non-negatively correlated, and the testing hypotheses are one-sided. Kling (2005) proved that the bound on WFDR is valid under the same conditions.

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 $w_i = 1,\ i = 1, \ldots, m$. They prove that it controls the FDR at level q for independent test statistics. The generalization of their procedure, that is especially useful when $\sum_{i \in I_0} w_i/m$ 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 $\hat{m}_0(w) = m - \sum_{i=1}^{k} w_{(i)}$. Let $k_2 = \max\{j : p_{(j)} \leq (\sum_{i=1}^{j} w_{(i)}/\hat{m}_0(w))q'\}$. 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 two-stage procedure for unit weights controls the FDR at level q under the PRDS assumption. The argument can be extended under weighting as follows: approximate the set of weights by integer positive weights $\{W_i : i = 1, \ldots, m\}$, then produce a vector of $m_{new} = \sum_{i=1}^{m} W_i$ p-values by repeating W_i times the p-value p_i for every $i = 1, \ldots, m$, thus producing an extreme case of positive dependency between the m_{new} p-values.

A more rigorous result under more general dependency structures can be given in an asymptotic framework. In this framework, we assume the locations are measured

on a grid. Let the n^{th} set of spatial observations consist of a grid of locations D_n. As the number of locations goes to infinity, we consider the following two asymptotic settings: (1)increasing domain asymptotics, where the distance between locations remains fixed but the domain goes to infinity $D_1 \subset D_2 \subset \ldots \subset \Re^d$, and (2)infill asymptotics, where the domain remains fixed but the number of points increases to infinity $D_n = D \cap (Z/\Lambda_n)^d$, where $\{\Lambda_n\}$ is an increasing sequence of positive integers and $(Z/\Lambda_n)^d = \{k/\lambda_n : k \in Z^d\}$ is the $1/\lambda_n$ lattice. The first setting is common in spatial statistics. The second setting makes sense if we think of an improved measuring device (e.g. an MRI machine), so that as the resolution of the device increases the device is also able to differentiate better between nearby locations.

Let S and S_0 be the number of locations measured and the number of null locations respectively. Consider the following asymptotic conditions:

$$\lim_{S\to\infty} \frac{S_0}{S} = A_0 \text{ Exists and } A_0 < 1 \tag{3}$$

$$F_S = \frac{1}{S}\sum_{s=1}^{S} 1[p_s < t|H_{0s}] \underset{S\to\infty}{\overset{a.s.}{\to}} A_0 F(t), \quad F(t) \leq t \quad \forall t \in (0,1] \tag{4}$$

$$G_S = \frac{1}{S}\sum_{s=1}^{S} 1[p_s < t|H_{1s}] \underset{S\to\infty}{\overset{a.s.}{\to}} (1-A_0)G(t) \quad \forall t \in (0,1] \tag{5}$$

These conditions can be satisfied, for example, if we assume that the spatial dependence of the noise between locations is local (m-dependence). The two asymptotic settings are equivalent under this assumption of local dependence. Since the dependence among the location test statistics is finite, if condition (3) is satisfied, the empirical processes (4) and (5) converge by the lemma of Glivenko-Cantelli (which is valid under local dependence).

Let $F_m = \frac{1}{m}\sum_{i=1}^{m} 1[p_i < t|H_{0i}]$ and $G_m = \frac{1}{m}\sum_{i=1}^{m} 1[p_i < t|H_{1i}]$ be the empirical distribution functions of the p-values coming from the cluster null and alternative

hypotheses respectively.

Theorem 3.2. *Assume conditions (3)-(5) hold, that the number of clusters $m(S)$ satisfies $\lim_{S\to\infty} m(S) = \infty$ and that $\lim \frac{m_0}{m}$ exists. Suppose also that $\delta \equiv \sup\{t : \frac{t}{\lim_{m\to(F_m(t)+G_m(t))}} \leq q\} \in (0, 1]$. Then, as $S \to \infty$ Procedure 3.2 with unit weights controls the FDR asymptotically at level q.*

Proof. The threshold in procedure 3.2 is $t^*_m = \sup\{t : \frac{\frac{\hat{m}_0}{m}(1+q)t}{(F_m(t)+G_m(t))\vee\frac{1}{m}} \leq q\}$. Therefore

$$FDR = E(\frac{F_m(t^*_m)}{(F_m(t^*_m)+G_m(t^*_m))\vee\frac{1}{m}}) = E(\frac{\frac{\hat{m}_0}{m}(1+q)t^*_m}{(F_m(t^*_m)+G_m(t^*_m))\vee\frac{1}{m}} + \frac{F_m(t^*_m)-\frac{\hat{m}_0}{m}(1+q)t^*_m}{(F_m(t^*_m)+G_m(t^*_m))\vee\frac{1}{m}})$$
$$\leq q + sup_{t\geq\delta}\{\frac{(F_m(t)-\lim\frac{m_0}{m}t)}{(F_m(t)+G_m(t))\vee\frac{1}{m}}\} + sup_{t\geq\delta}\{\frac{(\lim\frac{m_0}{m}-\frac{\hat{m}_0}{m}(1+q))t}{(F_m(t)+G_m(t))\vee\frac{1}{m}}\} + I\{t^*_m < \delta\}$$

From (3)-(5) and the definition of δ, it follows that (a) the second term is asymptotically negative because these conditions guarantee that the variance of $F_m(t)$ is asymptotically zero so $\lim F_m(t) \leq \lim(\frac{m_0}{m})t$, (b) the third term is asymptotically negative because our estimate $\frac{\hat{m}_0}{m} = \frac{m-R_1}{m} \geq (1-\frac{V_1}{m_0})\frac{m_0}{m}$, where R_1 and V_1 are the number of hypotheses rejected and rejected falsely respectively in the first stage, has a lower bound since $E(V_1) \leq m_0q'$:
$\lim_{m\to\infty}\frac{\hat{m}_0}{m}(1+q) \geq \lim_{m\to\infty}(1-q')\frac{m_0}{m}(1+q) = \lim\frac{m_0}{m}$, and (c) the fourth term is 0. Therefore the asymptotic upper bound for the FDR is q. □

See Genovese and Wasserman (2004), Storey et al. (2004) and Ferreira and Zwinderman (2006) for similar results under slightly different conditions and for different estimates of the proportion of null hypothesis.

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.

For any chosen location-wise test statistic, let the location-wise p-value map be $\{\tilde{p}_{li}|l = 1, \ldots, c_i \; i = 1, \ldots, m\}$, where the number of locations in cluster i is c_i and $z_{li} = \Phi^{-1}(1 - \tilde{p}_{li})$ is the corresponding z-score for location l in cluster i. The regular FDR on locations takes the following form:

$$\mathrm{FDR_{LOC}} = \mathrm{E}[\frac{\sum_{\mathrm{i=1}}^{\mathrm{m}} \sum_{\mathrm{l=1}}^{\mathrm{c_i}} \mathrm{V_{li}}}{\sum_{\mathrm{i=1}}^{\mathrm{m}} \sum_{\mathrm{l=1}}^{\mathrm{c_i}} \mathrm{R_{li}}} \mathrm{I}_{\{\sum_{\mathrm{i=1}}^{\mathrm{m}} \sum_{\mathrm{l=1}}^{\mathrm{c_i}} \mathrm{R_{li}} > 0\}}]$$

where $R_{li} = 1$ if location l in cluster i is rejected and 0 otherwise, and $V_{li} = 1$ if location l in cluster i is rejected erroneously and 0 otherwise.

If the test statistic used at the location level is independent of the one used at the cluster level, the BH procedure (or its two stage variant) can be directly applied at the location level (for locations within rejected clusters) to control $\mathrm{FDR_{LOC}}$. This is also the case asymptotically if the number of locations within each cluster grows to infinity in such a way that the correlation between the location test statistic and that of its cluster goes to zero. But what if the location test statistic is correlated with the test statistic at the cluster level?

To answer this, we will first test the cluster hypothesis (2) using the standardized z-score average of its locations as the cluster test statistic. Then, we will test the location hypothesis (1) for every location included in a rejected cluster.

The relevant location 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: $\rho_{li} = corr(Z_{li}, \overline{Z}_i)$(e.g. $\rho_{li} = 1/\sqrt{c_i}$ if the measured

signal between locations is independent), $\sigma_{\overline{Z}_i}$ is the standard deviation of $\overline{Z}_i$, and $\mu_i = E_{H_{1i}}(\overline{Z}_i)/\sigma_{\overline{Z}_i}$ is the standardized expectation of the cluster test statistic under the alternative hypothesis H_{1i}. The conditional p-value $p_{li}(u_1, m_0/m, \mu_i, \rho_{li})$ of a location within a cluster that passed an initial fixed cut-off u_1 is

$$p_{li} = \frac{\int_{z_{li}}^{\infty}(\frac{m_0}{m})(\widetilde{\Phi}\,(\frac{\widetilde{\Phi}^{-1}(u_1)-\rho_{li}u}{\sqrt{1-\rho_{li}^2}})) + (1-\frac{m_0}{m})(\widetilde{\Phi}\,(\frac{\widetilde{\Phi}^{-1}(u_1)-\rho_{li}u-\mu_i}{\sqrt{1-\rho_{li}^2}}))\phi(u)du}{\frac{m_0}{m}u_1 + (1-\frac{m_0}{m})(\widetilde{\Phi}\,(\widetilde{\Phi}^{-1}(u_1) - \mu_i))} \tag{6}$$

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

Lemma 4.1. *Assume that the non-null clusters' standardized z-scores are normally distributed. Then p_{li} given in (6) is a valid p-value for testing whether location l in rejected cluster i contains signal.*

Proof. We have to show that $p_{li} \sim U(0,1)$ (or stochastically larger than $U(0,1)$) if location l in cluster i contains no signal. Assume, as in **?**, that the cluster hypothesis come from a Bernoulli distribution with marginal probability m_0/m of being a null hypothesis. Then

$$\begin{aligned}
\mathrm{p_{value}} &= P_0(Z_{li} \geq z_{li} | \overline{Z}_i/\sigma_{\overline{Z}_i} \geq \widetilde{\Phi}^{-1}(u_1)) = \frac{P_0(Z_{li} \geq z_{li}, \overline{Z}_i/\sigma_{\overline{Z}_i} \geq \widetilde{\Phi}^{-1}(u_1))}{P_0(\overline{Z}_i/\sigma_{\overline{Z}_i} \geq \widetilde{\Phi}^{-1}(u_1))} \\
&= \frac{\frac{m_0}{m}P_0(Z_{li} \geq z_{li}, \overline{Z}_i/\sigma_{\overline{Z}_i} \geq \widetilde{\Phi}^{-1}(u_1)|H_{0i}) + (1-\frac{m_0}{m})P_0(Z_{li} \geq z_{li}, \overline{Z}_i/\sigma_{\overline{Z}_i} \geq \widetilde{\Phi}^{-1}(u_1)|H_{1i})}{\frac{m_0}{m}u_1 + (1-\frac{m_0}{m})P(\overline{Z}_i/\sigma_{\overline{Z}_i} \geq \widetilde{\Phi}^{-1}(u_1)|H_{1i})}
\end{aligned}$$

where the subscript zero indicates that the probabilities are calculated under the null hypothesis. Since $\overline{Z}_i$ is normally distributed under H_{1i}, the p-value is exactly p_{li} because the joint distribution of a location z-score with its standardized cluster

average z-score is bivariate normal with unit variance, correlation ρ_{li}, zero location mean and zero or μ_i mean for the standardized cluster average, depending on whether the cluster null hypothesis is true or false. □

Note that the average of c_i independent (or m-dependent) random variables will be approximately Gaussian for c_i large enough. Still, some caution should be exercised in using (6) since it is sensitive to the Gaussian distribution assumption of the average z-score in the extreme right tail of the distribution for non-null clusters. So it is possible to take the more conservative approach of approximating the p-value by an upper bound that will be valid regardless of whether the Gaussian distribution assumption on non-null clusters holds. This upper bound occurs when $\mu_i \to 0$: $\widetilde{\Phi}(z_{li}) + \frac{1}{u_1}\{\int_{z_{li}}^{\infty}(\widetilde{\Phi}(\frac{\widetilde{\Phi}^{-1}(u_1)-\rho_{li}u}{\sqrt{1-\rho_{li}^2}}))\phi(u)du - \widetilde{\Phi}(z_{li})u_1\}$. Note that the smallest p-value is achieved when $\mu_i = \infty$: $\widetilde{\Phi}(z_{li}) + \frac{1}{u_1+\frac{m-m_0}{m_0}}\{\int_{z_{li}}^{\infty}(\widetilde{\Phi}(\frac{\widetilde{\Phi}^{-1}(u_1)-\rho_{li}u}{\sqrt{1-\rho_{li}^2}}))\phi(u)du - \widetilde{\Phi}(z_{li})u_1\}$, so this upper bound is a tight upper bound as the fraction of null clusters goes to 1.

In practice, the parameters necessary for the p-value calculation are unknown and need to be estimated from the data. Moreover, unless the locations have independent statistics the standard error of $\overline{Z}_i$, $\sigma_{\overline{Z}_i}$, is unknown and has to be estimated from the data. The first step towards estimation of ρ_{li} and $\sigma_{\overline{Z}_i}$ is the estimation of the correlations between locations that belong to the same clusters. We suggest estimating the correlations between locations by first estimating the variogram, $var(Z_{li} - Z(_{ki})) = 2\gamma(Z_{li}, Z_{ki})$. If the Z-score process satisfies the following "intrinsic stationarity on clusters" assumptions (which are satisfied if the Z-map is intrinsically stationary): $var(Z_{li} - Z_{ki}) = 2\gamma(\sqrt{s_{li} - s_{ki}})$ where s_{li} and s_{ki} are the locations in space of Z_{li} and Z_{ki} respectively, and $E(Z_{li}) = E(Z_{ki})$, then the variogram $2\gamma(h)$ can be empirically estimated. However, since $E(Z_{li}) = E(Z_{ki})$ holds in most pairs

of locations, the variogram is estimated robustly by (page 75 in Cressie (1991))

$$2\hat{\gamma}(h) = [med\{(Z_{ki} - Z_{li})^2 : (s_{ki}, s_{li}) \in N(h)\}]/0.455. \tag{7}$$

On the estimated p_{li}'s, we propose to use the two stage FDR procedure of Benjamini et al. (2006). The gain in power over the BH FDR procedure may be high since the potential number of true location 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 procedure:*

1. *Testing Stage:*

   (a) *For every location calculate its original p-value and let z_{li} be the corresponding z-score for location l in cluster i.*

   (b) *Use the robust variogram estimator (7) to estimate the correlation between the z-scores of locations within clusters.*

   (c) *Let the correlation between two locations be $\hat{\rho}^i_{l,k} = 1 - \hat{\gamma}(s_{li} - s_{ki})$. Then $\hat{\sigma}_{\overline{Z}_i} = \frac{1}{c_i}\sqrt{(c_i + 2\sum_{l=1}^{c_i}\sum_{k=1}^{l-1}\hat{\rho}^i_{l,k})}$.*

   (d) *Apply procedure 3.1 at level q on $\{\widetilde{\Phi}\,(\overline{Z}_i/\hat{\sigma}_{\overline{Z}_i}), i = 1, \ldots, m\}$. Let k be the number of clusters rejected and let u_1 be the cut-off point of the largest p-value rejected (e.g. $u_1 = kq/m$ for a BH procedure).*

2. *Trimming Stage:*

   (a) *Let $\hat{m}_0 = \frac{(m-k)}{1-q}$, $\hat{\mu}_i = (\frac{\sum_{i=1}^{m}\sum_{l=1}^{c_i} z_{li}}{\sum_{i=1}^{m} c_i})/\hat{\sigma}_{\overline{Z}_i}$, $\hat{\rho}_{li} = \frac{1+\sum_{k\neq l,k=1}^{c_i}\hat{\rho}^i_{l,k}}{c_i\hat{\sigma}_{\overline{Z}_i}}$.*

   (b) *Estimate the p-value by plugging in the estimated parameters above in equation (6) to get $\hat{p}_{li}$.*

(c) Apply the **two-stage procedure** *at level* q_2*:*

i. Sort the resulting m_2 *p-values* $\hat{p}_{li}$*:* $\hat{p}_{(1)} \leq \ldots \leq \hat{p}_{(m_2)}$*. Let* $q_2' = q_2/(1+q_2)$, $\quad k_1 = \max\{j : \hat{p}_{(j)} \leq (j/m_2)q_2'\}$.

ii. Let $\hat{m_{02}} = m_2 - k_1$ *and* $k_2 = max\{j : \hat{p}_{(j)} \leq (j/\hat{m}_{02})q_2'\}$*. Keep all locations with* $\hat{p}_{li} \leq (k_2/\hat{m}_{02})q_2'$ *in the clusters, trimming the others.*

Consider the following condition: $\lim_{m\to\infty} \frac{\sum_{i=1}^{m} c_i \overline{z}_{\cdot i}}{\sum_{i=1}^{m} c_i} \leq \min_{i:i\in I_1, \text{i rejected}} E(\overline{Z}_i)$. This condition is satisfied if the expected cluster average of every non-null cluster (that is rejected in the testing stage) is larger than the expected average of the entire z-score map. This is a fairly weak assumption, especially when m_0/m is close to 1.

Lemma 4.2. *Assume the conditions of theorem 3.2 and lemma 4.1 are satisfied. Moreover, assume that the correlations between the locations within the clusters are consistently estimated and the above additional condition. Then* $\hat{p_{li}}$ *converges to a r.v. that is stochastically larger than* p_{li}*.*

See appendix A.2 for a proof.

Lemma 4.2 guarantees that the p-values are asymptotically valid, in the sense that $\hat{p}_{li}^{\infty} \equiv \lim \hat{p}_{li}(m)$ is $U(0,1)$ or stochastically larger than the uniform under the null hypothesis. The conditions of theorem 3.2 guarantee that the dependency between the p-values in the set $\{\hat{p}_{li}^{\infty}\}$ is local, so that the two-stage procedure in 4.1 controls the FDR_{LOC} asymptotically at level q_2 when the asymptotic threshold is greater than zero.

5 A Simulation Study

We conducted a simulation study in order to (1) validate that the Cluster Testing and Trimming Procedure 4.1 controls the FDR_{LOC} at level q_2 under realistic settings, and (2) compare the performance of procedures 3.1 and 4.1 with that of a location-wise analysis on the entire data.

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 locations 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 location was either zero or positive. Active locations 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 location means within a cluster ranged from 0.2 to 2). White noise was added to each location 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).

For simulations with spatial dependence, a map of 32×32 points was considered with various cluster and signal configurations. White noise convolved with a Gaussian filter created spatially correlated noise. The filter width varied in the simulations.

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 FDR was estimated by averaging over the simulations the proportion of false units rejected among all rejected units. For $\text{FDR}_{\text{CLUST}}$ and FDR_{LOC} the units were clusters and locations respectively. Note that after the trimming stage a cluster is a discovery if at least one location within the cluster is rejected. If no clusters were rejected in the testing stage, the realized $\text{FDR}_{\text{CLUST}}$ and FDR_{LOC} 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: locations 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 locations after the trimming stage with $q_2 = 0.05$ and with $q_2 = 0.25$, and the power of a location-wise analysis on the entire data.

FDR Control For all signal configurations, even with variable cluster size, variable signal intensity μ for non-null locations and a variable percent of active locations within an active cluster h, control over $\text{FDR}_{\text{CLUST}}$, FDR_{s} and FDR_{LOC} 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. Figure 1 shows that the $\text{FDR}_{\text{CLUST}}$ after the testing stage is around 0.05. After the trimming stage, the $\text{FDR}_{\text{CLUST}}$ is much lower than 0.05 for small μ. Note that although the Procedure 4.1 with fixed q_2 does not guarantee that the FDR on clusters is preserved, it was preserved in all our simulations. The reason why it is low for small μ is that only very few locations were rejected, and they belonged mostly to true cluster rejections. The estimated $\text{FDR}_{\text{CLUST}}$ 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). $\mathrm{FDR}_{\mathrm{LOC}}$ is not preserved after the testing stage, yet for all signal and cluster configurations the $\mathrm{FDR}_{\mathrm{LOC}}$ 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 location 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 percent signal within cluster $h = 0.25$ and cluster size $c = 8$, the single location analysis has more power. This is so since $\sqrt{c}h < 1$ in this case.

For independent test statistics, we observed via simulations that the power is higher for testing clusters than a single location analysis if $h > 1/\sqrt{c}$, where h is the average percent of signal within clusters containing signal, and c is the average cluster size. The gain in power is due to the reduced number of hypotheses tested as well as the higher SNR per cluster than per location ($E(\overline{Z}_i/\sigma_{\overline{Z}_i}) \overset{H_{1i}}{=} \sqrt{c}h\mu/\sigma$ and $E(Z_i(s)) \overset{H_1(s)}{=} \mu/\sigma$). Note that although a cluster is considered a true discovery if at least one location within it contains signal, i.e $ch > 1$, the power is guaranteed to increase only if $\sqrt{c}h > 1$. As $\sqrt{c}h$ increases the advantage in power of our procedures over single location analysis is larger. The gain in power is of course even larger for $q_2 > 0.05$.

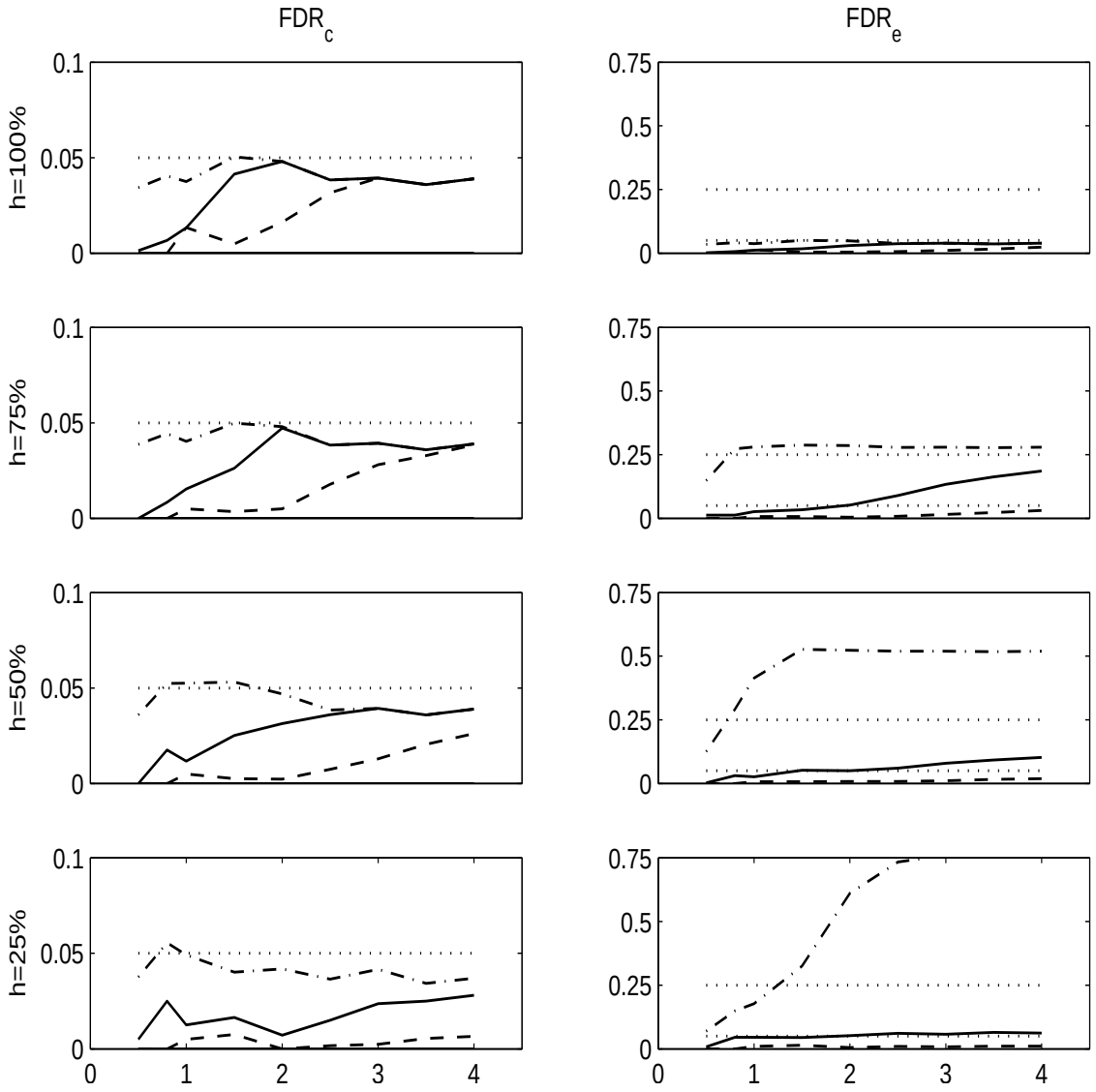


Figure 1: $\mathrm{FDR}_{\mathrm{CLUST}}$(left) and $\mathrm{FDR}_{\mathrm{LOC}}$(right) as a function of μ for (1) cluster-wise analysis (dash-dot line); (2)Cluster Testing and Trimming Procedure 4.1 with $q_2 = 0.05$ (dashed line); (3) Cluster Testing and Trimming Procedure 4.1 with $q_2 = 0.25$ (solid line). The $\mathrm{FDR}_{\mathrm{CLUST}}$ after the testing stage is around 0.05. After the trimming stage, the $\mathrm{FDR}_{\mathrm{CLUST}}$ is much lower than 0.05 for small μ. The $\mathrm{FDR}_{\mathrm{LOC}}$ is below q_2 after the trimming stage.

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 cognitive tasks. The

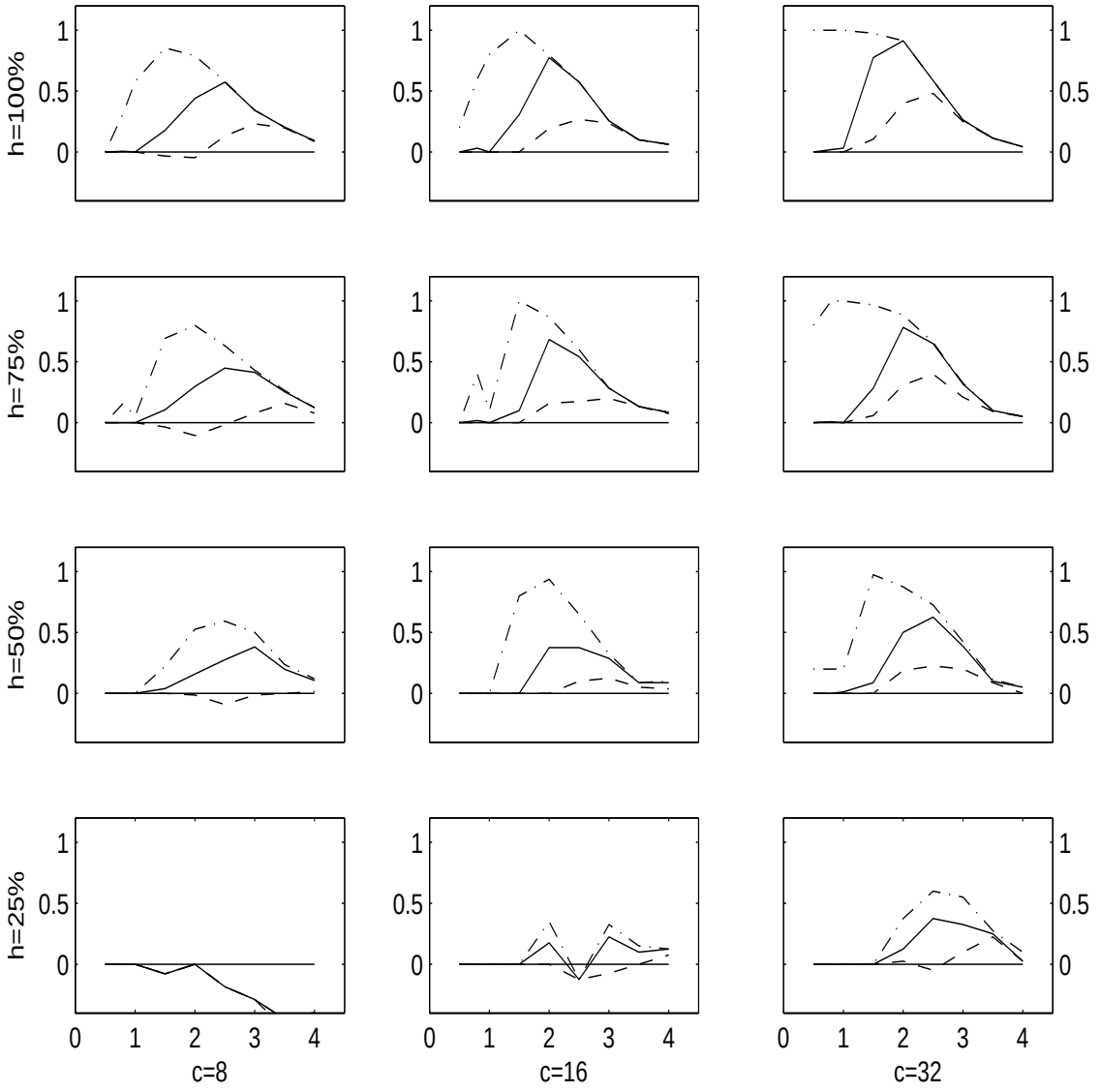


Figure 2: Power difference from location-wise analysis as a function of μ for (1) cluster-wise analysis (dash-dot line); (2)Cluster Testing and Trimming Procedure 4.1 with $q_2 = 0.05$ (dashed line); (3) Cluster Testing and Trimming Procedure 4.1 with $q_2 = 0.25$ (solid line). The power advantage is largest when μ is not too small and not too large. As $\sqrt{c}h$ increases the advantage is larger. Disadvantage only when $\sqrt{c}h < 1$.

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. (2006) for details. From the preparatory scan we also defined a limited region of interest (ROI, see Heller et al. (2006) for details) that contained 232 voxels grouped into 20 clusters.

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, we applied procedure 4.1 within the ROI with the modification that the variogram was estimated from brain regions outside the ROI. Procedure 4.1 is valid in this setting since (1) intrinsic stationarity is a good approximation for brain imaging data (see e.g. Worsley and Friston (2000)) (2) we only need this assumption for distances within clusters (3) according to Genovese et al. (2002) the correlations between locations are local and tend to be positive, so moving to clusters the correlations will also be positive and local.

Testing clusters at an FDR level of 0.05 within the region of interest, we found 195 voxels grouped into 14 clusters. Trimming further at an FDR level of 0.25 and 0.05 discovered 118 voxels within the 14 clusters and 34 voxels within 12 clusters respectively. For comparison, single voxel testing at an FDR level of 0.05 (using the two-stage FDR procedure) found 36 voxels. See figure 3 for the estimated location p-values within detected clusters in slices 9-10. From the estimated location p-values within rejected clusters we can see, for example, that for the light blue cluster most of the cluster contains signal except at the edges in slice 10. The largest p-value rejected when trimming at an FDR level of 0.25 and 0.05 was 0.23 and 0.01 respectively.

If the investigator wants control over FDR on locations at the 0.05 level, the Cluster Testing and Trimming Procedure 4.1 discovers almost the same voxels as single voxel testing. However, if the investigator is willing to control the FDR_{LOC} at a higher level (say 0.25) as long as he already has control over $\text{FDR}_{\text{CLUST}}$ at the 0.05 level, then many more voxels are discovered.

Remark 6.1. Typically the experiment is run on more than one subject and a group analysis is performed per voxel. We can apply our clustering algorithm on the averaged or concatenated multi-subject time series to find clusters common to all subjects. Once the cluster units are defined our suggested testing procedures can be easily applied. For example, we can apply the Cluster Testing and Trimming Procedure 4.1 on the p-value map that is produced from a standard fMRI group analysis per voxel. The difficult question is how to define common cluster units for the group that take into account the variability between the subjects.

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 or the WFDR on clusters. We generalized the two-stage FDR procedure in Benjamini et al. (2006) to arbitrary weights, and showed that it controlled the WFDR under independence of test statistics. When the test statistics are not independent, we know that procedure 3.1 controls the WFDR if the PRDS property is satisfied. We believe that, as argued in Benjamini et al. (2006), the

two-stage weighted procedure 3.2 controls the WFDR under dependency, but this remains to be proved. We showed that the procedure is valid for local dependencies asymptotically.

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 $E\Theta_\tau(T)$ where $\Theta_\tau(T) = \#\{1 \leq k \leq m_T : \lambda_0(C_i)/\lambda(C_i) \geq \tau\}/m_T$, 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 $C_1, \ldots, C_{m_T}$ defines the set of clusters rejected. Thus their procedure is based on each location's individual test statistic, and no advantage is taken of the SNR increase that an aggregation of locations within a cluster can offer. In this paper we showed the advantage of such averaging, in terms of power, over single location analysis. The partition of Pacifico et al. (2004) 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 but this is quite often feasible in spatial applications. An advantage of Pacifico et al. (2004) is that the clusters detected are separate regions, whereas our method may find a region that is comprised of several detected clusters. An interesting point for further research is to combine the two approaches, so that we gain power from cluster testing as well as control the FDR on separate regions.

The gain in power when testing clusters rather than individual locations depends on the quality of the partition. For independent test statistics, we showed via simulations 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. 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 locations 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 the Cluster Testing and Trimming procedure to indicate where the signal is within the detected cluster. This procedure extends the theory on hierarchical FDR controlling procedures in Yekutieli et al. (2006) 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, but this is not necessary. For example, the investigator may be satisfied in knowing that the detected clusters with no signal at all comprise no more that 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.

Although the p-value calculation for the trimming stage is exact, taking into account that the cluster average has passed an initial cut-off, the estimated trimming stage p-values are conservative due to the conservative estimation of the unknown parameters. If these parameters were known, the FDR controlling procedure would have been even more powerful. This is a point for further research.

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*, 93 (3):491–507.

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.

Cressie, N. (1991). *Statistics for Spatial Data*. Wiley, New York.

Ferreira, J. and Zwinderman, A. (2006). On the benjamini-hochberg method. *The Annals of Statistics*, 34 (4):1827–1849.

Genovese, C., Lazar, N., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15:870–878.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The annals of statistics*, 32 (3):1035–1061.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., and Benjamini, Y. (2006). Cluster based analysis of fmri data. *NeuroImage*, 33(2):599–608.

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.

Worsley, K. and Friston, K. (2000). A test for conjunction. *Statistics and probability letters*, 47:135–140.

Yekutieli, D., Reiner, A., Elmer, G., Kafkafi, N., Letwin, N., Lee, N., and Benjamini,

Y. (2006). Approaches to multiplicity issues in complex research in microarray analysis. *Statistica Neerlandica*, 60 (4):414–437.

A Appendix: Proofs of Theorems

A.1 Proof of theorem 3.1

Let a be the sum of weights rejected. Note that a need not be an integer, but that its range is finite since we have a finite number of weights. The $WFDR$ is

$$WFDR = E[\frac{\sum_{i\in I_0} w_i R_i}{\sum_{i=1}^{m} w_i R_i} I_{\{\sum_{i=1}^{m} R_i > 0\}}] = \sum_{a>0} \frac{1}{a} E[\sum_{i\in I_0} w_i R_i I_{\{\sum_{i=1}^{m} w_i R_i = a\}}]$$

$$= \sum_{i\in I_0} w_i \sum_{a>0} \frac{1}{a} E[R_i I_{\{\sum_{i=1}^{m} w_i R_i = a\}}] = \sum_{i\in I_0} w_i \sum_{a>0} \frac{1}{a} P(R_i = 1 \ \cap \ \sum_{i=1}^{m} w_i R_i = a)$$

$$= \sum_{i\in I_0} w_i \sum_{a} \frac{1}{a} P(R_i = 1 \ \cap \sum_{j=1, j\neq i}^{m} w_j R_j = a - w_i)$$

$$= \sum_{i\in I_0} w_i E_{P^{(i)}}[\frac{1}{a} \sum_{a>0} P(R_i = 1 \ \cap \sum_{j=1, j\neq i}^{m} w_j R_j = a - w_i | P^{(i)})] = \sum_{i\in I_0} w_i E_{P^{(i)}} Q(P^{(i)})$$

where $P^{(i)}$ a vector of p-values of the $m-1$ hypotheses excluding H_{0i}.

Let P_{0i} is the p-value associated with H_{0i}, and for each value of $P^{(i)}$ let $r(P_{0i})$ be the sum of weights rejected and $l(P_{0i})$ be the indicator that H_{0i} has been rejected as a function of P_{0i} . Since the p-values are independently distributed (specifically, P_{0i} is independent of $P^{(i)}$), for every $i \in I_0$ we can express $Q(P^{(i)})$ as

$$Q(P^{(i)}) = E_{P_{0i}}(\frac{l(P_{0i})}{r(P_{0i})} | P^{(i)}) = \int \frac{l(p)}{r(p)} dF_{01}(p) \tag{8}$$

where F_{01} is the cumulative distribution function of P_{01}.

Note that $\hat{m_0} = \hat{m_0}(P_{0i}, P^{(i)})$ is non decreasing in P_{0i} for fixed $P^{(i)}$, and it can

have two values depending on whether P_{0i} is rejected in the first stage:

$m_{01}(i) = m - \sum_{l=1}^{j_1} w_{(l)} - w_i$ if $P_{0i} \leq (\sum_{l=1}^{j_1} w_{(l)} + w_i)q'/m$ and $m_{02}(i) = m - \sum_{k=1}^{j_2} w_{(k)}$ otherwise, where $j_1 = \arg\max_{1\leq k\leq m-1}\{P_{(k)}^{(i)} \leq (\sum_{l=1}^{k} w_{(l)} + w_i)\frac{q'}{m}\}$ and $j_2 = \arg\max_{1\leq k\leq m-1}\{P_{(k)}^{(i)} \leq (\sum_{l=1}^{k} w_{(l)})\frac{q'}{m}\}$.

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 (8) we get

$$\begin{aligned} Q(P^{(i)}) &= \frac{1}{\sum_{l=1}^{r_1} w_{(l)} + w_i} \frac{\sum_{l=1}^{j_1} w_{(l)} + w_i}{m} q' \\ &+ \frac{1}{\sum_{l=1}^{r_2} w_{(l)} + w_i} \left(\frac{\sum_{l=1}^{r_2} w_{(l)} + w_i}{m_{02}(i)} q' - \frac{\sum_{l=1}^{j_1} w_{(l)} + w_i}{m} q'\right) \\ &\leq \frac{1}{\sum_{l=1}^{r_2} w_{(l)} + w_i} \left(\frac{\sum_{l=1}^{r_2} w_{(l)} + w_i}{m_{02}(i)} q'\right) = \frac{q'}{m_{02(i)}} \end{aligned} \tag{9}$$

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

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. *If $Y_j \sim B(1, 1 - q'), j = 1, \ldots, m_0$ are independent, then* $\sum_{i\in I_0} w_i E(\frac{1}{w_i + \sum_{j=1, j\neq i}^{m_0} w_j Y_j}) = \frac{1}{1-q'}(1 - (q')^{m_0})$.

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

$$\begin{aligned} &\sum_{i\in I_0} w_i E\left(\frac{1}{w_i + \sum_{j=1, j\neq i}^{m_0} w_j Y_j}\right) = \sum_{i\in I_0} \sum_{k=0}^{m_0-1} \sum_{s\in S(k,i)} \frac{w_i}{w_i + \sum_{j\in s} w_j} (1 - q')^k (q')^{m_0-1-k} \\ = &\sum_{k=0}^{m_0-1} (1 - q')^k (q')^{m_0-1-k} \sum_{i\in I_0} \sum_{s\in S(k,i)} \frac{w_i}{w_i + \sum_{j\in s} w_j} \\ = &\sum_{k=0}^{m_0-1} (1 - q')^k (q')^{m_0-1-k} \sum_{s\in S(k+1)} \sum_{j\in s} \frac{w_j}{\sum_{j\in s} w_j} = \sum_{k=0}^{m_0-1} (1 - q')^k (q')^{m_0-1-k} \binom{m_0}{k+1} \end{aligned}$$

□

The result is immediate

$$WFDR \leq q' \sum_{i \in I_0} w_i E_{P^{(i)}} \frac{1}{m_{02}(i)} \leq q' \sum_{i \in I_0} w_i E(\frac{1}{w_i + \sum_{j=1, j \neq i}^{m_0} w_j Y_j}) \leq \frac{q'}{1-q'} = q.$$

A.2 Proof of theorem 4.2

Proof. Let $A_0 = m_0/m$ and $A_1 = 1 - A_0$. Let us first show, under the conditions of theorem 4.2, the asymptotic properties of our estimators:

Property 1. Let $A_0^\infty = \lim_{m \to \infty} \hat{m}_0/m$. Then $A_0^\infty \geq A_0$ with probability 1:

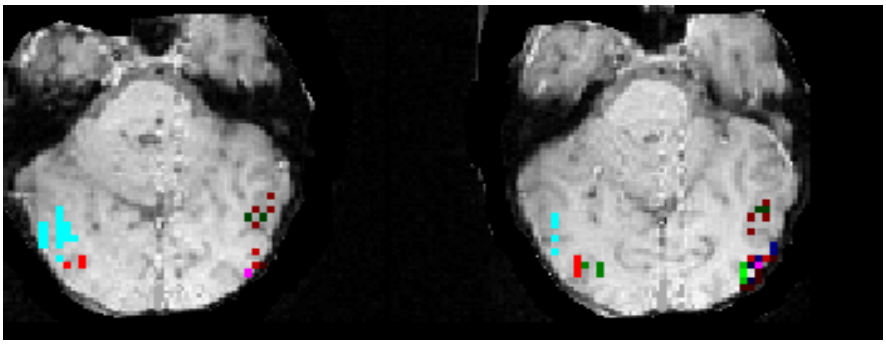
$$\frac{\hat{m_0}}{m} = \frac{(m - R_1)}{m(1-q)} \geq \frac{(m_0 - V_1)}{m(1-q)} = \frac{1 - \frac{V_1}{m_0}}{1-q} A_0$$

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 the p-value has to be less than q so $\lim_{m \to \infty} V_1/m_0 < q$.

Property 2. Let $\mu_c^\infty = \lim_{m \to \infty} \frac{\sum_{i=1}^m \sum_{l=1}^{c_i} z_{li}}{\sum_{i=1}^m c_i}$. Then $\mu_c^\infty \leq E(\overline{Z}_i)$ with probability 1: this follows from the condition on the cluster means in the theorem.

Property 3. The cluster testing procedure corresponds to a positive asymptotic threshold procedure with threshold δ (defined in theorem 3.2).

The asymptotic p-value of a location in the trimming stage is, using the notation in (6), $p_{li}^\infty = p_{li}(\delta, A_0, \mu_i, \rho_{li})$. The estimated p-value is $p(u_1, \hat{A}_0, \hat{\mu}_i, \hat{\rho}_{li})$. By Slutsky's theorem $\hat{p}_{li}^\infty = \lim_{m \to \infty} p(u_1, \hat{A}_0, \hat{\mu}_i, \rho_{li}) \stackrel{d}{=} p(\delta, A_0^\infty, \mu_c^\infty, \rho_{li})$. Property 1 guarantees that $p(\delta, A_0^\infty, \mu_c^\infty, \rho_{li}) \geq p(\delta, A_0, \mu_c^\infty, \rho_{li})$. Property 2, together with the fact that $P_0(Z_{li} \geq z_{li} | \overline{Z}_i/\sigma_{\overline{Z}_i} \geq \tilde{\Phi}^{-1}(u_1), H_{1i})$ is decreasing in μ_i (since the joint density of $(Z_{li}, \overline{Z}_i/\sigma_{\overline{Z}_i} - \mu_i)$ is totally positive of order 2 and therefore $P_0(Z_{li} \geq t | \overline{Z}_i/\sigma_{\overline{Z}_i} - \mu_i \geq x, H_{1i})$ is increasing in x for every fixed t), guarantees that $p(\delta, A_0, \mu_c^\infty, \rho_{li}) \leq p(\delta, A_0, \mu_i, \rho_{li})$. Therefore $\hat{p}_{li}^\infty \geq p_{li}^\infty$. □



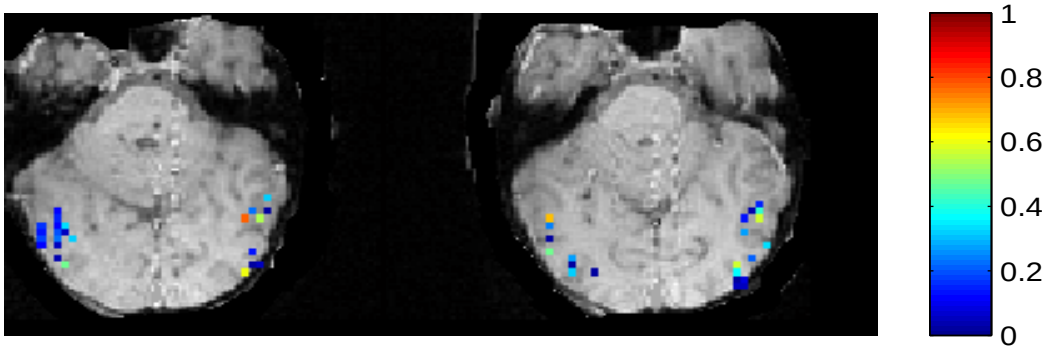


Figure 3: Results of the cluster testing and trimming procedure in brain slices 9-10. Top panel, clusters detected in the testing stage (every cluster in a different color). Bottom panel, estimated location p-values within detected clusters . The largest p-value rejected when trimming at an FDR level of 0.25 and 0.1 was 0.18 and 0.03 respectively.