FALSE DISCOVERY RATE CONTROL FOR SPATIAL DATA

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Alexandra Chouldechova

August 2014

Abstract

In many modern applications the aim of the statistical analysis is to identify 'interesting' or 'differentially behaved' regions from noisy spatial measurements. From a statistical standpoint the task is both to identify a collection of regions which are likely to be non-null, and to associate to this collection a measure of uncertainty. Viewing this task as a large scale multiple testing problem, we present methods for controlling the *clusterwise false discovery rate*, defined as the expected fraction of reported regions that are in truth null. Our methods extend the recent work of Siegmund, Zhang, and Yakir [41], and can be applied whenever the high level excursions of the noise process are well approximated by a (potentially inhomogeneous) Poisson process.

Borrowing ideas from the Poisson clumping heuristic literature, we show that the widely used *pointwise procedure* generally fails to control the clusterwise FDR. We also draw connections between the proposal of Siegmund et al. [41], random field-based familywise error rate control methods, and the STEM procedure introduced by Schwartzman, Gavrilov, and Adler [39].

As one of our extensions we describe a general framework for incorporating various measures of cluster significance into the clusterwise false discovery control procedure. We show that incorporating cluster size can result in a significant increase in power. In particular, we show that the augmented procedure can have better power than even the pointwise procedure, while still controlling the clusterwise false discovery rate.

Acknowledgements

This thesis marks the end of a five-yearlong journey. With deepest gratitude and sincerity, I would like to thank those who have helped me along the way.

First and foremost I would like to thank my advisors, Emmanuel Candes and Rob Tibshirani. I am much indebted to them for the years of patience, guidance and support. It has truly been a privilege.

I would not be where I am today were it not for the early encouragement of my Statistics professors at the University of Toronto. Thank you to Andrey Feuerverger, Nancy Reid and Jeff Rosenthal for sharing with me their enthusiasm for the field. Thank you also to David Brenner, whose introductory class first put Statistics on my radar.

Coming to Stanford gave me the opportunity to interact with and learn from some of the brightest minds and kindest people I have ever known. My first year cohort of Anirban, Matan, Max, Mike, Sumit, and Vec formed an amazing support system throughout the program. I also wish to thank Dennis, Jacob, Nike, Noah, Stefan, Stephen and Will for their friendship and for sharing in countless hours of stimulating conversation.

Sequioa Hall is home to many great educators whose classes have had a formative and lasting impact on my statistical thinking. A special thanks goes to Iain Johnstone, whose ability to impart understanding of even the most difficult concepts is beyond compare. Thank you also to Trevor Hastie and Art Owen, both of whom have had a great influence on how I approach applied problems.

I am deeply appreciative of the time and energy that Brad Efron and David Siegmund put into serving on my thesis committee. Brad and David have had an immense impact on how I approach the spatial inference problem. Many of the ideas in this thesis are directly inspired by their prior research, and I have been extremely fortunate to have them as part of my committee.

I would like to thank my parents, George and Yana, for inspiring in me a sense of intellectual curiosity, and for uprooting their lives to give me opportunities and freedoms that I otherwise would never have. Thank you for believing in me and encouraging me every step of the way.

Lastly, I wish to thank Max for his love, support and constant companionship. The Ph.D. journey may now be over, but ours is just beginning.

Alexandra Chouldechova
Pittsburgh, PA
August 2014

Contents

List of Figures

Chapter 1

Introduction

In many modern applications the aim of the statistical analysis is to identify 'interesting' or 'differentially behaved' regions from noisy spatial measurements. Some examples include (a) copy number variation studies where the goal is to identify contiguous regions of gain or loss from measurements obtained at ordered probe locations; (b) neuroimaging studies seeking to identify activated regions of the brain from measurements recorded at individual voxels; and (c) spatial epidemiology studies, such as those aiming to identify geographical clusters of elevated disease incidence from health information obtained at hospitals/clinics or from survey data. Figure 1.1 shows examples of the kind of data that may be encountered in these application areas.

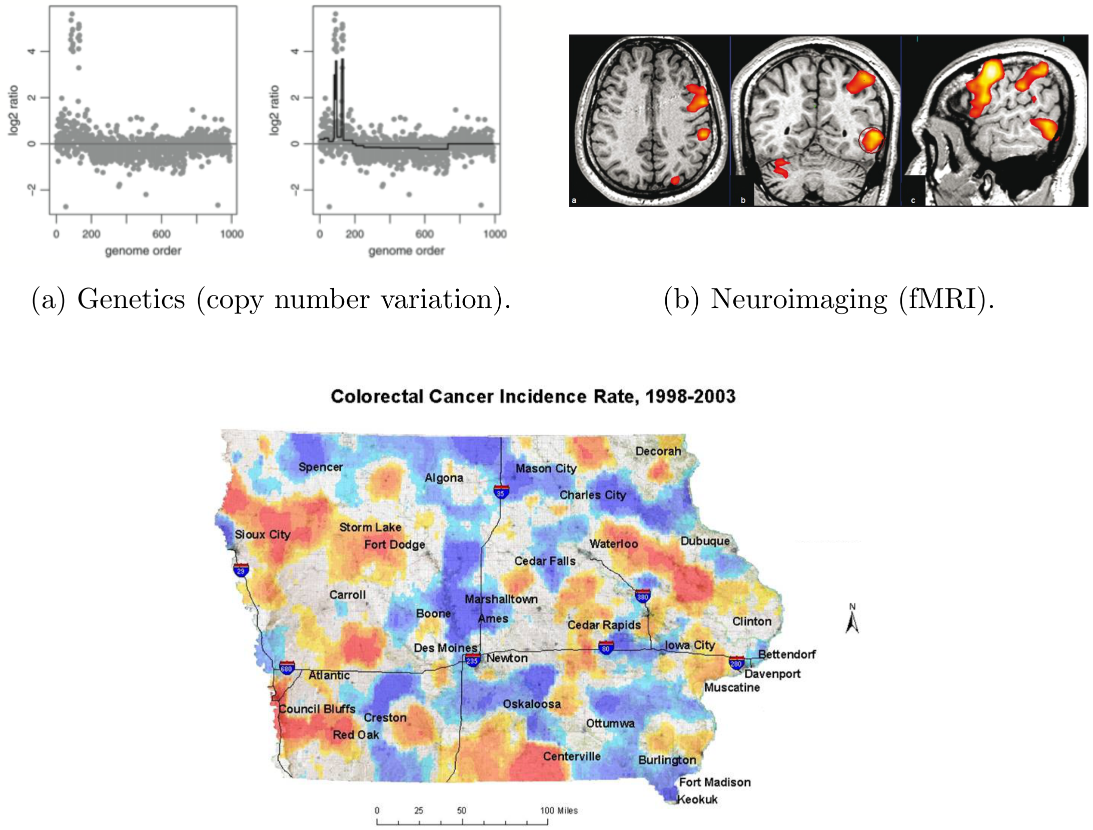
The scientific interest in each of these cases lies in identifying spatial *regions* or *clusters* that deviate from the expected null behaviour. Note that in the settings we consider, the regions of interest are *not* delineated in advance. From a statistical standpoint the task is therefore both to identify a collection of regions which are likely to be non-null, and to associate to this collection a measure of uncertainty. This task is commonly viewed as a large scale multiple testing problem, for which the false discovery rate (FDR) is often the error criterion of choice.

A widely used approach to this type of spatial inference problem is to apply an

FDR controlling procedure (e.g., Benjamini-Hochberg) to p-values calculated at each measurement location. This line of analysis, which we term the *pointwise procedure*, was introduced in the neuroimaging literature through the highly influential paper of Genovese, Lazar, and Nichols [21]. The approach has since been applied in hundreds of studies spanning a broad range of application areas, including but not limited to the three mentioned above. As we will see shortly, there are two main issues with the pointwise procedure: (i) the FDR is being controlled with respect to the wrong signal support; and (ii) given that the units of inference are regions, it isn't clear that the pointwise procedure is even attempting to control a meaningful error criterion.

Chapter outline

We begin in Section 1.1 by laying out the basic notation that will be used throughout the dissertation. In Section 1.2 we give a more precise statement of the pointwise procedure, and discuss the two key issues with this line of analysis. Having discussed the deficiencies of the pointwise procedure, we proceed in Section 1.3 to describe *mosaic processes*, which are widely applicable models for the occurrence of falsely detected regions under a broad range of noise distributions. We then summarize the recent work of Siegmund, Zhang, and Yakir [41], in which the authors present a *clusterwise* estimation and control procedure that can be applied when false discoveries are well modeled by a mosaic process. The introduction concludes with a review of relevant literature.



(a) Genetics (copy number variation).

(b) Neuroimaging (fMRI).

(c) Spatial epidemiology.

Figure 1.1: This Figure shows examples of the kinds of applications in which the spatial inference problem arises. Figure (a) originally appeared in Tibshirani and Wang [46], in which the authors present a fused lasso approach to detecting hot spots in CGH data. Figure (b) comes from Beyer and Rushton [12], and Figure (c) is taken from a review paper by Vaghela et al. [47].

1.1 Statement of the problem

We begin here by giving a more formal statement of the spatial inference problem. We also summarize some of the notation and key definitions that will be used throughout this dissertation.

We assume that we observe data $y(t)$ corresponding to a noisy version of a sparse signal $\mu(t)$ defined on an observation region D. The noise process will be denoted by $\epsilon(t)$, and it will often be convenient to think of the noise as being additive, in the sense that $y(t) = \mu(t) + \epsilon(t)$. While the additivity assumption is not necessary, it is satisfied in a broad range of settings. The region D is assumed to be a subset of Euclidean space, $\mathbb{R}^d$, and can be either discrete or continuous. When $d = 1$, we will typically denote $D = \{1, 2, \ldots, T\}$ or $D = [0, T]$.

The region D will be thought of as being partitioned into two disjoint sets, $D = D_0 \cup D_1$. D_0 is the set of 'null locations', which are locations that do not contain signal ($D_0 = \{t \in D : \mu(t) = 0\}$). D_1 is the set of 'non-null locations', which is the support of the signal $\mu(t)$ ($D_1 = \{t \in D : \mu(t) \neq 0\}$). For example, in the leftmost panels of Figure 1.4, D_1 is the union of the five shaded squares, and D_0 is all the white space. It will generally be assumed that $\mu(t) \geq 0 \; \forall t \in D$, and hence that we are interested in identifying regions where the signal is positive. The case where $\mu(t)$ can be both positive and negative can often be reduced to the positive case by taking absolute values, squaring or similar operations. We revisit this point in Section 1.5.2.

We define a cluster $C \subset D$ to be a *false discovery* or *false rejection* if $C \subset D_0$. A cluster is a *true discovery* or *correct rejection* if $C \cap D_1 \neq \emptyset$. This is a rather weak notion of true discovery as it only asks that C overlaps with the support D_1, and does not require that the overlap exceed some minimal size. As we discuss in Section 1.7, criteria requiring minimal overlap do appear in several other places in the literature.

When discussing pointwise error rates, we will adopt the natural definition that a location $t \in D$ is a *false discovery* if $t \in D_0$. Correspondingly, a location $t \in D$ is a *true discovery* if $t \in D_1$.

In studying the false discovery properties of a discovery procedure, we will use the following notation.

Key notation.

$$\begin{aligned}
\mathcal{V} &= \text{ The set of false rejections} \\
\mathcal{S} &= \text{ The set of correct rejections} \\
V &= \#\text{ of false rejections} = |\mathcal{V}| \\
S &= \#\text{ of correct rejections} = |\mathcal{S}| \\
R &= \text{total } \#\text{ of rejections} = V + S \\
\text{FDP} &= V/R = \text{False discovery proportion (FDP)} \\
\text{FDR} &= \mathbb{E}[V/R; R > 0] = \text{False discovery rate (FDR)}
\end{aligned}$$

This notation applies whether a discovery is considered to be a single location $t \in D$ or a cluster $C \subset D$. For the most part we will be referring to clusterwise quantities. In general it should be clear from the context whether the notation is being used to refer to pointwise or clusterwise quantities. When there is potential for confusion, subscripts C and P will be used to refer to the clusterwise and pointwise quantities, respectively.

1.2 Standard approach: Pointwise procedure

We begin by giving an overview of the *pointwise procedure*, which can be summarized as follows.

Pointwise procedure.

(i) At each location t, apply a local smoother to obtain $\tilde{y}(t)$ based on $y(t)$ and $\{y(s)\}$ for s in a neighbourhood of t.

(ii) For each t, calculate a p-value $p_t = \mathbb{P}(\tilde{\epsilon}(0) \geq \tilde{y}(t))$.[1]

(iii) Apply an FDR control procedure (e.g., Benjamini-Hochberg) to the set of p-values $\{p_t : t \in D\}$ to obtain a cutoff z_α that controls the FDR at level α.

(iv) Report as discoveries the spatial clusters of the set $\mathcal{E}_{z_\alpha} = \{t : \tilde{y}(t) > z_\alpha\}$

[1] Here $\tilde{\epsilon}(0)$ is the marginal distribution of the smoothed noise process. We are thinking of $\tilde{y}(t)$ as fixed, and calculating the probability that the random variable $\tilde{\epsilon}(0)$ exceeds level $\tilde{y}(t)$.

As previously mentioned, there are two key issues with this line of analysis: (i) the FDR is being controlled with respect to the wrong signal support; and (ii) given that the units of inference are regions, it isn't clear that the pointwise procedure is even attempting to control a relevant error criterion. We now elaborate on both of these points.

Failure to control pointwise FDR. For concreteness, consider a setting in which the observation $y(t) = \mu(t) + \epsilon(t)$ is smoothed with a linear filter to form $\tilde{y}(t) = \tilde{\mu}(t) + \tilde{\epsilon}(t)$. Even if the original noise $\epsilon(t)$ is uncorrelated, the smoothed noise $\tilde{\epsilon}(t)$ will be (positively) auto-correlated. This implies that the test statistics $\{\tilde{y}(t)\}_{t \in D}$ will generally be positively dependent, and thus some care should be taken to ensure that the FDR control procedure being applied is robust to positive dependence.

In the case of Benjamini-Hochberg, Benjamini and Yekutieli [11] establish that the BH procedure conservatively controls the FDR provided that the test statistics satisfy

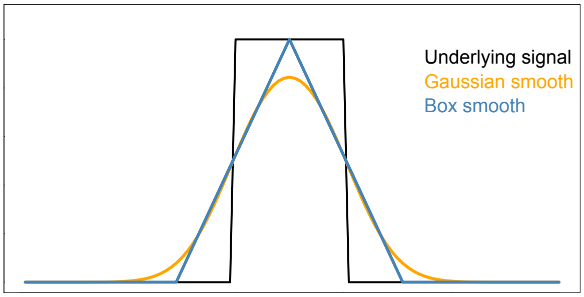
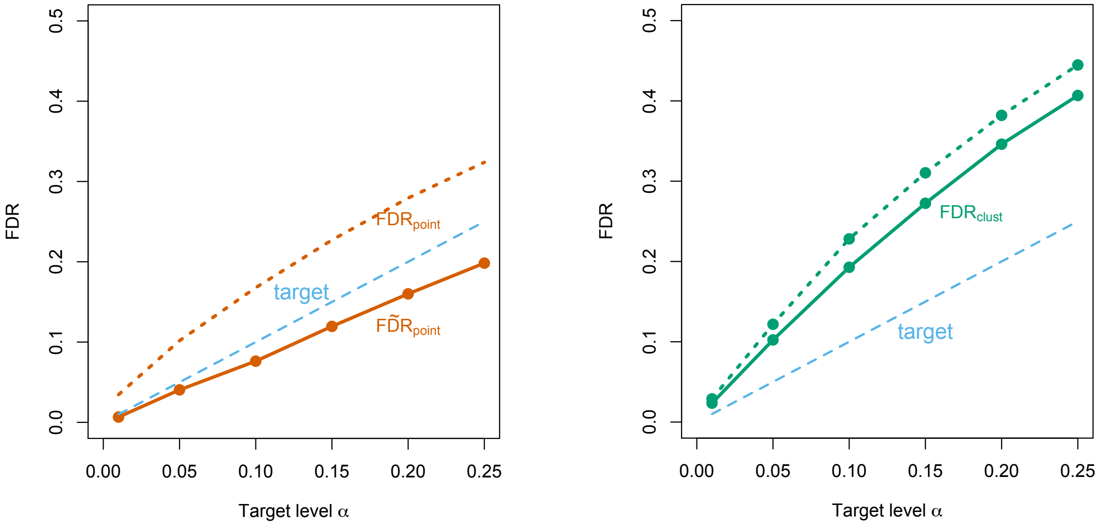


Figure 1.2: Illustration of the effect of smoothing on the support of the signal. The black curve shows the original box function signal. The two colored curves show the kernel-smoothed signal for two choices of kernel, both of which have considerably larger support compared to the original box function.

a certain positive dependence condition called PRDS. This condition is satisfied if, for instance, $\tilde{\epsilon}(t)$ is multivariate Gaussian with non-negative correlation. Even if the PRDS condition cannot be verified in a particular problem instance, the authors show that a modification of the BH procedure controls the FDR under arbitrary dependence. This modification is to carry out the BH procedure with α replaced with the smaller quantity, $\alpha/(\sum_{i=1}^{m}\frac{1}{i})$, where m is the total number of tests. Since m is large in all the settings we consider, one can simply take $\alpha/\log(m)$.

The real issue, however, is that smoothing 'smears' the original signal, meaning that even though the control procedure may be valid, it is conducted with respect to the smoothed signal $\tilde{\mu}(t)$. This phenomenon is illustrated in Figure 1.2, which shows the effect of applying a kernel smoother to a box function. Smoothing has the effect of enlarging the support of the signal. Thus when the BH procedure is applied to p-values derived from $\{\tilde{y}(t)\}$, any significant location in $\text{supp}(\tilde{\mu}) \supset \text{supp}(\mu)$ gets treated as a true detection. In other words, the BH procedure gives pointwise FDR control for detecting locations in $\text{supp}(\tilde{\mu})$, but not for the more restricted target set, $\text{supp}(\mu)$.

The left panel of Figure 1.3 shows the results of a small simulation study conducted to investigate how close the BH(α) procedure comes to controlling the pointwise FDR with respect to $\text{supp}(\mu)$. Details of the simulation setup are described in the Figure caption. The observed pointwise FDR measured with respect to $\text{supp}(\tilde{\mu})$ is also shown for reference. As dictated by the theory, the observed FDR curve with respect to supp($\tilde{\mu}$) is lower than the target for all values of α. Despite this, however, we see that for all values of the target FDR level, α, the observed pointwise false discovery rate is considerably higher than the target.



(a) Pointwise FDR. The solid curve shows observed pointwise FDR measured with respect to supp($\tilde{\mu}$). Dashed curve shows observed pointwise FDR measured with respect to supp(μ). The pointwise procedure fails to provide FDR control for detecting the support of $\mu(t)$.

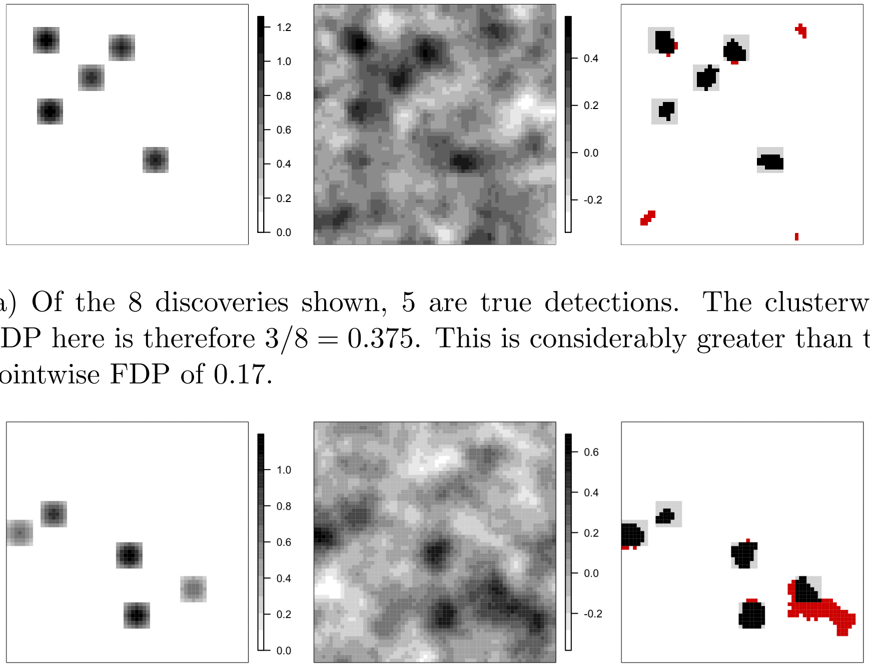
(b) Clusterwise FDR. The solid curve shows the observed clusterwise FDR for the BH(α) procedure. Observed clusterwise FDR is nearly two times higher than the target, indicating the the pointwise procedure fails to control clusterwise FDR. The still higher dashed curve shows clusterwise FDR when each box function in the support of $\mu(t)$ has a 5% chance of having width $5w$. The pointwise procedure becomes even more anti-conservative in the latter setting.

Figure 1.3: These plots present the results of a small simulation conducted to investigate the pointwise and clusterwise FDR control properties of the pointwise procedure. The underlying signal, $\mu(t)$ is a sparse train of 10 box functions of equal amplitude, and the noise is generated iid Gaussian. Except as indicated in the caption of plot (b), the signal regions all have equal width, w. A box kernel smoother with bandwidth w is applied to the data to produce $\tilde{y}(t)$. Both panels show the results of applying the pointwise procedure with BH(α) at the value of α indicated on the horizontal axis.

Failure to control clusterwise FDR. While the failure to control pointwise FDR with respect to supp(μ) is problematic, our focus is on a different deficiency of the pointwise procedure. The main issue stems from the fact that the scientific interest in the problems we consider is in identifying and making inference on spatial *regions*. Despite the widespread use of the pointwise procedure, the connection between pointwise and clusterwise false discovery control remains largely uninvestigated. Notably, a recent paper of Chumbley and Friston [15] discussing this issue has received considerable attention within the neuroimaging community.

As Figure 1.4 illustrates, there can be a large discrepancy between pointwise and clusterwise false discovery proportions. In general, pointwise FDR control fails to provide any assurance that the expected proportion of falsely discovered clusters is similarly controlled. Figure 1.3(b) shows the observed clusterwise FDR of the pointwise procedure applied using BH(α). We see that the observed clusterwise FDR can be more than twice the target level, and that performance further degrades in the presence of a small proportion of large signal regions. Using machinery introduced in the next section, we can show that the pointwise procedure is in general anti-conservative. We present a mathematical characterization of the discrepancy between pointwise and clusterwise inference in Section 2.1.

Summary. The main takeaway of the present discussion is that the pointwise procedure is simply counting the wrong thing. When the scientific interest is in identifying differentially behaved regions, a pointwise statistical analysis is inappropriate and can even give misleading results. This issue is already receiving attention from the broader scientific community. Lastly, even if one is content with measuring error pointwise, the standard procedure ensures FDR control only with respect to the support of the smoothed signal.



(a) Of the 8 discoveries shown, 5 are true detections. The clusterwise FDP here is therefore $3/8 = 0.375$. This is considerably greater than the pointwise FDP of 0.17.

(b) All 5 discoveries overlap the support of the signal, so the clusterwise FDP is zero. However, due to the large number of location-wise false detections in the bottom-right cluster, the pointwise FDP is very large (0.4).

Figure 1.4: The examples shown here illustrate that there isn't a clear correspondence between clusterwise and pointwise false discovery proportions. Shown here are 2-dimensional examples where a region with five clusters is corrupted with Gaussian white noise. Center panels show data smoothed via local averaging. Right panels show locations from center panel whose values exceeded a certain cutoff. Black: correct detection; Red: false detection; Gray: false non-detection.

1.3 Poisson clumping heuristic and mosaic processes

We argued in the previous section that the main problem with the pointwise procedure is that it fails to correctly account for the occurrence of falsely detected clusters. In this section we proceed to describe a widely applicable model for the occurrence of clusters of the excursion set $\mathcal{E}_z = \{t : \tilde{\epsilon}(t) \geq z\}$ at high thresholds z. Characterizing these clusters provides us with a model for V_z, the number of falsely detected clusters at level z, under the global null. The key idea can be summarized as follows.

At high levels z, clusters of the excursion set $\mathcal{E}_z$ are well modeled as random sets centered at points of a Poisson process.

This idea forms the basis of the Poisson clumping heuristic (PCH), which is an effective and broadly applicable method for approximating probabilities of rare events (such as high-level excursions) associated with random sequences, processes, and fields. Aldous [2] gives an extensive overview of many of the most interesting cases in which the PCH applies. For our purposes, we will rely heavily on the following characterization from the PCH literature: Clusters of $\mathcal{E}_z$ are well approximated by *mosaic processes*[1].

Definition (Mosaic process)**.** Let F be a distribution on sets in $\mathbb{R}^d$. Think of F as generating small sets located near the origin 0. A *mosaic* or *mosaic processes* is described by the following procedure

1. Generate points $x_1, x_2, \ldots$ according to a Poisson process with rate λ on $\mathbb{R}^d$.

2. Generate random sets $A_1, A_2, \ldots$ iid from F

[1] Also called *mosaics* for short.

3. Output the random set
$$A = \bigcup_i x_i \oplus A_i$$
which is the union of the sets A_i shifted to be centred at the points x_i.[1]

[1] Given a point y and a set B, $y \oplus B = \{y + b : b \in B\}$ is the translation of the set B by y

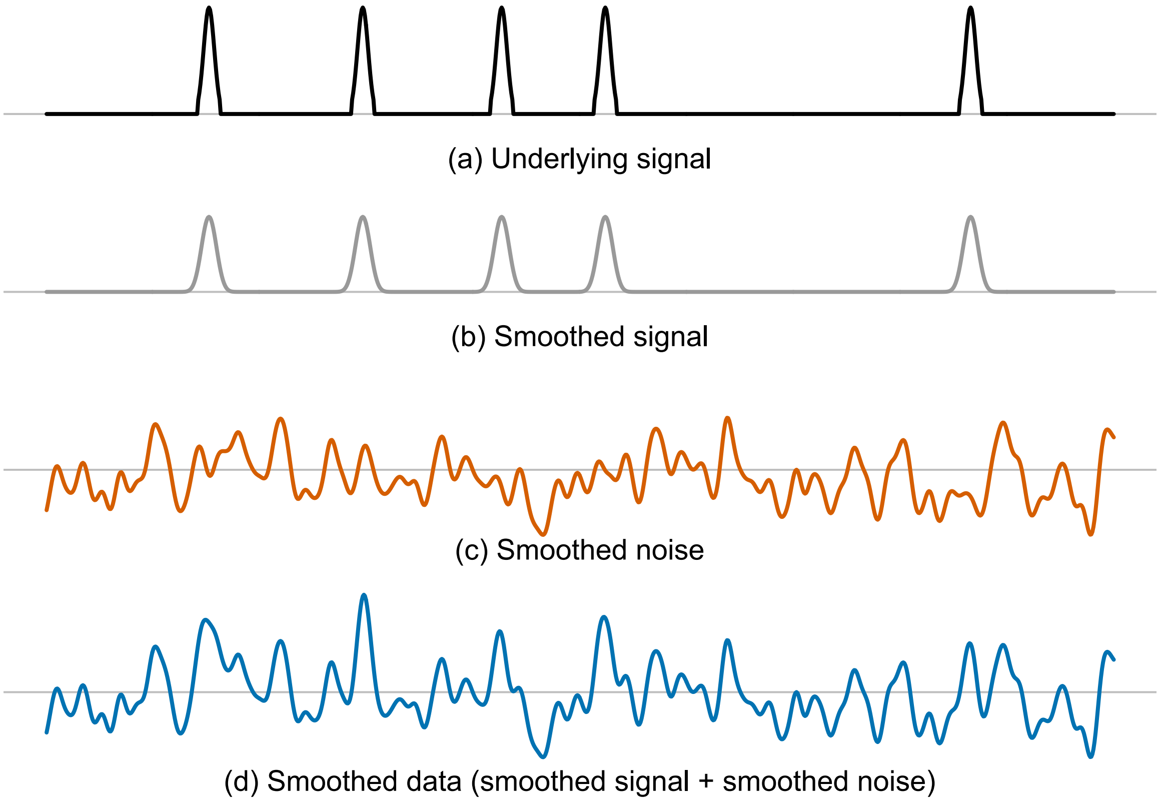
When the mosaic approximation holds we therefore have that $V_z \sim \text{Poisson}(\lambda_z)$, for some mean parameter λ_z.[2] Moreover, we may also be able to incorporate into our analysis other properties of the clusters (e.g., size) by understanding F. We further develop this idea in Section 3.3. As a starting point, we will show that we are able to get a lot of mileage just out of the mean parameter λ_z. We defer our discussion of when the mosaic approximation applies and how to calculate or estimate λ_z until Section 1.6. In the interim, we hope the reader is encouraged by the following excerpt.

"It turns out that [the] 'sparse mosaic limit' behavior for rare events is as ubiquitous as the Normal limit for sums; essentially, it requires only some condition of 'no long range dependence'. " pp. 6, Aldous [2]

1.4 Presence of signal

The mosaic process model characterizes the excursion set of the smoothed noise process $\tilde{\epsilon}(t)$. When there is no signal present, this is the same as the smoothed observation $\tilde{y}(t)$. In the presence of signal, however, the smoothed observation $\tilde{y}(t)$ is comprised of the smooth signal $\tilde{\mu}(t)$ and the smoothed noise $\tilde{\epsilon}(t)$. Since the main application of the current work is in cases where multiple signal regions are expected to exist, it is important to understand how the occurrence of null excursion sets is affected by the presence of signal.

[2] It will be more convenient to parameterize the Poisson in terms of a mean instead of a rate, so unless stated otherwise we take $\lambda = \lambda(D)$ to refer to the Poisson *mean* parameter.



(a) Underlying signal

(b) Smoothed signal

(c) Smoothed noise

(d) Smoothed data (smoothed signal + smoothed noise)

Figure 1.5: This figure illustrates the individual components comprising the observed smoothed data, $\tilde{y}(t)$. In this example the underlying signal, $\mu(t)$ consists of there are 5 regions, which are the supports of the bumps in plot (a). Plot (b) shows the smoothed data $\tilde{\mu}(t)$, which is obtained by applying a gaussian kernel smoother to the signal $\mu(t)$. Plot (c) shows the smoothed noise process $\tilde{\epsilon}(t)$. The noise in this instance is generated iid Gaussian. Lastly, plot (d) shows the observed smooth data for this problem instance, which is the sum of the smoothed signal and smoothed noise components: $\tilde{y}(t) = \tilde{\mu}(t) + \tilde{\epsilon}(t)$.

Consider a simple additive model in which we observe $y(t) = \mu(t) + \epsilon(t)$, with ϵ assumed to be stationary. Suppose that the observed data is linearly smoothed using a compactly supported kernel, and that the mosaic process approximation holds for high-level excursions of the smoothed noise $\tilde{\epsilon}(t)$.

Figure 1.5 shows the smoothed signal and smoothed noise processes that comprise the observed process $\tilde{y}(t) = \tilde{\mu}(t) + \tilde{\epsilon}(t)$. Given a threshold $z > 0$, the excursion set $\mathcal{E}_z$ can be decomposed into three types: (i) the true detections, which are intervals that intersect D_1, the support of the underlying signal; (ii) the borderline detections, which are intervals that intersect the support of $\tilde{\mu}(t)$, which we will denote by $\tilde{D}_1$, but not the support of $\mu(t)$ itself; and (iii) intervals that do not intersect the support $\tilde{\mu}(t)$.

Assuming that the underlying signal and smoother are sufficiently well behaved, components of type (ii) are highly unlikely occur. Thus to understand the false discovery process we really need only consider components of type (iii). Type (iii) components are simply components of the excursion set of the smoothed noise process $\tilde{\epsilon}(t)$. In other words, in the presence of signal, the occurrence of false detections is well modeled by a mosaic process on the complement of $\tilde{D}_1$.

This observation translates into a simple calculation of the mean parameter λ_z in the presence of signal. Define $\tilde{\pi}_0 = 1 - |\tilde{D}_1|/|D|$ to be the fraction of the observation region that does not contain smoothed signal. Let λ_z^ϵ denote the expected number of clusters comprising the excursion set of $\tilde{\epsilon}(t)$. The preceding argument implies that in the presence of signal, the expected number of false detections is approximately given by $\lambda_z = \tilde{\pi}_0 \lambda_z^\epsilon$. That is, in the presence of signal, we expect that $V \sim \text{Poisson}(\tilde{\pi}_0 \lambda_z^\epsilon)$.

Remark. Just as in the standard multiple testing setting, it may be advantageous to estimate the quantity π_0 and to incorporate it into the estimation and control procedures. Note that since π_0 is not a clusterwise quantity (i.e., it is the same regardless of whether inference is conducted pointwise or clusterwise), the standard estimators can still be used. A simple approach might be to use the now-standard estimator proposed in Storey [42], or the empirical Bayes upper bound described in Efron et al. [17].

1.5 Proposal of Siegmund, Zhang and Yakir

Now that we have a good model for the occurrence of false discoveries, we can describe a clusterwise FDR estimation and control procedure that relies on the Poisson distribution of V. This procedure is due to a recent paper of Siegmund, Zhang, and Yakir [41] (SZY), in which the authors pursue the line of investigation that we advocated above. Their construction can be thought of as a parallel to Storey et al. [43] for the case where $V \sim \text{Poisson}(\lambda)$. We devote this section to describing their proposal. The remainder of the thesis largely takes up the task of formalizing and extending the proposal of SZY.

As in the pointwise procedure, we begin with data $y(t)$, and at each location t evaluate a statistic $\tilde{y}(t)$ based on $y(t)$ and $y(s)$ for s in a neighbourhood of t. Just as before, given a cutoff z, the clusters that comprise the excursion set $\mathcal{E}_z = \{t : \tilde{y}(t) > z\}$ are reported as the discoveries (see Figure 1.6). The departure from the pointwise procedure comes at the inference stage, where instead of treating the $\tilde{y}(t)$ as unordered test statistics and relying on a pointwise FDR procedure to select the cutoff z or to estimate the FDR, the FDR is estimated directly in a clusterwise manner.

The SZY approach to clusterwise inference can be stated fairly simply. From the mosaic approximation we expect that the excursion set $\mathcal{E}_z$ on average contains λ_z null clusters. Thus if for a given realization we make R discoveries (i.e., observe R clusters in $\mathcal{E}_z$), it is reasonable to estimate that λ_z of the R are due to noise, and hence estimate the FDR by λ/R. It turns out that this estimator is biased, but a simple modification of it, where the denominator is replaced by $R + 1$, works out to be unbiased.

Having laid out some intuition, we now provide the details as given in [41]. The validity of the clusterwise FDR control and estimation procedure rests on the following two assumptions.

(1) The number of false discoveries, V, is distributed $V \sim \text{Poisson}(\lambda_z)$

(2) The number of correct discoveries, S, is independent of V

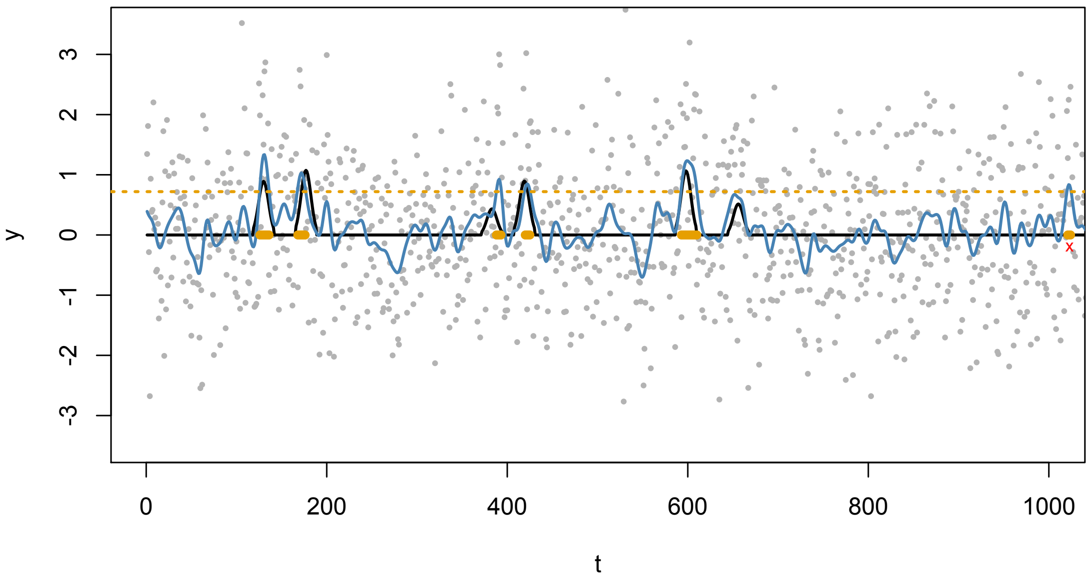


Figure 1.6: Underlying signal $\mu(t)$ is shown in black. D_1, the support of μ, is the union of the six intervals on which $\mu > 0$. Solid grey points are the observed data. The solid blue curve is $\tilde{y}(t)$, obtained by gaussian kernel smoothing the observed data. Dashed horizontal orange line is the cutoff z. Six intervals comprising the excursion set $\mathcal{E}_z = \{t : \tilde{y}(t) > z\}$ are shown in orange on the x-axis. The last of these is a false discovery, while the remaining five are true discoveries. There is a false negative around location 650, where the signal region fails to be detected by the procedure.

We have already provided justification for assumption (1) by invoking the mosaic process approximation and Poisson clumping heuristic. The second assumption will at least approximately be satisfied when the dependence in $\tilde{y}(t)$ is sufficiently short range. We begin by giving the estimation procedure.

Theorem (Estimation procedure. SZY Theorem 1)**.** *Under assumptions (1) and (2), the estimator*

$$\widehat{\mathrm{FDR}} = \frac{\lambda}{R+1}$$

is an unbiased estimator of the clusterwise FDR, in the sense that

$$E(\widehat{\mathrm{FDR}}) = \mathbb{E}(V/R; R > 0).$$

This estimator also appears in a slightly different context in Efron et al. [17].

Controlling the FDR in this setting entails selecting a value of the cutoff z, thought of as a data-dependent random variable, in such a way that $\mathbb{E}(V(z)/R(z)) < \alpha$ for the desired FDR level α. Here we are thinking of V and R as quantities that depend on z. As z varies, so does $\lambda = \lambda(z)$, so we can look at V_λ and R_λ as quantities depending on λ. Define

$$\Lambda = \max\{\lambda \leq \bar{\lambda} : R_\lambda \geq \lambda/\alpha\}$$

and let z_Λ be the corresponding cutoff.[3] The procedure that reports the R_Λ clusters comprising the excursion set E_{z_Λ} controls the FDR at level α.

Theorem (Control procedure. SZY Theorem 2)**.** *Suppose that V_λ is a Poisson process of rate 1 on $[0, \bar{\lambda}]$, and V_λ is independent of the process S_λ. Then, $\mathbb{E}(V_\Lambda/R_\Lambda) \leq \alpha$.*

We will frequently refer back to these procedures throughout the thesis.

1.5.1 A note on the independence assumption

In Chapter 3 we will describe several extensions of the base estimation and control procedures. Each extension will require that the corresponding false discovery process and the true discovery process are independent. While independence of S and V will not in itself be sufficient to establish independence of the quantities we later consider, the underlying argument is the same. We pause here to summarize the argument.

First, note that V is strictly a function of $\tilde{y}(t)$ for $y(t) \in D_0$, and S is a function of $\tilde{y}(t)$ for $y(t) \in \tilde{D}_1$. Assuming $\epsilon(t)$ does not exhibit long range dependence and the

[3] The mosaic approximation is only valid for values of z that are sufficiently large so as to ensure that the excursion sets of the smoothed noise process is well modelled by a mosaic, defined in §1.3. Since λ is a decreasing function of z, the restriction on z being sufficiently large translates into the restriction that λ be sufficiently small.

set $\tilde{D}_1$ is sparse, $\{\tilde{y}(t)\}_{t\in D_0}$ and $\{\tilde{y}(t)\}_{t\in\tilde{D}_1}$ will be largely independent. Consequently, V and S would be approximately independent as well.

We will refer back to this 'no long range dependence' argument throughout Chapter 3.

1.5.2 Detection of non-zero signal locations

Having introduced the key assumptions underlying the clusterwise FDR procedure, we can now revisit our earlier statement that the case where $\mu(t)$ is allowed to be both positive and negative can often be reduced to the positive case by taking absolute values, squaring or similar operations. The key observation is that the FDR procedures do not presuppose that the underlying model is additive, nor that the noise process is mean-0. The main requirement is that the excursion set of the noise process $f(\tilde{\epsilon}(t))$ is well approximated by a mosaic, where $f : \mathbb{R} \to \mathbb{R}_+$ is the positivity inducing function being used.

For concreteness, consider the case where $\tilde{y}(t) = \tilde{\mu}(t) + \tilde{\epsilon}(t)$ and $\tilde{\epsilon}(t)$ is stationary Gaussian noise. If we allow for the possibility that some non-zero components of $\mu(t)$ may be negative, we can consider basing our inference on locations where $\tilde{y}^2(t)$ is large. Under the null, $\tilde{y}^2(t) = \tilde{\epsilon}^2(t)$ is a χ^2 process, the high level excursions of which are well understood. The necessary Poisson approximation for high levels excursions is well established for χ^2 processes [7], and so the clusterwise FDR procedures continue to apply.

1.6 Poisson approximation and obtaining λ

There is an extensive literature describing conditions under which high level excursions of discrete and continuous processes are well modeled as occurring at points of a Poisson process [28, 2, 9, 31]. We begin this section by presenting a few of the settings in which the mosaic process approximation holds and for which good analytic approximations for the parameter λ_z are known. We then briefly discuss simulation-based approaches to estimating λ_z.

1.6.1 Moving averages of iid sequences

The first two results presented here appear in Aldous [2, §C4-C9]. We assume that $\{\epsilon(i)\}$ is an iid sequence and $\{c_i\}$ are constants such that the moving average $\tilde{\epsilon}(t) = \sum_{i=0}^{\infty} c_i \epsilon(t-i)$ is a stationary process.

Exponential tails. Suppose that for large z, $\mathbb{P}(\epsilon(t) > z) \sim A_2 e^{-az}$, and the c_i are such that $\mathbb{P}(\tilde{\epsilon}(t) > z) \sim A_1 e^{-az}$. Then,

$$\lambda_z \approx T A_1 e^{-az}.$$

Polynomial tails. Suppose that for large z, $\mathbb{P}(\epsilon(t) > z) = \mathbb{P}(\epsilon(t) < -z) \sim A z^{-\alpha}$. Then excursions of $\tilde{\epsilon}(t)$ above a high threshold z are generally due to a single large value of $\epsilon(t)$, and letting $c = \max c_i$ we have that,

$$\lambda_z \approx T A (c/b)^{\alpha}.$$

Gaussian. This next result is due to [40]. Here we suppose that the noise sequence $\{\epsilon(t)\}$ is iid Gaussian with variance 1, and

$$\tilde{\epsilon}(t) = \sum_{i=1}^{w} \frac{1}{\sqrt{w}} \epsilon(t+i-1).$$

Then,

$$\lambda_z \approx Tzw^{-1}\phi(z)\nu(z\sqrt{2/w}),$$

where $\nu(x) = (2/x)(\Phi(x/2) - 0.5)/[(x/2)\Phi(x/2) + \phi(x/2)]$.

1.6.2 Smooth Gaussian processes

The results here are borrowed from Aldous [2, §C23] and Lindgren [31, §8.1]. We assume that $\tilde{\epsilon}(t)$ is a mean-0 stationary differentiable Gaussian process with covariance function $r(t) \equiv \text{cov}(\tilde{\epsilon}(0), \tilde{\epsilon}(t))$. Supposing that $\tilde{\epsilon}(t)$ has variance $r(0) = \sigma^2$ setting $\omega_2 = r''(t) \equiv \text{var}(\tilde{\epsilon}'(t))$, Rice's formula gives that the expected number of upcrossings of level z by $\tilde{\epsilon}(t)$ is,

$$\rho_z = T\sqrt{\frac{\omega_2}{2\pi}}\phi(z/\sigma).$$

For high threshold levels z, upcrossings are rare isolated events, each of which can be associated with a cluster of the excursion set $\mathcal{E}_z$. Thus for large z we have that $\lambda_z \approx \rho_z$.

1.6.3 Simulation-based approaches

When analytic formulas are not available but the distribution of $\epsilon(t)$ is either known or can be well estimated, one can estimate λ_z via simulation. It would suffice to generate repeated realizations of the noise process $\epsilon(t)$, apply the smoother, and count the number of clusters in the excursion set $\mathcal{E}_z$ for a range of z values. When taking this approach it is good practice to also check that the number of clusters at high thresholds is indeed Poisson distributed.

There remain settings in which the noise distribution is both unknown and difficult to estimate. In such cases, permutation-based approaches might work, but they must be applied with care. For instance, if we know that $\epsilon(t)$ is iid, then under the global null the distribution of the process $\tilde{y}(t)$ is invariant under permutations of locations t. In the iid case is it also valid to conduct permutation inference even if there is signal present. Doing so will lead to an over-estimate of λ_z, and thus can only make the FDR procedures more conservative. However, if the noise $\epsilon(t)$ exhibits spatial dependence, then the null distribution of $\{\tilde{y}(t)\}_{t \in D}$ is no longer exchangeable in t, and permuting locations would lead to invalid inference.

Taking a step back, it is worth recalling that in many of the applications in which the spatial inference problem arises, the observed $y(t)$ corresponds to a difference between two groups calculated at location t. Unlike the locations t, the group labels may be exchangeable under the null. This allows one to estimate λ_z by repeatedly permuting the group labels and recalculating the process $y(t)$.

More precisely, suppose that we begin with data $x_1(t), \ldots, x_m(t), x_{m+1}(t), \ldots, x_{m+n}(t)$ observed for $t \in D$, where,

$$x_i(t) = \begin{cases} \mu_1(t) + \epsilon_i(t) & i = 1, \ldots, m \\ \mu_2(t) + \epsilon_i(t) & i = m+1, \ldots, m+n \end{cases},$$

and the ϵ_i are iid realizations of a mean-0 noise process (iid in i, but not necessarily in t). We'll suppose that $\mu_2(t) \geq \mu_1(t)\ \forall t$, and that we are interested in identifying locations where $\mu_2(t) > \mu_1(t)$. Indexes $\{1, \ldots, m\}$ can be thought of as coming from baseline or control measurements, while indexes $\{m+1, \ldots, m+n\}$ correspond to measurements in a stimulated or treated condition.

Let $Y : \mathbb{R}^{m+n} \to \mathbb{R}$ be a test statistic for testing for a mean difference between the two groups at location t (e.g., a 2-sample t-test). Setting,

$$y(t) = Y(x_1(t), \ldots, x_{m+n}(t)),$$

we can view the problem as one of identifying regions where $\mathbb{E}(y) = \mu_2 - \mu_1$ is positive. Since the ϵ_i are assumed to be iid, the group labels (equivalently, indexes) are exchangeable under the global null. We can therefore obtain a permutation null distribution for the process $y(t)$ by repeatedly permuting the indexes and recalculating the test statistics. From there we can estimate λ_z by counting the number of clusters observed in the smoothed process $\tilde{y}(t)$ for each permutation of the indexes.

1.7 Literature review

We conclude our introduction with an overview of some of the other existing literature related to the spatial inference problem. To better facilitate comparisons between existing work and the contributions of this thesis, we partition our review according to how the existing work differs from ours in terms of key goals or operating assumptions.

Regions of interest. In our work we *do not* assume that the clusters or regions of interest are known in advance, nor are the regions assumed to be constructed on an independent set of experimental data. Our goal is to present a method for identifying a collection of differentially behaved regions and for associating to this collection a measure of statistical uncertainty.

Several approaches have been proposed in the setting where the possible regions of interest are defined in advance, either independently of any data or based on an independent experiment. In, Yekutieli [53] the authors present a hierarchical FDR method motivated by a QTL application. The observation space is, in advance, partitioned along a tree with lower levels of the tree corresponding to finer partitions of the space. The proposed method provides FDR control both overall and within a given level of the tree. Heller et al. [24] propose an approach in the context of a brain imaging study in which a preliminary scan is used to select clusters by grouping highly correlated nearby voxels that are highly correlated. The recent work of Sun et al. [44] is also particularly relevant.

Note also that if we are simply interested in testing m pre-defined, disjoint regions, $A_1, A_2, \ldots, A_m$, then we are essentially back in the standard multiple testing setting. This problem reduces to forming test statistics for testing the hypotheses $\{H_i : \mu(t) = 0 \ \forall t \in A_i\}_{i=1}^m$, and then applying an appropriate FDR controlling procedure to the corresponding set of p-values.

Error criterion. The focus of this thesis is false discovery rate control. While family wise error rate controlling procedures have existed in the spatial inference setting for several decades, the interest in the false discovery rate control problem is fairly recent. Historically, many spatial FWER control methods were primarily motivated by increased interest in analyzing high resolution brain imaging data (see Nichols and Hayasaka [32] for a comparative review).

The familywise error rate of a multiple testing procedure is defined as the probability of making any false rejections (i.e., $\mathbb{P}(V > 0)$). Note that since $V_P = 0 \Rightarrow V_C = 0$, a procedure that gives pointwise FWER control will *a fortiori* also control the FWER clusterwise.[4] Thus standard FWER controlling procedures (Bonferonni, step-up/step-down test, etc.) can be—and are—applied in the spatial inference setting. However, because such procedures do not incorporate spatial information, they can be conservative and under-powered.

More powerful procedures for controlling the familywise error rate focus on the distribution of the *maximal statistic*, $M = \sup_{t \in D} \tilde{\epsilon}(t)$. The random field theory (RFT) approach pioneered in Worsley et al. [51] approximates the tail distribution of M under assumptions on the smoothness and distribution of the smoothed noise process, $\tilde{\epsilon}(t)$. There exists a rich body of literature surrounding the RFT approach, and it remains an important and active area of research [52, 45, 35, 33]. Permutation and resampling approaches such as those proposed in Nichols and Holmes [34] are also widely used. In addition to methods that assess significance based on peak height, there are several proposals in the fMRI and related literature that look instead at cluster size, or a combination of peak height and cluster size [15, 38, 23, 22, 54].

FWER controlling procedures all share one important drawback: They are do not adapt to the signal. Strong evidence of true discoveries has no effect on the FWER significance threshold. In contrast, the false discovery rate is adaptive in this sense. As we show in Section 2.2, adaptivity can result in large gains in power while still controlling a scientifically relevant error criterion.

Notion of true discovery. We define a cluster to be a true discovery if it has any overlap with the support of the signal: $C \cap D_1 \neq \emptyset$. Another common convention in the literature is to say that a cluster C is a true discovery if proportion at least τ of

[4] Recall that V_P and V_C refer to the pointwise and clusterwise quantities respectively; see §1.1. Note that the implication does not go the other way. In particular, the second example in Figure 1.4 shows a case where $V_C = 0$ but $V_P \gg 0$.

the cluster overlaps the signal:

$$\frac{|C \cap D_1|}{|C|} \geq \tau.$$

In the standard testing literature this criterion is typically referred to as a *partial conjunction hypothesis* [10]; and it has been appeared in the spatial inference setting in the work of Perone Pacifico et al. [36] and Heller et al. [24]. This definition of true discovery is not addressed by the methods presented in this thesis. We argue that the less stringent definition we employ is appropriate in many areas of application. It is particularly appropriate in settings where the experiment is being conducted in order to identify regions for further study.

Other literature. Schwartzman, Gavrilov, and Adler [39] propose an approach for 1-d smooth gaussian processes that's based on applying BH to p-values obtained at each local maximum. We discuss connections to their proposal in Section 2.3. Also of interest are the proposal of Jaffe et al. [25], who present permutation based method for detecting differentially methylated regions. The perspective on 'topological inference' presented in Chumbley and Friston [15] and Chumbley et al. [14] will be of interest to anyone working with neuroimaging data.

Chapter 2

Comparisons of Clusterwise FDR to other Methods

Chapter outline

This chapter delves into the connections between the clusterwise FDR control procedure introduced in Section 1.5 and several other spatial inference methods from the literature. We begin in Section 2.1 with a more in-depth study of the clusterwise FDR control properties of the pointwise procedure. Our analysis shows that the pointwise procedure behaves like a clusterwise procedure in which the number of rejections is given by, $R = V + \gamma S$, with $\gamma > 1$. Recall from Section 1.1 that V is the number of falsely rejected clusters, and S is the number of correctly rejected clusters. Our analysis shows that pointwise procedure effectively upweights each true detection by a factor of γ. This characterization helps to explain why the pointwise procedure is anti-conservative, and also allows us to obtain a bound on its clusterwise FDR control level.

In Section 2.2 we discuss a pair of random field-based approaches to familywise error rate control. Using simple Poisson clumping heuristic machinery we drive connections to the clusterwise FDR procedure and the standard Bonferroni procedure.

Our analysis suggests parallels between the standard multiple testing setting and the spatial one.

In Section 2.3 we discuss connections to the STEM procedure introduced by Schwartzman, Gavrilov, and Adler [39]. The STEM procedure can be used to obtain a form of clusterwise FDR control in the case where the smooth noise process, $\tilde{\epsilon}$, is a thrice differentiable stationary ergodic Gaussian process. We show that at high threshold levels (low-to-moderate values of α), the clusterwise procedure of Section 1.5 and the STEM procedure are essentially equivalent.

We conclude the chapter by presenting the results of a simulation study. Our experimental findings are in close agreement with the mathematical analyses presented in this chapter.

2.1 Comparison to pointwise FDR controlling procedure

We saw in the introduction that the pointwise procedure can be highly anti-conservative in terms of clusterwise FDR control. In this section we pursue a more formal analysis of pointwise procedure to develop a better understanding of its clusterwise FDR controlling properties. Using basic PCH machinery, we derive a key relation between the pointwise and clusterwise FDR controlling procedures in the large observation time regime. While our argument is presented in the 1-dimensional case in which $D = [0, T]$, it generalizes to higher dimensions.

Generative model. We will assume that our observed data takes the form,

$$y(t) = \sum_{k=1}^{K} h_k(t) + \epsilon(t), \tag{2.1.1}$$

where the h_k are compactly supported non-negative functions whose support is small relative to the size of the observation region, $D = [0, T]$. In this case, $D_1 = \bigcup_{k=1}^{K} \{t \in$

$[0, T] : h_k(t) > 0\}$.

For simplicity and to ensure that we have a reasonable limiting problem, we will further assume that the h_k themselves are generated iid from some distribution. More precisely, let H be a distribution on non-negative functions supported on the compact interval $[-M, M]$ for some small fixed M, and let $\{s_i\}$ denote points of a Poisson process of rate ν on $[0, \infty)$, with $1/\nu \gg 2M$. We will assume each h_k is an iid realization from H, translated to location s_k. The condition $1/\nu \gg 2M$ implies that expected distance between the signal components is considerably greater than their support, and hence that the supports of two h_k are unlikely to overlap.

We will assume that the smoother is chosen in such a way that the resulting noise term $\tilde{\epsilon}(t)$ is stationary and ergodic, and is such that the mosaic process approximation applies. That is, we assume that the excursions sets of $\tilde{\epsilon}(t)$ are well approximated by a mosaic process with clump rate λ_z/T.

As part of our analysis it will be helpful to consider the behaviour of the inferential procedure as the observation time, T, tends to infinity. We will therefore think of the model in (2.1.1) as being indexed by T.

Lastly, we introduce the quantities $\mathbb{E}(C_{0,z})$ and $\mathbb{E}(C_{A,z})$ to denote the expected cluster size under the null and alternative, respectively. Since both $\tilde{\epsilon}(t)$ and the signal process are assumed to be stationary and ergodic, these quantities are equal to the long-run averages,

$$\begin{aligned}\mathbb{E}(C_{0,z}) &= \lim_{T\to\infty} \frac{\sum_{C\in\mathcal{V}_z} |C|}{V(z)} \\ \mathbb{E}(C_{A,z}) &= \lim_{T\to\infty} \frac{\sum_{C\in\mathcal{S}_z} |C|}{S(z)},\end{aligned}$$

where $\mathcal{S}_z$ and $\mathcal{V}_z$ denote sets of true clusters and false clusters, respectively, at level z. The ratio $\gamma_z \equiv \mathbb{E}(C_{A,z})/\mathbb{E}(C_{0,z})$ will turn out to play a central role in our analysis.

Derivation. A useful approximation that comes out of the PCH literature is that,

$$p_z \approx \frac{\lambda_z \mathbb{E}(C_{0,z})}{T}. \tag{2.1.2}$$

where $p_z = \mathbb{P}(\tilde{\epsilon}(t) > z)$. This approximation is particularly useful here because it relates two quantities of interest: the p-value function, p_z, upon which the pointwise FDR controlling procedure is based, and the mean parameter function, λ_z, upon which the clusterwise procedure is based.

The intuition for this result is simple: If the expected number of clusters at threshold z is λ_z, and the expected cluster size is $\mathbb{E}(C_{0,z})$, then the expected number of locations at which $\tilde{\epsilon}(t) > z$ is given by $\mathbb{E}|\{t : \tilde{\epsilon}(t) > z\}| \approx \lambda_z \mathbb{E}(C_{0,z})$. Under ergodicity, dividing by the observation timespan T is thus a good approximation to the marginal probability $\mathbb{P}(\tilde{\epsilon}(t) > z)$.

Moving forward, let $R_P(z)$ denote the number of rejections at level z measured pointwise, and $R_C(z)$ denote the number of rejections measured clusterwise. Define the quantities $V_P(z)$, $V_C(z)$, $S_P(z)$ and $S_C(z)$ analogously.

Recall that the clusterwise FDR controlling procedure selects the cutoff z according to,

$$z_{clust} = \min\left\{ z : \frac{\lambda_z}{R_C(z)} \leq \alpha \right\}. \tag{2.1.3}$$

The Benjamini-Hochberg FDR controlling procedure amounts to selecting z according to,

$$z_{point} = \min\left\{ z : \frac{p_z T}{R_P(z)} \leq \alpha \right\}. \tag{2.1.4}$$

Using (2.1.2), the main argument appearing in (2.1.4) can be rewritten as,

$$\frac{p_z T}{R_P} = \frac{\lambda_z \mathbb{E}(C_{0,z})}{R_P(z)}.$$

We can view the pointwise rejections in $\mathcal{E}_z$ as being grouped into $R_C(z)$ clusters, of which $V_C(z)$ are false detections, and $S_C(z)$ are true detections. Denote the lengths of the false clusters by $C^z_{01}, C^z_{02}, \ldots, C^z_{0V_C(z)}$, and the lengths of true clusters by $C^z_{A1}, C^z_{A2}, \ldots, C^z_{AS_C(z)}$. Using this notation we can express $R_P(z)$ as,

$$\begin{aligned} R_P(z) &= \sum_{i=1}^{V_C(z)} C^z_{0i} + \sum_{i=1}^{S_C(z)} C^z_{Ai} \qquad (2.1.5) \\ &\approx V_C(z)\mathbb{E}(C_{0,z}) + S_C(z)\mathbb{E}(C_{A,z}). \end{aligned}$$

From the assumed ergodicity and stationary of $\tilde{\epsilon}$, we know that the approximation above is quite good when T is large. Using these expressions, we can rewrite the quantity of interest as,

$$\begin{aligned} \frac{p_z T}{R_P} &= \frac{\lambda_z \mathbb{E}_0 C_z}{\sum_{i=1}^{V_C(z)} C^z_{0i} + \sum_{i=1}^{S_C(z)} C^z_{Ai}} \qquad (2.1.6) \\ &\approx \frac{\lambda_z \mathbb{E}(C_{0,z})}{V_C(z)\mathbb{E}(C_{0,z}) + S_C(z)\mathbb{E}(C_{A,z})} \\ &= \frac{\lambda_z}{V_C(z) + S_C(z)\frac{\mathbb{E}(C_{A,z})}{\mathbb{E}(C_{0,z})}}. \qquad (2.1.7) \end{aligned}$$

Note that by keeping the smoother fixed while increasing the support and amplitude of the signal components, we can make $\gamma_z = \mathbb{E}(C_{A,z})/\mathbb{E}(C_{0,z})$ arbitrarily large, and thereby greatly inflate the denominator. Since we generally expect to have $\gamma_z > 1$, (2.1.7) implies,

$$\begin{aligned} \frac{p_z T}{R_P} &\lesssim \frac{\lambda_z}{V_C(z) + S_C(z)} \\ &= \frac{\lambda_z}{R_C(z)}. \end{aligned}$$

We see from this derivation that, at the same value of α, we generally have $z_{point} < z_{clust}$. This suggests that applying the pointwise BH(α) procedure will fail to control the clusterwise FDR at the target level α. Taking the approximation in (2.1.7) as a proxy for the pointwise control procedure, we find that the standard martingale argument establishes a much weaker control result.

Proposition 2.1. *For fixed $\gamma > 1$ and $\alpha \in [0,1)$, define the stopping rule*

$$\tilde{\Lambda} = \max\{\lambda \leq \bar{\lambda} : \lambda/(V_C(\lambda) + S_C(\lambda)\gamma) \leq \alpha\}.$$

Under the conditions of Section 1.5, the stopping rule $\tilde{\Lambda}$ controls the clusterwise FDR at level $\gamma\alpha$, in the sense that,

$$\mathbb{E}\left(\frac{V_C(\tilde{\Lambda})}{R_C(\tilde{\Lambda})}\right) \leq \min(1, \alpha\gamma).$$

Since the stopping rule $\tilde{\Lambda}$ depends on knowledge of the processes $V_C(z)$ and $S_C(z)$, the rule is not implementable in practice. Our interest in this proposition is due entirely to approximation (2.1.7), which suggests that, when the observation region is large, $\tilde{\Lambda}$ is a good proxy for the pointwise control procedure.

Proof (Proposition 2.1). All quantities here are clusterwise quantities, so we drop the subscript C for the duration of the proof. To begin, we rewrite the stopping rule as,

$$\tilde{\Lambda} = \max\left\{\lambda \leq \bar{\lambda} : \frac{\lambda}{R(\lambda) + (\gamma - 1)S(\lambda)} \leq \alpha\right\}.$$

Under the assumptions of Section 1.5, the process $V(\lambda)/\lambda$ is a mean-one continuous time backwards martingale with respect to the filtration $\mathcal{F}_\lambda = \sigma(V(t), S(t) : \lambda \leq t \leq \bar{\lambda})$. It is clear from the definition that $\tilde{\Lambda}$ is measurable with respect to the filtration $\mathcal{F}_\lambda$. By the optional stopping theorem, it follows that

$$\mathbb{E}\left(\frac{V(\tilde{\Lambda} \vee \lambda)}{\tilde{\Lambda} \vee \lambda}\right) = \mathbb{E}\left(\frac{V(\bar{\lambda})}{\bar{\lambda}}\right) = 1.$$

Note that for all λ, $\mathbb{1}(\tilde{\Lambda} < \lambda) = 1$ implies that

$$\frac{\lambda}{V(\lambda)} \geq \frac{\lambda}{R(\lambda) + (\gamma - 1)S(\lambda)} > \alpha.$$

Thus $\mathbb{1}(\tilde{\Lambda} < \lambda)V(\lambda)/\lambda$ is bounded above by $1/\alpha$. Since this quantity is bounded and converges to 0 as $\lambda \to 0$, by the dominated convergence theorem we get that,

$$\mathbb{E}\left(\frac{V(\tilde{\Lambda})}{\tilde{\Lambda}}; \tilde{\Lambda} > 0\right) = \mathbb{E}\left(\frac{V(\bar{\lambda})}{\bar{\lambda}}\right) = 1.$$

Since $V(\tilde{\Lambda})/R(\tilde{\Lambda})$ is defined to be 0 when $\tilde{\Lambda} = 0$, we also have that $\mathbb{E}(V(\tilde{\Lambda})/R(\tilde{\Lambda})) = \mathbb{E}(V(\tilde{\Lambda})/R(\tilde{\Lambda}); \tilde{\Lambda} > 0)$. Thus we obtain,

$$\mathbb{E}(V(\tilde{\Lambda})/R(\tilde{\Lambda})) = \mathbb{E}\left(\frac{V(\tilde{\Lambda})}{\tilde{\Lambda}} \cdot \frac{\tilde{\Lambda}}{R(\tilde{\Lambda})}; \tilde{\Lambda} > 0\right) \leq \gamma\alpha \cdot \mathbb{E}\left(\frac{V(\tilde{\Lambda})}{\tilde{\Lambda}}; \tilde{\Lambda} > 0\right) = \alpha\gamma.$$

where the inequality follows from the fact that $\tilde{\Lambda} > 0$ implies

$$\frac{\tilde{\Lambda}}{R(\tilde{\Lambda}) + (\gamma - 1)S(\tilde{\Lambda})} \leq \alpha \quad \Longrightarrow \quad \frac{\tilde{\Lambda}}{R(\tilde{\Lambda})} \leq \alpha\left(1 + \frac{S(\tilde{\Lambda})}{R(\tilde{\Lambda})}(\gamma - 1)\right) \leq \alpha\gamma.$$

□

Beyond the result established above, (2.1.6) and (2.1.7) give insight into key differences between the pointwise and clusterwise FDR controlling procedures. Looking at (2.1.6), we see that the presence of one or more large true clusters (values of C_{Ai}^z that are large relative to $\mathbb{E}_0 C_z$) can dramatically inflate the denominator. This makes pointwise inference particular ill-suited to cases where some differentially behaved regions are expected to be large in size (e.g., in genetics studies).

Even if the differentially behaved regions are expected to be well concentrated and moderate in size (i.e., C_{Ai} well concentrated around $\mathbb{E}(C_{A,z})$) we see from (2.1.7) that the pointwise procedure effectively up-weights each correctly rejected cluster by a factor of $\frac{\mathbb{E}(C_{A,z})}{\mathbb{E}(C_{0,z})} > 1$. This in turn inflates the denominator, and thereby allows rejections at lower values of z.

It is also worth noting that while the decomposition of R_P in (2.1.5) is valid, the two summands involved are generally *not* valid decompositions of V_P and S_P. This is because our definition of true cluster does not require that the entire cluster be contained in the support of the signal, D_1. At moderate values of the threshold z, many of the true clusters will intersect D_0, and locations in this intersection will contribute to V_P. We therefore have that,

$$V_P(z) \geq \sum_{i=1}^{V_C(z)} C_{0i}^z \quad \text{and} \quad S_P(z) \leq \sum_{i=1}^{S_C(z)} C_{Ai}^z,$$

with equality only when all true clusters are contained in D_1.

2.2 Connection to random field-based FWER controlling procedures

We briefly explore a connection to a couple of random-field based bounds used in the scientific literature to control the FWER. To control the FWER, one must select a threshold that satisfies,

$$z_{sup} = \min \left\{ z : \mathbb{P}\left(\sup_{[0,T]} \tilde{\epsilon}(t) > z \right) \leq \alpha \right\}.$$

We present here two approximations to the supremum probability, both of which are applied in the literature.

General upper bound. Rice's formula gives a general bound on the supremum probability that holds for all stationary differentiable random processes [1]. In our notation, this bound can be written as,

$$\mathbb{P}\left(\sup_{[0,T]} \tilde{\epsilon}(t) > z \right) \leq p_z + \mathbb{E}N_z,$$

where $p_z = \mathbb{P}(\tilde{\epsilon}(0) \geq z)$, and N_z is the number of up-crossings of level z by the process $\tilde{\epsilon}$. For large thresholds z, up-crossings are isolated events, each of which marks the start of an individual cluster of the excursion set $\mathcal{E}_z$. Thus we can typically equate

$\mathbb{E}N_z = \lambda_z$, and obtain the bound,

$$\mathbb{P}\left(\sup_{[0,T]} \tilde{\epsilon}(t) > z\right) \leq p_z + \lambda_z.$$

Requiring that $p_z + \lambda_z < \alpha$ is clearly a far more stringent condition than $\lambda_z / R(z) < \alpha$.

Poisson clumping heuristic approximation. There is a more direct approximation to the supremum probability that comes from applying the PCH. This approximation is simply,

$$\mathbb{P}\left(\sup_{D} \tilde{\epsilon}(t) > z\right) = \mathbb{P}(V_z > 0) \approx 1 - \exp(-\lambda_z) \leq \lambda_z. \tag{2.2.1}$$

This is slightly lower than the upper bound presented above, but in practice it gives effectively the same result.

Discussion. In the previous section we introduced approximation 2.1.2, which states that $p_z \approx \lambda_z \mathbb{E}C_{0,z}/|D|$. Plugging this into (2.2.1) gives,

$$\mathbb{P}\left(\sup_{D} \tilde{\epsilon}(t) > z\right) \approx \lambda_z \approx \frac{p_z |D|}{\mathbb{E}C_{0,z}}. \tag{2.2.2}$$

Note that the quantity on the right hand side takes the form of a Bonferroni correction to the p-value, p_z, with the effective number of tests given by $|D|/\mathbb{E}C_{0,z}$. This derivation illustrates the problem with controlling FWER by simply applying a pointwise Bonferroni correction to the $|D|$ test statistics. Doing so results in a p-value cutoff that is more stringent by a factor of $\mathbb{E}C_{0,z}$ compared to the random field-based approach. Depending on the problem setting, this may or may not translate into an appreciable difference in the corresponding z-thresholds or in the overall power.

We can also use this observation to build some intuition for the expected power gains to be had by using the clusterwise FDR procedure described in Section 1.5 over the random field-based FWER control procedure. Under appropriate calibration of signal strength, the gains should be similar to those attained in a standard multiple testing problem with $m \approx |D|/\mathbb{E}C_{0,z}$ hypotheses of which K are false.

A general comment on FDR vs. FWER. The typical argument for controlling FDR instead of FWER is that FDR controlling procedures can be considerably more powerful. When there are a large number of potential discoveries (i.e., when the number of signal regions, K, is large), and the signal strength is sufficient for detection, the power gains can be tremendous. This is certainly the case in many genetics studies, and also in pointwise analyses of imaging data. On the other hand, when K is expected to be small, there isn't much to be gained by controlling the FDR instead of the FWER. For instance, in fMRI studies it is generally expected that only a handful of distinct regions will be activated (or differentially activated) for any given task. Thus while much of the interest in spatial FDR control is coming from the neuroimaging community, it isn't clear that imaging is necessarily the best use case for the methodology.

2.3 Comparison to the STEM procedure in the case of smooth Gaussian data

As mentioned in the introduction, the recently introduced STEM procedure of Schwartzman, Gavrilov, and Adler [39] has close connections to our proposal.[1] As we will now argue, at high threshold levels the two procedures are essentially equivalent.

Description of the STEM procedure. The model considered by the STEM procedure is one in which we observe,

$$y(t) = \mu(t) + \epsilon(t), \qquad t \in D = [0, T],$$

where $\mu(t)$ is a sparse train of unimodal positive peaks, and $\epsilon(t)$ is stationary Gaussian noise. It is assumed that $\tilde{y}(t)$ is formed by convolution with a compactly supported unimodal kernel, and that the smoothed noise, $\tilde{\epsilon}(t)$, is a thrice differentiable stationary ergodic Gaussian process.

[1] STEM stands for Smoothing and TEsting of Maxima.

The main idea behind the STEM procedure is to conduct inference by testing the significance of the observed local maxima of $\tilde{y}(t)$. More precisely, the procedure is as follows.

STEM procedure.

1. Smooth the data to obtain $\tilde{y}(t)$.
2. Identify the locations of all of the local maxima of $\tilde{y}(t)$. Call this set $\mathcal{T}$.
3. For each $t \in \mathcal{T}$, compute a p-value $\tilde{p}(t)$ for testing the conditional hypothesis
$$H_0(t) : \mu(t) = 0 \quad \text{vs.} \quad H_A(t) : \mu(t) > 0,$$
conditional on $t \in \mathcal{T}$ (i.e., conditional on t being the location of a local maximum).
4. Apply the BH(α) procedure to the p-values $\{\tilde{p}(t)\}_{t \in \mathcal{T}}$, and report as discoveries all peaks (local maxima) corresponding to significant p-values.

The authors adopt the definition that local maximum is a true detection if it falls anywhere in the support of the underlying signal. Under some assumptions on minimal signal strength, the authors establish that the STEM procedure asymptotically controls the FDR in the limit where the observation period $T \to \infty$.

Let $\tilde{p}_z$ denote the conditional p-value function evaluated at z, let $\tilde{m} = |\mathcal{T}|$ denote the observed number of local maxima, and let $R(z)$ denote the number of local maxima whose height exceeds z. In this notation, Step 4. of the STEM procedure amounts to rejecting all local maxima whose height exceeds z_{STEM}, where,

$$z_{\text{STEM}} \equiv \min \left\{ z : \frac{\tilde{p}_z \tilde{m}}{R(z)} \leq \alpha \right\}. \tag{2.3.1}$$

Let $\tilde{D}_0 = [0, T] \backslash \tilde{D}_1$ denote the complement of the support of the smoothed signal, $\tilde{\mu}$.

The conditional p-value function, $\tilde{p}_z$ is formally a Palm distribution, defined by

$$\tilde{p}_z \equiv \mathbb{P}(\tilde{\epsilon}(t) > z | t \in \mathcal{T} \cap \tilde{D}_0) \equiv \frac{\mathbb{E}V(z)}{\mathbb{E}\tilde{m}_0},$$

where $V(z)$ is the number of local maxima of $\tilde{\epsilon}(t)$ in $\tilde{D}_0$ whose height is at least z, and $\tilde{m}_0$ is the total number of local maxima of $\tilde{\epsilon}(t)$ in $\tilde{D}_0$.

Since we are dealing with smooth Gaussian processes, we can identify high level local maxima with individual clusters (intervals) of the excursion set $\mathcal{E}_z$. At high thresholds z, local maxima of $\tilde{\epsilon}(t)$ are isolated events, each of which has a 1-to-1 correspondence to an isolated upcrossing of z by $\tilde{\epsilon}(t)$, which in turn marks the start of an individual cluster of the excursion set $\mathcal{E}_z$ [2, §C23]. In this case, we have that $\mathbb{E}V(z) = \lambda_z$, and so we can rewrite the argument of (2.3.1) as,

$$\frac{\tilde{p}_z \tilde{m}}{R(z)} = \frac{\lambda_z}{R_z} \frac{\tilde{m}}{\mathbb{E}\tilde{m}_0}.$$

We can further re-express the number of local maxima $\tilde{m}$ as $\tilde{m} = \tilde{m}_0 + \tilde{m}_A$, where $\tilde{m}_A$ is the number of local maxima of $\tilde{y}(t)$ in $\tilde{D}_1$. For large T, $\tilde{m}$ concentrates around its mean, so have approximately that, $\tilde{m} \approx \mathbb{E}\tilde{m}_0 + \mathbb{E}\tilde{m}_A$. This gives,

$$\frac{\tilde{p}_z \tilde{m}}{R(z)} \approx \frac{\lambda_z}{R_z} \frac{\mathbb{E}\tilde{m}_0 + \mathbb{E}\tilde{m}_a}{\mathbb{E}\tilde{m}_0}.$$

Now, the quantity $\frac{\mathbb{E}\tilde{m}_0}{\mathbb{E}\tilde{m}_0 + \mathbb{E}\tilde{m}_a}$ is the proportion of local maxima that are expected to be null. To emphasize the connection to the standard multiple testing problem, we denote this quantity by $\tilde{\pi}_0$.

Rewriting the argument in the final notation gives,

$$\frac{\tilde{p}_z \tilde{m}}{R(z)} \approx \frac{\lambda_z / \tilde{\pi}_0}{R_z}.$$

This is essentially equivalent to the argument in our clusterwise FDR control procedure, in the case where we calculate the null cluster parameter λ under the global null; i.e., $\lambda_z / \tilde{\pi}_0 \approx \mathbb{E}_0 V(z)$, where $\mathbb{E}_0$ denotes expectation in the case where $\tilde{D}_0 = D$.

Discussion. The preceding argument holds only for the range of threshold values at which the occurrence of local maxima are well approximated by a Poisson process on the line. Thus we can expect the STEM procedure and the clusterwise FDR control procedure of §1.5 to give the same result for low or moderate choices of target FDR level, α.

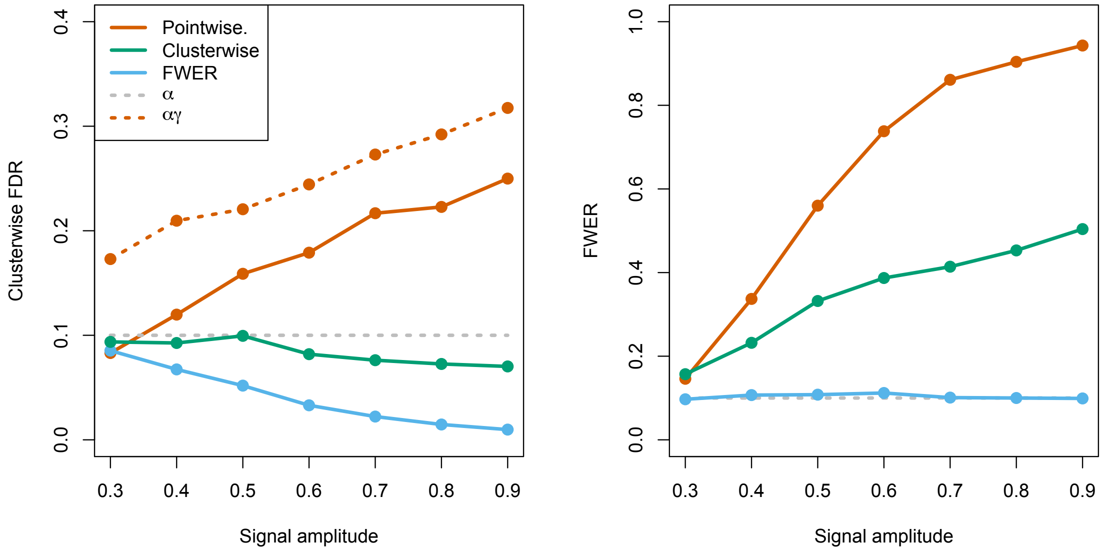
One advantage of the STEM procedure is that the Palm distribution p-value formula is exact for all threshold levels. This means that the STEM procedure can be freely applied at α levels close to 1. It is therefore an excellent procedure for spatial FDR control when the smooth noise process has the assumed structure; that is, when $\tilde{\epsilon}(t)$ is a 1-dimensional thrice differentiable stationary ergodic Gaussian process.

The mosaic process approximation applies for a much broader class of noise distributions, and is not limited to the 1-dimensional setting. Furthermore, as we will see in the next chapter, the methods of §1.5 have many useful extensions that can be obtained by taking advantage of the mosaic process approximation. The STEM procedure does not offer the same flexibility.

2.4 Experiments

In this section we present the results of a simulation study conducted in order to compare the pointwise procedure, FWER procedure and clusterwise procedure in terms of both FDR control and power. Our comparison is conducted under the recurring example of $K = 10$ box functions of equal amplitude and support size which are observed under iid Gaussian noise (see setup in Figure 1.3). We calculate λ_z using the analytic approximation presented in Section 1.6.

Figure 2.1 shows the observed FDR and FWER of the methods for various levels of signal strength in the non-smooth Gaussian noise example considered in the introduction. The target error level is set to $\alpha = 0.1$ for all of the procedures. Results for other values of α were also obtained, but are qualitatively the same and thus are not presented here. As previously observed, the pointwise procedure in general fails



(a) Observed FDR.

(b) Observed FWER.

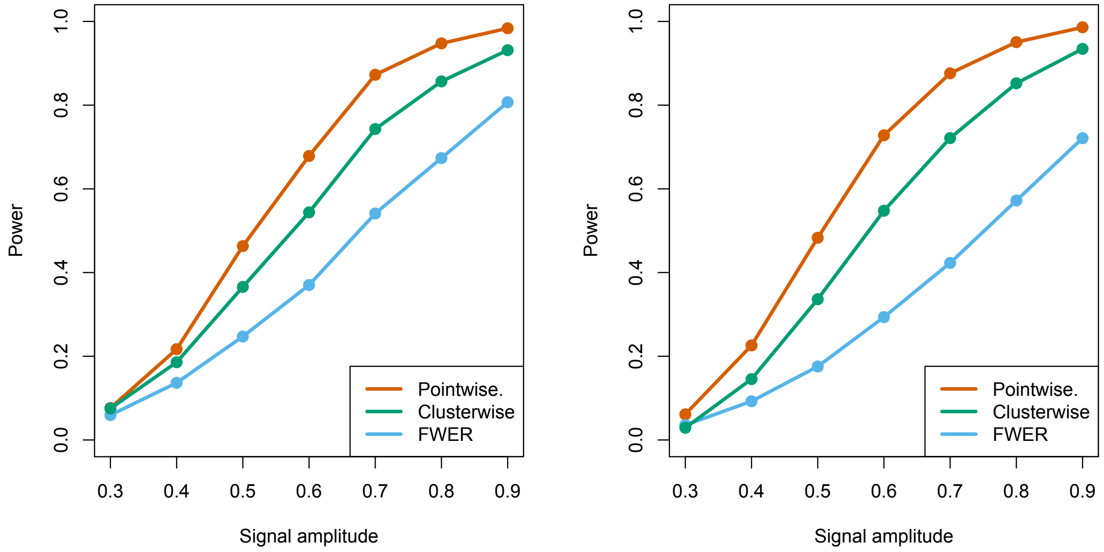
Figure 2.1: Observed error rates for the pointwise, clusterwise and FWER procedures for different choices of signal amplitude. $K = 10$, $T = 2000$, and target $\alpha = 0.1$ was used throughout. The pointwise procedure fails to provide clusterwise FDR control except at the lowest signal value, where the signal strength is too weak to be detected by any of the procedures. A curve corresponding to $\alpha \cdot \gamma_z$ is also shown; this is the upper bound on FDR control from Proposition 2.1.

to control the clusterwise FDR. The only exception is at the lowest value of signal strength, a regime in which the signal is too weak to be detected by any of the procedures. We observe that the clusterwise FDR of the pointwise procedure increases with signal strength. This is consistent with the observation that γ_z increases in signal strength.

Figure 2.2 shows the observed average power of the methods. Average power is defined as the fraction of the K underlying signal locations that were detected. More precisely, if we let $S_1, S_2, \ldots, S_K$ denote the support regions of the signal components h_k, the average power of a selection procedure $\hat{z}(\tilde{y})$ is defined as,

$$\mathbb{E}\left(\frac{\sum_{k=1}^{K} \mathbb{1}\{S_k \cap \mathcal{E}_{\hat{z}} \neq \emptyset\}}{K}\right).$$

Results are shown for two choices of problem size. The pointwise procedure has higher power than the clusterwise procedure, but as we have seen the pointwise procedure fails to provide clusterwise FDR control. Comparing the clusterwise procedure to the PCH FWER procedure, we see that controlling FDR results in considerably higher power. Moreover, as the problem size increases, the power gap between the FDR procedure and the FWER procedure widens considerably. FWER procedures do not scale well to large problems.



(a) $T = 2000$, $K = 10$.

(b) $T = 6000$, $K = 30$.

Figure 2.2: Power plots for the pointwise, clusterwise and FWER procedures. Power corresponds to the average fraction of the K signal locations that were detected. Target $\alpha = 0.1$ was used throughout. The pointwise procedure has higher power than the clusterwise procedure, but it does not control the clusterwise FDR. The FWER procedure has lower power than the clusterwise procedure, and the gap increases in the size of the problem. FDR control has significant power gains over FWER control when there are a large number of potential detections.

Chapter 3

Variations on a Theme

Chapter outline

This Chapter presents some new results that serve to formalize and extend the clusterwise FDR procedure of Siegmund et al. [41] as introduced in Section 1.5.

To begin, in Section 3.1 we consider the problem of identifying clusters when the clusters themselves are expected to be disconnected. The merge procedure is shown to result in valid clusterwise inference, and simulation results suggest that its performance is fairly insensitive to the choice of tuning parameter.

In Section 3.2 we discuss an alternative false discovery proportion-based error control criterion referred to as the False discovery exceedance (FDX). We present a general thresholding-based FDX control method that is based on the augmentation method introduced in van der Laan et al. [48]. While the control procedure itself does not rely on the mosaic process approximation, we show how the Poisson approximation can be applied in controlling the FDX.

In Section 3.3 we demonstrate how the basic clusterwise FDR procedure can be generalized via *thinning* to incorporate other measures of cluster significance. Our experiments indicate that aggressive thinning at high thresholds results in a more powerful clusterwise FDR control procedure. In the same simulation setup as in

Section 2.4, we find that incorporating cluster size leads to a clusterwise FDR control procedure with *better* power than the pointwise procedure.

Section 3.4 considers the case where the smoothed noise process $\tilde{\epsilon}(t)$ is non-stationary. We show how the clusterwise FDR procedures extend to the non-homogeneous setting in cases where the occurrence of false clusters is well approximated by a non-homogeneous form of the mosaic process. We discuss how non-homogeneity can arise due to non-stationarity of $\epsilon(t)$, or non-uniform sampling density of the observation points.

We conclude in Section 3.5 with a discussion of stratification. We present extensions of the FDR estimation and control procedures for settings where we may wish to allow different selection thresholds across different pre-defined strata.

3.1 Automated cluster determination

When working with smooth continuous time processes or similarly behaved discrete time processes on the line, the geometry of high level excursion sets is fairly well understood. In such cases the components of high level excursion sets are simply isolated random intervals. For non-smooth processes, or at lower thresholds, an upcrossing can be followed in close succession by a sequence of rapid upcrossings and downcrossings, which results in clusters that are themselves fragmented. Figure 3.1 gives an example of this behaviour.

Consider for now the 1-dimensional case. What we observe is that each cluster consists of a random number, $N_z \geq 1$, of nearby components. Letting ρ_z denote the expected number of upcrossings in time $[0, T]$, we get the following relation,

$$\lambda_z = \frac{\rho_z}{\mathbb{E}N_z}.$$

At high threshold levels, we have $\mathbb{E}N_z \approx 1$, and thus $\lambda_z \approx \rho_z$. If we wish only to consider very small values of α and hence very high thresholds z, we may not need to worry about merging. However, if we're careful in dealing with the case where

$\mathbb{E}N_z > 1$, we can apply the methods for a much greater range of threshold values.

In order to conduct valid inference when $\mathbb{E}N_z > 1$, we need to devise rules for identifying components of the excursion set that need to be merged into single clusters. This section presents one possible merging approach. We begin by discussing what is desired of a valid merging procedure.

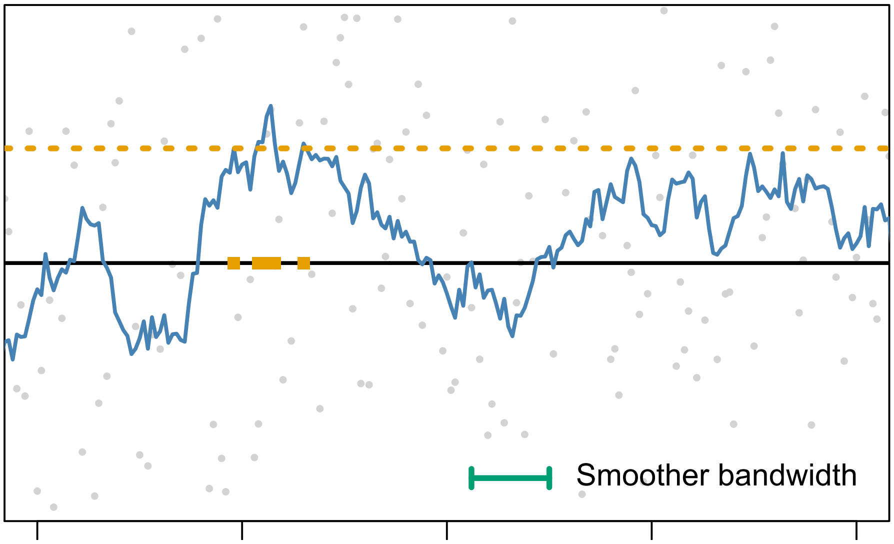


Figure 3.1: Example of a cluster that consist of several disconnected components.

First, we note that the estimation and control procedures do not rely on the distribution of S beyond assuming that $S \geq 0$ is independent of V. We can therefore think of S as being determined by the underlying signal, smoother, and the choice of merge rule. Under-merging true detections may result in multiply counting several detections that overlap with the same underlying component of the signal. While this is undesirable, it does not invalidate the inference.

Another important thing to observe is that the procedures are fairly robust to over-merging. To see this, consider merging two excursion set components E_1 and E_2. Let V and S denote the testing quantities when E_1 and E_2 are not merged, and let $V_\cup$ and $S_\cup$ denote the testing quantities when E_1 and E_2 are merged. There are three possibilities in this case.

1. $(E_1 \cup E_2) \cap D_1 = \emptyset$. If neither E_1 nor E_2 intersects the support of the signal, then $S_{\cup} = S$, while $V_{\cup} = V-1$. We therefore have that (a) $1/(V_{\cup}+S_{\cup}+1) > 1/(R+1)$, and (b) $V_{\cup}/R_{\cup} < V/R$. (a) implies that the the FDR estimate $\lambda/(R+1)$ can only be biased upward by merging. (b) implies that the FDR of the procedure with merging is smaller than the FDR without merging.

2. $E_1 \cap D_1 \neq \emptyset$ and $E_2 \cap D_1 = \emptyset$. When one of the components is a true detection while the other is not, the merged component $E_1 \cup E_2$ will still be a true detection. Thus we will once again have that $S_{\cup} = S$ and $V_{\cup} = V - 1$, which is the same as the previous case.

3. If $E_1 \cap D_1 \neq \emptyset$ and $E_2 \cap D_1 \neq \emptyset$, then V is unaffected by the merge. Thus if the estimation and control procedures were valid prior to merging, they remain valid after merging, though with a different distribution for S.

The main issue, therefore, is potentially under-merging false detections, which would in fact invalidate our inference. This is because the parameter λ_z in the mosaic process approximation is a good description of the false discovery process only after appropriate merging. In particular, counting each excursion of in a rapid sequence of crossings to be its own cluster clearly invalidates the Poisson process assumption. By under-merging false detections we inflate V, which makes the estimation and control methods anti-conservative. Our goal is to make sure that the merging procedure results in a V to which the methods continue to apply.

Consider the false discovery process $V_{\lambda(z)}$ under the global null as the process varies with the threshold parameter z. A minimal consistency requirement for the discovery procedure is that $V_{\lambda(z)}$ be non-increasing in z. To ensure monotonicity, we will require that the merge procedure have the following persistence property.

Definition 3.1. For $z < z'$, let $\mathcal{E}_z = \bigcup E_i$ and $\mathcal{E}_{z'} = \bigcup E'_j$. We say that a merging procedure is *persistent* if for any index set I such that $\{E_i\}_{i \in I}$ is merged into a single cluster at level z, the set $\{E'_j\}_{j : E'_j \subset E_i \text{ for some } i \in I}$ is merged at level z'.

This is a rather complicated way of saying that a discovery at threshold z should not be split into multiple discoveries at some higher threshold $z' > z$.

With these considerations in mind, we now present the details of our proposed merging procedure. We begin by discussing the 1-dimensional case, for which we can provide rigorous justification.

Horizontal information. We begin by summarizing some distributional facts regarding the components of high level excursion sets. Given a threshold z, we once again let $\mathbb{E}C_{0,z}$ denote the expected size of a false cluster at threshold z. We can rearrange (2.1.2) to give the approximation,

$$\mathbb{E}C_{0,z} \approx \frac{p_z T}{\lambda_z}.$$

Next, note that the Poisson process assumption on the occurrence of the null clusters implies that inter-clump distances are distributed approximately as $\text{Exp}(\lambda_z/T)$. The average inter-cluster distance is therefore T/λ_z, which is generally very large compared to $\mathbb{E}C_{0,z}$.

A reasonable approach is thus to select a value of τ satisfying

$$\mathbb{E}C_{0,z} \approx \frac{p_z T}{\lambda_z} < \tau \ll \frac{T}{\lambda_z}, \tag{3.1.1}$$

and to associate clusters with τ-upcrossing of level z, which are locations t such that $\tilde{y}(s) < z$ for all $s \in [t-\tau, t)$ and $\tilde{y}(t) \geq z$.

Vertical information. Just looking at horizontal distances isn't sufficient for the monotonicity condition to be satisfied for a given realization. We therefore propose to further merge candidate clusters that are contained within the same the connected component of $\mathcal{E}_{z_0}$ for a lower level $0 \leq z_0 < z$. In practice, the procedure does not appear to be particularly sensitive to the choice of z_0, as long as it's sufficiently small.

With these two considerations in mind, we now present our merging procedure

Merging procedure (1-d case).

- Fix $\tau > 0$, and a merging threshold $0 \leq z_0 \leq z_{min}$, where z_{min} is the lowest value to be considered for the FDR procedures.
- Let $t_1^z < \ldots < t_J^z$ denote the ordered τ-upcrossings of z by $\tilde{y}(t)$
- Let $\mathcal{S}_0 = \{S_i\}_{i=1}^I$ denote the connected components of $\mathcal{E}_{z_0}$.

(1) Define the initial clusters $C_1, C_2, \ldots, C_J$ according to

$$C_j = [t_j^z, t_{j+1}^z) \cap \mathcal{E}_z$$

(2) Merge C_j and $C_{j'}$ for any $j \neq j'$ such that $C_j \cup C_{j'} \subset S_i$ for some $S_i \in S_0$.

In order to state the main result of this section, we need to introduce some notation. Given a threshold $z > 0$ and an integer $\tau > 0$, define

(a) $X_t(z) = \mathbb{1}\{\tilde{\epsilon}(t) \geq z \text{ and } \tilde{\epsilon}(s) < z \; \forall \, t - \tau \leq s < t\}$

(b) $p_1(z) = \mathbb{P}(X_t(z) = 1) = \mathbb{E}(X_t(z))$

(c) $a_2(z) = \mathbb{E}\big|\mathbb{E}\left(X_t(z) - p_1(z) \mid X_s, |s-t| > \tau\right)\big|$

With these definition in hand, we can now state the main result of this section.

Theorem 3.1. *Set $D = [1, 2, \ldots, T]$, and fix a high threshold $z > 0$. Assume that the smoother is chosen so that $\tilde{\epsilon}(t)$ is a stationary sequence. Suppose that the merge procedure is applied with some choice of integer $\tau > 0$ and merge threshold $0 < z_0 < z$. Let V_z be the resulting number of false clusters at level z. Then $V_z \leq \tilde{V}_z$ where $\tilde{V}_z$ satisfies,*

$$\|\mathcal{L}(\tilde{V}_z) - \mathcal{L}(W)\|_{\mathrm{TV}} \leq 2\left(\frac{2\tau\lambda^2}{T} + Ta_2(z)\right) \tag{3.1.2}$$

for W a Poisson random variable with mean $\lambda = \lambda(z) \approx -\log(1 - p^*(z))$, where $p^*(z) = \mathbb{P}(\sup_{1 \leq t \leq T} \tilde{\epsilon}(t) \geq z)$.

We leave the proof of this result for Appendix A.

We can get a stronger result in the special case where the original noise sequence $\{\epsilon(t)\}$ is iid and the smoothed process is a moving average of the form,

$$\tilde{y}(t) = \sum_{i=0}^{M} c_i y(t-i),$$

with coefficients $c_i \geq 0$ and $M \leq \tau$. In this case, we have that $a_2(z) = 0$, and so the second term in (3.1.2) disappears (see Corollary A.1).

As for the first term in the bound (3.1.2), recall from (3.1.1) that we're looking to choose $\tau = \beta p_z T / \lambda_z$ where $\beta > 1$ is some (small) multiplicative factor. Plugging this into (3.1.2) gives,

$$\frac{2\tau\lambda_z^2}{T} = 2\beta\lambda_z p_z.$$

For the range of z values that we consider in our simulations, this quantity is typically no larger than 0.02, and is often much smaller. Thus we can view (3.1.2) as providing a useful bound for the range of z values that we're interested in.

Discussion. While τ-upcrossings do not generalize well to higher dimensions, the second step of the merge procedure continues to be applicable. Indeed, our experiments in the 1-dimensional case suggest that performing step (2) alone is sufficient to maintain the validity of the control and estimation procedures. This simplified procedure is outlined below.

Simpler merging procedure (general case).

- Fix a merging threshold $0 \leq z_0 \leq z_{min}$, where z_{min} is the lowest value to be considered for the FDR procedures.
- Let $\mathcal{S}_0 = \{S_i\}_{i=1}^{I}$ denote the connected components of $\mathcal{E}_{z_0}$.
- Let $\{C_j\}_{j=1}^{J}$ denote the connected components of $\mathcal{E}_z$

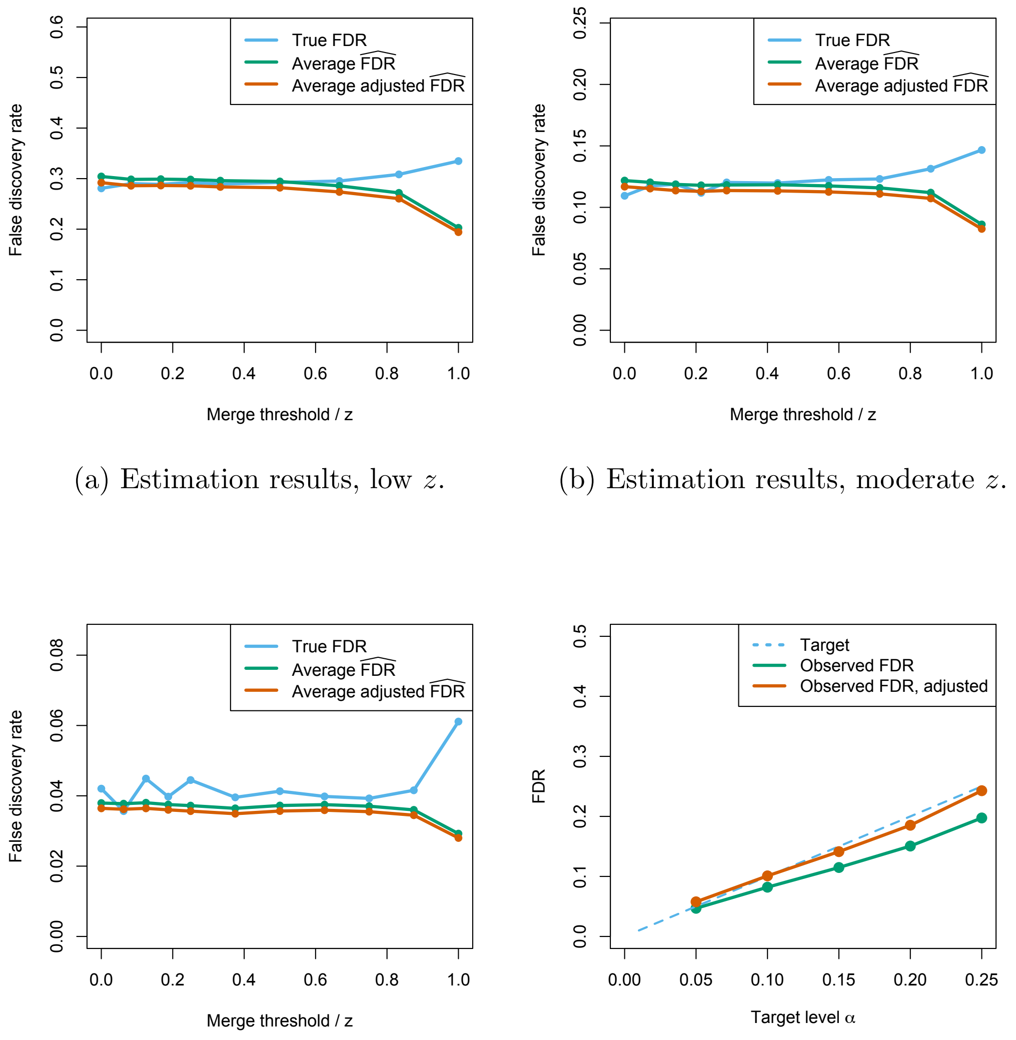
(∗) Merge C_j and $C_{j'}$ for any $j \neq j'$ such that $C_j \cup C_{j'} \subset S_i$ for some $S_i \in S_0$.

Though we do not pursue a rigorous justification for this procedure, we note that it is closely connected to the *conditioning on semi-local maxima* method for calculating the parameter λ_z [2, §C5, §J7].

3.1.1 Experiments

In this section we assess the performance of our merging procedure in the same problem setting as described in Figure 1.3. We saw in the introduction that in this example the pointwise procedure failed to control the clusterwise FDR at the target level. Both merging procedures yielded the same results in our experiment. Figure 3.2 summarizes our findings.

Figures 3.2(a-c) demonstrate the insensitivity of the merge procedures to the choice of merge threshold, z_0. For z_0 close to the selection threshold z, the FDR estimation procedure greatly underestimates the true FDR. However, for $0 \leq z_0 \leq 0.6z$, the estimation procedure is seen to perform uniformly well. Figure 3.2(d) shows that the FDR control procedure has good performance after merging. The merge threshold is chosen at $z_0 = 0.3z_{min}$, where z_{min} is the lowest selection threshold considered by the control procedure. Once we adjust for $\tilde{\pi}_0$, we get control at almost exactly the target level.



(a) Estimation results, low z.

(b) Estimation results, moderate z.

(c) Estimation results, high z.

(d) Control results.

Figure 3.2: Performance of clusterwise estimation and control procedure with merging. Figures (a), (b) and (c) summarize the estimation procedure for three choices of z. The horizontal axis corresponds to the ratio z_0/z, as the merge threshold z_0 varies in $[0, z]$. $z_0 = z$ corresponds to no merging. When no merging is done, the estimator greatly underestimates the true FDR. Performance is uniformly good for $z_0 \in [0, 0.6z]$. Figure (d) summarizes the performance of the FDR control procedure with $z_0 = 0.3z_{min}$, where z_{min} is the lowest value of z considered by the control procedure. The FDR is controlled below the target level. A $\tilde{\pi}_0$-adjusted FDR curve is also shown. We see that the adjusted FDR is extremely close to the target.

Simulation results not shown here suggest that our findings are fairly robust to changes in signal strength, extent, and bandwidth of smoother.

3.2 False discovery exceedance

The false discovery exceedance (FDX) is an FDP-related error criterion introduced in [48] and [19]. For $c, \alpha \in (0, 1]$, a (c, α)-*exceedance control* procedure is one that selects the rejection set to satisfy

$$\mathbb{P}(V/R \geq c) \leq \alpha.$$

In other words, instead of controlling the expected FDP, we seek to control the probability that the FDP exceeds a certain level. The quantity $\mathbb{P}(V/R \geq c)$ is the FDX.

Several procedures have been proposed for controlling the FDX in the standard multiple testing setting. [36] also give an exceedance control procedure for the spatial multiple testing problem they consider. Their approach uses a confidence envelope method they refer to as *inversion*. The control procedure we present in this section can be viewed as an extension of the *augmentation method* introduced in van der Laan et al. [48]. In [20] the authors show that, in the standard multiple testing setting, under general conditions there is a one-to-one correspondence between augmentation method and inversion method.

We require one further definition before stating the first result of this section. We recall that the k-FWER of a multiple testing procedure is defined as $\mathbb{P}(V > k)$. Note that the 0-FWER is simply the standard FWER. With this definition, our first result is stated below. The proof is given in Appendix A.

Theorem 3.2. *Suppose that the procedure $z_k(\alpha, \tilde{y})$ controls the k-FWER at level α. Take $z_{min} \leq z^* \leq z_k(\alpha, \tilde{y})$, and let A be the set of clusters rejected at level z^* that do not intersect clusters rejected at level $z_k(\alpha, \tilde{y})$. Then,*

$$\mathbb{P}\left(\frac{V_{z^*}}{R_{z^*}} > c\right) \leq \alpha$$

where $c = \frac{|A|+k}{R_{z^*}}$.

Note that this particular result does not directly rely on the Poisson clumping machinery that underlies the FDR estimation and control procedures. While FWER control in the spatial testing setting is generally not as simple as in the standard multiple testing setting, this result can be applied in conjunction with the 0-FWER control procedures discussed in Section 2.2. Furthermore, when the approximation $V_z \sim \text{Poisson}(\lambda_z)$ assumed throughout the paper does hold, we can easily and directly control the k-FWER by taking,

$$z_k(\alpha) = \min \left\{ z : 1 - \sum_{j=0}^{k} \frac{\lambda(z)^j e^{-\lambda(z)}}{j!} \leq \alpha \right\}.$$

Theorem 3.2 has one rather unsatisfying feature, which is that the quantity c at which the FDX is controlled is random. The more interesting case in practice is where c is given, and the threshold z^* is selected to control the FDX accordingly. At least in the case where $k = 0$, we can provide an improved result which permits a user-specified c. The proof is once again given in Appendix A.

Theorem 3.3. *Let $c \in (0,1)$ and $\alpha \in (0,1)$ be given. Suppose that the procedure $z_0(\alpha, \tilde{y})$ controls the clusterwise FWER at level α. Define z^* according to,*

$$z^* = \min \left\{ z \in [z_{min}, z_0] : \frac{|A_z|}{R_{z_0} + |A_z|} \leq c \right\}$$

where A_z is the set of clusters rejected at level z that do not intersect clusters rejected at level z_0. Then,

$$\mathbb{P}\left(\frac{V_{z^*}}{R_{z^*}} > c \right) \leq \alpha.$$

3.3 Incorporating cluster size

Thus far we have focussed on just one of the parameters in the mosaic process model of high level excursion sets, the Poisson mean parameter λ. In this section we propose a way of incorporating properties of the cluster size distribution itself. The methods presented here assume that the cluster size distribution is known, or can be reliably estimated through simulation. The distribution is known in the smooth Gaussian case, and a handful of others. When we do know the cluster size distribution, we can construct estimation and control procedures that screen out small excursion sets.

Spatial inference based on excursion set size have appeared elsewhere in the literature. Two references of note are the neuroimaging development described in [15], and the method for identifying differentially methylated regions of the genome proposed in [25]. In [15] the authors assess significance based on excursion set volume, and in [25] the authors look at the area under $\tilde{y}$ along the excursion set. We highlight these papers as their focus was on FDR control; several references for FWER control via cluster size inference are given in Section 1.7.

The intuition for our proposal is quite simple, and is based on the idea of Poisson thinning. First, recall that the mosaic process model for high-level excursion sets decomposes into two parts: (a) the Poisson distribution governing the cluster rate, and (b) the distribution F governing the cluster shape. The fact that the Poisson process centres x_i are independent of the clusters C_i allows us to establish the following useful result, which holds for general ambient dimension.

Proposition 3.1 (Mosaic thinning)**.** *Suppose that* $A = \bigcup_i x_i \oplus A_i$ *is a mosaic process with centres occurring at rate* λ *and sets* $A_i \overset{iid}{\sim} F$*. Let* $\ell \geq 0$ *be a threshold such that* $0 < \rho \equiv \mathbb{P}(|A_i| \geq \ell) \leq 1$*. Under these assumptions, random set defined by*

$$\bigcup_{i:|A_i|\geq\ell} x_i \oplus A_i$$

is a mosaic process with rate $\lambda\rho$ *and set distribution* $\mathcal{L}(A_i \mid |A_i| \geq \ell)$*.*

Proof. Since the sets A_i are independent of the Poisson process selecting the centres x_i, the thresholding procedure removes each x_i independently with probability $1-\rho = \mathbb{P}(|A_i| < \ell)$. This amounts to *Poisson thinning*, and standard theory tells us that the resulting process, $\{x_i : |A_i| \geq \ell\}$ is a Poisson process with rate $\lambda\rho$. By construction, the sets that survive thresholding are distributed according to $\mathcal{L}(A_i | A_i \geq \ell)$. □

In the context of the spatial inference problem, this result effectively states that the essential features of the null cluster process hold also for the thinned process. That is to say, the null clusters that survive thinning remain well approximated by a mosaic process, albeit with a lower rate.

While independence of S and V in the non-thinned case is not in itself sufficient to establish independence of these quantities after thinning, the argument discussed in Section 1.5.1 applies equally well to establish approximate independence of the thinned quantities. The same reasoning applies for independence of the processes $S(z)$ and $V(z)$. This motivates the following extensions of the procedures of Section 1.5. We begin with the estimation procedure.

Corollary 3.1 (Estimation with thinning)**.** *Fix a threshold $\ell \geq 0$, and let V_ℓ, S_ℓ and R_ℓ denote the number of (null, non-null, total, resp.) clusters whose size is at least ℓ. In the setting of Section 1.5, the estimator $\widehat{\mathrm{FDR}} = \lambda\rho/(R_\ell + 1)$ is unbiased for the FDR. More precisely,*

$$\mathbb{E}\left(\frac{\lambda\rho}{R_\ell + 1}\right) = \mathbb{E}\left(\frac{V_\ell}{R_\ell \vee 1}\right).$$

For the estimation procedure it was sufficient to consider a fixed thinning probability. In order to extend the control result, we need to view the thinning probability as a function of z. To this end, we let $\rho(z, \ell) \equiv \mathbb{P}_0(|C_{0,z}| \geq \ell)$ denote the probability that a cluster at level z is of size at least ℓ. The control result is as follows.

Theorem 3.4 (Control with thinning)**.** *Suppose that $\tilde{\epsilon}$ is stationary and ergodic.*[1] *Suppose also that the thinned processes $S_\ell(z)$ and $V_\ell(z)$ are independent. Define $\lambda_\ell(z) \equiv \lambda(z)\rho(z, \ell)$. Then the control procedure of Section 1.5 applied with $\lambda_\ell(z)$ controls the FDR at the target level α.*

We present the proof of this result in A. Our argument is a modification of the original control proof presented in Siegmund et al. [41].

[1] When $\tilde{\epsilon}(t)$ is a random field on $\mathbb{R}^d$, ergodicity is the requirement that all empirical averages limit to the corresponding population average as the observation space D grows to $\mathbb{R}^d$.

Note that the arguments advanced in this section go through effectively unchanged for a larger class of thinning procedures. This is summarized in the omnibus result below.

Theorem 3.5 (General thinning)**.** *Suppose that $\tilde{\epsilon}$ is a stationary ergodic process. Further suppose that the excursion set of the noise process, $\tilde{\epsilon}$ is a mosaic process with rate λ and set distribution F. Let $\delta = \delta(C, y) \in \{0, 1\}$ be a thinning procedure, taking the value 1 if cluster C is retained. If the thinning indicators $\{\delta(C_i, \tilde{\epsilon})\}$ are mutually independent, then,*

(a) The thinned process $\bigcup_{\delta(C_i,\tilde{\epsilon})=1} C_i$ is a mosaic process with rate $\lambda\rho_\delta$, with $\rho_\delta = \mathbb{P}(\delta(C, \tilde{\epsilon}) = 1)$ and set distribution $\mathcal{L}(C \mid \delta(C, \tilde{\epsilon}) = 1)$.

If also the conditions of Corollary 3.1 and Theorem 3.4 are satisfied by $\tilde{\epsilon}$, and $\{\delta(C_i, \tilde{y})\}_{C_i \in \mathcal{S}}$ are mutually independent of $\{\delta(C_i, \tilde{y})\}_{C_i \in \mathcal{V}}$, then,

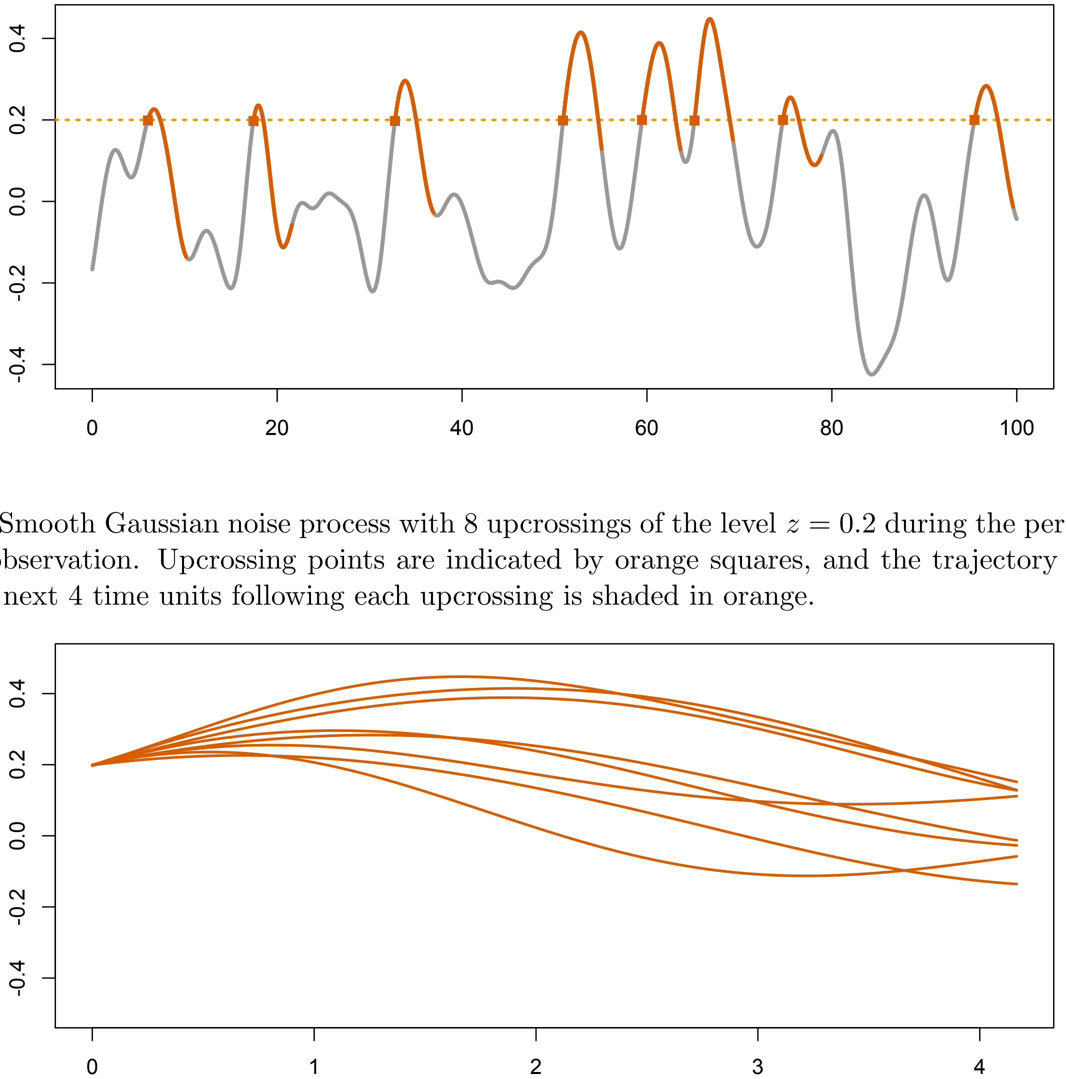
(b) Corollary 3.1 holds for thinning procedure δ and $\rho = \rho_\delta$.

(c) Corollary 3.4 holds for thinning procedure δ with $\rho = \rho(z, \delta) = \mathbb{P}(\delta(C_{0,z}, \tilde{\epsilon}) = 1)$.

Since we are assuming that $\tilde{\epsilon}$ does not exhibit long range dependence, independence of the thinning indicators can be expected to hold if the thinning procedure depends on the data $\tilde{y}$ in only a small neighbourhood of cluster C. The main difficulty in applying Theorem 3.5 with a more complex thinning procedure is that $\rho(z, \delta)$ may be analytically intractable. Of course, all of the necessary quantities can be estimated through simulation, so this does not pose a significant obstacle.

The main application for Theorem 3.5 is that it can be used to define more powerful control procedures that are better tailored to the expected signal structure. For instance, when some components have low amplitudes but large supports, considering peak height alone can lead to low power. By thinning out short regions or low-area regions at high thresholds, the procedure can get down to threshold levels at which the low amplitude signals start to become present.

3.3.1 A note on marked crossings and Slepian model processes



(a) Smooth Gaussian noise process with 8 upcrossings of the level $z = 0.2$ during the period of observation. Upcrossing points are indicated by orange squares, and the trajectory for the next 4 time units following each upcrossing is shaded in orange.

(b) Figure shows the 8 observed upcrossing trajectories, translated to the origin. These can be thought of as 8 realizations of a new process, which is defined by the conditional distribution of $\tilde{\epsilon}(t)$ following an upcrossing of level z.

Figure 3.3

Our discussion of thinning is closely connected to the notions of *marked crossings* and *Slepian model processes* that commonly arise in the study of random processes and associated point processes [31]. In stochastic process terminology, a *mark* is anything that can be observed in combination with an event of interest. The events of primary interest to us are (isolated) upcrossings of high thresholds, and the corresponding marks can be any feature of the post-upcrossing trajectory that we may wish to thin on.

More formally, given a stationary differentiable noise process $\tilde{\epsilon}(t)$, and a jointly stationary vector process $\xi(t) = (\xi_1(t), \ldots, \xi_p(t))$, the associated *mark process* is simply defined as $\mathbf{m}(t) \equiv (\tilde{\epsilon}'(t), \xi(t))$. We allow $\xi(t)$ to be infinite dimensional, in which case it takes the form $\xi(t) = (\xi_\beta(t))_{\beta \in I}$ for an possibly uncountably infinite index set I. Mark processes allow us to formally view thinning functions as indicators that the mark process at an upcrossing belongs to some Borel set. A couple of examples of marks and restrictions are given below.

- *Derivative at crossing.* $\mathbf{m}(t) = \tilde{\epsilon}'(t)$, with the restriction that $\tilde{\epsilon}'(t) \geq u > 0$.
- *Cluster length.* Here the relevant mark process is $\mathbf{m}(t) = (\tilde{\epsilon}'(t), \tilde{\epsilon}(t+s), 0 \leq t \leq \ell)$, with the restriction that $\tilde{\epsilon}'(t) > 0$ and $\min_{0<s<\ell} \tilde{\epsilon}(t+s) > z$.

In this framework, if B is a Borel set of continuous functions, and t_0 is an upcrossing time, then thinning essentially amounts to determining whether the translated process $\tilde{\epsilon}(t_0 + \cdot)$ belongs to A.

One way of computing thinning probabilities is therefore to carry out calculations with respect to the conditional distribution of $\tilde{\epsilon}(t_0 + \cdot)$, conditional on there being an upcrossing of z at time t_0.[2] Assuming that $\tilde{\epsilon}(t)$ is stationary and ergodic, and letting $t_1, t_2, \ldots$ denote the upcrossing locations, the thinning probability is formally defined as the long run limit,

$$\rho(z, A) \equiv \lim_{T \to \infty} \frac{\#\{t_k \in [0,T] \,:\, \tilde{\epsilon}(t_k + \cdot) \in A\}}{\#\{t_k \in [0,T]\}}. \tag{3.3.1}$$

[2] Since we're conditioning on a probability-0 event, the conditional distribution is formally treated via the theory of *Palm distributions*. We already encountered Palm distributions in Section 2.3.

Figure 3.3 shows a realization of $\tilde{\epsilon}(t)$ along with the translated processes $\tilde{\epsilon}(t_k + \cdot)$ that feature in the preceding definition. It is often straightforward to simulate the smoothed noise process $\tilde{\epsilon}(t)$ and to estimate (3.3.1) by calculating the ratio for a large value of T (or across multiple independent realizations of the process).

Furthermore, for certain kinds of processes, the conditional process $\tilde{\epsilon}(t_0 + \cdot)$ can be described explicitly via a *Slepian model* [31, §8.4]. In basic terms, a Slepian model is a simple and explicitly defined process that has the same distribution as the conditional process of interest. Slepian model processes are well understood for smooth stationary Gaussian (and related) processes [29, 30], non-stationary Gaussian processes [18], and recent work has also characterized the process in certain non-Gaussian settings [37]. A simple consequence of the Gaussian Slepian model is the high threshold cluster size distribution summarized in the Example below.

While calculating thinning probabilities via simulation is often the more expedient approach, Slepian model processes offer a principled analytic alternative. The analytic solution may be preferable in cases where the resulting procedure is to be implemented as part of a pipeline or package for analyzing a given type of data.

Example: *Simple Gaussian case* [2, §C23], [31, Example 8.1]. Suppose that $\tilde{\epsilon}(t)$ is a mean-zero smooth Gaussian process with variance 1 and correlation function r that satisfies

$$r(t) = \mathbb{E}\tilde{\epsilon}_0\tilde{\epsilon}_t \sim 1 - \frac{1}{2}\theta t^2$$

for small t. In this setting, the cluster lengths C_z of $\{t : \tilde{\epsilon}_t \geq z\}$ are approximately distributed as,

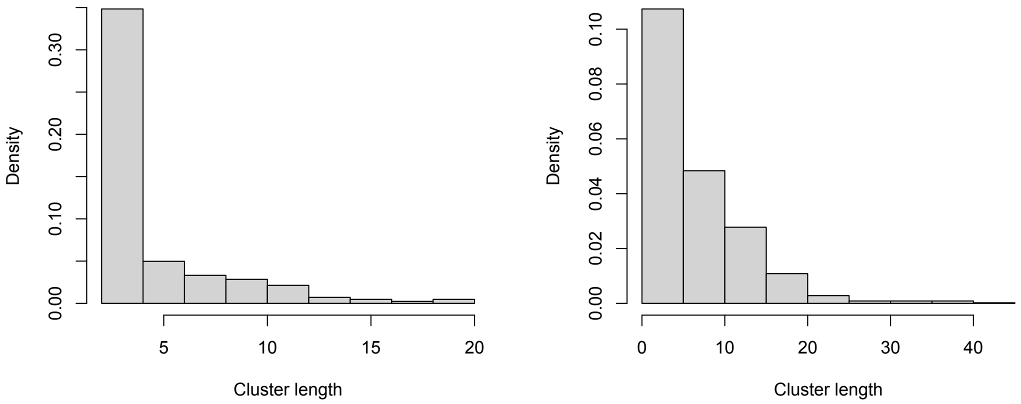
$$C_z \sim \frac{2}{z\sqrt{\theta}} C,$$

where C has the Raleigh distribution: $f_C(x) = xe^{-x^2/2}$, defined for $x > 0$.

3.3.2 Experiments

In this section we present the results of a simulation study conducted to investigate the performance of the FDR control procedure when thinning based on cluster size. All results shown correspond to a procedure that considers only those clusters whose length is $\geq \ell$. Note that the basic procedure of Section 1.5 is recovered by taking $\ell = 1$. We revisit the recurring example where $\mu(t)$ consists of $K = 10$ box functions of equal support and amplitude, and the noise is iid Gaussian. It is observed that thinning can lead to considerable increases in power over the base procedure.

Figure 3.4 shows histograms of the null cluster size distribution for two choices of selection threshold z. In our analysis we consider cluster size cutoffs from $\ell = 1$ (no thinning) to $\ell = 12$. Taking $\ell = 12$ results in thinning out 97% of clusters at the highest value of z that we consider, and 65% of the clusters at the lowest of value of z that we consider. The case $\ell = 12$ should therefore be thought of as an aggressive thinning strategy.

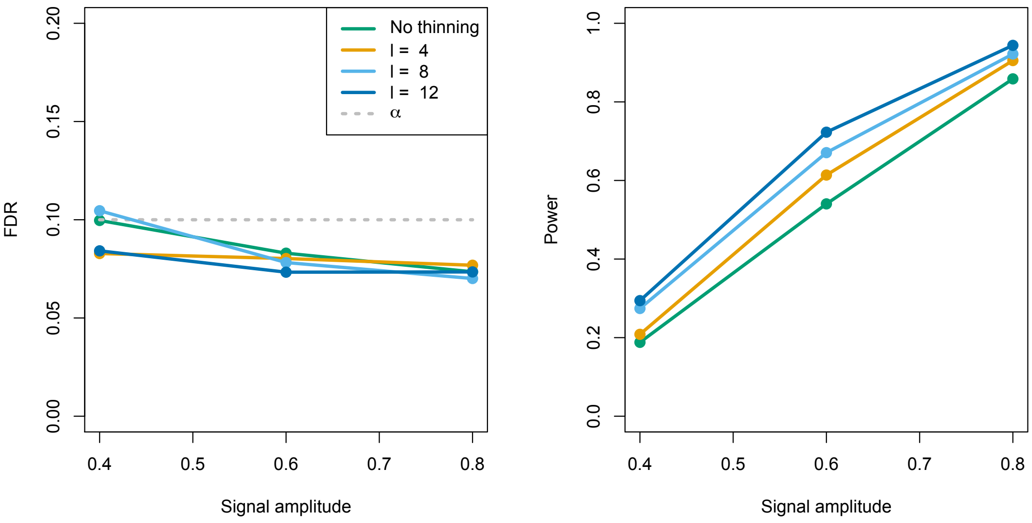


(a) High z threshold.

(b) Low z threshold.

Figure 3.4: Histograms showing the distribution of the null cluster size for two choices of threshold z.

Figure 3.5 shows the observed FDR and average power of the thinned procedure in the equal amplitude setting. All simulations shown are conducted taking $\alpha = 0.1$. Results for other choices of α were found to be qualitatively the same. We find that



(a) Observed clusterwise FDR.

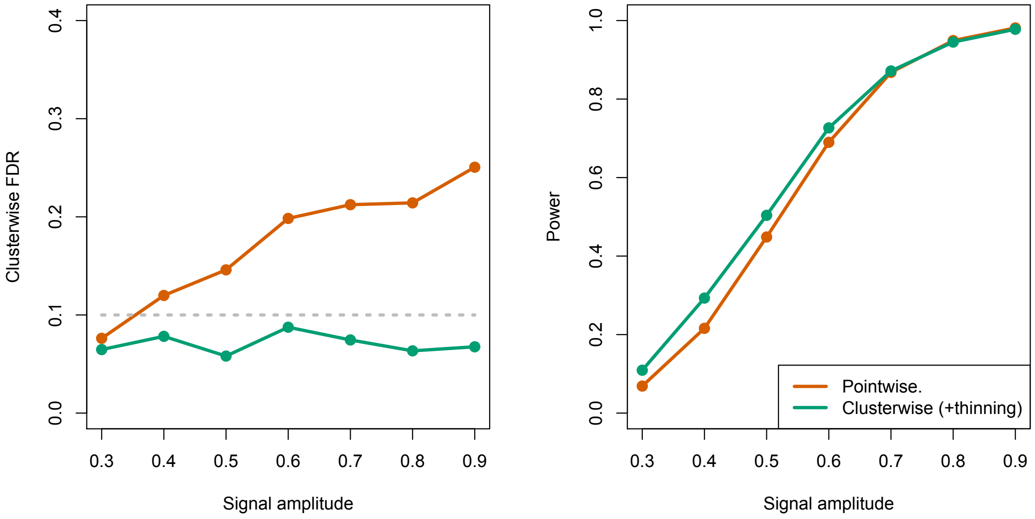
(b) Observed average power.

Figure 3.5: Observed clusterwise FDR and average power of the thinned procedure for the simple example where the components of the signal are box functions having equal amplitude and equal support. The four different curves represent four choices of length cutoff parameter, ℓ. We see that the FDR is controlled at the target level across all values of the signal amplitude and across all choices of length cutoff. We also find that thinning results in a more powerful procedure; the higher the cutoff ℓ, the greater the power.

the FDR is controlled at the target level for all choices of signal amplitude and all choices of length cutoff ℓ. Moreover, we observe that larger values of ℓ resulted in greater power gains.

Figure 3.6 shows how the clusterwise procedure with thinning at $\ell = 12$ compares in terms of FDR control and power to the pointwise procedure. The results of this simulation are extremely encouraging. Not only does the thinned procedure successfully control clusterwise FDR, but in this example it is also *at least as powerful* as the pointwise procedure. Indeed, for low-to-moderate values of signal strength, the thinned clusterwise procedure has *higher* power than the pointwise procedure.

In Section 2.4 we found that the pointwise procedure had higher power than the base clusterwise procedure. Now we see that by thinning we are able to both retain control of the clusterwise FDR, and to match (or exceed) the power of the pointwise procedure.



(a) Observed clusterwise FDR.

(b) Observed average power.

Figure 3.6: Comparison of clusterwise FDR and average power for the pointwise procedure and the thinned clusterwise procedure ($\ell = 12$). The power of the thinned clusterwise procedure is at least as high as that of the pointwise procedure. However, while the pointwise procedure is highly anti-conservative, the thinned clusterwise procedure does control the clusterwise FDR.

3.4 Extension to non-homogeneous cluster rates

In our discussion thus far we have been assuming that the smoothed noise process $\tilde{\epsilon}(t)$ is stationary. While there are many problem settings for which this is a reasonable assumption, it is useful to note that the methods we can discuss are applicable even when $\tilde{\epsilon}(t)$ in non-stationary. Such extensions are of interest in cases where the original noise process $\epsilon(t)$ is non-stationary, and also when the sampling density of the observation locations is non-uniform. The clusterwise FDR methods can be extended to both of these problems settings, provided that a non-homogeneous version of the Poisson clumping heuristic can be assumed to apply [see, e.g., 2, §C27, §D18].

In the non-homogeneous setting, the false cluster rate is viewed as being a function of both the threshold level z and the location $t \in D$. For the duration of this section we will adopt the notation $\lambda_z(t)$ to refer to the false cluster *rate* at threshold height z. Our model for the occurrence of false clusters in the non-homogeneous setting is stated below. Note that both the Poisson rate parameter and the cluster distribution are now allowed to depend on the location t.

Definition (Non-homogeneous mosaic process)**.** Let $\{F_t\}$ be a family of distributions on sets in $\mathbb{R}^d$. Think of each F_t as generating small sets located near the origin 0. A *non-homogeneous mosaic processes* is described by the following procedure

1. Generate points $x_1, x_2, \ldots$ according to a non-homogeneous Poisson process with rate function $\lambda(t)$ on $\mathbb{R}^d$.

2. Generate random sets $A_1, A_2, \ldots$ independently such that $A_i \sim F_{x_i}$

3. Output the random set
$$A = \bigcup_i x_i \oplus A_i,$$
which is the union of the sets A_i shifted to be centred at the points x_i.

Consider a region D for which the non-homogeneous mosaic process approximation holds, with false cluster rate given by $\lambda_z(t)$. At a given value of z, let $V(D)$, $S(D)$ and $R(D)$ denote the multiple testing summary statistics for the region D. Define the function Λ_z according to,

$$\Lambda_z(D) = \int_D \lambda_z(t)dt.$$

Under the global null, the non-homogeneous mosaic process approximation tells us that $V(D) \sim \text{Poisson}(\Lambda_z(D))$. In other words, the function $\Lambda_z(D)$ is the non-homogeneous generalization of the mean parameter function λ_z. The generalizations of the clusterwise FDR estimation and control procedures are given below.

3.4.1 Non-homogeneous FDR estimation and control procedures

Corollary 3.2 (Non-homogeneous estimation). *Given a thresholding $z > z_{min}$, suppose that $V(D)$ and $S(D)$ are independent, and $V(D) \sim \text{Poisson}\,(\Lambda_z(D))$. Then,*

$$\widehat{\text{FDR}}(D) = \frac{\Lambda_z(D)}{R(D) + 1}$$

is an unbiased estimate of the FDR, $\mathbb{E}(V(D)/R(D))$.

Corollary 3.3 (Non-homogeneous control). *Suppose that the processes $S_\Lambda(D)$ and $V_\Lambda(D)$, viewed as functions of Λ, are independent. Also suppose that $V_\Lambda(D)$ is a rate-1 Poisson process on $\Lambda \in [0, \bar{\Lambda}]$ for some $\bar{\Lambda} > 0$. Define $\Lambda^* = \max\{0 \leq \Lambda_z(D) \leq \bar{\Lambda} : \Lambda_z(D)/R(D) \leq \alpha\}$. The procedure that thresholds at level z^* corresponding to Λ^* controls the clusterwise FDR.*

3.4.2 Non-uniform sampling

As mentioned at the outset of this section, one way in which non-homogeneity can arise is if the observation locations t_i are not uniformly sampled across the observation region D. In this section we discuss some considerations that arise in such settings.

To help guide the forthcoming discussion, we begin by introducing two data examples where the measurement locations are generally very far from resembling a uniform sample or an evenly spaced grid in $\mathbb{R}^d$. While both examples are best motivated when $d > 1$, it will be helpful to keep in mind stylized 1-dimensional versions of the examples. Our primary goal here is to develop some heuristics for a reasonable smoothing strategy, and to gain some understanding of how the smoothing strategy affects the noise $\tilde{\epsilon}(t)$.

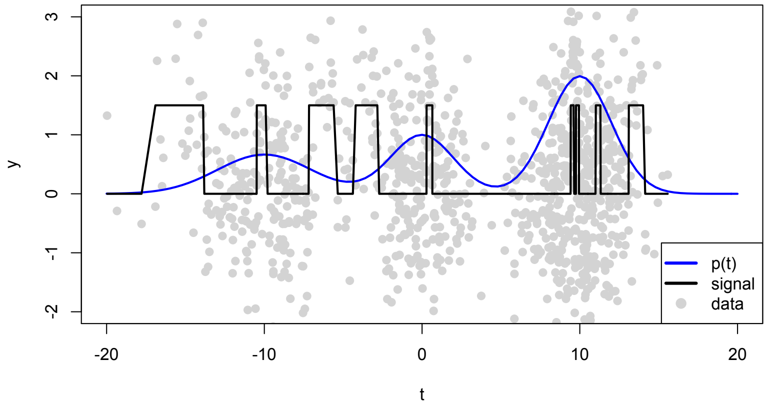


Figure 3.7: 1-dimensional version of Example 1 of §3.4. Blue curve $p(t)$ is the (scaled) underlying density of the sampled points. Regions with higher density are more densely sampled than regions of low density. The support of the signal is a union of 9 intervals, each of which is 21 time points wide. The intervals are therefore longer in sparsely sampled areas and shorter in densely sampled areas.

Example 1: Health survey data. Consider a study in which we obtain information on non-infectious disease incidence at the granularity level of, say, 5-digit zipcodes. The goal may be to identify regions where the disease incidence is significantly elevated. We can expect that densely populated regions will be more densely sampled than rural regions. Furthermore, it may be reasonable to assume that clusters will be smaller in the densely populated regions than in in the rural ones. This reflects the fact that, from an epidemiological standpoint, two people who live one mile apart in a city are likely to have dramatically different exposure, while two people living the same distance apart in a rural area are likely to have very similar exposure.

A good smoothing strategy in this setting would therefore be to use a smaller bandwidth in densely populated areas and a larger bandwidth in sparsely populated areas. Following this strategy, we would expect that the Poisson rate $\lambda_z(t)$ will be higher in more densely populated regions, and also that the set distribution F_t will be concentrated on small sets for t in densely populated regions, and large sets for t in sparsely populated regions.

Example 2: Sensor network. Consider a setting where a sensor network is monitoring the level of some factor for signs of anomalies. One such example might be a water sensor network monitoring levels of a particular contaminant in a large body of water. We might expect that the sensors are randomly but non-uniformly located throughout the region, and that the density of sensors in an area is not related to the size of anomaly we expect to see in that area.

A reasonable smoothing strategy in this case may be to assign to each location t a value equal to the local average across all sensors within a given radius, or to consider a distance-weighted average of all sensors. In this case we likely expect $\lambda_z(t)$ to vary with location, but the set distribution F_t may remain fairly stable.

3.4.3 Simple 1-d model of Example 1

It is instructive to consider a simple model 1-dimensional model having the structure of Example 1. For this model we are able to give an explicit calculation of the rate parameter $\lambda_z(t)$ in terms of a homogeneous parameter in a related problem.

We suppose that we observe a signal corrupted by white noise at observation points $t(i) \sim p(t)$, $i = 1, \ldots, T$. In order to guide our choice of scan statistic, we further assume that the size (length) of the bumps we are seeking to detect is inversely proportional to the sampling density $p(t)$. This roughly equates to assuming that the support of each bump spans a fixed number of observation points.

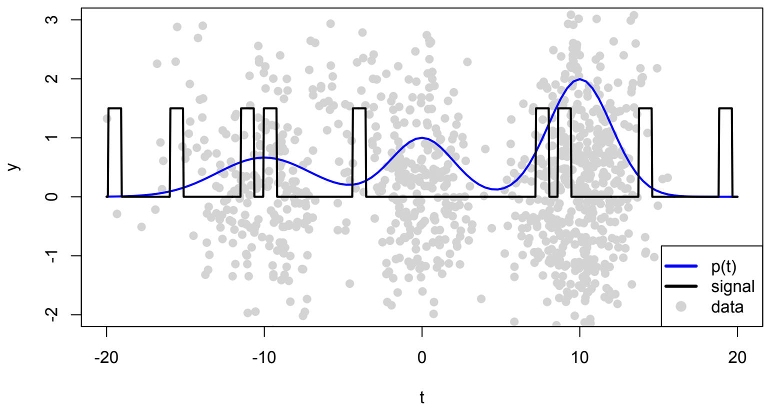


Figure 3.8: 1-dimensional version of Example 2 of §3.4. Blue curve $p(t)$ is the (scaled) underlying density of the sampled points. Regions with higher density are more densely sampled than regions of low density. The support of the signal is a union of 9 intervals, each of which is 0.84 time units wide. The size of the cluster in this case is independent of the underlying density $p(t)$.

More precisely, we consider the model

$$y(t) = \mu(t) + \epsilon(t),$$

where $\epsilon(t)$ is stationary independent noise, and the signal μ is a sparse train of unimodal positive peaks described by,

$$\mu(t) = \sum_{k=1}^{K} a_k h_k(t).$$

The number of peaks K and the amplitudes $a_k > 0$ are assumed unknown. Our assumption that the length of the bumps is inversely proportional to the sampling density $p(t)$ can be summarized by writing,

$$h_k(t) = h\left(\frac{t - t_k}{b/p(t_k)}\right),$$

where h is a compactly supported function. For instance, if h is the indicator function of the interval $[-1, 1]$, this is saying that each h_k is the indicator of an interval having equal probability measure under p. Equivalently, each h_k is supported on roughly the same number of observation points t_i.

A common choice of scan statistic in this setting is some form of local averaging. To keep things simple, we'll take as our scan statistic the simple $2w$-nearest neighbour average,

$$\tilde{y}(t_i) = \frac{1}{2w+1} \sum_{i-w}^{i+w} Y(t_i), \qquad i = 1, \ldots, T.$$

For the time being we'll assume that we have a good guess at what w should be.

Question: What is the false cluster rate for the smoothed noise process $\tilde{\epsilon}(t)$ constructed using this choice of smoother?

We will show that $\lambda(t) \propto p(t)$. To see this, let $P(t) = \int_{-\infty}^{t} p(s)ds$ and, fixing a threshold, denote by $\Lambda(t)$ the expected number of false clusters in the interval $[0, t]$. If we rescale time according to $\tilde{t}_i = TP(t_i)$, then the observation points $\tilde{t}_i$ are uniformly distributed on the interval $[0, T]$. In this time scaling, the false cluster rate of the scan statistic turns out to be constant. Denoting the false cluster rate in the rescaled domain by $\tilde{\lambda}$, we get that

$$\Lambda(t) = \int_0^t \lambda(s)ds = \int_0^{TP(t)} \tilde{\lambda}ds = \tilde{\lambda}TP(t).$$

Thus the local false cluster rate in the original domain is given by,

$$\lambda(t) = \frac{d}{dt}\Lambda(t) = \tilde{\lambda}Tp(t).$$

Integrating over the whole region of observation, we get back to the familiar result that the number of false clusters is Poisson distributed with mean $\tilde{\lambda}T$ (more precisely, $\pi_0\tilde{\lambda}T$).

3.5 Stratification

Suppose that the observation region is partitioned into J strata, such as in Figure 3.9. We can think of these strata as corresponding to predefined regions of interest (e.g., individual chromosomes, US states, functional regions of the brain, etc.). Each stratum should be large relative to the expected cluster size, but can be small relative to the full observation space. When working with strata, it may be of interest to allow a different selection threshold within each stratum. In this section we show how the FDR estimation and control procedures can be extended to the stratified setting.

A strong case for performing false discovery analyses separately within each stratum is made in Efron [16, Chapter 10]. One setting in which a separated analysis is to be preferred is when some strata contain a considerably higher rate of signal components. To be concrete, consider a setting in which there are two equally sized strata, one of which contains regions where $\mu(t) > 0$, while the other does not. All of the true discoveries would therefore come from just one of the strata, while both would contribute equally to the false detection count. By treating the strata separately we would effectively cut the null parameter λ in half, thereby enabling ourselves to make more discoveries in the signal-enriched stratum.

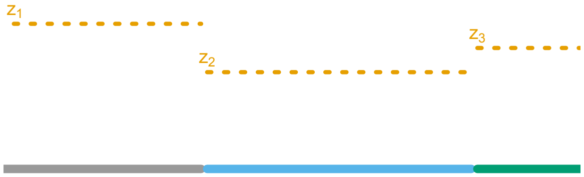


Figure 3.9: Schematic of three different strata across which the values of z_i, $i \in \{1, 2, 3\}$ can differ.

3.5.1 FDR estimation

Suppose that in stratum j we threshold at level z_j to obtain R_j discoveries. Let V_j denote the number of false discoveries in stratum j. We will assume that the $\{S_j\}$ and

$\{V_j\}$ are independent. The mosaic process assumption tells us that $V_j \sim \text{Pois}(\lambda_j)$, with $\lambda_j = \lambda(z_j, |\text{stratum}_j|)$. In this setup we can obtain unbiased stratrum-specific estimates in addition to an overall estimate of FDR. A version of this result appears in Efron [16, §10.4] in the context of combining analyses across strata within an empirical Bayes framework.

Theorem 3.6. *Under the assumptions above,*

$$\widehat{\text{FDR}}_j = \frac{\lambda_j}{R_j + 1}$$

is an unbiased estimate of the FDR within stratum j. Furthermore, if the V_j are independent, then

$$\widehat{\text{FDR}} = \frac{\sum_{j=1}^{J} \lambda_j}{1 + \sum_{j=1}^{J} R_j}$$

is an unbiased estimate of the overall FDR for the entire observation region.

Proof. The first claim follows by applying the estimation result of Section 1.5 directly to V_j and S_j. To establish the second result, note that under independence $V = \sum_{j=1}^{J} V_j \sim \text{Poisson}(\sum \lambda_j)$. By assumption, $S = \sum_{j=1}^{J} S_j$ and V are independent, so the same estimation result applies. □

We note here that the finer assumption of independence between the $\{S_j\}$ and $\{V_j\}$ should hold whenever the strata are large relative to the expected cluster size and the the aggregate quantities S and V can reasonably be assumed to be independent. As discussed in Section 1.5.1, the typical argument for independence of S and V relies on sparsity of the signal region D_1 and short-range dependence between $\tilde{y}(t)$ and $\tilde{y}(t')$ for $t \in D_0$ and $t' \in D_1$. The same type argument applies equally well to establish approximate independence of the stratum-specific quantities.

3.5.2 FDR control

FDR control is a bit more nuanced because in general controlling the FDR at level α in each stratum does not imply overall control at level α. However, the fact that $z \mapsto \lambda(z)$ is a 1-to-1 invertible function allows us to derive a simple procedure for controlling overall FDR while varying the thresholds across the strata. The proposed approach is effectively to tether together thresholds $(z_1, \ldots, z_J)$ and thus reduce the multivariate problem to a univariate one. *Note: For the purpose of this discussion we will think of $\lambda(z)$ being the Poisson process* <u>*rate*</u> *parameter, instead of the mean parameter.*

Given weights $w_j > 0$ with $\sum_{j=1}^J w_j = 1$ and overall rate λ, we can select z_j in each stratum so that $V_j \sim \text{Poisson}(w_j|D|\lambda)$. This is done by setting,

$$z_j(\lambda) = \lambda^{-1}\left(\lambda w_j \frac{|D|}{|\text{stratum}_j|}\right). \tag{3.5.1}$$

Given a rate λ, the aggregated false discovery process is given by,

$$V = \sum_{j=1}^J V_j \sim \text{Poisson}\left(\sum_{j=1}^J \mathbb{E}V_j\right),$$

where, by construction, we have that,

$$\begin{aligned}\sum_{j=1}^J \mathbb{E}V_j &= \sum_{j=1}^J |\text{stratum}_j|\lambda_j \\ &= \sum_{j=1}^J |\text{stratum}_j|\lambda w_j \frac{|D|}{|\text{stratum}_j|} \\ &= \lambda|D|\sum_{j=1}^J w_j \\ &= \lambda|D|.\end{aligned}$$

We can therefore apply the standard control procedure to the overall rate λ to determine cutoffs $(z_1, \ldots, z_J)$. This is summarized in the following result.

Theorem 3.7. *Let* $\Lambda = \max\{\lambda \leq \bar{\lambda} : \lambda|D|/R_\lambda \leq \alpha\}$*. Under the assumptions of Section 3.5, thresholding within stratum j at level $z_j^* = z_j(\Lambda)$ defined in* (3.5.1) *controls the overall FDR at level α.*

This procedure will result in improved power if we assign large w_j to regions that are expected to be enriched. We may also wish to control FDR at different levels when the real cost of false negatives and false positives varies across different strata.

Chapter 4

Conclusion

4.1 Summary

In this thesis we studied the spatial inference problem of identifying the support regions of a noisily observed sparse signal while controlling the clusterwise false discovery rate. Borrowing ideas from the Poisson clumping heuristic literature, we showed that the widely used *pointwise procedure* generally fails to control the clusterwise FDR.

We presented several extensions of the clusterwise FDR control procedure introduced by Siegmund, Zhang, and Yakir [41]. Our methods may be applied whenever the excursion set of the smoothed noise process is well approximated by a (potentially inhomogeneous) mosaic process. As one of our extensions we described a general framework for incorporating various measures of cluster significance into the FDR procedure. We showed that by incorporating cluster size we can obtain a significant increase in power. In particular, we showed that the augmented procedure can have better power than even the pointwise procedure, while still controlling the clusterwise FDR.

4.2 Future directions

4.2.1 Detection and power

One of the assumptions implicit in the spatial inference problem is that the signal-to-noise ratio (SNR) is too low to be able to directly estimate the support of $\mu(t)$ via a support recovery algorithm. An interesting problem is therefore to understand when detection is feasible via the clusterwise FDR procedure and its extensions, and how the FDR detection threshold compares to that which is necessary for support recovery. There is a growing body of literature investigating detection thresholds for problems related to the one we consider here [3, 5, 4, 6]. The recent work of Cai and Yuan [13] is particularly relevant.

A related problem is that of power. The key question is as follows: Given that a signal with multiple support components is detectable, what is the procedure that maximizes (average) power? Our interest here is in understanding how we should choose the smoother and thinning procedure to maximize the number of signal components detected subject to a bound on the FDR of the procedure. Schwartzman et al. [39] address a simpler version of this problem within the context of the STEM procedure. The work of Walther et al. [49, 50] provides a different perspective.

4.2.2 Higher dimensions

While much of the methodology developed in this thesis theoretically applies when the observation region D is a subset of $\mathbb{R}^d$ for $d > 1$, the analysis becomes considerably more difficult in practice. Analytical approximations for the parameter λ_z become more difficult to obtain. It also becomes less clear what a cluster should/will look like.

Signal shape starts to play a bigger role in the analysis. As an example, isotropic smoothers that blur edges will have great difficulty detecting filamentary signal components. When discussing detection and power for higher dimensional spatial signals, the theory can be very different depending on the assumed shape of the signal components.

4.2.3 Spatial scan statistic

One of the most popular methods for anomaly and cluster detection is the spatial scan statistic pioneered by Kulldorff [26]. In highly simplified terms, the typical analysis proceeds as follows.

1. Begin with data $y(t)$ for $t \in D$. (This could be e.g., count data, presence/absence indicators, or real-valued measurements.)
2. Select a set of scanning windows $\{W_i\}_{i \in I}$. (Typically these are rectangles or circles of possibly varying size, centered at observation locations t.)
3. For each W_i, compute a test statistic T_i.
4. Obtain a p-value, p_i for each test statistic T_i.
5. If any $p_i < \alpha$, report the window W_{i^*} as a discovery, where i^* is the index of the smallest (most significant) p_i. This is often called the MLC (most likely cluster).
6. If any $p_j < \alpha$ for W_j that do not overlap W_{i^*}, report W_{j^*} as a discovery, where j^* is the index of the smallest p_j among windows j that do not overlap W_{i^*}.

The final step is repeated until all significant p-values remaining correspond to windows that overlap with one of the previously reported windows. This procedure controls the FWER under the global null. We give two examples of commonly used window types and test statistics below.

Example 1. Gaussian scan statistic for continuous data. In Kulldorff et al. [27], the authors introduce a scan statistic based on a Gaussian probability model for the measurements $y(t_i)$ observed at each spatial location $t_i \in D$ with D a finite index set of observation locations. They consider windows W_i which are taken to be circles of a fixed radius centered at at the points observation locations t_i. In the notation introduced above, this corresponds to taking $I = D$, and $W_i = \{t_j \in D : \|t_j - t_i\|_2\}$.

The test statistic T_i is taken to the the log-likelihood ratio for comparing the mean in circle W_i to the mean outside of W_i. Letting $N = |D|$ denote the number of locations at which measurements were observed, the authors show that the likelihood ratio statistic for a circle W_i reduces to,

$$T_i = -N \log(\sqrt{2\pi}) - N \log(\sqrt{\hat{\sigma}^2(W_i)}) - N/2,$$

where $\hat{\sigma}^2(W_i)$ is the MLE for the common variance in the model where the $y(t_j)$ are generated independently with distribution,

$$y(t_j) \sim \begin{cases} N(\mu_1, \sigma^2(W_i)) & \text{for } t_j \in W_i \\ N(\mu_2, \sigma^2(W_i)) & \text{for } t_j \notin W_i \end{cases}.$$

Example 2. Bernoulli model for identifying spatial clusters of events. In Kulldorff [26], the author presents a spatial scan statistic for identifying spatial clusters of events. He we observe locations t_i at which a certain type of event occurred, and it is of interest to determine if there are any regions where events appear to 'cluster' or group together in a manner that is not consistent with the baseline event rate. This type of problem arises, for instance, if one is interested in identifying forest regions that have a high occurrence of a particular species of plant or wildlife. Alternatively, one may be interested in geographical clusters of disease that are not consistent with chance fluctuations.

The author proposes several possibilities for the choice of windows. These include, (a) circles of radius $\leq r$ centered at all points of a fixed grid; and (b) all rectangles of a fixed size. The test statistic in this case is given by the likelihood ratio for comparing the event rate in window W_i to that outside of W_i. Details on the particular form of the likelihood ratio are given in [26, §3.1].

Discussion. A different way of thinking about the spatial scan statistic is in terms of a variable bandwidth smoother applied to the data $y(t)$. Instead of looking at a bag of test statistics $\{T_i\}$, we could look at a smoothed version of the data given by,

$$\tilde{y}(t) = \max_{i:t \in W_i} T_i.$$

Assuming that the maximum window size is small relative to the size of the observation region D, the mosaic process approximation will apply, and so would the FDR methodology developed in this thesis. It would be interesting to study how the clusterwise FDR methodology carries over to the variable bandwidth problem, and to investigate the potential power gains over the FWER control approach.

Appendix A

Proofs

Proposition A.1 (Global null.)**.** *Take $D = \{1, 2, \ldots, T\}$, and suppose $\mu(t) \equiv 0$. Let $\tau > 0$ be an integer. Suppose that the smoother is chosen so that $\tilde{\epsilon}(t)$ is a stationary sequence. Given a threshold $z > 0$, define*

(a) $X_t(z) = \mathbb{1}\{\tilde{\epsilon}(t) \geq z \text{ and } \tilde{\epsilon}(s) < z \; \forall \, t - \tau \leq s < t\}$

(b) $p_1(z) = \mathbb{P}(X_t(z) = 1) = \mathbb{E}(X_t(z))$

(c) $a_2(z) = \mathbb{E} \left| \mathbb{E} \left(X_t(z) - p_1(z) \mid X_s, |s - t| > \tau \right) \right|$

Let W be a Poisson random variable with mean $\lambda(z) = Tp_1(z)$, and set $\tilde{V}_z = \sum_{i=1}^{T} X_t(z)$. $\tilde{V}_z$ satisfies,

$$\begin{aligned} \|\mathcal{L}(\tilde{V}_z) - \mathcal{L}(W)\|_{\mathrm{TV}} &\leq 2T(2\tau p_1^2(z) + a_2(z)) \\ &= 2\left(\frac{2\tau\lambda(z)^2}{T} + Ta_2(z)\right) \end{aligned} \tag{A.0.1}$$

Proof. We use here the notation of Arratia et al. [8], taking as our neighbourhood $B_t = \{s : |s - t| \leq \tau\}$. By stationarity, we have that $b_1 = 2T\tau p_1(z)$, and $b_3 = Ta_2(z)$. Lastly, since for $s \in B_t$ $X_t(z) = 1 \Rightarrow X_s(z) = 0$ and $X_s(z) = 1 \Rightarrow X_t(z) = 0$, we get $b_2 = 0$. Theorem 1 of [8] gives that the TV distance is bounded by $b_1 + b_2 + b_3$, which we have now shown to be equal to (A.0.1). □

Corollary A.1 (Moving average of an iid sequence.)**.** *In addition to the assumptions of Prop. A.1, suppose that the underlying noise $\epsilon(t)$ is iid, and $\tilde{\epsilon}(t)$ is a moving average taking the form,*

$$\tilde{\epsilon}(t) = \sum_{i=0}^{M} c_i \epsilon(t-i)$$

and constants $c_i \geq 0$ and $M \leq \tau$. Then,

$$\|\mathcal{L}(V_z) - \mathcal{L}(W)\|_{\text{TV}} \leq \frac{4\tau\lambda(z)^2}{T}$$

Proof. By the iid assumption, $\tilde{\epsilon}(t)$ is independent of $\tilde{\epsilon}(s)$ for $|s-t| > M$. Since $M \leq \tau$, we thus have that $X_t(z)$ is independent of $X_s(z)$ for $|s-t| > \tau$, and hence $a_2(z) = 0$. □

Proof of Theorem 3.1. Note that the number of false clusters observed in the presence of signal is bounded by the number observed under the global null. Since Step (2) of the merging procedure can only further decrease the number of false clusters, we have that $V_z \leq \#$ of τ-upcrossings of z by $\tilde{\epsilon}(t) = \tilde{V}_z$, with $\tilde{V}_z$ as defined in Prop A.1. This established (3.1.2).

It remains to show that $EW \equiv Tp_1(z) \approx -\log(1-p^*(z))$, where $p^*(z) = \mathbb{P}(\sup_{1\leq t\leq T} \tilde{\epsilon}(t) \geq z)$. The following argument establishes a bound on $|p^*(z) - e^{-\lambda(z)}|$.

First, note that

$$\left\{ \sup_{1\leq t\leq T} \tilde{\epsilon}(t) < z \right\} \iff \{X_t(z) = 0 \ \forall t\} \iff \{\tilde{V}_z = 0\}.$$

Thus $p^*(z) = \mathbb{P}(\tilde{V}_z = 0)$. By the second part of [8, Theorem 1],

$$\left|p^*(z) - e^{-\lambda(z)}\right| < (1 \vee [1/\lambda(z)]) \left(\frac{2\tau\lambda(z)^2}{T} + Ta_2(z) \right)$$

□

Proof of Theorem 3.2. Let $\mathcal{R}_k$ denote the set of clusters rejected at level $z_k(\tilde{y})$, and let $\mathcal{R}_*$ denote the set of clusters rejected at level z^*. By assumption, we can partition $\mathcal{R}_*$ as $\mathcal{R}_* = \tilde{\mathcal{R}} \cup A$, where we define $\tilde{\mathcal{R}} = \{C \in \mathcal{R}_* : C \cap C' \neq \emptyset \text{ for some } C' \in \mathcal{R}_k\}$.

Since $z^* \leq z_k$, each $C' \in \mathcal{R}_k$ is a subset of some $C \in \tilde{\mathcal{R}}$. Recall that a cluster is a true discovery if it has any intersection with the support of the signal. Thus if $C' \in \mathcal{R}_k$ is a true discovery, then any $C \in \tilde{\mathcal{R}}$ such that $C' \subset C$ is also a true discovery. This implies that,

$$V_{z^*} \leq V_{z_k} + |A|.$$

We therefore deduce that,

$$\begin{aligned}
\mathbb{P}\left(\frac{V_{z^*}}{R_{z^*}} > c\right) &= \mathbb{P}\left(\frac{V_{z^*}}{R_{z^*}} > \frac{|A| + k}{R_{z^*}}\right) \\
&= \mathbb{P}(V_{z^*} > |A| + k) \\
&\leq \mathbb{P}(V_{z_k} > k) \\
&\leq \alpha
\end{aligned}$$

where the final inequality follows from the fact that at threshold z_k the k-FWER is assumed to be controlled at level α. □

Proof of Theorem 3.3. Let $\mathcal{R}_0$ denote the set of clusters rejected at level z_0, and let $\mathcal{R}_*$ denote the set of clusters rejected at level z^*. Also, let A_* denote the set of clusters rejected at level z that do not intersect clusters rejected at level z_0. By assumption, we can partition $\mathcal{R}_*$ as $\mathcal{R}_* = \tilde{\mathcal{R}} \cup A_*$, where we define $\tilde{\mathcal{R}} = \{C \in \mathcal{R}_* : C \cap C' \neq \emptyset \text{ for some } C' \in \mathcal{R}_0\}$.

Since $z^* \leq z_0$, each $C' \in \mathcal{R}_0$ is a subset of some $C \in \tilde{\mathcal{R}}$. Recall that a cluster is a true discovery if it has any intersection with the support of the signal. Thus if $C' \in \mathcal{R}_0$ is a true discovery, then any $C \in \tilde{\mathcal{R}}$ such that $C' \subset C$ is also a true discovery. This implies that,

$$V_{z^*} \leq V_{z_0} + |A_*|.$$

By construction, $R_{z^*} = R_{z_0} + |A_*|$, so also have that,

$$\frac{V_{z^*}}{R_{z^*}} \leq \frac{V_{z_0} + |A_*|}{R_{z_0} + |A_*|}.$$

Observe that on the event $\{V_{z_0} = 0\}$,

$$\frac{V_{z_0} + |A_*|}{R_{z_0} + |A_*|} = \frac{|A_*|}{R_{z_0} + |A_*|} \leq c,$$

where the inequality follows from the definition of z^*.

This allows us to conclude,

$$\begin{aligned}
\mathbb{P}\left(\frac{V_{z^*}}{R_{z^*}} > c\right) &\leq \mathbb{P}\left(\frac{V_{z_0} + |A_*|}{R_{z_0} + |A_*|} > c\right) \\
&= \mathbb{P}\left(\frac{V_{z_0} + |A_*|}{R_{z_0} + |A_*|} > c \middle| \mathbb{1}_{\{V_{z_0}=0\}} = 1\right)\mathbb{P}(V_{z_0} = 0) \\
&\quad + \mathbb{P}\left(\frac{V_{z_0} + |A_*|}{R_{z_0} + |A_*|} > c \middle| \mathbb{1}_{\{V_{z_0}=0\}} = 0\right)\mathbb{P}(V_{z_0} > 0) \\
&\leq 0 \cdot \mathbb{P}(V_{z_0} = 0) + \mathbb{P}(V_{z_0} > 0) \\
&\leq \alpha,
\end{aligned}$$

where the final inequality follows from the assumption that at z_0 the FWER is controlled at level α. □

Proof of Theorem 3.4. Let $\pi_\ell(z_1, z_2)$ denote the conditional probability that a cluster of size $\geq \ell$ at level z_1 contains a cluster of size $\geq \ell$ at level $z_2 \geq z_1$. Since we're assuming that our merge procedure is persistent (see Definition 3.1), a cluster at level z_1 can contain at most one cluster at level $z_2 \geq z_1$. By stationarity and ergodicity of the noise process we can therefore identify the conditional probability of interest with the Palm distribution,

$$\begin{aligned}\pi_\ell(z_1, z_2) &= \frac{\mathbb{E}(\#\{C_{z_2} : |C_{z_2}| \geq \ell\})}{\mathbb{E}(\#\{C_{z_1} : |C_{z_1}| \geq \ell\})} \\ &= \frac{\lambda(z_2)\rho(z_2, \ell)}{\lambda(z_1)\rho(z_2, \ell)} \\ &= \frac{\lambda_\ell(z_2)}{\lambda_\ell(z_1)}.\end{aligned}$$

Next, consider the process $V_{\lambda_\ell(z)}/\lambda_\ell(z)$ on $[z(\bar{\lambda}), \infty)$, where $z(\bar{\lambda})$ is the lowest z threshold at which the Poisson approximation holds for the unthinned process. We show that this process is a martingale on $[z(\bar{\lambda}), \infty)$ with respect to the filtration $\mathcal{F}_z = \sigma(V_\ell(z'), S_\ell(z') : z(\bar{\lambda}) \leq z' \leq z)$.

Observe that, by independence of $S_\ell(z)$ and $V_\ell(z)$ we have that,

$$\mathbb{E}\left(\frac{V_\ell(z_2)}{\lambda_\ell(z_2)}\middle| F_{z_1}\right) = \mathbb{E}\left(\frac{V_\ell(z_2)}{\lambda_\ell(z_2)}\middle| V_\ell(z_1)\right),$$

and hence that,

$$\begin{aligned}\mathbb{E}\left(\frac{V_\ell(z_2)}{\lambda_\ell(z_2)}\middle| F_{z_1}\right) &= \frac{V_{\ell(z_1)}}{\lambda_\ell(z_2)}\pi_\ell(z_1, z_2) \\ &= \frac{V_\ell(z_1)}{\lambda_\ell(z_2)}\frac{\lambda_\ell(z_2)}{\lambda_\ell(z_1)} \\ &= \frac{V_\ell(z_1)}{\lambda_\ell(z_1)}\end{aligned}$$

This establishes that $V_{\lambda_\ell(z)}/\lambda_\ell(z)$ is a martingale. By construction, the stopping time Λ is measurable respect to the filtration $\mathcal{F}_z$, and hence the argument in Siegmund et al. [41, Proof of Theorem 2] goes through to establish control. □

Bibliography

[1] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*, volume 115. Springer, 2009.

[2] David J Aldous. *Probability approximations via the Poisson clumping heuristic.* Springer-Verlag New York, 1989.

[3] Ery Arias-Castro, Emmanuel J Candès, Hannes Helgason, and Ofer Zeitouni. Searching for a trail of evidence in a maze. *The Annals of Statistics*, pages 1726–1757, 2008.

[4] Ery Arias-Castro, Emmanuel J Candès, Arnaud Durand, et al. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011.

[5] Ery Arias-Castro, Emmanuel J Candès, Yaniv Plan, et al. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

[6] Ery Arias-Castro, Geoffrey R Grimmett, et al. Cluster detection in networks using percolation. *Bernoulli*, 19(2):676–719, 2013.

[7] Michael Aronowich and Robert J Adler. Extrema and level crossings of χ 2 processes. *Advances in Applied Probability*, pages 901–920, 1986.

[8] Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, pages 9–25, 1989.

[9] Jean-Marc Azaïs and Mario Wschebor. *Level sets and extrema of random processes and fields*. John Wiley & Sons, 2009.

[10] Yoav Benjamini and Ruth Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008.

[11] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[12] Kirsten MM Beyer and Gerard Rushton. Peer reviewed: Mapping cancer for community engagement. *Preventing chronic disease*, 6(1), 2009.

[13] T Tony Cai and Ming Yuan. Rate-optimal detection of very short signal segments. *arXiv preprint arXiv:1407.2812*, 2014.

[14] J Chumbley, Keith Worsley, Guillaume Flandin, and K Friston. Topological fdr for neuroimaging. *Neuroimage*, 49(4):3057–3064, 2010.

[15] Justin R Chumbley and Karl J Friston. False discovery rate revisited: Fdr and topological inference using gaussian random fields. *Neuroimage*, 44(1):62–70, 2009.

[16] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

[17] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[18] Tamar Gadrich and Robert J Adler. Slepian models for non-stationary gaussian processes. *Journal of applied probability*, pages 98–111, 1993.

[19] Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, pages 1035–1061, 2004.

[20] Christopher R Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476): 1408–1417, 2006.

[21] Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

[22] Satoru Hayasaka and Thomas E Nichols. Combining voxel intensity and cluster extent with permutation test framework. *Neuroimage*, 23(1):54–63, 2004.

[23] Satoru Hayasaka, K Luan Phan, Israel Liberzon, Keith J Worsley, and Thomas E Nichols. Nonstationary cluster-size inference with random field and permutation methods. *Neuroimage*, 22(2):676–687, 2004.

[24] Ruth Heller, Damian Stanley, Daniel Yekutieli, Nava Rubin, and Yoav Benjamini. Cluster-based analysis of fmri data. *NeuroImage*, 33(2):599–608, 2006.

[25] Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–209, 2012.

[26] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[27] Martin Kulldorff, Lan Huang, and Kevin Konty. A scan statistic for continuous data based on the normal probability model. *International journal of health geographics*, 8(1):58, 2009.

[28] M Ross Leadbetter, Georg Lindgren, and Holger Rootzén. Extremes and related properties of random sequences and processes. 1989.

[29] Georg Lindgren. Use and structure of slepian model processes for prediction and detection in crossing and extreme value theory. In *Statistical extremes and applications*, pages 261–284. Springer, 1984.

[30] Georg Lindgren. Slepian models for χ2-processes with dependent components with application to envelope upcrossings. *Journal of Applied Probability*, pages 36–49, 1989.

[31] Georg Lindgren. *Stationary Stochastic Processes: Theory and Applications*. CRC Press, 2012.

[32] Thomas Nichols and Satoru Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5):419–446, 2003.

[33] Thomas E Nichols. Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage*, 62(2):811–815, 2012.

[34] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1): 1–25, 2002.

[35] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images*. Academic Press, 2011.

[36] M Perone Pacifico, C Genovese, I Verdinelli, and L Wasserman. False discovery control for random fields. *Journal of the American Statistical Association*, 99 (468):1002–1014, 2004.

[37] K Podgórski, Igor Rychlik, and J Wallin. Slepian models for moving averages driven by a non-gaussian noise. 2014.

[38] Jean-Baptiste Poline, Keith J Worsley, Alan C Evans, and Karl J Friston. Combining spatial extent and peak intensity to test for activations in functional imaging. *Neuroimage*, 5(2):83–96, 1997.

[39] Armin Schwartzman, Yulia Gavrilov, and Robert J Adler. Multiple testing of local maxima for detection of peaks in 1d. *Annals of statistics*, 39(6):3290, 2011.

[40] David Siegmund and Benjamin Yakir. *The statistics of gene mapping.* Springer, 2007.

[41] DO Siegmund, NR Zhang, and B Yakir. False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985, 2011.

[42] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[43] John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[44] Wenguang Sun, Brian J Reich, T Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014.

[45] Jonathan E Taylor and Keith J Worsley. Detecting sparse signals in random fields, with an application to brain mapping. *Journal of the American Statistical Association*, 102(479):913–928, 2007.

[46] Robert Tibshirani and Pei Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29, 2008.

[47] Viratsinh Vaghela, Chandrasekharan Kesavadas, Bejoy Thomas, et al. Functional magnetic resonance imaging of the brain: A quick review. *Neurology India*, 58(6):879, 2010.

[48] Mark J van der Laan, Sandrine Dudoit, and Katherine S Pollard. Multiple testing. part iii. procedures for control of the generalized family-wise error rate and proportion of false positives. 2004.

[49] Guenther Walther et al. Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics*, 38(2):1010–1033, 2010.

[50] Guenther Walther et al. The average likelihood ratio for large-scale multiple testing and detecting sparse mixtures. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 317–326. Institute of Mathematical Statistics, 2013.

[51] Keith J Worsley, Alan C Evans, S Marrett, P Neelin, et al. A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, 12:900–900, 1992.

[52] KJ Worsley, M Andermann, T Koulis, D MacDonald, and AC Evans. Detecting changes in nonisotropic images. *Human brain mapping*, 8(2-3):98–101, 1999.

[53] Daniel Yekutieli. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008.

[54] Hui Zhang, Thomas E Nichols, and Timothy D Johnson. Cluster mass inference via random field theory. *Neuroimage*, 44(1):51–61, 2009.