Scan Clustering: A False Discovery Approach

M. Perone Pacifico, C. Genovese, I. Verdinelli, L. Wasserman
Carnegie Mellon University and Università di Roma "La Sapienza"

June 25, 2004

We present a method that scans a random field for localized clusters while controling the fraction of false discoveries. We use a kernel density estimator as the test statistic and correct for the bias in this estimator by a method we introduce in this paper. We also show how to combine information across multiple bandwidths while maintaining false discovery control.

1 Introduction

A problem that arises in a wide variety of applications is to identify unusual clusters among events scattered over space or time. Astronomers, for example, look for clustering in the position of objects on the sky to distinguish real groupings from happenstance alignments. Epidemiologists look for clustering in the incidence of a disease to detect outbreaks. What constitutes an event and a cluster varies with each application, but from a mathematical perspective, we consider data as drawn from spatial or temporal point process with a cluster corresponding to a region of high intensity. In this paper, we consider the problem of finding clusters from point process data, extending the method in Perone Pacifico, Genovese, Verdinelli and Wasserman (2004), henceforth denoted by PGVW.

Let $X = (X_1, \ldots, X_N)$ be a realization of a point process with intensity function $\nu(s)$ defined on a compact set $S \subset \mathbb{R}^d$. We assume that conditional on $N = n$, $X = (X_1, \ldots, X_n)$ is an IID sample from the density $f(s) = \nu(s)/\int_S \nu(u)\,du$. We also assume that $\nu(s) = \nu_0$ for all s in an unknown subset $S_0 \subset S$ and that $\nu(s) > \nu_0$ for $s \notin S_0$. The connected components of $S_1 = S_0^c$ are called *clusters*.

An important method for cluster detection is based on scan statistics (Glaz, Naus, and Wallenstein 2001, Patil and Taillie 2003). The usual approach begins with the number of points N_s observed in a fixed window

(such as a rectangle or circle) centered at each $s \in S$. The null hypothesis that there are no clusters is tested via the statistic $T = \sup_{s \in S} N_s$, where the null is rejected if T is large enough. The p-value for T is computed under the uniform distribution on S, and the threshold is designed to control family-wise type I error over S. Finding ways to compute the p-value is an area of active interest; see, for example, Naiman and Priebe (2001).

Controlling familywise error provides a strong guarantee, but it can be conservative in the sense of low power. PGVW presented a method that instead controls the False Discovery Proportion (FDP): the area of false rejections divided by the area of rejections. (For more on false discovery proportions, see Benjamini and Hochberg 1995, Genovese and Wasserman 2002, 2004 and Storey, Taylor, and Siegmund 2003.) Using a kernel density estimator as a test statistic, PGVW tested the set of local null hypotheses:

$$H_{0s} : s \in S_0 \qquad \text{versus} \qquad H_{1s} : s \notin S_0, \tag{1}$$

for every $s \in S$, and used the results of these tests to devise a threshold $T(X)$ such that the random set $L_T = \{s \in S : \ X(s) \geq T(X)\}$ approximates S_1 with a specified error bound. PGVW left open two problems: (i) how to choose the bandwidth of the density estimator and (ii) how to correct for the fact that density estimators are biased. The present paper addresses both problems.

Specifically, we perform our test using a set of bandwidths. We then adjust the rejection region of the test – by appropriately reducing the size of the rejection region – to account for smoothing bias. Very small bandwidths yield low power because of the test statistic's high variance while large bandwidths yield low power because they require large bias adjustment. Between these extremes lie bandwidths with higher power. We show how to combine across bandwidths while maintaining control of the false discovery proportion. We also show that the validity of the Gaussian approximation underlying our test statistic is preserved over the range of bandwidths.

2 The Test Statistic

Testing the null hypotheses in (1) is equivalent to testing $H_{0s} : f(s) = \nu_0 / \int_S \nu(s)ds$. The value of the integral $\int_S \nu(s)ds$ is not known, but $\int_S \nu(s)ds \geq \nu_0 \cdot \lambda(S)$, where $\lambda(\cdot)$ denotes Lebesgue measure. Thus, a conservative test can be obtained by testing

$$H_{0s} : f(s) \leq \overline{\nu}_0 \qquad \text{versus} \qquad H_{1s} : f(s) > \overline{\nu}_0. \tag{2}$$

where

$$\overline{\nu}_0 = \frac{1}{\lambda(S)}. \tag{3}$$

Remark 1. Actually, we do have some information about $\int_S \nu(s)\, ds$ through the total number of observed points N. Under more specific assumptions, such as a Poisson distribution for N, we could construct a confidence interval for $\int_S \nu(s)\, ds$ that is consistent with the constraint $\int_S \nu(s)ds \geq \nu_0 \cdot \lambda(S)$. Nonetheless, in this paper, we use the simpler and more general approach described above. □

We use the kernel density estimator

$$\widehat{f}_H(s) = \frac{1}{n} \sum_{i=1}^{n} K_H(s - X_i) \tag{4}$$

as a test statistic, where the kernel K_H, based on some d-dimensional density φ, is defined for any $s \in S$ and for any bandwidth matrix H by

$$K_H(s) = \frac{1}{\det H}\, \varphi\left(H^{-1}s\right), \tag{5}$$

and $\det H$ denotes the determinant of the matrix H. We take H to be diagonal, but, with the possible exception of Lemma 3, the theory holds for any symmetric, positive definite matrix. In one-dimensional cases, H denotes a positive real number.

Let the smoothed density f_H be defined by

$$f_H(s) \equiv \mathbb{E}[\widehat{f}_H(s)] = \int K_H(s - x) f(x)\, dx \neq f(s). \tag{6}$$

For one dimension, the asymptotic distribution of $\widehat{f}_H - f_H$ as $n \to \infty$, was first derived in Bickel and Rosenblatt (1973). We have been unable to find a similar result for higher dimensions that holds uniformly over a set of bandwidths. The following theorem provides such a result; the proof is in Section 9.

Theorem 1 *Suppose the kernel K_H satisfies (53). Given a decreasing sequence of constants $c_n \downarrow 0$, define the sets of bandwidth matrices*

$$\mathcal{H}_n = \{H : \ H \text{ is a diagonal bandwidth matrix with } \det H \geq c_n\}.$$

Let

$$r_{d,n} = \begin{cases} \frac{(\log n)^d}{\sqrt{n}} & d = 1, 2 \\ \frac{(\log n)^{3/2}}{n^{1/(2d+2)}} & d \geq 3. \end{cases} \tag{7}$$

For each $\delta \in [0, 1]$, there exists mean 0 Gaussian processes $A_n(s, H)$ over $\mathbb{R}^d$, indexed by $H \in \mathcal{H}_n$, with covariance

$$\mathbb{C}(A_n(s, H), A_n(r, L)) = \\ (\det H \cdot \det L)^{\delta} \left(\int K_H(s - x) K_L(r - x) dF(x) - f_H(s) f_L(r) \right) \tag{8}$$

such that

$$\sup_{s \in \mathbb{R}^d, H \in \mathcal{H}_n} \left| (\det H)^{\delta} \sqrt{n} \left(\widehat{f}_H(s) - f_H(s) \right) - A_n(s, H) \right| = O\left(\frac{r_{d,n}}{c_n^{1-\delta}} \right) \ a.s. \tag{9}$$

Theorem 1 states that

$$\sqrt{n} \left(\widehat{f}_H(s) - f_H(s) \right) \tag{10}$$

converges to a mean zero Gaussian process as $n \to \infty$ and that this convergence is uniform for H in an appropriate class $\mathcal{H}_n$ of bandwidth matrices. This resolves an open question raised in Chaudhuri and Marron (2000), which required a fixed lower bound on the bandwidth to get convergence.

In light of this result, we use the test statistic process

$$Z_H(s) = \frac{\widehat{f}_H(s) - \overline{\nu}_0}{\sigma_H(s)} \tag{11}$$

where $\sigma_H(s) = \sqrt{\mathbb{V}(\widehat{f}_H(s))}$. Under the null hypothesis H_{0s}, $f_H(s) \leq \overline{\nu}_0$ and $Z_H(s)$ is approximately a normal random variable with mean less than or equal to 0. The variance $\sigma_H(s)$ does depend on the unknown density and can be estimated from the data, but for many clustering problems, departures from the null occur only in small localized regions. In such cases it suffices to use, as an approximation, the variance under the global null hypothesis $f_H(s) = \frac{1}{\lambda(S)}$, which is

$$\sigma_H^2 \approx \frac{1}{\lambda(S)} \int K_H(s-x)^2 dx - \frac{1}{\lambda(S)^2}. \tag{12}$$

We use this approximation in our examples.

A complication is that nonparametric density estimates are biased, that is, $f_H(s) \neq f(s)$. This bias can lead to excessive rejections. Put another way, a test based on $\widehat{f}_H$ does not really test (2), rather it tests the *biased hypotheses*

$$H_{0s} : f_H(s) \leq \overline{\nu}_0 \qquad \text{versus} \qquad H_{1s} : f_H(s) > \overline{\nu}_0. \tag{13}$$

We address this problem in Section 4. But first, we discuss the general problem of testing the mean of a Gaussian process using false discovery methods. This is the subject of the next section.

3 False Discovery Control for Gaussian Processes

Let $Z(s)$ be a Gaussian process on S with known covariance function. Let $\mu(s) = \mathbb{E}(Z(s))$ and suppose that $\mu(s) \leq 0$ for $s \in S_0 \subset S$ and $\mu(s) > 0$ for $s \in S_1 = S_0^c$. Consider testing the set of hypotheses

$$H_{0s} : \mu(s) \leq 0 \quad \text{versus} \quad H_{1s} : \mu(s) > 0. \tag{14}$$

Suppose we reject H_{0s} for all $s \in B \subset S$. Define the false discovery proportion (FDP) of B by

$$\Gamma(B) = \frac{\lambda(B \cap S_0)}{\lambda(B)} \tag{15}$$

where the ratio is defined to be zero when the denominator is zero. The idea of controlling the mean of the FDP in multiple testing problems is due to Benjamini and Hochberg (1995). Omnibus tests for Gaussian random fields are discussed, for example, in Worsley (1994, 1995).

Given $t \in \mathbb{R}$, define the level set

$$L_t = \{s \in S : \ Z(s) > t\}. \tag{16}$$

PGVW proposed a rejection region L_T based on a data-dependent threshold T that controls the false discovery exceedance (FDX),

$$\text{FDX} \equiv \mathbb{P}(\Gamma(L_T) > \gamma) \leq \alpha \tag{17}$$

for given α and γ. This procedure – which we call inversion – is based on first finding a confidence superset U that contains S_0 with probability $1-\alpha$:

$$\mathbb{P}(U \supset S_0) \geq 1 - \alpha. \tag{18}$$

PGVW give an algorithm to compute U. The algorithm is based on inverting the class of tests

$$H_0 : A \subset S_0 \quad \text{versus} \quad H_1 : A \not\subset S_0 \tag{19}$$

for every subset $A \subset S$, using the test statistic $\sup_{s \in A} Z(s)$.

The confidence superset U can be described as follows. Let $\mathbb{P}$ denote the law of the Gaussian process Z and let $\mathbb{P}_0$ denote the law of a mean zero Gaussian process with the same covariance. Then,

$$U = \bigcup \left\{ A \subset S : \mathbb{P}_0 \left(\sup_{s \in A} Z(s) > \sup_{s \in A} z(s) \right) \geq \alpha \right\} \tag{20}$$

where $z(s)$ is the observed value of the process $Z(s)$. Since $\mathbb{P}(U \supset S_0) \geq 1-\alpha$,

$$\overline{\Gamma}(B) \equiv \frac{\lambda(U \cap B)}{\lambda(B)} \tag{21}$$

is a confidence envelope for $\Gamma(B)$, meaning that

$$\mathbb{P}\left(\Gamma(B) \leq \overline{\Gamma}(B) \ \text{ for all } B\right) \geq 1 - \alpha. \tag{22}$$

Then PGVW chose

$$L_T = \{s \in S : Z(s) \geq T\}$$

where

$$T = \inf\left\{t \in \mathbb{R} : \ \overline{\Gamma}(L_t) \leq \gamma\right\}. \tag{23}$$

This guarantees FDX control as in (17).

Remark 2. The tail probability $\mathbb{P}_0\left(\sup_{s\in A} Z(s) > \sup_{s\in A} z(s)\right)$ in (20) can be approximated with the formulas in, for example, Adler (1981, 1990, 2000), Piterbargh (1996) and Worsley (1994, 1995). □

A different method for exceedence control, called augmentation, is proposed in van der Laan, Dudoit and Pollard (2004), hereafter referred to as VDP. Their method was defined for finite S, however, it is easy to see that it works for the random field setting as well. Let $R \subset S$ be any (random) rejection region that controls the familywise error rate in the sense that

$$\mathbb{P}(R \cap S_0 \neq \emptyset) \leq \alpha. \tag{24}$$

Define

$$\text{aug}_\gamma(R) = \begin{cases} \emptyset & \text{if } R = \emptyset \\ R \cup A & \text{otherwise} \end{cases} \tag{25}$$

where A is any set such that $A \cap R = \emptyset$ and

$$\frac{\lambda(A)}{\lambda(R) + \lambda(A)} \leq \gamma. \tag{26}$$

Then the augmented rejection set $\text{aug}_\gamma(R)$ controls FDX. More formally:

Theorem 2 *(VDP) If R satisfies (24), then $\mathbb{P}(\Gamma(\text{aug}_\gamma(R)) > \gamma) \leq \alpha$. Also, $\overline{\Gamma}(B) = \lambda((\text{aug}_\gamma(R))^c \cap B)/\lambda(B)$ is a confidence envelope.*

It is not difficult to see that the superset U in (20) is a continuous version of a stepdown testing method. Specifically, note that

$$U = \{s \in S : \ Z(s) < Q\}$$

with

$$Q = \inf \left\{ t \in \mathbb{R} : \ \mathbb{P}_0 \left(\sup_{\{z(s) \leq t\}} Z(s) > t \right) < \alpha \right\}. \tag{27}$$

Clearly $T < Q$, hence the rejection region L_T can be written as

$$L_T = R \cup A$$

where

$$R = U^c = \{ s \in S : \ Z(s) \geq Q \} \tag{28}$$

and $A = \{ s : \ T \leq Z(s) < Q \}$.

This gives an explanation of the procedure in PGVW in terms of VDP. More precisely, the above calculations prove the following result, described in more detail in Genovese and Wasserman (2004b).

Theorem 3 *Inversion and augmentation yield the same procedure, that is,* $L_T = \text{aug}_\gamma(R)$.

Although our focus is on false discovery methods, it is worth noting that the region R in (28) provides a new familywise test, more powerful than the commonly used test based on $\sup_{s \in S} Z(s)$, namely,

$$W = \{ s \in S : \ Z(s) \geq Q' \} \tag{29}$$

with

$$Q' = \inf \left\{ t \in \mathbb{R} : \ \mathbb{P}_0 \left(\sup_{s \in S} Z(s) > t \right) < \alpha \right\}. \tag{30}$$

Theorem 4 *Let R and W be the rejection regions (28) and (29) respectively. Then, $W \subset R$.*

Proof: Since

$$\sup_{\{z(s) \leq t\}} Z(s) \leq \sup_{s \in S} Z(s),$$

then for all t,

$$\mathbb{P}_0 \left(\sup_{\{z(s) \leq t\}} Z(s) > t \right) \leq \mathbb{P}_0 \left(\sup_{s \in S} Z(s) > t \right).$$

Hence, $Q \leq Q'$ and thus $W \subset R$. ■

4 Bias Correction

In this section we let $S_0 = \{s : \ f(s) \leq \overline{\nu}_0\}$ denote the set of points satisfying the true null hypothesis in (2) and let

$$S_{0,H} = \{s : \ f_H(s) \leq \overline{\nu}_0\} \tag{31}$$

denote the set of points satisfying the biased null hypothesis in (13). Let $\text{aug}_\gamma(R_H) = R_H \cup A_H$ denote the rejection set giving exceedance control for the biased null $S_{0,H}$. Here, R_H controls familywise error for the biased null:

$$\mathbb{P}(R_H \cap S_{0,H} \neq \emptyset) \leq \alpha. \tag{32}$$

Our goal is to adjust $\text{aug}_\gamma(R_H)$ to give exceedance control for S_0.

Figure 1 is an illustration of the bias problem in cluster detection. Assume there are only three clusters as shown in the figure. The true density, the mean of a kernel estimator and typical realizations of the estimator are shown for increasing bandwidths. For large bandwidth, $\widehat{f}_H$ is close to its mean but the mean distorts the clusters. Specifically, $\{s : \widehat{f}_H(s) > t\}$ is larger than $\{s : f(s) > t\}$ for some values of t, leading to an excess in false discoveries. We want to correct for this kernel smoothing bias. In general, correcting the bias of a kernel density estimator is difficult. This is because the pointwise bias of $\widehat{f}_H(s)$ is, asymptotically, proportional to $f''(s)$ and derivative estimation is harder than estimating f. However, in our case, we need only correct the bias of the edges of the level sets $\{s \in S : \ \widehat{f}_H(s) > t\}$ rather than the density estimate itself. To illustrate this point, Figure 2 shows the rejection regions, both bias-corrected (called shaved, panel B) and not bias-corrected (called unshaved, panel A) as a function of the bandwidth, for the previous example.

We now define the bias correction method – which we call shaving – in detail. The Minkowski sum of two sets A and B is

$$A \oplus B = \{a + b : \quad a \in A, b \in B\}.$$

The Minkowski difference is

$$A \ominus B = \{s : \ s + B \subset A\} = (A^c \oplus -B)^c$$

where $-B = \{-s : \ s \in B\}$. Let C_H denote the support of the kernel K_H, with bandwidth matrix H. We assume that C_H is a connected, compact set and that C_H is symmetric: $-C_H = C_H$.

The bias corrected procedure replaces $\text{aug}_\gamma(R_H)$ with

$$\text{aug}_\gamma(\text{sh}(R_H)), \tag{33}$$

where

$$\text{sh}(R_H) = (R \ominus C_H) \tag{34}$$

is the shaved version of R_H. Schematically, the procedure is as follows:

$$R_H \stackrel{\text{shave}}{\longrightarrow} \text{sh}(R_H) \stackrel{\text{augment}}{\longrightarrow} \text{aug}_\gamma(\text{sh}(R_H)) \tag{35}$$

To show that $\text{aug}_\gamma(\text{sh}(R_H))$ controls the FDX, we need to make some assumptions about $S_1 = S_0^c$. The key assumption is the following separation condition:

$$\text{for every } s \in (S_1 \oplus C_H) - S_1, \qquad (s \oplus C_H) \bigcap (S_1 \oplus C_H)^c \neq \emptyset. \tag{36}$$

This condition precludes clusters from being very close together. See Figure 3.

The proof of the following lemma is straightforward and is omitted.

Lemma 1 *A sufficient condition for the separation condition is that S_1 is the union of finitely many connected, compacts sets $C_1, \ldots, C_k$ such that*

$$\min_{i \neq j} \inf_{s \in C_i, t \in C_j} d(s,t) > w_H \tag{37}$$

and

$$\min_{i} \inf_{s \in C_i, t \in \partial S} d(s,t) > w_H \tag{38}$$

where w_H is the diameter of C_H, d is Euclidean distance, and ∂S is the boundary of S.

Theorem 5 *Suppose that K_H has compact, symmetric support, and that the separation condition (36) holds. Then,* $\mathrm{sh}(R_H)$ *controls familywise error for* S_0 *at level* α *and* $\mathrm{aug}_\gamma(\mathrm{sh}(R_H))$ *controls the FDP for* S_0 *at level* γ *with probability at least* $1-\alpha$.

Proof: From Theorem 2 it suffices to show that $\mathbb{P}(\mathrm{sh}(R_H) \cap S_0 \neq \emptyset) \leq \alpha$ where $\mathrm{sh}(R_H) = R_H \ominus C_H$. First, because $\{s:\ f_H(s) > \overline{\nu}_0\} = \{s:\ Z_H(s) > 0\} \subset (S_1 \oplus C_H)$, we have, using the symmetry of C_H, that

$$S_0 \ominus C_H = (S_0^c \oplus -C_H)^c = (S_1 \oplus C_H)^c \subset \{s:\ f_H(s) \leq \overline{\nu}_0\} = S_{0,H}. \quad (39)$$

Next we show that

$$\mathrm{sh}(R_H) \cap S_0 \neq \emptyset \quad \text{implies that} \quad R_H \not\subset S_1 \oplus C_H. \quad (40)$$

Suppose that $R_H \subset S_1 \oplus C_H$. Let $s \in S_0$. Consider two cases: (i) $s \in R_H^c$ and (ii) $s \in R_H$. For case (i), clearly $s \notin \mathrm{sh}(R_H)$ For case (ii), argue as follows. If $s \in R_H \cap S_0$, then $s \in (S_1 \oplus C_H) - S_1$. From the separation condition, there exists $y \in (S_1 \oplus C_H)^c \subset R_H^c$ such that $y \in s \oplus C_H$. Therefore, $s \notin R_H \ominus C_H = \mathrm{sh}(R_H)$. This establishes

$$R_H \subset S_1 \oplus C_H \quad \text{implies that} \quad \mathrm{sh}(R_H) \cap S_0 = \emptyset. \quad (41)$$

and (40) thus follows. Now $R_H \not\subset S_1 \oplus C_H$ implies that $R_H \cap (S_1 \oplus C_H)^c \neq \emptyset$. But $(S_1 \oplus C_H)^c = (S_0^c \oplus C_H)^c = (S_0^c \oplus -C_H)^c = (S_0 \ominus C_H)$. So we have that

$$\mathrm{sh}(R_H) \cap S_0 \neq \emptyset \quad \text{implies that} \quad R_H \cap (S_0 \ominus C_H) \neq \emptyset. \quad (42)$$

Finally,

$$\begin{aligned} \mathbb{P}(\mathrm{sh}(R_H) \cap S_0 \neq \emptyset) &\leq \mathbb{P}(R_H \cap (S_0 \ominus C_H) \neq \emptyset) && \text{from (42)} \\ &\leq \mathbb{P}(R_H \cap S_{0,H} \neq \emptyset) && \text{from (39)} \\ &\leq \alpha && \text{from (32).} \end{aligned}$$

That $\mathrm{aug}_\gamma(\mathrm{sh}(R_H))$ controls the FDP for S_0 at level γ with probability at least $1-\alpha$ follows by construction. ■

Remark 3. The theorem above applies to kernels with bounded support. In practice, it is sometimes convenient to use Gaussian kernels, which have unbounded support. Without kernels of compact support, the previous theorem is no longer true, but our numerical experience is that the procedure still works well, by taking C_H to be a compact set with high probability under K_H. Also, if the separation condition fails then clusters that are too close together will get blended together. □

Remark 4. A similar procedure is used by Taylor (2004) for a different purpose. He shows that by replacing $Z_H(s)$ with a new test statistic, one can remove small, isolated portions of the rejection region while still preserving false discovery control. Moreover, the rejection region for the new statistic seems to be related to the shaving operation. Also, Walther (1997) uses similar tools for optimal level set estimation. □

5 Power and Bandwidth Selection

Now we consider the problem of choosing a bandwidth H. In density estimation, one usually tries to choose an H that balances bias and variance to optimize mean squared error. But this is not our goal here. Indeed, Figure 4 shows that the density estimator based on a bandwidth that is optimal for "testing" (shown in the left panel) is different from the density estimator using a bandwidth that is optimal for estimation (right panel).

First, some notation. Define the realized power of a rejection region B by

$$\pi(B) = \frac{\lambda(B \cap S_1)}{\lambda(S)}.$$

Given a set of possible bandwidths $\mathcal{H}_n$, define

$$\pi^*(\alpha) = \sup_{H \in \mathcal{H}_n} \pi(\text{aug}_\gamma(\text{sh}(R_H(\alpha))))$$

which is the power of the best, single bandwidth procedure. Rather than trying to find this best bandwidth, our proposal is to combine rejection regions over the bandwidths in $\mathcal{H}_n$ as follows.

Take $\mathcal{H}_n$ to be a finite set of bandwidths. We combine the shaved rejection regions from the individual bandwidths and augment. Define

$$\Delta = B \oplus \Lambda_\epsilon = B \cup A_\epsilon \tag{43}$$

where

$$B = \left(\bigcup_{H \in \mathcal{H}_n} \text{sh}\left(R_H\left(\frac{\alpha}{m}\right)\right)\right), \tag{44}$$

m is the number of elements in $\mathcal{H}_n$, Λ_ϵ is a sphere of radius ϵ and $A_\epsilon = (B \oplus \Lambda_\epsilon) - B$. Here, ϵ is the largest number such that

$$\frac{\lambda(A_\epsilon)}{\lambda(A_\epsilon) + \lambda(B)} \leq \gamma. \tag{45}$$

Notice that Δ is just an augmentation of B. Here is a summary of the steps:

$$R_H \stackrel{\text{shave}}{\longrightarrow} \text{sh}(R_H) \stackrel{\text{combine}}{\longrightarrow} B = \bigcup_H \text{sh}(R_H) \stackrel{\text{augment}}{\longrightarrow} \Delta = B \oplus \Lambda_\epsilon$$

Remark 5. One could of course use other augmentations although this augmentation is simple and does not increase the number of clusters. □

The set Δ controls FDP and has power close to the optimal with high probability.

Theorem 6 *We have that* $\mathbb{P}(\Gamma(\Delta) > \gamma) < \alpha$ *and*

$$\mathbb{P}\left(\pi(\Delta) \geq \pi^*(\alpha/m) - \frac{\gamma}{1-\gamma}\right) \geq 1 - \alpha.$$

Proof: Without loss of generality, take $\lambda(S) = 1$. By Bonferroni's inequality, B controls familywise error at level α. Hence, the augmented set Δ controls FDP at level γ with probability at least $1-\alpha$. Let $\overline{R}_H = \text{aug}_\gamma(\text{sh}(R_H(\alpha/m)))$.

For each H we have $\lambda(A_H) \leq \frac{\gamma}{1-\gamma}\lambda(\text{sh}(R_H))$, since $\lambda(A_H)/(\lambda(A_H)+\lambda(\text{sh}(R_H))) \leq \gamma$. Hence,

$$\begin{aligned}
\pi(\Delta) &= \lambda(\Delta \cap S_1) \geq \lambda(B \cap S_1) = \lambda(B) - \lambda(B \cap S_0) \\
&\geq \lambda(B) \qquad \text{with probability at least } 1-\alpha \\
&\geq \lambda(\text{sh}(R_H)), \qquad \text{for every } H \in \mathcal{H}_n \\
&= \lambda(\overline{R}_H) - \lambda(A_H) \geq \lambda(\overline{R}_H \cap S_1) - \lambda(A_H) \\
&\geq \lambda(\overline{R}_H \cap S_1) - \frac{\gamma}{1-\gamma}\lambda(R_H) \\
&\geq \lambda(\overline{R}_H \cap S_1) - \frac{\gamma}{1-\gamma} = \pi(\overline{R}_H) - \frac{\gamma}{1-\gamma}.
\end{aligned}$$

This completes the proof. ■

Remark 6. Regarding the choice of $\mathcal{H}_n$, there are several possibilities. In one dimension, we recommend choosing m equally spaced points in the interval $[c_n, h_{OS}]$ where

$$c_n = \frac{\widehat{\sigma}(\log n)^3}{n}, \tag{46}$$

h_{OS} is the oversmoothing bandwidth from Scott (1992), and $\widehat{\sigma}$ is the sample standard deviation. Thus, the minimum bandwidth c_n satisfies $r_{1,n}^2/c_n \to 0$ where $r_{1,n}$ is defined in (7). The condition $r_{1,n}^2/c_n \to 0$ is needed for Theorem 1 to apply. The maximum bandwidth is h_{OS} commonly recommended as an upper bound for the bandwidth. Our experience suggests that the choice of m is not crucial; one can even let m increase with n, for example, $m = n$. For d-dimensional data, we suggest taking

$$H = h \begin{pmatrix} \widehat{\sigma}_1 & 0 & \cdots & 0 \\ 0 & \widehat{\sigma}_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \widehat{\sigma}_d \end{pmatrix}$$

where $\widehat{\sigma}_j$ is the standard deviation of the j^{th} variable. Then, h is allowed to vary in a finite set as in the one-dimensional case. However the set of bandwidth matrices is constructed, the smallest determinant c_n over the set

of bandwidth matrices should again satisfy $r_{d,n}^2/c_n \to 0$ where $r_{d,n}$ is defined in (7). □

Remark 7. An alternative that eliminates the need to use control at level $1 - \alpha/m$ is data-splitting. Randomly split the data into two sets of equal size. Choose $\widehat{H}$ to maximize $\lambda(R_H(\alpha))$ using the first half of the data. Now apply the procedure to the second half of the data using bandwidth $\widehat{H}$. This controls FDX conditionally (on the first half) and hence marginally. □

Remark 8. Our work may also be viewed as a contribution to the scale-space approach to smoothing esposed by Chaudhuri and Marron (2000). They consider finding modes of a density f by finding points where $f'_H(x) = 0$ and then plotting the results as a function of H. In our setting, we could similarly display the significant clusters as a function of H. Viewed this way, our method fits nicely in their framework, the main differences being our focus on FDR and on clusters rather than modes. Indeed, Figure 1 can be thought of as a scale-space representation of clustering. We believe that the scale-space approach could be quite useful in some applications. But in other cases it is desirable to correct bias and combine information across bandwidths. □

6 A One-Dimensional Example

In this section, we report the results of a simulation for a one dimensional example. We draw a sample of $n = 1,000$ observations from a uniform density over $[0, 1]$ with 3 clusters of different heights. The true density (shown in the left panels of Figure 1) is

$$f(s) = \frac{4}{9} \times \begin{cases} 3 & s \in \text{cluster 1} \\ 6 & s \in \text{cluster 2} \\ 9 & s \in \text{cluster 3} \\ 1 & \text{elsewhere.} \end{cases} \tag{47}$$

The density estimation has been performed using the `R` function `density` with a Gaussian kernel. The estimate has been evaluate over a grid of 1,024

equally spaced points over $[0, 1]$. Figure 1 shows the bias as a function of bandwidth for this example.

The exceedence control procedure (with $\alpha = 0.05$ and $\gamma = 0.1$) was applied using 50 different bandwidths between 0.0001 and the approximate oversmoothing bandwidth,

$$h_{OS} = 1.1 \times \left(\frac{4}{3n}\right)^{\frac{1}{5}} \sigma$$

suggested in Scott (1992, page 181). Here, σ is the standard deviation which, in practice, is estimated using the sample standard deviation or a robust estimate of scale. Figure 2 A shows the clusters identified without any bias correction procedure ($\text{aug}_\gamma(R_H)$) as the bandwidth varies, similarly clusters obtained after shaving ($\text{aug}_\gamma(\text{sh}(R_H))$) are shown in plot B of the same figure.

The increasing bias in the non-shaved clusters is evident. Shaving is effective at reducing the bias. Cluster 1 is hard to detect; its height is $4/3$ and is barely higher than $1/\lambda(S) = 1$. Panel A in figure 5 compares the width of shaved and non shaved rejection regions. Except for extremely small bandwidths, the area of non-shaved rejected regions is increasing and this is basically due to bias. If one looks at the area of shaved regions, there is a local maximum (which could be used as a single-bandwidth procedure).

Figure 5 B compares the behavior of the FDP for shaved and non-shaved rejected regions. The improvement due to shaving is evident. Conversely, shaving causes a loss of power. However, as shown in Figure 5 C, the loss of power does not seem to be comparable to what was gained in terms of FDP. Figure 6 shows the set Δ. The corresponding FDP and power are 0.0474 and 0.196. In this case Δ is more powerful than even $\pi^*(\alpha) = 0.186$.

The simulation was repeated 1,000 times drawing different samples from density (47). Plots in Figure 7 show that the behavior of FDP and power is almost the same for all simulations. In all of the simulations, the power was greater than $\pi^*(\alpha) - \frac{\gamma}{1-\gamma}$.

7 A Two Dimensional Example

Figure 8 B shows the clusters detected using our proposed procedure with a sample of $n = 15,000$ observations from the density shown in Figure 8 A.

The true density is a mixture of uniforms over subsets of $[0,1]^2$

$$f(s) = \frac{256}{466} \times \begin{cases} 3 & s \in \text{cluster 1 and 6} \\ 6 & s \in \text{cluster 2 and 5} \\ 9 & s \in \text{cluster 3 and 4} \\ 1 & \text{elsewhere} \end{cases} \tag{48}$$

where the clusters are enumerated clockwise from top-left.

The density estimation was performed using the `R` package `MASS` with a Gaussian kernel and the estimate was evaluated over a grid of 256×256 equally spaced points.

We applied the exceedence control procedure (with $\alpha = 0.05$ and $\gamma = 0.1$) using 20 different bandwidths ranging between the pixel size and the oversmoothing bandwidth,

$$h_{OS} = 1.1 \times \sigma \cdot n^{-\frac{1}{6}}.$$

Figure 9 A, C, E show the clusters identified without any bias correction procedure ($\text{aug}_\gamma(R_H)$) for very small, intermediate and large bandwidth respectively, panels B, D, F in the same figure show the clusters obtained after shaving ($\text{aug}_\gamma(\text{sh}(R_H))$).

Figure 10 A shows the behavior of the area of the clusters obtained with and without shaving. Figure 10 B compares the behavior of FDP for shaved and non shaved rejected regions, as the bandwidth varies. In this case too, the loss of power due to shaving is small respect to the reduction of FDP.

The final set has null FDP (there are no false rejections) and power 0.098, which is again higher than $\pi^*(\alpha) = 0.073$.

8 Asymptotic Mean Control

Our main interest is in exceedance control. However, for completeness, we also discuss mean control. There are at least two methods for obtaining mean

control. The first method is from PGVW, Theorem 4b. It relies on the simple fact that $\mathbb{P}(\Gamma(B) > \beta) < \alpha$ implies that $\mathbb{E}(\Gamma(B)) \leq \gamma = \beta + (1-\beta)\alpha$.

Lemma 2 *Let $\gamma \in (0,1)$. Choose any $\beta \in (0,\gamma)$ and let T be a (β,α) confidence threshold with $\alpha = (\gamma-\beta)/(1-\beta)$. Then,*

$$\text{FDR} = \mathbb{E}(\Gamma(L_T)) \leq \gamma.$$

The second method is asymptotic. While it gives up exact control, it appears to often have higher power. Define

$$T = \inf\left\{ z \in \mathbb{R} : \ \frac{\lambda(S)(1-\Phi(z))}{\lambda(\{s \in S : \ Z_H(s) > z\})} \leq \gamma \right\} \tag{49}$$

where Φ is the CDF of a standard Normal. Now suppose we reject the null when $Z_H(s) > T$. As we now explain, this controls, asymptotically, the FDR.

Theorem 7 *Suppose that $\lambda(\partial S) = \lambda(\partial S_0) = 0$ and that the equation*

$$\frac{\mathbb{E}(\lambda(\{s : \ Z_H(s) > t\}))}{\lambda(S)} - \frac{1-\Phi(t)}{\gamma} = 0 \tag{50}$$

has a unique root for all large n. Let T be defined as in (49). Then, for testing the biased null,

$$\mathbb{E}[\Gamma(L_T)] \leq \frac{\lambda(S_0)}{\lambda(S)}\,\gamma + o(1) \leq \gamma + o(1)$$

as $n \to \infty$, uniformly for $H \in \mathcal{H}_n$.

The proof is in the next section.

Remark 9. Condition (50) will hold with reasonable regularity conditions on f. □

9 Theoretical Background

9.1 Asymptotics for the Density Estimator

For any bandwidth matrix H and any $s \in \mathbb{R}^d$, the kernel density estimator $\widehat{f}_H(s)$ and its expectation $f_H(s)$ are

$$\widehat{f}_H(s) = \frac{1}{n}\sum_{i=1}^{n} K_H(s - X_i) = \int K_H(s-x)\, d\widehat{F}_n(x), \tag{51}$$

$$f_H(s) = \int K_H(s-x)\, dF(x) \tag{52}$$

where $\widehat{F}_n$ and F are the empirical and the true distribution function respectively. We will suppose that the kernel is of the form:

$$K_H(s) = \frac{1}{\det H}\varphi(H^{-1}s) = \frac{1}{\det H}\ \left[b_1 W_1(H^{-1}s) - b_2 W_2(H^{-1}s)\right] \tag{53}$$

where b_1 and b_2 are two positive constants and W_1 and W_2 are two CDFs over $\mathbb{R}^d$.

Remark 10. In the univariate case, condition (53) requires the kernel to be right-continuous and to have bounded variation. Right-continuity is not an issue, since one can always "adjust" a density over a set with zero Lebesgue measure. All the most common univariate kernels have bounded variation, including all the options for the `R` function `density`. Unfortunately there is no straightforward extension of the notion of bounded variation to higher dimensions. For a discussion on this topic see, for example, Koenker and Mizera (2004). In any case, condition (53) is satisfied by many multivariate densities, in particular by products of right-continuous univariate kernels with bounded variation. □

Before proving Theorem 1 we state four lemmas.

Lemma 3 *If the kernel K_H satisfies (53), H is diagonal, and W is a* CDF*, then*

$$\int K_H(s-x)\, dW(x) = \int W(s-x)\, dK_H(x). \tag{54}$$

Proof: Write K_H as in (53) and let X, X_1, and X_2 be drawn independently from W, W_1, and W_2 respectively. Because H is positive definite, the functions $x \mapsto W(H^{-1}x)$, $x \mapsto W_1(H^{-1}x)$, and $x \mapsto W_2(H^{-1}x)$ are all CDFs. (of HX, HX_1, and HX_2 respectively). The integrals in (54) can be written

$$\begin{aligned}\frac{b_1}{\det H}\int W_1(H^{-1}(s-x))\,dW(x) \quad - \quad \frac{b_2}{\det H}\int W_2(H^{-1}(s-x))\,dW(x)\\ = \frac{b_1}{\det H}\int W(s-x)\,dW_1(H^{-1}x) - \frac{b_2}{\det H}\int W(s-x)\,dW_2(H^{-1}x),\end{aligned}$$

and the corresponding terms on both sides are equal, representing the convolutions of independent random variables. ■

Lemma 4 below summarizes a number of results, all reported in Massart (1989).

Lemma 4 *Let $\widehat{G}_n = \sqrt{n}(\widehat{F}_n - F)$ be the centered empirical process over $\mathbb{R}^d$ and define $r_{d,n}$ as in (7). There exists a sequence G_n of centered Gaussian processes with covariance*

$$\mathbb{C}(G_n(s), G_n(r)) = F(s \wedge r) - F(s)F(r)$$

such that

$$\sup_{s\in\mathbb{R}^d} |\widehat{G}_n(s) - G_n(s)| = O(r_{d,n}) \qquad a.s.$$

Note that the distribution of the processes G_n does not depend on n. For multidimensional spaces, the expression $s \wedge d$ in the covariance is the componentwise minimum.

Lemma 5 *If the kernel K_H satisfies (53) and W is bounded over $\mathbb{R}^d$, then*

$$\left|\int W(s)\; dK_H(s)\right| \leq \sup_{s\in\mathbb{R}^d} |W(s)|\; \frac{b_1 + b_2}{\det H}.$$

Proof:

$$\begin{aligned}
&\left|\int W(s)\, dK_H(s)\right| = \\
&= \left|\int \frac{1}{\det H} W(s)\, d\varphi(H^{-1}s)\right| = \frac{1}{\det H}\left|\int W(Ht)\, d\varphi(t)\right| \\
&= \frac{1}{\det H}\left|b_1 \int W(Hs)\, dW_1(s) - b_2 \int W(Hs)\, dW_2(s)\right| \\
&\leq \frac{1}{\det H}\left(b_1 \left|\int W(Hs)\, dW_1(s)\right| + b_2 \left|\int W(Hs)\, dW_2(s)\right|\right) \\
&\leq \frac{1}{\det H}\left(b_1 \sup_{s\in\mathbb{R}^d} |W(s)| + b_2 \sup_{s\in\mathbb{R}^d} |W(s)|\right) = \sup_{s\in\mathbb{R}^d} |W(s)| \frac{b_1+b_2}{\det H}.
\end{aligned}$$

■

Lemma 6 *The function $F(s \wedge t)$ is a cumulative distribution function over $\mathbb{R}^{2d}$ and for each function $w : \mathbb{R}^{2d} \mapsto \mathbb{R}$ we have*

$$\iint_{\mathbb{R}^{2d}} w(s,t)\, dF(s \wedge t) = \int_{\mathbb{R}^d} w(s,s)\, dF(s). \tag{55}$$

Proof: Let X be a random variable in $\mathbb{R}^d$ with cumulative distribution function F and let Y be such that $Y = X$ almost surely. The joint cumulative distribution function of (X, Y) is

$$F_{X,Y}(s,r) = \mathbb{P}(X \leq s, Y \leq r) = \mathbb{P}(X \leq s, X \leq r) = \mathbb{P}(X \leq s \wedge r) = F(s \wedge r).$$

The left hand side of (55) can be viewed as the expectation of $w(X, Y)$

$$\mathbb{E}(w(X,Y)) = \iint w(s,r)\, dF_{X,Y}(s,r) = \iint w(s,r)\, dF(s \wedge r)$$

but, since $Y = X$ almost surely, $w(X, Y) = w(X, X)$ and

$$\mathbb{E}(w(X,Y)) = \mathbb{E}(w(X,X)) = \int w(s,s)\, dF(s)$$

that gives (55). ■

Proof of Theorem 1: As a consequence of Lemma 3 we can write

$$\begin{aligned}\widehat{f}_H(s) &= \int K_H(s-x)\, d\widehat{F}_n(x) = \int \widehat{F}_n(s-x)\, dK_H(x)\\ f_H(s) &= \int K_H(s-x)\, dF(x) = \int F(s-x)\, dK_H(x).\end{aligned}$$

Let $G_n = \sqrt{n}(\widehat{F}_n - F)$ be the process in Lemma 4 and

$$A_n(x,H) = (\det H)^{\delta} \int G_n(s-x)\, dK_H(x);$$

we have

$$\begin{aligned}&(\det H)^{\delta}\sqrt{n}\left(\widehat{f}_H(s) - f_H(s)\right) =\\ &= (\det H)^{\delta} \int \sqrt{n}\left(\widehat{F}_n(s-x) - F(s-x)\right)\, dK_H(x)\\ &= (\det H)^{\delta} \int \widehat{G}_n(s-x)\, dK_H(x)\\ &= A_n(s,H) + (\det H)^{\delta} \int \left(\widehat{G}_n(s-x) - G_n(s-x)\right)\, dK_H(x).\end{aligned}$$

From Lemmas 4 and 5 it follows, almost surely

$$\begin{aligned}&\left|(\det H)^{\delta} \int (\widehat{G}_n(s-x) - G_n(s-x))\, dK_H(x)\right| \leq\\ &\leq (\det H)^{\delta} \sup_{s\in\mathbb{R}^d}\left|\widehat{G}_n(s) - G_n(s)\right| \frac{b_1+b_2}{\det H}\\ &\leq (\det H)^{\delta} \frac{O(r_{d,n})}{\det H} \leq \frac{O(r_{d,n})}{c_n^{1-\delta}}\end{aligned}$$

that gives (9).

Since G_n is a centered Gaussian process, A_n is also Gaussian with zero

mean and covariance

$$
\begin{aligned}
&\mathbb{C}(A_n(s,H), A_n(r,L)) = \\
&\quad = (\det H \det L)^{\delta} \iint \mathbb{C}(G_n(s-x), G_n(r-y))\, dK_H(x) dK_L(y) \\
&\quad = (\det H \det L)^{\delta} \iint [F((s-x)\wedge(r-y)) - F(s-x)F(r-y)]\, dK_H(x) dK_L(y) \\
&\quad = (\det H \det L)^{\delta} \left[\iint F((s-x)\wedge(r-y))\, dK_H(x) dK_L(y) \right. \\
&\qquad \left. - \int F(s-x)\, dK_H(x) \int F(r-y)\, dK_L(y) \right].
\end{aligned}
$$

The last term in the covariance is

$$
\int F(s-x)\, dK_H(x) \int F(r-y)\, dK_L(y) = f_H(s) f_L(r).
$$

Since $K_H(x) \cdot K_L(y)$ is right-continuous with bounded variation over $\mathbb{R}^{2d}$ and, from Lemma 6, $F(x \wedge y)$ is a cumulative distribution function, we can use (54) and (55) and obtain

$$
\begin{aligned}
\iint F((s-x)\wedge(r-y))\, dK_H(x) dK_L(y) &= \iint K_H(s-x) K_L(r-y)\, dF(x \wedge y) \\
&= \int K_H(s-x) K_L(r-x)\, dF(x)
\end{aligned}
$$

from which the covariance in (8) is obtained. ■

9.2 Proof of Theorem 7

We will use a result analogous to the one proved in Benjamini and Yekutieli (2001) for discrete problems. To be consistent with the notation of their paper, we switch to the p-value scale. Hence we consider the process $p : S \mapsto [0,1]$ defined as $p(s) = 1 - \Phi(Z_H(s))$. The p-value, $p(\cdot)$ is continuous as long as Z is continuous. For $t \in [0,1]$, define

$$
G(t) = \frac{\lambda(\{s \in S : p(s) \le t\})}{\lambda(S)} \qquad H(t) = \frac{\lambda(\{s \in S_0 : p(s) \le t\})}{\lambda(S)}.
$$

Using the threshold T in (49) and rejecting all $Z(s) \geq T$ is equivalent to rejecting all $p(s) \leq T$, where

$$T = \sup \left\{ t \in [0,1] : G(t) - \frac{t}{\gamma} \geq 0 \right\}.$$

On the p-value scale, the false discovery proportion at each t is

$$\Lambda(t) \equiv \frac{\lambda(\{s:\ p(s) \leq t\} \cap S_0)}{\lambda(\{s:\ p(s) \leq t\})} = \begin{cases} \frac{H(t)}{G(t)} & \text{if } G(t) > 0 \\ 0 & \text{if } G(t) = 0. \end{cases}$$

Thus $\Lambda(t)$ corresponds to $\Gamma(L_t)$ on the test statistic scale.

To use the result by Benjamini and Yekutieli (2001), we consider a sequence of discrete problems converging to the continuous problem at hand. Thus, for each m, partition S into N_m subsets, all with the same measure $\lambda(S)/N_m$. The partitions must be nested and degenerating in the sense of PGVW. By choosing one point from each element of the partition, we select N_m points $s_1, \cdots, s_{N_m}$ and we put on each of them mass $\lambda(S)/N_m$. For each Borel set $A \subset S$, consider the measure λ_m:

$$\lambda_m(A) = \frac{\lambda(S)}{N_m} \sum_{s_j \in A} I_A(s_j),$$

so to define discrete analogous of G, H, and Λ as follows:

$$G_m(t) = \frac{\lambda_m(\{s \in S : p(s) \leq t\})}{\lambda(S)}, \qquad H_m(t) = \frac{\lambda_m(\{s \in S_0 : p(s) \leq t\})}{\lambda(S)},$$

and

$$\Lambda_m(t) = \begin{cases} \frac{H_m(t)}{G_m(t)} & \text{if } G_m(t) > 0 \\ 0 & \text{if } G_m(t) = 0. \end{cases}$$

The following lemma shows uniform convergence (denoted as $\xrightarrow{u}$) of all the discrete functions defined above as $m \to \infty$, for fixed n.

Lemma 7 *Under the hypotheses of Theorem 7, $G_m \xrightarrow{u} G$ and $H_m \xrightarrow{u} H$, almost surely. Moreover, for every $\delta > 0$, $\Lambda_m \xrightarrow{u} \Lambda$, almost surely, over the random set $\{t \in \mathbb{R} : G(t) \geq \delta\}$.*

Proof: Weak convergence of λ_m to λ is easy to prove. Hence, if Y_m and Y are random vectors over S, with distribution

$$\frac{\lambda_m(\cdot)}{\lambda_m(S)} \qquad \text{and} \qquad \frac{\lambda(\cdot)}{\lambda(S)}$$

respectively, then $Y_m \to Y$ in distribution.

The continuous mapping theorem guarantees almost sure convergence of the distribution of $p(Y_m)$ to $p(Y)$, because the process p is continuous almost surely. Since G_m is the CDF of $p(Y_m)$ and G the CDF of $p(Y)$, then $G_m \to G$ at each continuity point for G. Continuity of G ensures almost sure pointwise convergence. With a proof analogous to that of Glivenko-Cantelli Theorem (see, for instance, van der Vaart (1998), page 266) we obtain uniform convergence.

Uniform convergence of H_m to H can be proved similarly, but considering respectively

$$\frac{\lambda_m(\cdot \cap S_0)}{\lambda_m(S_0)} \qquad \text{and} \qquad \frac{\lambda(\cdot \cap S_0)}{\lambda(S_0)}$$

as distributions of Y_m and Y. Uniform convergence of Λ_m to Λ is straightforward (but note that each path converge on a different set). ■

Proof of Theorem 7: From the limiting Normal approximation and the form of the covariance function of Z_H, the p-value process satisfies the positive dependence condition of Benjamini and Yekutieli (2001) a.s. for all large n. Hence, from their main result, we have that

$$\mathbb{E}(\Lambda_m(T_m)) \leq \frac{\lambda_m(S_0)}{\lambda_m(S)}\,\gamma$$

with

$$T_m = \sup\left\{t \in [0,1] : G_m(t) - \frac{t}{\gamma} \geq 0\right\}.$$

For each ω such that uniform convergence of g_n to g holds, we have (omitting the dependence on ω from the notation) that:

- T is the unique solution of equation $G(t) - \frac{t}{\gamma} = 0$ and $G(t) - \frac{t}{\gamma}$ is strictly positive for all $t < T$. Thus there exists m_t such that $G_m(t) - \frac{t}{\gamma} > 0$ for all $m \geq m_t$. Hence $t \leq \inf_{m \geq m_t} T_m \leq \liminf T_n$. It follows $T \leq \liminf T_m$.

- For each $x > T$, we have $\max_{t \in [x,1]}(G(t) - \frac{t}{\gamma}) < 0$. Hence, from uniform convergence of G_m to G, there exists m_x such that, for all $m \geq m_x$, $\sup_{t \in [x,1]}(G_m(t) - \frac{t}{\gamma}) < 0$. Then $T_m < x$ for all $m \geq m_x$ and $x \geq \limsup T_m$. It follows that $T \geq \limsup T_m$. This proves that $T = \lim T_m$.

- If $G(T) > 0$, then continuity of Λ and uniform convergence of Λ_m give $\Lambda_m(T_m) \to \Lambda(T)$. If $G(T) = 0$, then $\Lambda(T) = 0$.

  In either case, $\Lambda(T) \leq \liminf \Lambda_m(T_m)$.

All the above hold almost surely, hence $\Lambda(T) \leq \liminf \Lambda_m(T_m)$ almost surely. By Fatou's Lemma

$$\begin{aligned} \mathbb{E}(\Lambda(T)) &\leq \mathbb{E}(\liminf \Lambda_m(T_m)) = \liminf \mathbb{E}(\Lambda_m(T_m)) \\ &\leq \liminf \left[\frac{\lambda_m(S_0)}{\lambda_m(S)} \, \gamma \right] = \frac{\lambda(S_0)}{\lambda(S)} \, \gamma. \end{aligned}$$

■

10 Discussion

We have presented a method for finding clusters in a spatial process that controls proportion of false discoveries. As shown in PGVW, such methods can be adapted to control the fraction of false clusters, instead of false proportion. The same can be done with the method in this paper although we do not pursue it here.

An open question is whether there exists some optimal way to choose the finite candidate set of bandwidths $\mathcal{H}_n$. There is a tradeoff in power by taking $\mathcal{H}_n$ too large (making α/m small) or too small (making the set of rejection regions being combined small). Our experience suggests, however, that the choice of the size of $\mathcal{H}_n$ is not crucial in practice.

Another open question is the relationship between the bias correction method used here and the new testing method proposed by Taylor (2004). The contexts are quite different: we are reducing bias due to smoothing while he begins with a Gaussian process and derives new test statistics to eliminate small, insignificant clusters. However, both involve set reduction via Minkowski subtraction so it is possible that there is a connection between the procedures.

11 References

Adler, R.J. (1981). *The Geometry of Random Fields*, Wiley. New York.

Adler, R.J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, Institute of Mathematical Statistics. Hayward.

Adler, R.J. (2000). On excursion sets, tube formulas and maxima of random fields. *The Annals of Applied Probability*, **10**, 1-74.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300.

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

Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, **1**, 1071-1095.

Chaudhuri, P. and Marron, J.S. (2000). Scale space view of curve estimation. *The Annals of Statistics*, **28**, 408–428.

Genovese, C. and Wasserman L. (2002). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society, Series B*, **64**, 499-517.

Genovese, C. and Wasserman L. (2004). A stochastic process apporoach to false discovery control. *The Annals of Statistics*, in press.

Genovese, C. and Wasserman L. (2004b). Exceedance Control of the False Discovery Proportion, Technical Report, Department of Statistics, Carnegie Mellon University.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer, New York.

Koenker, R. and Mizera, I. (2004). Penalized triograms: total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society, Series B*, **66**, 145-163.

Massart, P. (1989). Strong approximation for multivariate empiriocal and related processes, via KMT constructions. *The Annals of Probability*, **17**, 266–291.

Naiman, D.Q. and Priebe, C.E. (2001). Computing scan statistic p values using importance sampling, with applications to genetics and medical image analysis, *Journal of Computational and Graphical Statistics*, **10**, 296–328.

Patil, G.P. and Taillie, C. (2003). Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science*, **18**, 457-465.

Perone Pacifico, M., Genovese, C., Verdinelli, I. and Wasserman, L. (2004). False discovery control for random fields. In press: *Journal of the American Statistical Association.*

Piterbargh, V.I. (1996). *Asymptotic methods in the theory of Gaussian processes and fields*, American Mathematical Society. Providence.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley. New York.

Storey J.D., Taylor J.E. and Siegmund D. (2003). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approahc. *Journal of the Royal Statistical Society, Series B*, **66**, 187-205.

Taylor, J. (2004). Seminar presented at Carnegie Mellon University.

van der Laan, M., Dudoit, S. and Pollard, K. (2004). Multiple testing Part III. Procedures for control of the generalized familywise error rate and proproton of false positives. Working paper 141, Department of Biostatistics, Berkeley.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press. Cambridge.

Walther, G. (1997). Granulometric smoothing. *The Annals of Statistics*, 25, 2273–2299.

Worsley, K.J. (1994). Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. *Advances in Applied Probability*, **26**, 13–42.

Worsley, K.J. (1995). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *The Annals of Statistics*, **23**, 640–669.

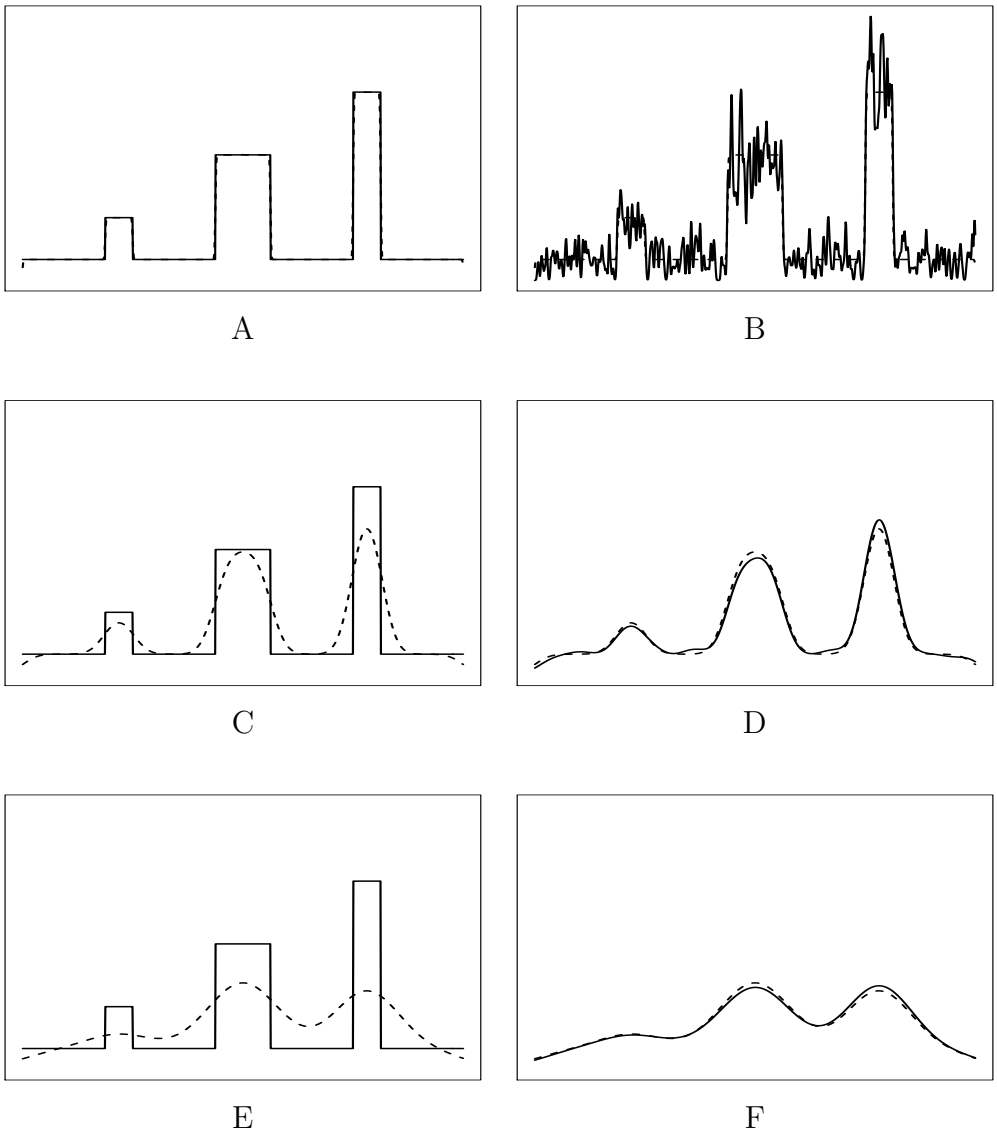


Figure 1: In plots A, C and E the solid line is the true density and the dashed line is the mean of the kernel density estimator for a small bandwidth (A), medium bandwidth (C) and large bandwidth (E). The plots B, D and E show the mean (dashed line) and typical kernel estimates (solid line).

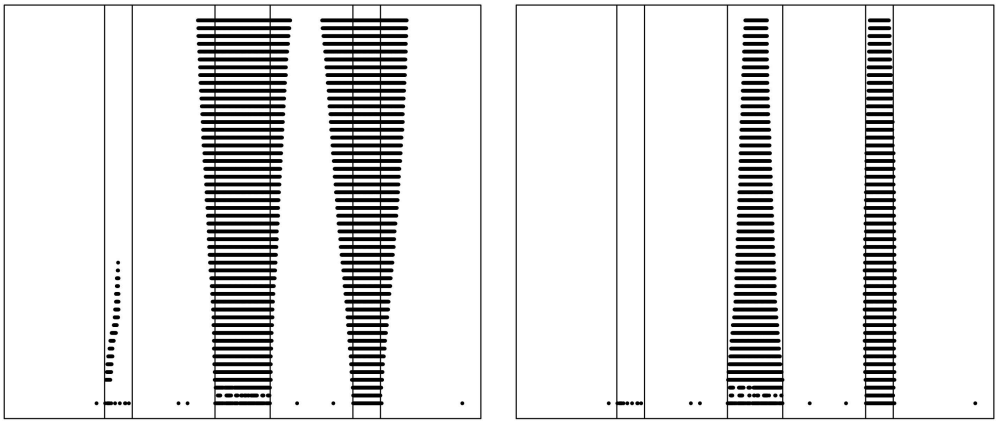


Figure 2: Rejection regions (A: non-shaved, B: shaved) for different bandwidths. Vertical lines delimit the true clusters. As the bandwidth H (vertical axis) increases, the size of the rejection region increases (left panel). This is due to the increasing bias of the density estimate. This results in extra false discoveries not necessarily controlled by the testing procedure. The shaved rejected region is shown in the right panel. The extra false rejections have been eliminated.

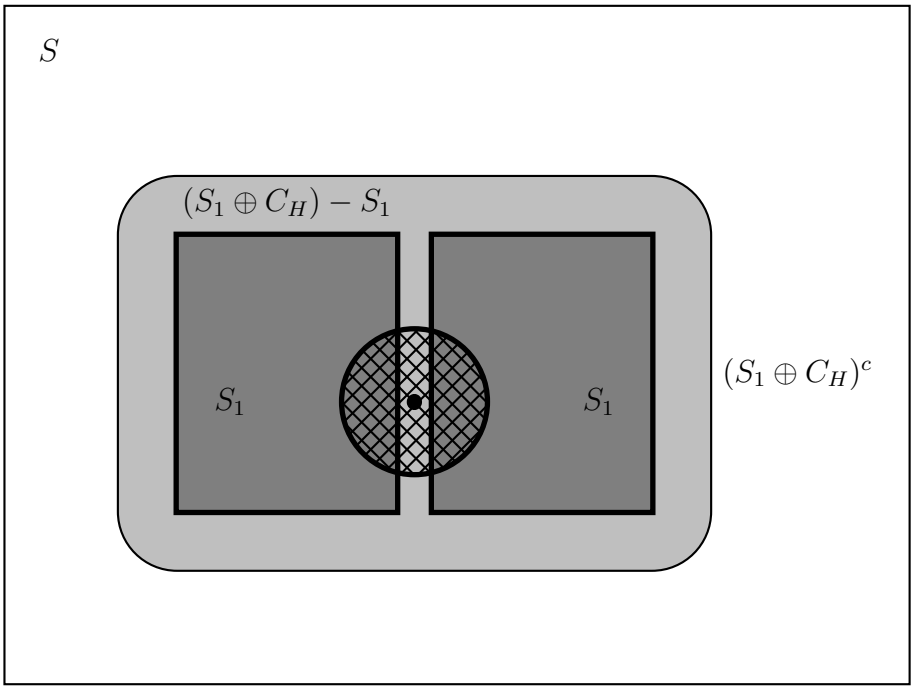


Figure 3: The separation condition (36) fails. S_1 consists of two clusters (the two dark rectangles) and $(S_1 \oplus C_H) - S_1$ is the light gray area. The black dot is a point $s \in (S_1 \oplus C_H) - S_1$. The hatched circle is $(s \oplus C_H)$. Note that $(s \oplus C_H) \cap (S_1 \oplus C_H)^c = \emptyset$ because the two clusters are close together.

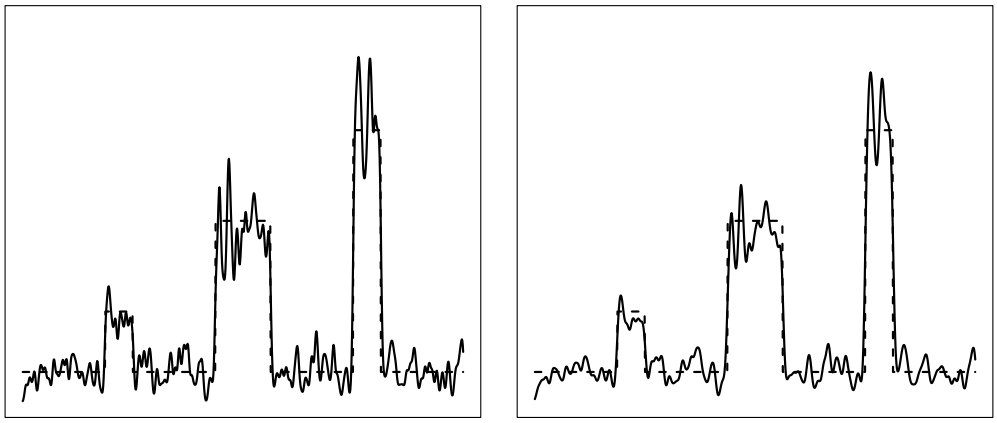


Figure 4: Density estimates with bandwidth chosen for testing (left) by the exceedance control method and bandwidth chosen for estimation to minimize the integrated mean squared error (right).

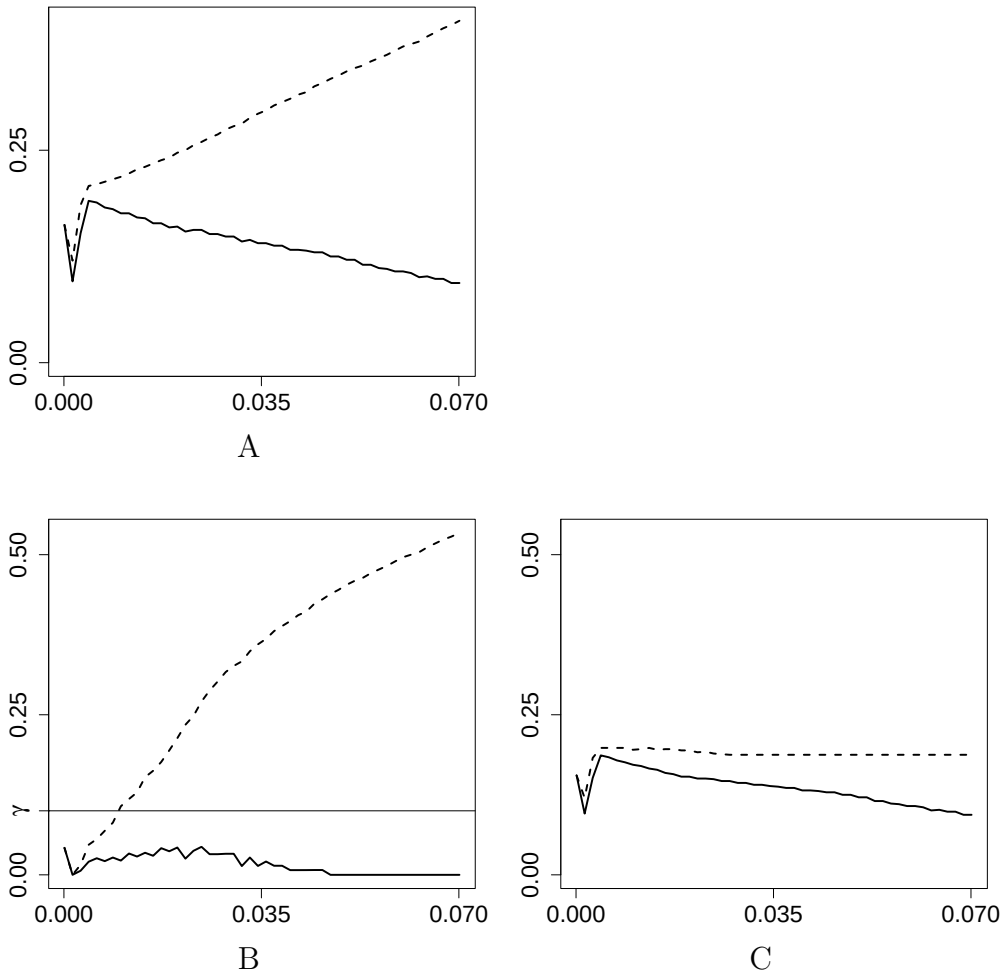


Figure 5: Area (A), FDP (B) and power (C) of non shaved (dashed line) and shaved (solid) rejected regions as functions of the bandwidth. Note that shaving keeps the FDP below the nominal level but without sacrificing too much power.

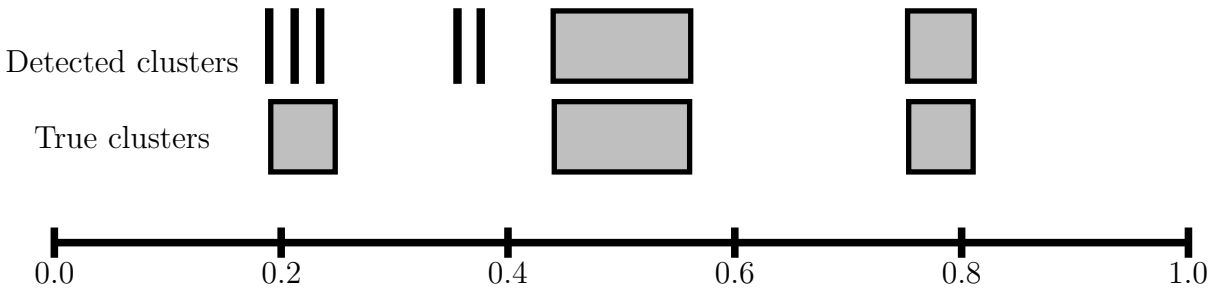


Figure 6: The true clusters and the detected clusters (the set Δ).

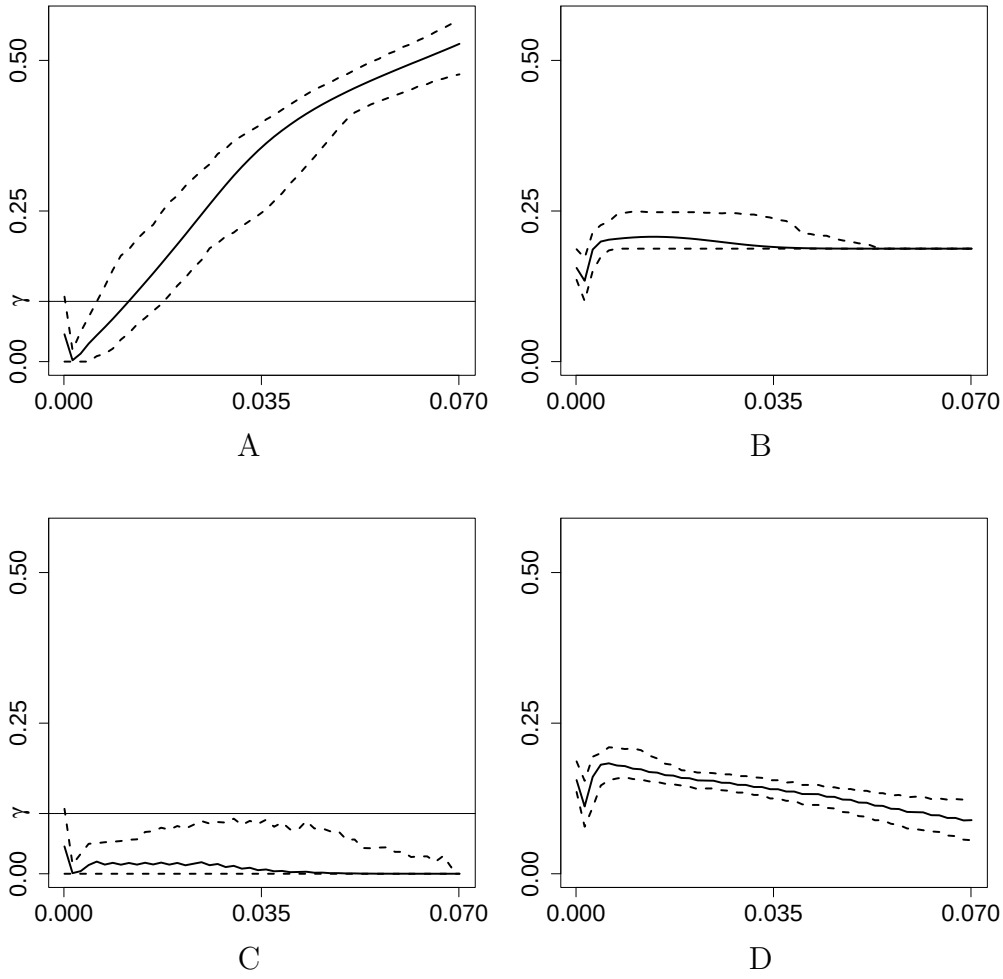
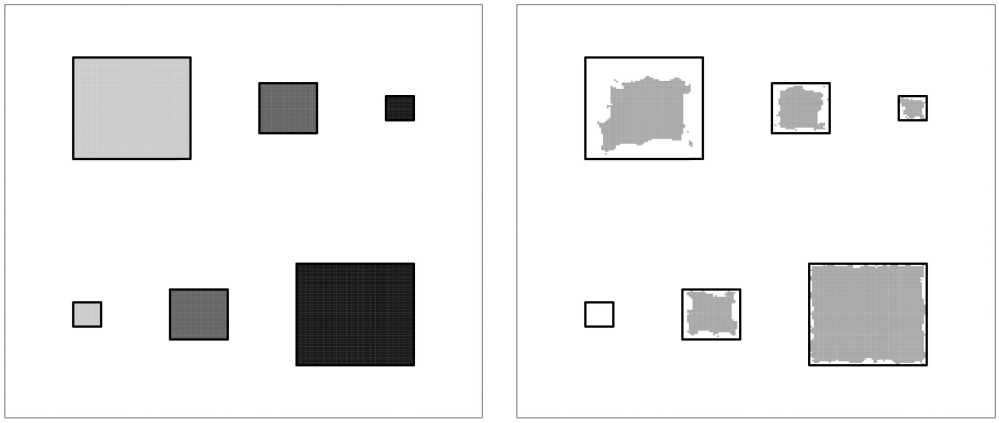


Figure 7: FDP of non shaved (A) and shaved (C) detected clusters, power of non shaved (B) and shaved (D) detected clusters. Minimum, mean and maximum in 1000 simulations.



A B

Figure 8: True density (A) and detected clusters (B).

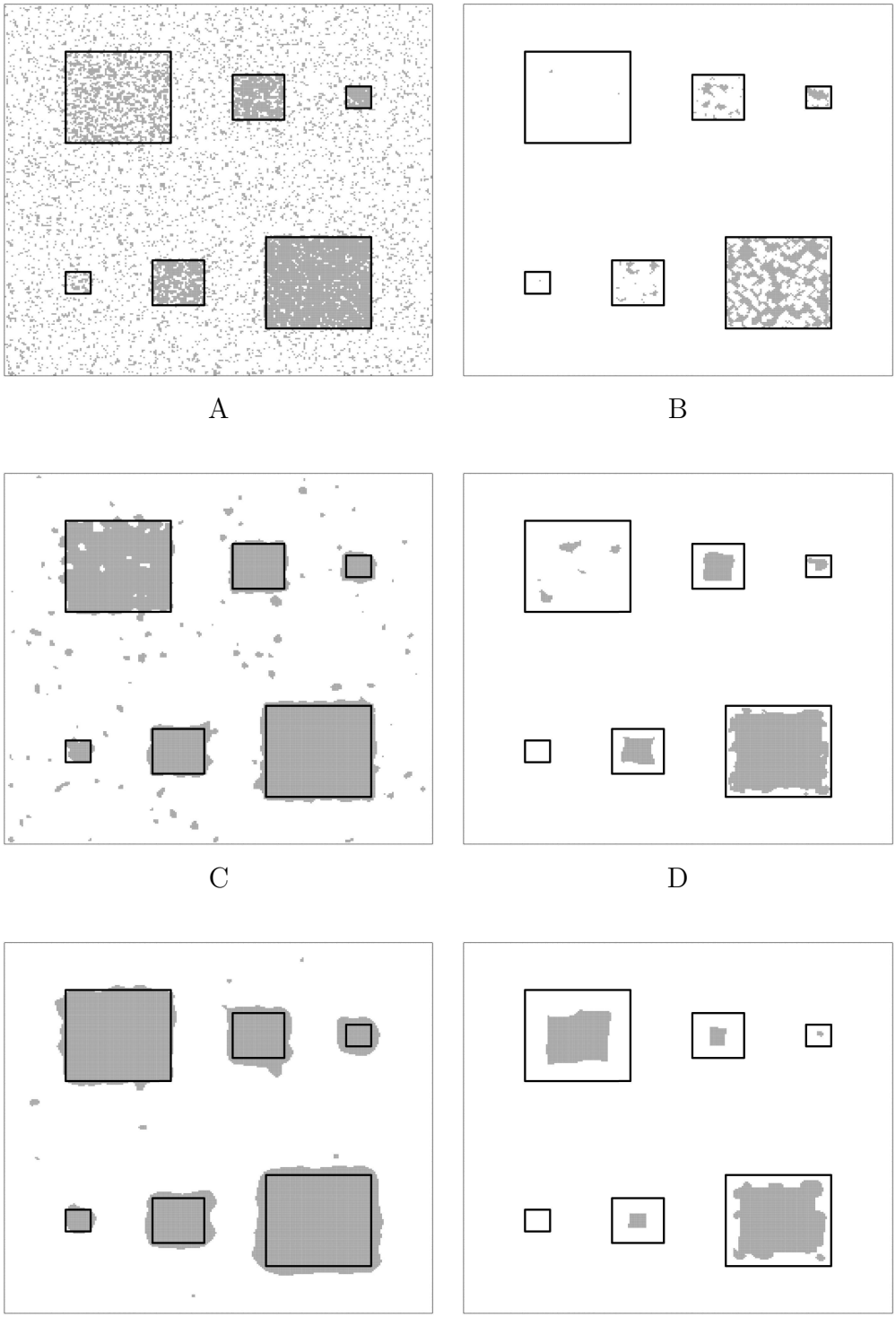


Figure 9: Rejection regions (left: without shaving $\text{aug}_\gamma(R_H)$, right: shaving the clusters $\text{aug}_\gamma(\text{sh}(R_H))$) for small, intermediate and large bandwidths.

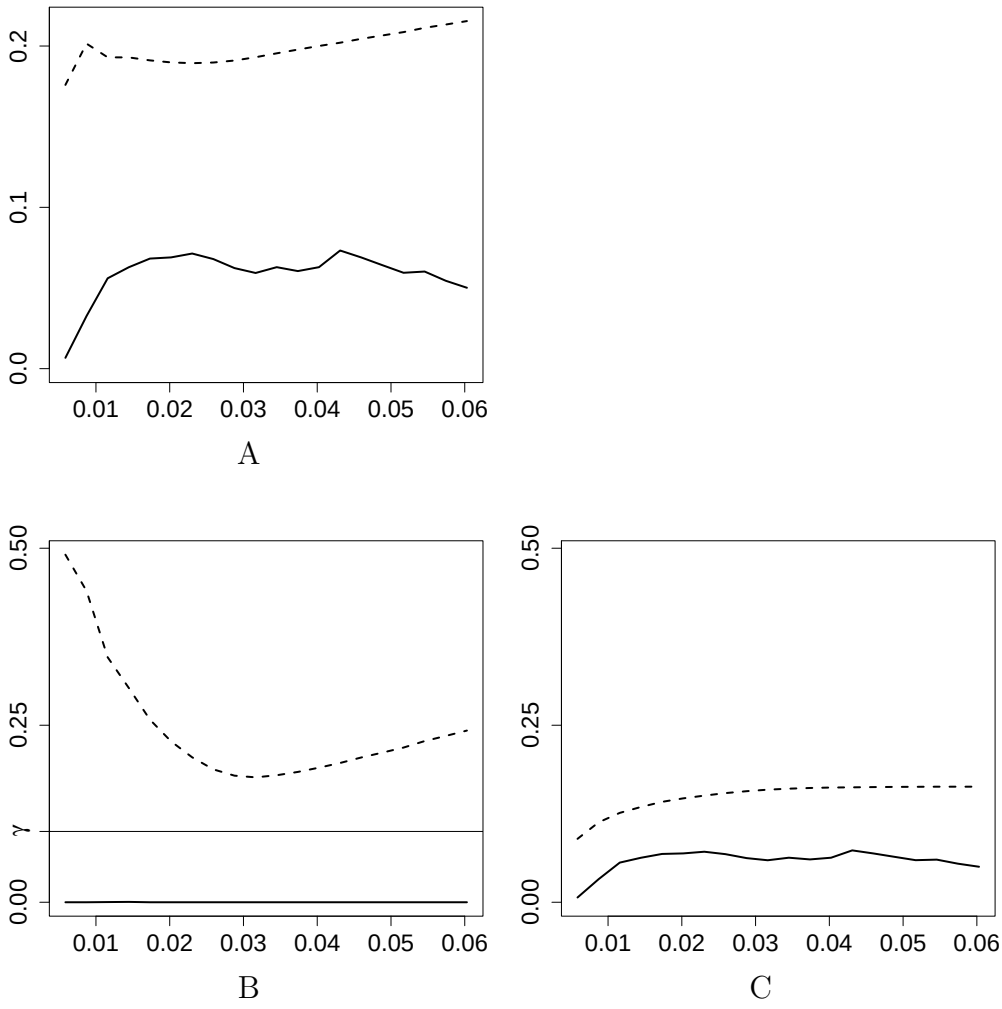


Figure 10: Area (A), FDP (B) and power (C) of non shaved (dashed line) and shaved (solid) rejected regions as functions of the bandwidth.