The Annals of Statistics
2020, Vol. 48, No. 2, 655–678
https://doi.org/10.1214/18-AOS1806

MULTIDIMENSIONAL MULTISCALE SCANNING IN EXPONENTIAL FAMILIES: LIMIT THEORY AND STATISTICAL CONSEQUENCES

BY CLAUDIA KÖNIG*, AXEL MUNK** AND FRANK WERNER†

*Institute for Mathematical Stochastics, University of Goettingen, * claudia.koenig@stud.uni-goettingen.de; ** munk@math.uni-goettingen.de; † f.werner@math.uni-goettingen.de*

We consider the problem of finding anomalies in a d-dimensional field of independent random variables $\{Y_i\}_{i\in\{1,\dots,n\}^d}$, each distributed according to a one-dimensional natural exponential family $\mathcal{F} = \{F_\theta\}_{\theta\in\Theta}$. Given some baseline parameter $\theta_0 \in \Theta$, the field is scanned using local likelihood ratio tests to detect from a (large) given system of regions $\mathcal{R}$ those regions $R \subset \{1,\dots,n\}^d$ with $\theta_i \neq \theta_0$ for some $i \in R$. We provide a unified methodology which controls the overall familywise error (FWER) to make a wrong detection at a given error rate.

Fundamental to our method is a Gaussian approximation of the distribution of the underlying multiscale test statistic with explicit rate of convergence. From this, we obtain a weak limit theorem which can be seen as a generalized weak invariance principle to nonidentically distributed data and is of independent interest. Furthermore, we give an asymptotic expansion of the procedures power, which yields minimax optimality in case of Gaussian observations.

1. Introduction. Suppose we observe an independent, d-dimensional field Y of random variables

$$Y_i \sim F_{\theta_i}, \quad i \in I_n^d := \{1,\dots,n\}^d, \tag{1}$$

where each observation is drawn from the same given one-dimensional natural exponential family model $\mathcal{F} = \{F_\theta\}_{\theta\in\Theta}$, but with potentially different parameters θ_i. Prominent examples include Y_i with varying normal means μ_i or a Poisson field with varying intensities λ_i. Given some baseline parameter $\theta_0 \in \Theta$ (e.g., all $\mu_i = 0$ for a Gaussian field), we consider the problem of finding anomalies (hot spots) in the field Y, that is, we aim to identify those regions $R \subset I_n^d$ where $\theta_i \neq \theta_0$ for some $i \in R$. Here, R runs through a given family of candidate regions $R \in \mathcal{R}_n \subset \mathcal{P}(I_n^d)$ where $\mathcal{P}(A)$ denotes the power set of a set A. For simplicity, we will suppress the subindex n whenever it is clear from the context, that is, write $\mathcal{R} = \mathcal{R}_n$ in what follows. Such problems occur in numerous areas of application ranging from astronomy and biophysics to genetics engineering; specific examples include detection in radiographic images (Kazantsev et al. (2002)), genome screening (Jiang et al. (2016)) and object detection in astrophysical image analysis (Friedenberg and Genovese (2013)), to mention a few. Our setting includes the important special cases of Gaussian (Arias-Castro, Donoho and Huo (2005), Kou (2017), Sharpnack and Arias-Castro (2016), Cheng and Schwartzman (2017)), Bernoulli (Walther (2010)) and Poisson random fields (Kulldorff et al. (2005), Rivera and Walther (2013), Tu (2013), Zhang et al. (2016)). Extensions to models without exponential family structure as well as replacing the baseline parameter θ_0 by a varying field of known baseline intensities can be treated as well (cf. Remark 2.8 below), but to keep the presentation simple, we restrict ourselves to the aforementioned setting.

Received March 2018; revised August 2018.
MSC2010 subject classifications. Primary 60F17, 62H10; secondary 60G50, 62F03.
Key words and phrases. Exponential families, multiscale testing, invariance principle, scan statistic, weak limit, familywise error rate.

1.1. *Methodology.* Inline with the above mentioned references (see also Section 1.4 for a more comprehensive review), the problem of finding hot spots is regarded as a multiple testing problem, that is, many "local" tests on the regions $\mathcal{R}$ are performed simultaneously, while keeping the overall error of wrong detections controllable. For a fixed region $R \in \mathcal{R}$, the likelihood ratio test (LRT) for the testing problem

$$(H_{R,n}) \qquad \forall i \in R : \theta_i = \theta_0$$

versus

$$(K_{R,n}) \qquad \exists i \in R \text{ s.t. } \theta_i \neq \theta_0,$$

is a powerful test in general, and often known to have certain optimality properties (depending on the structure of R; see, e.g., Lehmann and Romano (2005)). Therefore, the LRT will always be considered throughout this paper as the "local" test. We stress, however, that our methodology could also be used for other systems of local tests, provided they obey a sufficiently well behaving asymptotic expansion (see Remark 2.8). The LRT is based on the test statistic

$$T_R(Y, \theta_0) := \sqrt{2 \log\Big(\frac{\sup_{\theta \in \Theta} \prod_{i \in R} f_\theta(Y_i)}{\prod_{i \in R} f_{\theta_0}(Y_i)} \Big)}, \tag{2}$$

where f_θ denotes the density of F_θ, and $H_{R,n}$ is rejected when $T_R(Y, \theta_0)$ is too large. As it is not known a priori which regions R might contain anomalies, that is, for which $R \in \mathcal{R}$ the alternative $(K_{R,n})$ might hold true, it is required to control the familywise error arising from the multiple test decisions of the local tests based on $T_R(Y, \theta_0)$, $R \in \mathcal{R}$. Obviously, without any further restriction on the complexity of $\mathcal{R}$ this error cannot be controlled. To this end, we will assume that the regions R can be represented as a sequence of discretized regions in

$$\mathcal{R} = \mathcal{R}_n := \big\{ R \subset I_n^d | R = I_n^d \cap nR^* \text{ for some } R^* \in \mathcal{R}^* \big\} \tag{3}$$

for some system of subsets (e.g., all hypercubes) of the unit cube $\mathcal{R}^* \subset \mathcal{P}([0,1]^d)$, to be specified later. This gives rise to the sequence of multiple testing problems

$$H_{R,n} \text{ versus } K_{R,n} \qquad \textit{simultaneously} \text{ over } \mathcal{R}_n, \quad n \in \mathbb{N}. \tag{4}$$

The aim of this paper is to provide methodology to control (asymptotically) the familywise error rate (FWER) $\alpha \in (0,1)$ when (4) is considered as a multiple testing problem, that is, to provide a sequence of multiple tests $\Phi := \Phi_n$ (see, e.g., Dickhaus (2014)) for (4) such that

$$\sup_{R \in \mathcal{R}_n} \mathbb{P}_{H_{R,n}} \big[\Phi \text{ rejects any } H_{R',n} \text{ with } R' \subset R \big] \leq \alpha + o(1) \tag{5}$$

as $n \to \infty$. In words, this ensures that the probability of making any wrong detection is controlled at level α, as $n \to \infty$.

This task has been the focus of several papers during the last decades; for a detailed discussion, see Section 1.4. We contribute to this field by providing a general theory for a unifying method in the model (1) including Gaussian, Poisson and Bernoulli observations. In view of (Arias-Castro, Candès and Durand (2011)), where observations from exponential families as in (1) are also discussed, but the local tests are always as in the Gaussian case, we emphasize that our local tests are of type (2), hence exploiting the likelihood in the exponential family. This will result in improved power and better finite sample accuracy (see Frick, Munk and Sieling (2014) for $d = 1$). Our main technical contribution is to prove a weak limit theorem for the asymptotic distribution of our test statistic for general exponential family models as in (1) and arbitrary dimension d. This can be viewed as a "multiscale" weak invariance principle for independent but not necessarily identically distributed r.v.'s. Further, we will provide

an asymptotic expansion of the test's power which leads to minimax optimal detection of the test in specific models.

Throughout the following, we consider tests of scanning-type, controlling the FWER by the maximum over the local LRT statistics in (2), that is,

$$T_n \equiv T_n(Y, \theta_0, \mathcal{R}_n, v) := \max_{R \in \mathcal{R}_n} \big[T_R(Y, \theta_0) - \mathrm{pen}_v(|R|)\big]. \tag{6}$$

Here, $|R|$ denotes the number of points in R. The values

$$\mathrm{pen}_v(r) := \sqrt{2v\big(\log\big(n^d/r\big) + 1\big)}, \tag{7}$$

where log denotes the natural logarithm and act as a scale penalization; see also (Dümbgen and Spokoiny (2001), Dümbgen and Walther (2008), Frick, Munk and Sieling (2014), Walther (2010)). This penalization with proper choice of v guarantees optimal detection power on all scales simultaneously as it prevents smaller regions from dominating the overall test statistic (see Section 2.3). To obtain an a.s. bounded distributional limit for T_n in (6), the constant v in (6) can be any upper bound of the complexity of $\mathcal{R}^*$ measured in terms of the packing number (see Assumption 3 below). For example, whenever $\mathcal{R}^*$ has finite VC-dimension $\nu(\mathcal{R}^*)$, we can choose $v = \nu(\mathcal{R}^*)$. However, we will see that the test has better detection properties if v is as small as possible (see Section 2.3). Hence, from this point of view it is advantageous to know exactly the complexity of $\mathcal{R}^*$ in terms of the packing number, a topic which has received less attention than computing VC-dimensions. Therefore, we compute the packing numbers for three important examples of $\mathcal{R}^*$, namely hyperrectangles, hypercubes and halfspaces explicitly in Appendix A of the Supplementary Material (König, Munk and Werner (2020)).

1.2. *Overview over the results.* To construct a test which controls the FWER (5), we have to find a sequence of universal global thresholds $q_{1-\alpha,n}$ such that

$$\mathbb{P}_0[T_n > q_{1-\alpha,n}] \le \alpha + o(1), \tag{8}$$

where $\mathbb{P}_0 := \mathbb{P}_{H_{I_n^d,n}}$ corresponds to the case that no anomaly is present. Such a threshold suffices, as it can be readily seen from (6) that

$$\begin{aligned}\sup_{R \in \mathcal{R}_n} \mathbb{P}_{H_{R,n}}\big[\Phi \text{ rejects any } H_{R',n} \text{ with } R' \subset R\big] &\le \sup_{R \in \mathcal{R}_n} \mathbb{P}_{H_{R,n}}[\Phi \text{ rejects } H_{R,n}] \\ &\le \mathbb{P}_0[\Phi \text{ rejects } H_{I_n^d,n}].\end{aligned}$$

Given $q_{1-\alpha,n}$, the multiple test will reject whenever $T_n \ge q_{1-\alpha,n}$, and each local test rejects if $T_R(Y, \theta_0) \ge q_{1-\alpha,n} + \mathrm{pen}_v(|R|)$. Due to (5) and (8), this will not be the case with (asymptotic) probability $\le \alpha$ for any $R \in \mathcal{R}_n$ such that $H_{R,n}$ holds true.

To obtain the thresholds $q_{1-\alpha,n}$, we provide a Gaussian approximation of the scan statistic (6) under $\mathbb{P}_0$ given by

$$M_n \equiv M_n(\mathcal{R}_n, v) := \max_{R \in \mathcal{R}_n} \left[|R|^{-1/2} \left|\sum_{i \in R} X_i\right| - \mathrm{pen}_v(|R|)\right] \tag{9}$$

with i.i.d. standard normal r.v.'s X_i, $i \in I_n^d$. We also give a rate of convergence of this approximation (Theorem 2.5), which is determined by the smallest scale in $\mathcal{R}_n$. Based on these results, we obtain the $\mathbb{P}_0$-limiting distribution of T_n as that of

$$M \equiv M\big(\mathcal{R}^*, v\big) := \sup_{R^* \in \mathcal{R}^*} \left[\frac{|W(R^*)|}{\sqrt{|R^*|}} - \mathrm{pen}_v\big(n^d|R^*|\big)\right] < \infty \quad \text{a.s.}, \tag{10}$$

where W is white noise on $[0,1]^d$ and (with a slight abuse of notation) $|R^*|$ denotes the Lebesgue measure of $R^* \in \mathcal{R}^*$. This holds true as soon as $\mathcal{R}^*$ and $\mathcal{R}_n$ have a finite complexity, $\mathcal{R}^*$ consists of sets with a sufficiently regular boundary (see Assumption 2(b) below), and the smallest scales $|R_n|$ of the system $\mathcal{R}_n$ are restricted suitably; see (12) below and the discussion there.

In case of $\mathcal{R}^*$ being the subset of all hypercubes, we will also give an asymptotic expansion of the above test's power, which allows to determine the necessary average strength of an anomaly such that it will be detected with asymptotic probability 1. This is only possible due to the penalization in (6), as otherwise the asymptotic distribution is not a.s. finite. If the anomaly is sufficiently small, we show that the anomalies which can be detected with asymptotic power one by the described multiscale testing procedure are the same as those of the oracle single scale test, which knows the size (scale) of the anomaly in advance. This generalizes findings of Sharpnack and Arias-Castro (2016) to situations where not only the mean of the signal is allowed to change, but its whole distribution. Furthermore, if the observations are Gaussian, and $\mathcal{R}^*$ is the system of squares, our test with the proper choice $v = 1$ (see Example 2.3) achieves the asymptotic optimal detection boundary, that is, no test can have larger power in a minimax sense, asymptotically.

1.3. *Computation*. Note that the weak limit M of T_n in (10) does not depend on any unknown quantity, and hence can be, for example, simulated generically in advance for any given system $\mathcal{R}$ as soon as a bound for the complexity of $\mathcal{R}^*$ can be determined. If the system $\mathcal{R}$ has special convolution-type structure, we discuss an efficient implementation using fast Fourier transforms in Section 3.1 with computational complexity $\mathcal{O}(d\#\text{ scales } n^d \log n)$ for a single evaluation of T_n or M_n. Once the quantiles are precomputed, this allows for fast processing of incoming data sets.

1.4. *Literature review and connections to existing work*. Scan statistics and scanning-type procedures based on the maximum over an ensemble of local tests have received much attention in the literature over the past decades.

To determine the quantile, a common option is to approximate the tails of the asymptotic distribution suitably, as done, for example, by Fang and Siegmund (2016), Naus and Wallenstein (2004), Pozdnyakov et al. (2005), Siegmund and Venkatraman (1995), Siegmund and Yakir (2000) for $d = 1$, by Haiman and Preda (2006) for $d = 2$ and by Jiang (2002) in arbitrary dimensions. If the random field is sufficiently smooth (in contrast to the setting here), the Gaussian kinematic formula or similar tools can be employed; see, for example, Adler (2000), Taylor and Worsley (2007), Schwartzman, Gavrilov and Adler (2011), Cheng and Schwartzman (2017). We also mention Alm (1998), who considers the situation of a fixed rectangular scanning set in two and three dimensions. In all of these papers, no penalization has been used, which automatically leads to a preference for small scales of order $\log(n)$ (see, e.g., Kabluchko and Munk (2009)) and to an extreme value limit, in contrast to the weak invariance principle type limit (10). Arias-Castro et al. (2018) study the case of an unknown null distribution and propose a permutation based approximation, which is shown to perform well in the natural exponential family setting (1), however, only for $d = 1$. Technically, mostly related to our work are weak limit theorems for scale penalized scan statistics, which have, for example, been obtained by Frick, Munk and Sieling (2014) and Sharpnack and Arias-Castro (2016). However, these results are either limited to special situations such as Gaussian observations, or to $d = 1$. If a limit exists, the quantiles of the finite sample statistic can be used to bound the quantiles of the limiting ones as, for example, done by Datta and Sen (2018), Dümbgen and Spokoiny (2001), Rivera and Walther (2013).

Our results can be interpreted in both ways as we provide a Gaussian approximation of the scan statistic in (6) by (9) *and* that we obtain (10) as a weak limit.

Weak limits for T_n as in (10) are immediately connected to those for partial sum processes. Classical KMT-like approximations (see, e.g., Komlós, Major and Tusnády (1976), Massart (1989), Rio (1993)) provide in fact a strong coupling of the whole process $(T_R(Y, \theta_0))_{R\in\mathcal{R}_n}$ to a Gaussian version. Results of this form have been employed for $d = 1$ previously in Frick, Munk and Sieling (2014), Schmidt-Hieber, Munk and Dümbgen (2013). Proceeding like this for general d will restrict the system $\mathcal{R}_n$ to scales r_n s.t. $|R| \geq r_n$ where

$$n^{d-1}\log(n) = o(r_n) \tag{11}$$

as $n \to \infty$, which is unfeasibly large for $d \geq 2$. Therefore, we take a different route and employ a coupling of the maxima in (6) and (9), which relies on recent results by Chernozhukov, Chetverikov and Kato (2014); see also Proksch, Werner and Munk (2018). However, in contrast to the present paper, they do not consider the local LR statistic and require that $|R| = o(n^d)$ for all R. This excludes large scales and leads to an extreme value type limit in contrast to (10) which incorporates all (larger) scales. To make use of Chernozhukov et al.'s (2014) coupling results in our general setting, we provide a symmetrization-like upper bound for the expectation of the maximum of a partial sum process by a corresponding Gaussian version; cf. Lemma 4.2. Doing so we are able to approximate the distribution of T_n in (6) by (9) as soon as we restrict ourselves to $R \in \mathcal{R}_n$ with $|R| \geq r_n$ where the smallest scales only need to satisfy the lower scale bound (LSB)

$$\log^{12}(n) = o(r_n) \quad \text{as } n \to \infty, \tag{12}$$

which compared to (11) allows for considerably smaller scales whenever $d \geq 2$. Note that (12) does not to depend on d. However, as we consider the discretized sets in I_n^d here, the corresponding lower bound a_n for sets in $\mathcal{R}^* \subset \mathcal{P}([0, 1]^d)$ is $n^{-d}\log^{12}(n) = o(a_n)$, which in fact depends on d as now the volume of the largest possible set has been standardized to one (see (3) and Theorem 2.9 below) and coincides with the sampling rate n^{-d} up to a poly-log-factor. In contrast, (11) gives $n^{-1}\log(n) = o(a_n)$, independent of d, which only for $d = 1$ achieves the sampling rate n^{-d}. Under (12), we also obtain $O_{\mathbb{P}}((\log^{12}(n)/r_n)^{1/10})$ as rate of convergence of this approximation (see (16) below).

Also the asymptotic power of scanning-type procedures has been discussed in the literature. An early reference is Arias-Castro, Donoho and Huo (2005), who provide a test for $d = 1$ achieving optimal detection power on the smallest scale. However, to obtain optimal power on all scales, a scale dependent penalization is necessary. We mention Walther (2010), who achieves this for the detection of spatial clusters in a two-dimensional Bernoulli field by scale adaptive thresholding of local test statistics. Butucea and Ingster (2013) for $d = 2$ and Kou (2017) for general d provide optimality of scanning procedures for Gaussian fields. Based on Kabluchko (2011), Sharpnack and Arias-Castro (2016) provide asymptotic power expansions for the multiscale statistic in (6) with a slightly different penalization, yielding minimax optimality in case of d-dimensional Gaussian fields. Inspired by their, however incomplete proof, we are able to generalize these results in case of $\mathcal{R}^*$ being the set of all hypercubes to the exponential family model, (1), despite the fact that under the alternative the whole distribution in (1) might change, whereas for Gaussian fields typically only the mean changes. Doing so we obtain sharp detection boundaries, which are known to be minimax in the Gaussian situation, if the parameter v in the penalization (7) is chosen to be equal to the packing number of the system of hypercubes. In contrast, if v is chosen to be the VC-dimension, the detection power turns out to be suboptimal. This emphasizes the importance of knowledge of the packing number explicitly; for an illustration, cf. Example 2.7.

Finally, we also mention weaker error measures such as the false discovery rate (FDR) as a potential alternative to FWER control, and hence more sensitive tests are to be expected (see, e.g., Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), Li, Munk and Sieling (2016)). However, this is a different task and beyond the scope of our paper.

2. Theory. In this section, we will summarize our theoretical findings. In Section 2.1, we give an overview and details on our precise setting and present our assumptions on the set of candidate regions $\mathcal{R}^*$. Section 2.2 provides the validity of the Gaussian approximation in (9) and determines the $\mathbb{P}_0$-limiting distribution of T_n. In Section 2.3, we derive an asymptotic expansion of the detection power. Throughout this paper, the constants appearing might depend on d.

2.1. *Setting and assumptions.* In the following, we assume that $\mathcal{F} = \{F_\theta\}_{\theta\in\Theta}$ in (1) is a one-dimensional exponential family, which is regular and minimal, that is, the ν-densities of F_θ are of the form $f_\theta(x) = \exp(\theta x - \psi(\theta))$, the natural parameter space

$$\mathcal{N} = \left\{\theta \in \mathbb{R}^d : \int_{\mathbb{R}^d} \exp(\theta x)\,\mathrm{d}\nu(x) < \infty\right\}$$

is open and the cumulant transform ψ is strictly convex on $\mathcal{N}$. Then the moment generating function exists and the random variables Y_i have subexponential tails; see Casella and Berger (1990) and Brown (1986) for details. We further assume that $\operatorname{Var} Y_i > 0$.

EXAMPLE 2.1. Let us discuss three important examples of the model (1).

1. Gaussian fields: Let $Y_i \sim \mathcal{N}(\theta, \sigma^2)$ where the variance $\sigma^2 > 0$ is fixed. In this case, $\psi(\theta) = \frac{1}{2}\theta^2$, and

$$T_R(Y, \theta_0) = \sqrt{|R|}\frac{|\overline{Y}_R - \theta_0|}{\sigma}.$$

2. Bernoulli fields: Let $Y_i \sim \mathrm{Bin}(1, p)$ with $p \in (0, 1)$. Note, that w.l.o.g. the cases $p = 0$ and $p = 1$ are excluded as in these cases one would screen the field correctly, anyway. The natural parameter is $\theta = \log(p/(1-p))$, and using $\psi(\theta) = \log(1 + \exp(\theta))$ we compute

$$T_R(Y, \theta_0) = \sqrt{2|R|\left[\overline{Y}_R \log\left(\frac{\overline{Y}_R}{\frac{\exp(\theta_0)}{1+\exp(\theta_0)}}\right) + (1 - \overline{Y}_R)\log\left(\frac{1-\overline{Y}_R}{\frac{1}{\exp(\theta_0)+1}}\right)\right]}.$$

3. Poisson fields: Let $Y_i \sim \mathrm{Poi}(\lambda)$ with $\lambda \in \mathbb{R}$. Again, $\lambda = 0$ has to be excluded, but this case is again trivial. The natural parameter is $\theta = \log(\lambda)$, and using $\psi(\theta) = \exp(\theta)$ we compute

$$T_R(Y, \theta_0) = \sqrt{2|R|\left[\overline{Y}_R \log\left(\frac{\overline{Y}_R}{\exp(\theta_0)}\right) - (\overline{Y}_R - \exp(\theta_0))\right]}.$$

To derive the Gaussian approximation (9) of T_n in (6), we need to restrict the cardinality of $\mathcal{R}_n$.

ASSUMPTION 1 (Cardinality of $\mathcal{R}_n$). There exist constants $c_1, c_2 > 0$ such that

$$\#(\mathcal{R}_n) \le c_1 n^{c_2}. \tag{13}$$

To furthermore control the supremum in (10), we have to restrict the system of regions $\mathcal{R}^*$ suitably. To this end, we introduce some notation.

For a set $R^* \in \mathcal{R}^*$ and $x \in [0, 1]^d$, we define $d(x, \partial R^*) := \inf_{y\in\partial R^*} \|x - y\|_2$ where ∂R^* denotes the topological boundary of R^*, that is, $\partial R^* = \overline{R^*} \setminus (R^*)^\circ$. Furthermore, we define the ϵ-annulus $R^*(\epsilon)$ around the boundary of R^* for some $\epsilon > 0$ as

$$R^*(\epsilon) := \left\{x \in [0, 1]^d | d\big(x, \partial R^*\big) < \epsilon\right\}.$$

In the following, we will consider the symmetric difference

$$R_1^* \triangle R_2^* := (R_1^* \cup R_2^*) \setminus (R_1^* \cap R_2^*), \quad R_1^*, R_2^* \in \mathcal{R}^*$$

and the corresponding metric

$$\rho^*(R_1^*, R_2^*) := \sqrt{|R_1^* \triangle R_2^*|}, \quad \text{for } R_1^*, R_2^* \in \mathcal{R}^*. \tag{14}$$

To derive the weak limit of T_n, we need to restrict the system $\mathcal{R}^*$ further. Recall the VC-dimension (see, e.g., van der Vaart and Wellner (1996)).

ASSUMPTION 2 (Complexity and regularity of $\mathcal{R}^*$).

(a) The VC-Dimension $\nu(\mathcal{R}^*)$ of the set $\mathcal{R}^*$ is finite.
(b) There exists some constant $C > 0$ such that $|R^*(\epsilon)| \le C\epsilon$ for all $\epsilon > 0$ and all $R^* \in \mathcal{R}^*$ with the Lebesgue measure $|\cdot|$.

Finally, to ensure a.s. boundedness of the limit in (10), we will furthermore require that v in (6) is chosen appropriately. To this end, we introduce the packing number $\mathcal{K}(\epsilon, \rho, \mathcal{W})$ of a subset $\mathcal{W}$ of $\mathcal{R}^*$ w.r.t. a metric ρ, which is given by the maximum number m of elements $W_1, \dots, W_m \in \mathcal{W}$ s.t. $\rho(W_i, W_j) > \epsilon$ for all $i \ne j$, that is, by the largest number of ϵ-balls w.r.t. ρ which can be packed inside $\mathcal{W}$; see, for example, van der Vaart and Wellner ((1996), Definition 2.2.3).

ASSUMPTION 3 (Choice of v). The constant v in (6) and (7) is chosen such that there exist constants $k_1, k_2 > 0$ such that

$$\mathcal{K}\big((\delta u)^{1/2}, \rho^*, \{R \in \mathcal{R}^* : |R| \le \delta\}\big) \le k_1 u^{-k_2} \delta^{-v} \tag{15}$$

for all $u, \delta \in (0, 1]$ with ρ^* as in (14).

Let us briefly comment on the above assumptions.

REMARK 2.2.

- Assumption 1 will allow us to apply recent results by Chernozhukov, Chetverikov and Kato (2014) to couple the process in (6) with a Gaussian version as in (9). Note that Assumption 2(a) immediately implies Assumption 1.
- We stress that the Assumption 2(b) is satisfied whenever $\mathcal{R}^*$ consists of regular Borel sets R^* only, that ism each $R^* \in \mathcal{R}^*$ is a Borel set and $|\partial R^*| = 0$ for all $R^* \in \mathcal{R}^*$.
- Note that Assumption 2(a) also implies that $v = \nu(\mathcal{R}^*)$ is a valid choice in the sense of Assumption 3. This basically follows from the relationship between capacity and covering numbers and a bound on covering numbers from van der Vaart and Wellner ((1996), Theorem 2.6.4). However, (15) might also be satisfied for considerably smaller numbers v (see the examples below).

EXAMPLE 2.3.

1. Consider the set $\mathcal{S}^*$ of all hyperrectangles in $[0, 1]^d$, that is, each $S^* \in \mathcal{S}^*$ is of the form $S^* = [s, t] := \{x \in [0, 1]^d | s_i \le x_i \le t_i \text{ for } 1 \le i \le d\}$. Obviously, the corresponding discretization $\mathcal{S}_n$ consists of hyperrectangles in I_n^d, which are determined by their upper left and lower right corners, that is, $\#(\mathcal{S}_n) \le n^{2d}$, which proves Assumption 1. According to van der Vaart and Wellner ((1996), Example 2.6.1), we have $\nu(\mathcal{S}^*) = 2d$, and as $\mathcal{S}^*$ consists only of regular Borel sets; also Assumption 2 is satisfied. In Appendix A of the Supplementary

Material, we give a simple argument that Assumption 3 holds true whenever $v > 2d - 1$. Employing more refined computations, it can even be shown that $v = 1$ is a valid choice if we allow for additional powers of $(-\log(\delta))$ on the right-hand side of (15); see Theorem 1 in Walther (2010) or Lemma 2.1 in Datta and Sen (2018).

2. We may also consider the (smaller) set $\mathcal{Q}^*$ of all hypercubes in $[0,1]^d$, that is, each $Q^* \in \mathcal{Q}^*$ is of the form $[t, t+h]$ with $t \in [0,1]^d$ and $0 < h \le 1 - \max_{1\le i\le d} t_i$. As $\mathcal{Q}^* \subset \mathcal{S}^*$, Assumptions 1 and 2 are satisfied. Refined computations in Appendix A of the Supplementary Material show that $v = 1$ is a valid choice in the sense of Assumption 3, independent of d (as opposed to the VC-dimension $\nu(\mathcal{Q}^*) = \lfloor \frac{3d+1}{2} \rfloor$ according to Despres (2014)).

3. Let $\mathcal{H}^*$ be the set of all half-spaces in $[0,1]^d$, that is,

$$\mathcal{H}^* := \{H_{a,\alpha} | \alpha \in \mathbb{R}, a \in \mathbb{S}^{d-1}\}, \qquad H_{a,\alpha} := \{x \in [0,1]^d | \langle x, a\rangle \ge \alpha\}.$$

The VC-dimension of $\mathcal{H}^*$ is $\le d+1$ (see, e.g., Devroye and Lugosi (2001), Corollary 4.2), which proves that Assumptions 1 and 2 are satisfied. On the other hand, we prove in Appendix A of the Supplementary Material that $v = 2$ satisfies Assumption 3.

REMARK 2.4. As discussed in the Introduction, we will show in the case of hypercubes that a smaller value of v in Assumption 3, and hence in (7) will lead to a better detection power. More precisely, only for $v = 1$ we will obtain minimax optimality in a certain sense (see Section 2.3 below). In the case of hyperrectangles, this is more involved, but it can, however, be argued along Walther (2010) that for $d = 2$ the choice $v = 1$ yields minimax optimality also in this situation for specific sequences of rectangles.

2.2. *Limit theory.* Now we are in position to show that the quantiles of the multiscale statistic in (6) can be approximated uniformly by those of the Gaussian version in (9), and furthermore that $M_n(\mathcal{R}_n, v)$ in (9) converges to a nondegenerate limit whenever v satisfies Assumption 3. For the former, we require a lower bound on the smallest scale as given in (12). Given a discretized set of candidate regions $\mathcal{R}_n \subset \mathcal{P}(I_n^d)$ and $c > 0$ we introduce

$$\mathcal{R}_{n|c} := \{R \in \mathcal{R}_n | |R| \ge c\}.$$

With this notation, we can formulate our main theorems.

THEOREM 2.5 (Gaussian approximation). *Let Y_i, $i \in I_n^d$ be a field of random variables as in (1), let $\mathcal{R}^*$ be a set of candidate regions satisfying Assumption 1 and let $(r_n)_n \subset (0,\infty)$ be a sequence such that the LSB (12) holds true. Let $v \in \mathbb{R}$ be fixed.*

(a) *Then under $\mathbb{P}_0$*

$$T_n(Y, \theta_0, \mathcal{R}_{n|r_n}, v) - M_n(\mathcal{R}_{n|r_n}, v) = O_{\mathbb{P}}\left(\left(\frac{\log^{12}(n)}{r_n}\right)^{1/10}\right) \tag{16}$$

as $n \to \infty$ with M_n as in (9).

(b) *For all $q \in \mathbb{R}$, we have*

$$\lim_{n\to\infty} \left|\mathbb{P}_0\big[T_n(Y, \theta_0, \mathcal{R}_{n|r_n}, v) > q\big] - \mathbb{P}\big[M_n(\mathcal{R}_{n|r_n}, v) > q\big]\right| = 0. \tag{17}$$

Note that M_n does not depend on any unknown quantities and can, for example, be simulated for fixed n; see Section 3 for details. Beyond this, we can now also derive a weak limit of T_n.

THEOREM 2.6 (Weak $\mathbb{P}_0$ limit). *Under the Assumptions of Theorem 2.5, suppose that also Assumption 2 is satisfied. Then it holds for any fixed $\upsilon \in \mathbb{R}$ under $\mathbb{P}_0$ that*

$$T_n(Y, \theta_0, \mathcal{R}_{n|r_n}, \upsilon) \xrightarrow{\mathcal{D}} M(\mathcal{R}^*, \upsilon) \quad \text{as } n \to \infty, \tag{18}$$

with $M(\mathcal{R}^*, \upsilon)$ as in (10). If υ furthermore satisfies Assumption 3, then $M(\mathcal{R}^*, \upsilon)$ is almost surely finite and nondegenerate.

Note that our proof of Theorem 2.6 explicitly requires the VC-dimension $\nu(\mathcal{R}^*)$ to be finite, and it is not clear if this assumption could be dropped.

EXAMPLE 2.7 (Gaussian approximation in the hyperrectangle/hypercube case). Recall Example 2.3 and let $\mathcal{S}^*$ be the set of all hyperrectangles and $\mathcal{Q}^*$ be the set of all hypercubes in $[0,1]^d$. Then for any sequence r_n satisfying the LSB (12) the approximation (16) holds under $\mathbb{P}_0$ for $\mathcal{S}_{n|r_n}$ and $\mathcal{Q}_{n|r_n}$, respectively. Monte Carlo simulations (by means of (9) with $n = 128$ and $d = 2$) of the densities of M_n with different values of υ are shown in Figure 1. The smallest possible values of υ which we may choose according to Example 2.3 are $\upsilon = 3 + \epsilon$ and $\upsilon = 1$, respectively. The corresponding results are depicted in the first picture of Figure 1 with $\epsilon = 0$ for simplicity. Alternatively, we may use the VC-dimensions $\nu(\mathcal{S}^*) = 4$ and $\nu(\mathcal{Q}^*) = 3$ respectively, which lead to the simulated densities of M_n shown in the second picture of Figure 1. Note that the distributions of $M_n(\mathcal{S}^n, 4)$ and $M_n(\mathcal{Q}_n, 3)$ are extremely close, which somewhat contradicts the intuition that detection in the less complex system of squares should be notably easier than detection in the system of all rectangles. The explanation for this is that $\upsilon = 3$ clearly overpenalizes the system $\mathcal{Q}_n$ of squares. In contrast, if the penalization is chosen according to the smallest possible values satisfying Assumption 3 (which allows for minimax detection in the system of squares; cf. Corollary 2.11 below), then the densities differ substantially.

REMARK 2.8 (Beyond exponential families).

(a) Obviously, $\theta_0 \in \Theta$ can be replaced by a field $(\theta_i)_{i \in I_n^d}$ of known baseline parameters.

(b) The proofs of Theorem 2.5 and Theorem 2.6 rely on a third-order Taylor expansion of T_R and on the subexponential tails of the random variables Y_i, but not explicitly on the exponential family structure. Therefore, if in more general models corresponding assumptions are posed (see also Arias-Castro et al. (2018), Section 2.2), our results do immediately generalize to the case that the observations are not drawn from an exponential family as in (1). As an

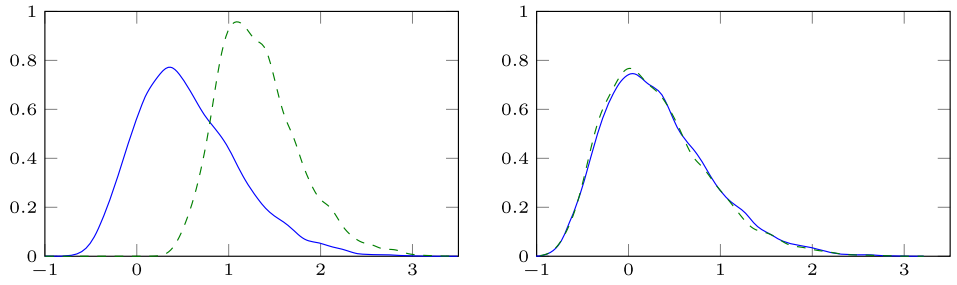

FIG. 1. *Simulated densities of the Gaussian approximations, displayed by a standard kernel estimator obtained from 10^4 runs of the test statistic (9) ($M_n(\mathcal{S}^n, \upsilon)$ (——), $M_n(\mathcal{Q}_n, \upsilon)$ (- - -)). Left: optimal calibration with the covering number $\upsilon = 3$ and $\upsilon = 1$, respectively. Right: alternative calibration using the VC-dimension $\nu(\mathcal{S}^*) = 4$ and $\nu(\mathcal{Q}^*) = 3$.*

example, suppose our observations are drawn from the Weibull distribution with fixed scale parameter $\lambda > 0$ and variable shape parameter $\theta > 0$, that is,

$$f_\theta(x) = \left(\frac{\theta}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\theta-1} \exp\left(-\left(\frac{x}{\lambda}\right)^{\theta}\right), \quad x \geq 0. \tag{19}$$

It is well known that $\{f_\theta\}_{\theta>0}$ is not an exponential family. However, it is clear from (19) that the likelihood-ratio test statistics T_R are arbitrary smooth, that is, a third-order Taylor expansion is valid. If we restrict to $\theta \geq 1$ (nondecreasing failure rate), we immediately obtain subexponential tails, the MLE is unique and for $\theta \geq 2$ one also has asymptotic normality (see, e.g., Farnum and Booth (1997), Smith (1985)). As a consequence, a similar coupling result as in Theorem 4.3 below is possible, which would yield analogs to Theorems 2.5 and 2.6 also in this nonexponential family situation. We emphasize that also Theorem 2.9 below can be generalized accordingly.

2.3. *Asymptotic power.* In this section, we will analyze the power of our multiscale testing approach in the hypercube case. The detection power clearly depends on the size and strength of the anomaly. To describe the latter, we will frequently employ the functions

$$m(\theta) := \psi'(\theta) = \mathbb{E}[Y], \qquad v(\theta) := \psi''(\theta) = \mathbb{V}[Y]$$

for $Y \sim F_\theta$.

Heuristics. The key point for the following power considerations is that the observations in (1) can be approximated as

$$\frac{Y_i - m(\theta_0)}{\sqrt{v(\theta_0)}} = \frac{m(\theta_i) - m(\theta_0)}{\sqrt{v(\theta_0)}} + \frac{\sqrt{v(\theta_i)}}{\sqrt{v(\theta_0)}} \frac{Y_i - m(\theta_i)}{\sqrt{v(\theta_i)}}, \tag{20}$$

that is, as "signal" $v(\theta_0)^{-1/2}(m(\theta_i) - m(\theta_0))$, which is nonzero on the anomaly only, plus a standardized noise component $(Y_i - m(\theta_i))/\sqrt{v(\theta_i)}$ which is scaled by a factor $v_i := \sqrt{v(\theta_i)/v(\theta_0)}$. In case of Gaussian observations with variance 1, one has $v_i \equiv 1$ and recovers the situation considered by Sharpnack and Arias-Castro (2016). Whenever the "signal" part in (20) is strong enough, the anomaly should be detected. In the following, we will make this statement mathematically precise and also give a comparison of the multiscale testing procedure with an oracle procedure.

Considered alternatives. Consider a given family $(Q_n^*)_{n\in\mathbb{N}}$ of hypercube anomalies $Q_n^* \subset [0,1]^d$ with Lebesgue measure $|Q_n^*| = a_n \in (0,1)$. The corresponding discretized anomalies $Q_n := I_n^d \cap nQ_n^* \subset I_n^d$ have size $|Q_n| \sim n^d a_n$. We will consider alternatives $K_{i,n}$ in (4) where $\theta^n \in \Theta^{n^d}$ s.t.

$$\theta_i^n = \theta_1^n \mathbb{I}_{Q_n} + \theta_0 \mathbb{I}_{Q_n^c}. \tag{21}$$

The parameters θ_1^n determine the total strength of the anomaly, which is given by

$$\mu^n(Q_n) := \sqrt{|Q_n|}\frac{m(\theta_1^n) - m(\theta_0)}{\sqrt{v(\theta_0)}}.$$

Clearly, any anomaly with fixed size or strength can be detected with asymptotic probability 1. Therefore, we will consider vanishing anomalies in the sense that

$$a_n \searrow 0, \qquad \mu^n(Q_n) \nearrow \infty, \quad \text{as } n \to \infty. \tag{22}$$

Furthermore, we will restrict to parameters θ_1^n in (21) which yield uniformly bounded variances and uniform subexponential tails for the standardized version, this is

$$\mathbb{E}_\theta\left[\exp\left(s\frac{Y-m(\theta)}{\sqrt{v(\theta)}}\right)\right] \le C \quad \text{for all } 0 \le s \le t \text{ and } \theta \in \{\theta_0\} \cup \bigcup_{n\in\mathbb{N}} \{\theta_1^n\}, \tag{23}$$

$$\underline{v} \le \sqrt{\frac{v(\theta_1^n)}{v(\theta_0)}} \le \bar{v} \quad \text{for all } n \in \mathbb{N} \tag{24}$$

for $Y \sim F_\theta$ with constants $t > 0$, $C > 0$ and $0 < \underline{v} < \bar{v} < \infty$.

In case of Gaussian observations with variance σ^2, (23) and (24) are obviously satisfied, for a Poisson field this means that the intensities are bounded away from zero and infinity.

Oracle and multiscale procedure. Recall that $\mathcal{Q}^*$ is the set of all hypercubes in $[0,1]^d$ (cf. Example 2.3), and $\mathcal{Q}^n$ its discretization (cf. (3)).

If the size a_n of the anomaly is known, but its position is still unknown, then one would naturally restrict the set of candidate regions to $\mathcal{R}^*_{\mathrm{O}} := \{Q^* \in \mathcal{Q}^* | |S^*| = a_n\}$, and consequently scan only over (cf. (3))

$$\mathcal{R}^{\mathrm{O}}_n := \{Q \subset I_n^d | Q = I_n^d \cap nQ^* \text{ for some } Q^* \in \mathcal{R}^*_{\mathrm{O}}\}.$$

As for the true anomaly $Q^* \in \mathcal{R}^*_{\mathrm{O}}$, its discretized version Q_n also satisfies $Q_n \in \mathcal{R}^{\mathrm{O}}_n$. This gives rise to an oracle test, which rejects whenever $T_n(Y, \theta_0, \mathcal{R}^{\mathrm{O}}_n, v) > q^{\mathrm{O}}_{1-\alpha,n}$ where $q^{\mathrm{O}}_{1-\alpha,n}$ is the $(1-\alpha)$-quantile of $M_n(R^{\mathrm{O}}_n, v)$ as in (9). Similar as in Theorem 2.5, one can show that this quantile sequence ensures the oracle test to have asymptotic level α. The asymptotic power of this oracle test can be seen as a benchmark for any multiscale test.

To obtain a competitive multiscale procedure, let us choose some r_n satisfying the LSB (12), and furthermore assume that $r_n = o(n^d a_n)$, as otherwise the multiscale procedure will never be able to detect the true anomaly (as it is not contained in the set of candidate regions which we scan over). As now position and size of the anomaly are unknown, we consider all such sets in $\mathcal{R}^*_{\mathrm{MS}} = \mathcal{Q}^*$ as candidate regions, and consequently scan over

$$\mathcal{R}^{\mathrm{MS}}_{n|r_n} := \{Q \subset I_n^d | Q = I_n^d \cap nQ^* \text{ for some } Q^* \in \mathcal{Q}^* \text{ and } |Q| \ge r_n\}.$$

Clearly, the true anomaly Q^* satisfies $Q^* \in \mathcal{R}^*_{\mathrm{MS}}$, and by $r_n = o(n^d a_n)$ its discretized version Q_n also satisfies $Q_n \in \mathcal{R}^{\mathrm{MS}}_{n|r_n}$. This gives rise to a multiscale test, which rejects whenever $T_n(Y, \theta_0, \mathcal{R}^{\mathrm{MS}}_{n|r_n}, v) > q^{\mathrm{MS}}_{1-\alpha,n}$ where $q^{\mathrm{MS}}_{1-\alpha,n} := q_{1-\alpha}^{M_n(\mathcal{R}^{\mathrm{MS}}_{n|r_n}, v)}$ is the $(1-\alpha)$-quantile of $M_n(\mathcal{R}^{\mathrm{MS}}_{n|r_n}, v)$ as in (9). Theorem 2.5 ensures that the multiscale test has asymptotic level α.

Now, due to Theorem 2.6 $q^*_{1-\alpha} := q_{1-\alpha}^{M(\mathcal{Q}^*, v)} < \infty$ whenever v satisfies Assumption 3 (which corresponds to $v \ge 1$ here), it holds that

$$q^{\mathrm{O}}_{1-\alpha,n} \le q^{\mathrm{MS}}_{1-\alpha,n} \le q^*_{1-\alpha} < \infty$$

for all $n \in \mathbb{N}$.

Asymptotic power. We will now show that the multiscale procedure described above (which requires no a priori knowledge on the scale of the anomaly) asymptotically detects the same anomalies with power 1 as the oracle benchmark procedure for a known scale. Hence, the penalty choice to calibrate all scales as in (6) (where $\mathcal{R}^* = \mathcal{Q}^*$), renders the adaptation to all scales for free, at least asymptotically. This can be seen as a structural generalization of (Sharpnack and Arias-Castro (2016), Theorems 2 and 4), as under the alternative the whole distribution in (1) and not just its mean might change. Also the power considerations in Proksch, Werner and Munk (2018) restrict to this simpler case.

THEOREM 2.9. *In the setting described above, let $a_n \searrow 0$ be a sequence of scales such that $(\log n)^{12}/n^d = o(a_n)$ as $n \to \infty$. Denote by*

$$F(x, \mu, \sigma^2) := \Phi\left(-\frac{x+\mu}{\sigma}\right) + \Phi\left(\frac{\mu - x}{\sigma}\right), \quad x \geq 0$$

the survival function of a folded normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$. Let furthermore $v \geq 1$. If (22) is satisfied, then the following holds true:

(a) *The single scale procedure has asymptotic power*

$$\begin{aligned}&\mathbb{P}_{\theta^n}\big[T_n(Y, \theta_0, \mathcal{R}_n^{\mathrm{O}}, v) > q_{1-\alpha,n}^{\mathrm{O}}\big] \\ &\quad = \alpha + (1-\alpha) F\left(q_{1-\alpha,n}^{\mathrm{O}} + \sqrt{2v \log\left(\frac{1}{a_n}\right)}, \right. \\ &\qquad \left. n^{d/2}\sqrt{a_n}\frac{m(\theta_1^n) - m(\theta_0)}{\sqrt{v(\theta_0)}}, \frac{v(\theta_1^n)}{v(\theta_0)}\right) + o(1).\end{aligned}$$

(b) *If $a_n = o(n^{\beta-d})$ with $\beta > 0$ sufficiently small, then the multiscale procedure has asymptotic power*

$$\begin{aligned}&\mathbb{P}_{\theta^n}\big[T_n(Y, \theta_0, \mathcal{R}_{n|r_n}^{\mathrm{MS}}, v) > q_{1-\alpha,n}^{\mathrm{MS}}\big] \\ &\quad \geq \alpha + (1-\alpha) F\left(q_{1-\alpha,n}^{\mathrm{MS}} + \sqrt{2v \log\left(\frac{1}{a_n}\right)}, \right. \\ &\qquad \left. n^{d/2}\sqrt{a_n}\frac{m(\theta_1^n) - m(\theta_0)}{\sqrt{v(\theta_0)}}, \frac{v(\theta_1^n)}{v(\theta_0)}\right) + o(1).\end{aligned}$$

REMARK 2.10. In Sharpnack and Arias-Castro (2016), a similar result in case of Gaussian observations is shown. However, the proof of (Sharpnack and Arias-Castro (2016), Theorem 4) is incomplete and we require the additional condition that $a_n = o(n^{\beta-d})$ with $\beta > 0$ sufficiently small for our proof. In Proksch, Werner and Munk (2018), it suffices to assume $a_n \searrow 0$, as large scales have been excluded s.t. the maximum tends to a Gumbel limit.

The above theorem allows us to explicitly describe those anomalies which will be detected with asymptotic power 1.

COROLLARY 2.11. *Under the setting in this section, the assumptions of Theorem 2.9 and if v satisfies Assumption 3, any such anomaly is detected with asymptotic power 1 either by the single scale or the multiscale testing procedure if and only if*

$$\frac{\sqrt{2v \log(\frac{1}{a_n}) v(\theta_0)} - n^{d/2}\sqrt{a_n}|m(\theta_1^n) - m(\theta_0)|}{\sqrt{v(\theta_1^n)}} \to -\infty \tag{25}$$

as $n \to \infty$.

REMARK 2.12. Equation (25) implies that a smaller value of v makes more anomalies detectable. However, this is limited by Assumption 3, which requires v to be an upper bound of the complexity of $\mathcal{Q}^*$ in terms of the packing number. As we compute in Appendix A of the Supplementary Material, this yields $v = 1$ as the optimal choice.

EXAMPLE 2.13.

1. In case of Gaussian observations $Y_i \sim \mathcal{N}(\Delta_n \mathbb{I}_{Q_n}, \sigma^2)$ with variance σ^2, where the baseline mean is 0 and Δ_n the size of the anomaly, this yields detection if and only if
$$|\Delta_n| n^{d/2} \sqrt{a_n} \gtrsim \sigma \sqrt{2v \log\left(\frac{1}{a_n}\right)} \quad \text{as } n \to \infty.$$
If we calibrate the statistic with $v = 1$ (cf. Example 2.3), then this coincides with the well-known asymptotic detection boundary for hypercubes; see, for example, Arias-Castro, Donoho and Huo (2005), Frick, Munk and Sieling (2014) for $d = 1$, Butucea and Ingster (2013) for $d = 2$ or Kou (2017) for general d.
2. For Bernoulli r.v.'s $Y_i \sim \text{Ber}(p_0 \mathbb{I}_{Q_n^c} + p_n \mathbb{I}_{Q_n})$ with $p_0, p_n \in (0, 1)$ s.t. $p_0 + p_n \leq 1$, the condition (25) reads as follows:
$$\frac{\sqrt{2v p_0(1-p_0) \log(\frac{1}{a_n})} - n^{d/2} \sqrt{a_n} |p_n - p_0|}{\sqrt{p_n(1-p_n)}} \to -\infty.$$
Note that the minimax detection rate is unknown in this case to best of our knowledge.
3. For a Poisson field $Y_i \sim \text{Poi}(\lambda_0 \mathbb{I}_{Q_n^c} + \lambda_n \mathbb{I}_{Q_n})$ with $\lambda_0, \lambda_n > 0$, Theorem 2.9 and Corollary 2.11 can only be applied if λ_n is a bounded sequence. In this case, (25) reduces to
$$\frac{\sqrt{2v\lambda_0 \log(\frac{1}{a_n})} - n^{d/2} \sqrt{a_n} |\lambda_n - \lambda_0|}{\sqrt{\lambda_n}} \to -\infty.$$
Again, the minimax detection rate is unknown in this case to best of our knowledge.

3. Numerical simulations. In this section, we provide an implementation of the suggested multiscale testing procedure and discuss its computational complexity. Furthermore, we explore the influence of the penalization parameter v in (6) on the finite sample power, the speed of convergence in (18) and the influence of the LSB r_n on the distribution of T_n in (6).

3.1. *Implementation and computational complexity.* To evaluate the statistic T_n in (6) in general, all local statistics T_R have to be computed separately. Therefore, the computational complexity will in general be of the order $\mathcal{O}(\#\mathcal{R}_n \cdot n^d)$. Note that for the situations mentioned in Example 2.1, each T_R is given by a function of the local mean $\bar{Y}_R$, which already reduces the computational effort.

However, if the system of candidate regions $\mathcal{R}^*$ has a special convolution-type structure, a more efficient evaluation is possible. Therefore, assume that there is a global shape $B \subset I_n^d$ such that for every $R \in \mathcal{R}_n$ there exist $t, h \in I_n^d$ with $t_i + h_i \leq n$ for all $1 \leq i \leq d$ and $\mathbf{1}_R(x) = \mathbf{1}_B((x-t)/h)$. This is, for example, the case for the system of hyperrectangles or the system of hypercubes. In this special situation, we may use the fast Fourier transform (FFT). If we denote by $*$ a discrete convolution, then it holds that
$$\bar{Y}_R = Y * \mathbf{1}_B\left(\frac{\cdot - t}{h}\right) = \mathcal{F}^{-1}\left(\mathcal{F}(Y) \cdot \mathcal{F}\left(\mathbf{1}_B\left(\frac{\cdot - t}{h}\right)\right)\right).$$
Consequently, for a fixed scale h, all corresponding values T_R can be computed by means of 3 FFTs. Note that no zero-padding is necessary here as for an inverse problem (see Proksch, Werner and Munk (2018)). This gives a computational complexity of $\mathcal{O}(d\#\text{scales}\, n^d \log n)$ for a single evaluation of the test statistic T_n in (6). In the hyperrectangle and hypercube case, using all possible scales, this yields $\mathcal{O}(dn^{2d} \log n)$ and $\mathcal{O}(dn^{d+1} \log n)$ respectively.

Compared the naive implementation described at the beginning, which yield complexities $\mathcal{O}(n^{3d})$ and $\mathcal{O}(n^{2d+1})$, respectively, this is a significant improvement.

We also briefly mention a possible implementation using cumulative sums, which is also possible for hyperrectangles and hypercubes. Once the cumulative sum of all observations has been computed, each local mean $\bar{Y}_R$ can be computed summing or subtracting 2^d values. Hence, this implementation gives in in general a computational complexity of $\mathcal{O}(n^d + 2^d \cdot \#\mathcal{R}_n)$, which yields $\mathcal{O}(2^d n^{2d})$ and $\mathcal{O}(2^d n^{d+1})$ for hyperrectangles and hypercubes, respectively. Compared to the implementation using FFT described above, this differs by a factor $d2^{-d} \log n$, which reveals the FFT implementation to be more efficient for large d.

Note that in many applications, a priori information is available, which allows to select a (small) subset of scales instead of using all, which clearly reduces the computational effort further.

We emphasize that the quantiles $q_{1-\alpha,n}$ of the approximating Gaussian version (9) can be universally precomputed and stored as long as n and the system $\mathcal{R}_n$ do not change, and for large n the asymptotic values can be used in a universal manner (cf. Section 3.3 below). Even for small values of α, the above implementation allows to simulate the $(1-\alpha)$-quantile of the Gaussian approximation (9) efficiently. This makes fast computations on incoming data sets in an "online" fashion possible, which is important in many applications. In contrast, permutation based methods as considered in (Arias-Castro et al. (2018)) require to simulate the unknown null distribution separately for every given problem instance.

3.2. *Influence of v on the power.* To study the influence of the penalization parameter v on the power of the procedure, we turn to the setting of Section 2.3. Let $n = 512$ and $d = 2$. For simplicity, we consider a Gaussian model, that is, $F_\mu = \mathcal{N}(\mu, 1)$ in (1) and choose $\mu_i = \mu \mathbb{I}_Q$ with $\mu \in \{1, 1.2\}$ and $|Q| \in \{6^2, 7^2\}$. Afterwards, we simulate the empirical power from 1000 repetitions. This procedure is performed for the VC-based choice $v = 3$ and for the capacity-based choice $v = 1$, which is asymptotically minimax optimal (see Corollary 2.11). The results are depicted in Table 1.

We find that the power for $v = 1$ is substantially larger than the one obtained by using the VC-dimension for calibration. This is in line with our findings from Example 2.7.

3.3. *Speed of convergence in* (18). To investigate the speed of convergence in (18), we consider the system of hypercubes $\mathcal{R}_n = \mathcal{Q}_n$ as in Section 3.2. Figure 2 shows estimated densities of M_n for different values of n in dimensions $d = 1$ and $d = 2$.

We find that the speed of convergence of M_n toward the weak limit M in (10) decreases with increasing d, but we can however conclude that the distribution of M_n stabilizes already at moderate values of n. This is especially helpful in situations, where data with significantly larger sample size n is given, such that the distribution of T_n cannot be simulated anymore.

TABLE 1
Empirical power of the investigated testing procedure for different choices of v in different Gaussian settings determined by μ (columns) and $|Q|$ (rows)

	$v = 1$			$v = 3$	
$\|Q\|$ and μ	1	1.2	$\|Q\|$ and μ	1	1.2
6^2	0.429	0.817	5^2	0.104	0.182
7^2	0.809	0.983	6^2	0.187	0.577

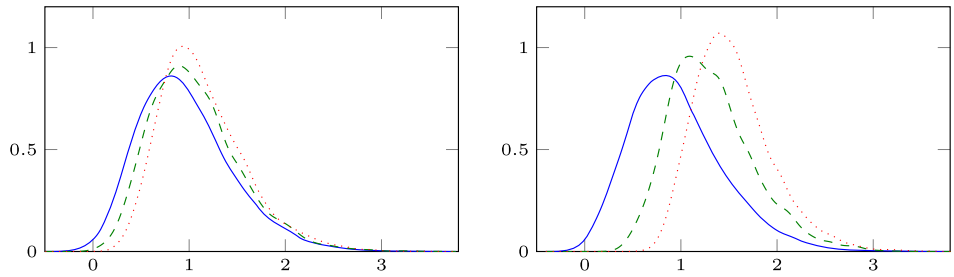

FIG. 2. *Simulated densities of the Gaussian approximations, displayed by a standard kernel estimator obtained from* 10^4 *runs of the test statistic* (9), *with different values of* n *and* d. *Left*: $d = 1$ *and* $M_n(\mathcal{Q}_n, 1)$ *with* $n = 2^{10}$ (——), $n = 2^{12}$ (- - -) *and* $n = 2^{14}$ (·····). *Right*: $d = 2$ *and* $M_n(\mathcal{Q}_n, 1)$ *with* $n = 2^5$ (——), $n = 2^7$ (- - -) *and* $n = 2^9$ (·····).

3.4. *Influence of the lower scale bound* r_n. Let us again consider $n = 512$, $d = 2$ and the system of hypercubes $\mathcal{R}_n = \mathcal{Q}_n$ as in Section 3.2. Let $Y_i \sim \mathrm{Bin}(1, \theta)$ and $\theta_0 = 0.5$. Figure 3 shows the simulated densities of $T_n(Y, \theta_0, \mathcal{R}_{n|r_n}, v)$ for different values of r_n.

In conclusion, we find that the distribution of $T_n(Y, \theta_0, \mathcal{R}_{n|r_n}, v)$ is surprisingly robust w.r.t. the choice of r_n even below the LSB (12).

4. Auxiliary results. In this section, we will present the main ingredients needed for our proofs, which might be of independent interest. The key steps of our proofs will be sketched in Section 5, and details can be found in Appendix B of Supplement A. One tool is a coupling result which allows us to replace the maximum over partial sums of standardized NEF r.v.'s by a maximum over a corresponding Gaussian version. This can be obtained from recent results by Chernozhukov, Chetverikov and Kato (2014) as soon as certain moments can be controlled, which is the purpose of the following two lemmas, which generalize known bounds for sub-Gaussian random variables to subexponential ones. In what follows, the letter $C > 0$ denotes some constant, which might change from line to line.

The following lemma gives an upper bound for the maximum of uniformly subexponential random variables.

LEMMA 4.1. *Let* W_i, $i = 1, 2, \ldots$ *be independent subexponential random variables s.t. there exist* $k_1 > 1$ *and* $k_2 > 0$ *s.t.*

$$\mathbb{P}\big[|W_i| > t\big] \le k_1 \exp(-k_2 t) \tag{26}$$

for all i. *Then for all* $m \in \mathbb{N}$ *there exists a constant* C, *s.t. for all* $N \ge 2$,

$$\mathbb{E}\Big[\max_{1 \le i \le N} |W_i|^m\Big] \le C(\log N)^m.$$

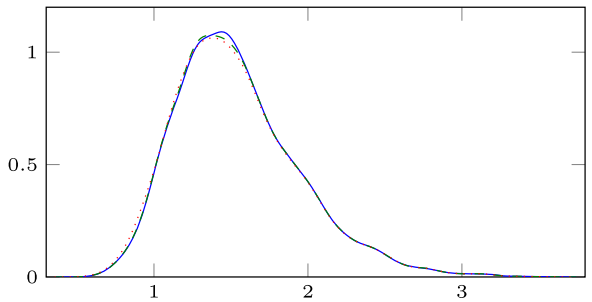

FIG. 3. *Simulated densities of the test statistic* $T_n(Y, \theta_0, \mathcal{R}_{n|r_n}, v)$ *in* (6) *in case of i.i.d. Bernoulli observations with* $p = 1/2$ *for different values of* r_n: $r_n = 2^3$ (——), $r_n = 2^4$ (- - -), $r_n = 2^5$ (·····).

Lemma 4.1 might be of independent interest, as it generalizes the well-known bound

$$\mathbb{E}\Big[\max_{1\le i\le N}|X_i|\Big]\le C\sqrt{\log N} \tag{27}$$

for sub-Gaussian random variables to subexponential random variables.

Now we will show that the maximum over the partial sum process of independent random variables can be bounded by the maximum over the corresponding Gaussian version. The latter can be controlled as in (27) by exploiting the fact that a maximum over dependent Gaussian random variables is always bounded by a maximum over corresponding independent Gaussian random variables (see, e.g., Šidák (1967))

$$\mathbb{E}\left[\max_{I\in\mathcal{I}}\frac{|X_I|}{\sqrt{|I|}}\right]\le C\sqrt{\log\big(\#(\mathcal{I})\big)} \tag{28}$$

with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $X_I := \sum_{i\in I} X_i$. This allows us to prove the following.

LEMMA 4.2. *Let $(Z_i)_{i=1,\dots,N}$ be independent random variables with $\mathbb{E}[Z_i]=0$ and denote $Z_I := \sum_{i\in I} Z_i$. If $\mathcal{I}$ is an arbitrary index set of sets $\{I\}_{I\in\mathcal{I}}$, then there exists a constant $C>0$ independent of $\mathcal{I}$ s.t.*

$$\mathbb{E}\left[\max_{I\in\mathcal{I}}\frac{|Z_I|}{\sqrt{|I|}}\right]\le C\sqrt{\log\big(\#(\mathcal{I})\big)}\,\mathbb{E}\Big[\max_{1\le i\le N}|Z_i|\Big].$$

THEOREM 4.3 (Coupling). *Let $Z_i, i\in I_n^d$ independent, $\mathbb{E}[Z_i]=0$, $\mathbb{V}[Z_i]=1$, such that (26) is satisfied for all i with uniform constants $k_1>1$ and $k_2>0$. Let furthermore $a_i, i\in I_n^d$ with $0<\inf a_i\le\sup a_i<\infty$ independent of i and n, and $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $i=1,\dots,n^d$, and $\mathcal{R}_n$, s.t. inequality (13) holds. Then*

$$\max_{\substack{R\in\mathcal{R}_n:\\ |R|\ge r_n}}|R|^{-1/2}\sum_{i\in R}a_iZ_i-\max_{\substack{R\in\mathcal{R}_n:\\ |R|\ge r_n}}|R|^{-1/2}\sum_{i\in R}a_iX_i=O_{\mathbb{P}}\left(\left(\frac{\log^{10}(n)}{r_n}\right)^{1/6}\right).$$

REMARK 4.4. Note that Theorem 4.3 requires only $\log^{10}(n)=o(r_n)$ for convergence in probability, whereas we require an exponent of 12 in the LSB (12). The reason is that Theorem 4.3 yields a coupling for the unpenalized partial sums, whereas Theorems 2.5 and 2.6 work with penalized partial sums. Including the penalty term requires an additional slicing argument, which results in an additional 2 in the exponent (see the proof of Theorem 5 in the the Supplementary Material).

5. Proofs. In this section, we sketch most of the proofs. Details and missing proofs are all given in Appendix B of the Supplementary Material. In the following, we will denote by p_n the cardinality of $\mathcal{R}_n$, that is, $p_n := \#(\mathcal{R}_n)$, which by (13) satisfies $\log(p_n)\sim\log n$. Recall that C denotes a generic constant which might differ from line to line.

5.1. *Proof of the auxiliary results.* We start with proving the auxiliary statements from Section 3. The proof of Lemma 4.1 is straightforward and can be obtained by using integration by parts; see Appendix B of the Supplementary Material for details.

PROOF OF LEMMA 4.2 (SKETCH). Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and r_i be i.i.d. Rademacher random variables, that is, they take the values ±1 with probability $1/2$. It follows from Lemma 4.5 of Ledoux and Talagrand (1991) that

$$\mathbb{E}_r\left[\max_I\frac{1}{\sqrt{|I|}}\left|\sum_{i\in I}r_i\right|\right]\le\sqrt{\frac{\pi}{2}}\mathbb{E}\left[\max_I\frac{1}{\sqrt{|I|}}\left|\sum_{i\in I}X_i\right|\right]. \tag{29}$$

With independent copies $(Z_i')_{1\le i\le N}$ and $\widetilde{Z}_i := Z_i - Z_i'$, $\widetilde{Z}_I := \sum_{i\in I}(Z_i - Z_i')$, we derive

$$\mathbb{E}\Big[\max_I \frac{1}{\sqrt{|I|}}|Z_I|\Big] \le 2\mathbb{E}\Big[\mathbb{E}_r\Big[\max_I \frac{1}{\sqrt{|I|}}\Big|\sum_{i\in I}|\widetilde{Z}_i|r_i\Big|\Big]\Big].$$

The contraction principle (see, e.g., Ledoux and Talagrand (1991), Theorem 4.4) implies

$$\mathbb{E}\Big[\max_I \frac{1}{\sqrt{|I|}}|Z_I|\Big] \le \sqrt{8\pi}\mathbb{E}\Big[\max_{1\le i\le N}|Z_i|\Big]\mathbb{E}\Big[\max_I \frac{|X_I|}{\sqrt{|I|}}\Big]$$

which yields the claim. □

PROOF OF THEOREM 4.3 (SKETCH). Enumerate each region in $\mathcal{R}_n$ by $j, 1 \le j \le p_n$ and define

$$X_{ij} := \frac{a_i}{\sqrt{|R_j|}} Z_i \mathbb{I}_{\{i\in R_j\}}\mathbb{I}_{\{|R_j|\ge r_n\}}, \tag{30}$$
$$X_i := (X_{ij})_{j=1,\dots,p_n}, \quad i = 1,\dots,N = n^d,$$

for some sequence r_n. Then $Z := \max_{1\le j\le p_n}\sum_{i=1}^N X_{ij}$ satisfies

$$Z \overset{\mathcal{D}}{=} \max_{\substack{R\in\mathcal{R}_n:\\ |R|\ge r_n}} \frac{1}{\sqrt{|R|}}\sum_{i\in R} a_i Z_i.$$

Recall that $\log(p_n) \lesssim \log(n)$. According to Chernozhukov, Chetverikov and Kato ((2014), Corollary 4.1), we find that for every $\delta > 0$ there exists a Gaussian version $\tilde{Z} := \max_{1\le j\le p_n}\sum_{i=1}^N a_i N_{ij}$ with independent random vectors $N_1,\dots,N_n$ in $\mathbb{R}^{p_n}$, $N_i \sim N(0, \mathbb{E}[X_i X_i^t])$, $1 \le i \le N$, such that

$$\mathbb{P}\big[|Z - \tilde{Z}| > 16\delta\big] \lesssim \delta^{-2}\big\{B_1 + \delta^{-1}(B_2 + B_4)\log(n)\big\}\log(n) + \frac{\log(n)}{n^d},$$

where

$$B_1 := \mathbb{E}\Big[\max_{1\le j,k\le p_n}\Big|\sum_{i=1}^N\big(X_{ij}X_{ik} - \mathbb{E}[X_{ij}X_{ik}]\big)\Big|\Big],$$

$$B_2 := \mathbb{E}\Big[\max_{1\le j\le p_n}\sum_{i=1}^N |X_{ij}|^3\Big],$$

$$B_4 := \sum_{i=1}^N \mathbb{E}\Big[\max_{1\le j\le p_n}|X_{ij}|^3\mathbb{I}_{\{\max_{1\le j\le p_n}|X_{ij}|>\delta/\log(p_n\vee n)\}}\Big].$$

Using Lemma 4.2 and 4.1, we derive

$$B_1 \lesssim \frac{\sqrt{\log(n)}}{\sqrt{r_n}}\big(\overline{a}^2 C\log(N)^2 + \overline{a}^2\big) \lesssim \Big(\frac{\log^5(n)}{r_n}\Big)^{1/2},$$

where $\overline{a} := \sup a_i$. For B_2 it follows again from Lemma 4.1 that

$$B_2 \lesssim \Big(\frac{\log^6(n)}{r_n}\Big)^{1/2}.$$

Finally, we can bound B_4 by

$$B_4 \le \frac{3k_1\overline{a}^3}{k_2^3}\frac{n^d}{(r_n)^{3/2}}n^{-d} = \frac{3k_1\overline{a}^3}{k_2^3}\frac{1}{(r_n)^{3/2}},$$

which yields the claim. □

5.2. *Proofs of Section 2.2.* Exploiting a Taylor expansion of T_R and Theorem 4.3 (see Appendix B of the Supplementary Material for a detailed proof), the following can be shown.

LEMMA 5.1. *Let $\mathcal{R}_n$ be a collection of sets s.t. (13) holds, $\epsilon > 0$ and $(r_n)_n \subset (0, \infty)$ be a sequence, s.t. $(\log n)^{10+\epsilon}/r_n \to 0$. Suppose $Y_i \sim F_{\theta_0} \in \mathcal{F}$, $i \in I_n^d$, are i.i.d. random variables, and recall that for $R \in \mathcal{R}_n$ we denote $\overline{Y}_R := |R|^{-1} \sum_{i \in R} Y_i$. Then it holds that*

$$\max_{\substack{R \in \mathcal{R}_n: \\ |R| \geq r_n}} \left| T_R(Y, \theta_0) - |R|^{\frac{1}{2}} \frac{|\overline{Y}_R - m(\theta_0)|}{\sqrt{v(\theta_0)}} \right| = O_{\mathbb{P}}\left(\left(\frac{\log^3(n)}{r_n} \right)^{1/4} \right)$$

as $n \to \infty$.

Now we are in position to prove Theorem 2.5. So far, we have only shown that the maximum over the local likelihood ratio statistics can be approximated by Gaussian versions, but we did not include the scale penalization $\mathrm{pen}_v(|R|)$ in (7). To include this in the approximation result, we will slice the maximum into scales, where the penalty term is almost constant. Then we show that we may bound the maximum over all scales by the sum of the maximum over these families. The price to pay is an additional $\log(n)$ factor on the smallest scale.

PROOF OF THEOREM 2.5. (a) Lemma 5.1 implies that

$$\left| \max_{\substack{R \in \mathcal{R}_n: \\ |R| \geq r_n}} \left(T_R(Y, \theta_0) - \mathrm{pen}_v(|R|) \right) - \max_{\substack{R \in \mathcal{R}_n: \\ |R| \geq r_n}} \left(|R|^{1/2} \left| \frac{\overline{Y}_R - m(\theta_0)}{\sqrt{v(\theta_0)}} \right| - \mathrm{pen}_v(|R|) \right) \right|$$
$$= O_{\mathbb{P}}\left(\left(\frac{\log^3(n)}{r_n} \right)^{1/4} \right).$$

Define

$$Y^R := |R|^{-1/2} \sum_{i \in R} \left(\frac{Y_i - m(\theta_0)}{\sqrt{v(\theta_0)}} \right),$$

$$X^R := |R|^{-1/2} \sum_{i \in R} X_i, \quad X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1).$$

With this notation and a symmetry argument, we find from the proof of Theorem 4.3 with $a_i \equiv 1$ that

$$\mathbb{P}\left[\left| \max_{\substack{R \in \mathcal{R}_n: \\ |R| \geq r_n}} |Y^R| - \max_{\substack{R \in \mathcal{R}_n: \\ |R| \geq r_n}} |X^R| \right| > \delta \right] \lesssim \delta^{-3} \left(\frac{\log^{10}(n)}{r_n} \right)^{1/2}.$$

Let $\delta_n := ((\log^{12}(n)/r_n)^{1/10} \searrow 0$. Now define $\epsilon_j := j\delta_n$, $j \in \mathbb{N}$ and

$$\mathcal{R}_{n,j} := \{ R \in \mathcal{R}_n \,|\, \exp(\epsilon_j) < |R| < \exp(\epsilon_{j+1}) \}.$$

Then the set of candidate regions $\mathcal{R}_n$ can be written as

$$\mathcal{R}_{n|r_n} = \bigsqcup_{j \in J} \mathcal{R}_{n,j}, \quad J := \left\{ \frac{1}{\delta_n} \log(\log^{12}(n)), \ldots, \frac{1}{\delta_n} \log(n^d) \right\}$$

with $|J| \leq \frac{\log(n^d)}{\delta_n}$. If we abbreviate

$$\mathrm{pen}_j := \mathrm{pen}_v(\exp(\epsilon_j)) = \sqrt{2v\left(\log\left(\frac{n^d}{\exp(\epsilon_j)} \right) + 1 \right)},$$

then the slicing above implies

$$\operatorname{pen}_{j+1} \le \operatorname{pen}_v(|R|) \le \operatorname{pen}_j, \quad \text{for all } R \in \mathcal{R}_{n,j}.$$

Using $\sqrt{a} - \sqrt{b} = (a-b)/(\sqrt{a} + \sqrt{b})$, we get

$$\begin{aligned} 0 &\le \operatorname{pen}_j - \operatorname{pen}_{j+1} \\ &= \frac{2v(\epsilon_{j+1} - \epsilon_j)}{\sqrt{2v[\log(n^d) + 1 - \epsilon_j]} + \sqrt{2v[\log(n^d) + 1 - \epsilon_{j+1}]}}. \end{aligned}$$

The largest index in J is $\frac{1}{\delta_n}\log(n^d)$ and, therefore, the maximal value of ϵ_i is given by $\bar{\epsilon} = \log(n^d)$ and $\log(n^d) + 1 - \bar{\epsilon} = 1$. Therefore,

$$0 \le \operatorname{pen}_j - \operatorname{pen}_{j+1} \le \frac{2v(\epsilon_{j+1} - \epsilon_j)}{2\sqrt{2v}} = \delta_n \sqrt{\frac{v}{2}}.$$

This means that for $n \to \infty$ the penalty terms $\operatorname{pen}_v(|R|)$, $R \in \mathcal{R}_{n,j}$ can be considered as constant. Therefore, by straightforward computations, $|J| \le \delta_n^{-1}\log(n^d)$ and choosing $\delta_n \le \frac{\epsilon}{2}$ we derive

$$\begin{aligned} &\mathbb{P}\Big[\Big|\max_{\substack{R\in\mathcal{R}_n:\\ |R|\ge r_n}} \big(|Y^R| - \operatorname{pen}_v(|R|)\big) - \max_{\substack{R\in\mathcal{R}_n:\\ |R|\ge r_n}} \big(|X^R| - \operatorname{pen}_v(|R|)\big)\Big| \ge \epsilon\Big] \\ &\le \mathbb{P}\Big[\max_{j\in J}\Big|\max_{R\in\mathcal{R}_{n,j}} |Y^R| - \max_{R\in\mathcal{R}_{n,j}} |X^R|\Big| \ge \frac{\epsilon}{2}\Big] \\ &\le \sum_{j\in J} \mathbb{P}\Big[\Big|\max_{R\in\mathcal{R}_{n,j}} |Y^R| - \max_{R\in\mathcal{R}_{n,j}} |X^R|\Big| \ge \frac{\epsilon}{2}\Big] \\ &\le |J|\frac{\delta_n^2}{\log(n^d)} = \delta_n \searrow 0, \quad n \to \infty. \end{aligned}$$

(b) This is a direct consequence of (a). □

We will now continue with the proof of Theorem 2.6. Taking into account the result of Theorem 2.5, the main statement can be proven by exploiting an invariance principle and the continuous mapping theorem. The a.s. boundedness and nondegenerateness of $M(\mathcal{R}^*, v)$ follows from Dümbgen and Spokoiny ((2001), Theorem 6.1). For details, we refer to Appendix B of the Supplementary Material.

5.3. *Proofs of Section 2.3.* Let us introduce some abbreviations to ease notation. We set $q^* := q^{\mathrm{O}}_{1-\alpha,n}$, $q := q^{\mathrm{MS}}_{1-\alpha,n}$ and denote the total signal on $Q \in \mathcal{Q}^n$ by

$$\mu^n(Q) := |Q|^{-1/2} \sum_{i\in Q} \frac{m(\theta_i^n) - m(\theta_0)}{\sqrt{v(\theta_0)}} = \frac{|Q \cap Q_n|}{\sqrt{|Q|}} \frac{m(\theta_1^n) - m(\theta_0)}{\sqrt{v(\theta_0)}}. \tag{31}$$

For brevity, introduce the Gaussian process

$$\gamma(Q) := \Big|\mu^n(Q) + |Q|^{-\frac{1}{2}} \sum_{i\in Q} v_i X_i\Big| - \operatorname{pen}_v(|Q|), \quad Q \in \mathcal{Q}^n$$

with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $v_i = \sqrt{v(\theta_i)/v(\theta_0)}$.

Let us now start with the analysis of the oracle procedure. As a preparation, we require to leave out a suitable subset of hypercubes close to the true anomaly Q_n. Therefore, choose a

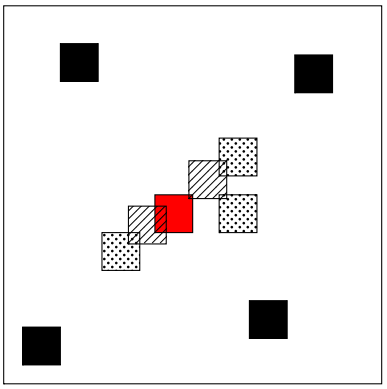

FIG. 4. *Exemplary elements of the sets $\mathcal{U}_n, \mathcal{U}$ and $\mathcal{T}$ in $d = 2$: The anomaly is shown in red, the hatched cubes belong to $\mathcal{U}_n$, the dotted cubes to $\mathcal{U}$ and all black cubes belong to $\mathcal{T}$. By definition, for all $Q \in \mathcal{U}_n$ and $Q' \in \mathcal{T}$ it holds $Q \cap Q' = \varnothing$, which implies independence of $\{\gamma(Q)\}_{Q\in\mathcal{T}}$ and $\{\gamma(Q)\}_{Q\in\mathcal{U}_n}$.*

sequence ε_n such that $\varepsilon_n \searrow 0$ but $\varepsilon_n \mu^n(Q_n) \to \infty$ and denote the set of all hypercubes which are close to the anomaly by

$$\mathcal{U}_n := \{Q \in \mathcal{Q}^n(a_n) | \mu^n(Q) \geq \mu^n(Q_n)(1 - \varepsilon_n)\}.$$

Furthermore, define the extended neighborhood of the anomaly by

$$\mathcal{U} := \{Q \in \mathcal{Q}^n(a_n) | Q \cap Q' \neq \varnothing \text{ for some } Q' \in \mathcal{U}_n\},$$

its complement by $\mathcal{T} := \mathcal{Q}^n(a_n) \setminus \mathcal{U}$. By definition, $\{\gamma(Q)\}_{Q\in\mathcal{T}}$ and $\{\gamma(Q)\}_{Q\in\mathcal{U}_n}$ are independent, which will allow us to compute the asymptotic power of the single-scale procedure. For a sketch of $\mathcal{U}_n$ and $\mathcal{U}$, see Figure 4.

The following lemma can be proven by bounding the covering number of $\mathcal{U}$ (see Appendix B of the Supplementary Material for details).

LEMMA 5.2. *Consider the setting from Section 2.3 and recall that q^* is the $(1-\alpha)$-quantile of $M_n(\mathcal{Q}^n(a_n))$ as in (9). Then*:

(a) $\max_{Q\in\mathcal{U}} |Q|^{-\frac{1}{2}} |\sum_{i\in Q} v_i X_i| = O_{\mathbb{P}}(1)$ *as* $n \to \infty$;
(b) $\lim_{n\to\infty} \mathbb{P}[\max_{Q\in\mathcal{T}} \gamma(Q) \leq q^*] = 1 - \alpha$.

With this lemma at hand, we are now in position to derive the asymptotic power of the oracle procedure.

PROOF OF THEOREM 2.9(A) (SKETCH). To analyze $\mathbb{P}_{\theta^n}[T_n(Y, \theta_0, \mathcal{Q}^n(a_n)) > q^*]$, we start with showing a $\geq$ in the statement of Theorem 2.9(a). By Lemma 5.1 and Theorem 4.3, we find

$$\mathbb{P}_{\theta^n}\big[T_n\big(Y, \theta_0, \mathcal{Q}^n(a_n)\big) > q^*\big] = \mathbb{P}\Big[\max_{Q\in\mathcal{Q}^n(a_n)} \gamma(Q) > q^*\Big] + o(1).$$

Now we derive

$$\mathbb{P}\Big[\max_{Q\in\mathcal{Q}^n(a_n)} \gamma(Q) > q^*\Big] = \mathbb{P}\Big[\max_{Q\in\mathcal{T}} \gamma(Q) \leq q^*\Big]\mathbb{P}[\gamma(Q_n) > q^*] + \mathbb{P}\Big[\max_{Q\in\mathcal{T}} \gamma(Q) > q^*\Big],$$

where we exploited $Q_n \in \mathcal{U}$ and independence of $\{\gamma(Q)\}_{Q\in\mathcal{T}}$ and $\gamma(Q_n)$. Lemma 5.2(b) states that $\mathbb{P}[\max_{Q\in\mathcal{T}} \gamma(Q) \le q^*] = 1-\alpha+o(1)$, and hence

$$\mathbb{P}_{\theta^n}\big[T_n\big(Y,\theta_0,\mathcal{Q}^n(a_n)\big) > q^*\big] \ge \alpha + (1-\alpha)\mathbb{P}\big[\gamma(Q_n) > q^*\big] + o(1).$$

As $\gamma(Q_n) + \mathrm{pen}_v(|Q_n|)$ follows a folded normal distribution with parameters $\mu = \mu^n(Q_n)$ and $\sigma^2 = |Q_n|^{-1}\sum_{i\in Q_n} v_i^2$, this yields the proposed lower bound. For the upper bound (i.e., $\le$ in the statement of Theorem 2.9(a)), we proceed as before and obtain

$$\begin{aligned}&\mathbb{P}_{\theta^n}\big[T_n\big(Y,\theta_0,\mathcal{Q}^n(a_n)\big) > q^*\big]\\ &\quad= \alpha + (1-\alpha)\mathbb{P}\Big[\max_{Q\in\mathcal{U}_n}\gamma(Q) > q^*\Big] + \mathbb{P}\Big[\max_{Q\in\mathcal{U}\setminus\mathcal{U}_n}\gamma(Q) > q^*\Big] + o(1).\end{aligned}$$

If $\mu^n(Q_n) - \sqrt{2\log(a_n^{-1})} \to C \in [-\infty,\infty)$, then $\mathbb{P}[\max_{Q\in\mathcal{U}\setminus\mathcal{U}_n}\gamma(Q) > q^*] = o(1)$ by Lemma 5.2(a), and otherwise nothing has to be shown. Hence

$$\begin{aligned}\mathbb{P}_{\theta^n}\big[T_n\big(Y,\theta_0,\mathcal{Q}^n(a_n)\big) > q^*\big] &\le \alpha + (1-\alpha)\mathbb{P}\Big[\max_{Q\in\mathcal{U}_n}\gamma(Q) > q^*\Big] + o(1)\\ &= \alpha + (1-\alpha)\mathbb{P}\big[\gamma(Q_n) + o_{\mathbb{P}}(1) > q^*\big] + o(1),\end{aligned}$$

and hence the claim is proven. □

Now we turn to the multiscale procedure. As here different scales are considered, the set $\mathcal{U}$ is not large enough any more. Especially, we cannot construct a subset $\mathcal{V}$ such that $\{\gamma(Q)\}_{\mathcal{V}^c}$ and $\gamma(Q_n)$ are independent and $\max_{Q\in\mathcal{V}}\gamma(Q)$ is still negligible. Due to this, the corresponding proof in Sharpnack and Arias-Castro (2016) is incomplete. To overcome this difficulty, we follow the idea to distinguish if the anomaly Q_n has asymptotically an effect on $\gamma(Q)$ or not. Whenever Q is sufficiently large compared to Q_n, the impact will asymptotically be negligible.

For some sequence $\epsilon_n \searrow 0$ with $\epsilon_n = O(|Q_n|^{-\gamma})$ with some $\gamma > 0$, we introduce

$$\begin{aligned}\delta_n &:= \epsilon_n \max\left\{\mu^n(Q_n), \log(n)\sqrt{\frac{|Q_n|}{r_n}}\right\}^{-1},\\ \mathcal{V} &:= \big\{Q \in \mathcal{Q}^{\mathrm{MS}}_{n|r_n} \,|\, \mu^n(Q) \ge \delta_n \mu^n(Q_n)\big\}\end{aligned} \tag{32}$$

and its complement $\mathcal{T}' := \mathcal{Q}^{\mathrm{MS}}_{n|r_n} \setminus \mathcal{V}$. For a sketch, see Figure 5.

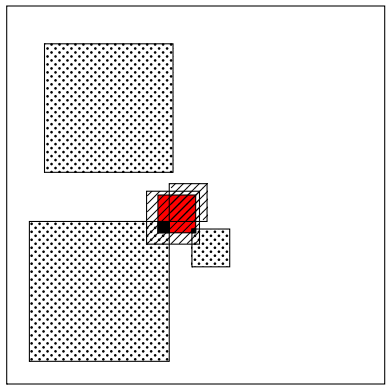

FIG. 5. *Exemplary elements of the sets $\mathcal{V}$ and $\mathcal{T}'$ in $d = 2$: The anomaly is shown in red, the hatched cubes belong to $\mathcal{V}$ and the dotted cubes to $\mathcal{T}'$. However, the intersections marked in black are small enough such that they have asymptotically no influence on $\gamma(Q)$.*

Opposed to the oracle procedure, we do not have independence of $\{\gamma(Q)\}_{Q\in\mathcal{T}'}$ and $\gamma(Q_n)$. However, asymptotically a similar property holds true as shown in the following lemma, which can again be proven by estimating the covering number (see Appendix B of the Supplementary Material for details).

LEMMA 5.3. *Consider the setting from Section 2.3 and recall that q is the $(1-\alpha)$-quantile of $M_n(\mathcal{Q}^{\mathrm{MS}}_{n|r_n})$ as in (9). Then the following statements hold true as $n\to\infty$:*

(a) $\max_{Q\in\mathcal{V}} ||Q|^{-\frac{1}{2}}\sum_{i\in Q} v_i X_i| = O_{\mathbb{P}}(\sqrt{\ln(|Q_n|)} + \sqrt{\ln(-\ln(m(\theta_1^n)-m(\theta_0)))})$;
(b) $\max_{Q\in\mathcal{T}'} ||Q|^{-\frac{1}{2}}\sum_{i\in Q\cap Q_n} v_i X_i| = o_{\mathbb{P}}(1)$;
(c) $\mathbb{P}[\max_{Q\in\mathcal{T}'}\gamma(Q)\le q] = 1-\alpha+o(1)$.

PROOF OF THEOREM 2.9(B). For the multiscale procedure, we have to compute a lower bound for $\mathbb{P}_{\theta^n}[T_n(Y,\theta_0,\mathcal{Q}^{\mathrm{MS}}_{n|r_n})>q]$. Similar to the proof of Theorem 2.9(a), we obtain

$$\begin{aligned}&\mathbb{P}_{\theta^n}\big[T_n\big(Y,\theta_0,\mathcal{Q}^{\mathrm{MS}}_{n|r_n}\big)>q\big]\\&\quad\ge \mathbb{P}\Big[\Big\{\max_{Q\in\mathcal{T}'}\gamma(Q)\le q\Big\}\cap\{\gamma(Q_n)>q\}\Big] + \mathbb{P}\Big[\max_{Q\in\mathcal{T}'}\gamma(Q)>q\Big]+o(1).\end{aligned}$$

By Lemma 5.3(b), we furthermore get

$$\begin{aligned}&\mathbb{P}\Big[\max_{Q\in\mathcal{T}'}\gamma(Q)\le q\Big]\\&\quad=\mathbb{P}\Big[\max_{Q\in\mathcal{T}'}\Big[\Big|\mu^n(Q)+\frac{1}{\sqrt{|Q|}}\sum_{i\in Q}v_iX_i\Big|-\mathrm{pen}_v(|Q|)\Big]\le q\Big]\\&\quad=\mathbb{P}\Big[\max_{Q\in\mathcal{T}'}\Big[\Big|\mu^n(Q)+\frac{1}{\sqrt{|Q|}}\sum_{i\in Q\setminus Q_n}v_iX_i\Big|-\mathrm{pen}_v(|Q|)\Big]\le q\Big]+o(1),\end{aligned}$$

which shows by independence that

$$\begin{aligned}&\mathbb{P}\Big[\Big\{\max_{Q\in\mathcal{T}'}\gamma(Q)\le q\Big\}\cap\{\gamma(Q_n)>q\}\Big]\\&\quad=\mathbb{P}\Big[\max_{Q\in\mathcal{T}'}\gamma(Q)\le q\Big]\mathbb{P}[\gamma(Q_n)>q]+o(1).\end{aligned}$$

Now the proof can be concluded as the one of Theorem 2.9(a). □

Acknowledgments. We thank Katharina Proksch for helpful comments on the proof of Theorem 2.5 as well as Guenther Walther and three anonymous referees for several constructive comments which helped us to substantially improve the presentation of the paper.

Financial support by the German Research Foundation DFG of CRC 755 A04 is acknowledged.

SUPPLEMENTARY MATERIAL

Supplement to "Multidimensional multiscale scanning" (DOI: 10.1214/18-AOS1806SUPP; .pdf). This supplementary material contains the explicit computation of the packing numbers for three important examples of $\mathcal{R}^*$, namely hyperrectangles, hypercubes and halfspaces, and detailed proofs of all statements.

REFERENCES

ADLER, R. J. (2000). On excursion sets, tube formulas and maxima of random fields. *Ann. Appl. Probab.* **10** 1–74. MR1765203 https://doi.org/10.1214/aoap/1019737664

ALM, S. E. (1998). Approximation and simulation of the distributions of scan statistics for Poisson processes in higher dimensions. *Extremes* **1** 111–126. MR1652932 https://doi.org/10.1023/A:1009965918058

ARIAS-CASTRO, E., CANDÈS, E. J. and DURAND, A. (2011). Detection of an anomalous cluster in a network. *Ann. Statist.* **39** 278–304. MR2797847 https://doi.org/10.1214/10-AOS839

ARIAS-CASTRO, E., DONOHO, D. L. and HUO, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* **51** 2402–2425. MR2246369 https://doi.org/10.1109/TIT.2005.850056

ARIAS-CASTRO, E., CASTRO, R. M., TÁNCZOS, E. and WANG, M. (2018). Distribution-free detection of structured anomalies: Permutation and rank-based scans. *J. Amer. Statist. Assoc.* **113** 789–801. MR3832227 https://doi.org/10.1080/01621459.2017.1286240

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. MR0882001

BUTUCEA, C. and INGSTER, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli* **19** 2652–2688. MR3160567 https://doi.org/10.3150/12-BEJ470

CASELLA, G. and BERGER, R. L. (1990). *Statistical Inference. The Wadsworth & Brooks/Cole Statistics/Probability Series*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA. MR1051420

CHENG, D. and SCHWARTZMAN, A. (2017). Multiple testing of local maxima for detection of peaks in random fields. *Ann. Statist.* **45** 529–556. MR3650392 https://doi.org/10.1214/16-AOS1458

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461 https://doi.org/10.1214/14-AOS1230

DATTA, P. and SEN, B. (2018). Optimal inference with a multidimensional multiscale statistic. Preprint. Available at arXiv:1806.02194.

DESPRES, C. J. (2014). The Vapnik–Chervonenkis dimension of norms on $\mathbb{R}^d$. Preprint. Available at arXiv:1412.6612.

DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation. Springer Series in Statistics*. Springer, New York. MR1843146 https://doi.org/10.1007/978-1-4613-0125-7

DICKHAUS, T. (2014). *Simultaneous Statistical Inference*. Springer, Heidelberg. MR3184277 https://doi.org/10.1007/978-3-642-45182-9

DÜMBGEN, L. and SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124–152. MR1833961 https://doi.org/10.1214/aos/996986504

DÜMBGEN, L. and WALTHER, G. (2008). Multiscale inference about a density. *Ann. Statist.* **36** 1758–1785. MR2435455 https://doi.org/10.1214/07-AOS521

FANG, X. and SIEGMUND, D. (2016). Poisson approximation for two scan statistics with rates of convergence. *Ann. Appl. Probab.* **26** 2384–2418. MR3543900 https://doi.org/10.1214/15-AAP1150

FARNUM, N. R. and BOOTH, P. (1997). Uniqueness of maximum likelihood estimators of the 2-parameter Weibull distribution. *IEEE Trans. Reliab.* **46** 523–525. https://doi.org/10.1109/24.693786.

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728 https://doi.org/10.1111/rssb.12047

FRIEDENBERG, D. A. and GENOVESE, C. R. (2013). Straight to the source: Detecting aggregate objects in astronomical images with proper error control. *J. Amer. Statist. Assoc.* **108** 456–468. MR3174633 https://doi.org/10.1080/01621459.2013.779829

HAIMAN, G. and PREDA, C. (2006). Estimation for the distribution of two-dimensional discrete scan statistics. *Methodol. Comput. Appl. Probab.* **8** 373–381. MR2329304 https://doi.org/10.1007/s11009-006-9752-1

JIANG, T. (2002). Maxima of partial sums indexed by geometrical structures. *Ann. Probab.* **30** 1854–1892. MR1944008 https://doi.org/10.1214/aop/1039548374

JIANG, Y., QIU, Y., MINN, A. J. and ZHANG, N. R. (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. USA* **113** E5528–E5537.

KABLUCHKO, Z. (2011). Extremes of the standardized Gaussian noise. *Stochastic Process. Appl.* **121** 515–533. MR2763094 https://doi.org/10.1016/j.spa.2010.11.007

KABLUCHKO, Z. and MUNK, A. (2009). Shao's theorem on the maximum of standardized random walk increments for multidimensional arrays. *ESAIM Probab. Stat.* **13** 409–416. MR2554963 https://doi.org/10.1051/ps:2008020

KAZANTSEV, I. G., LEMAHIEU, I., SALOV, G. I. and DENYS, R. (2002). Statistical detection of defects in radiographic images in nondestructive testing. *Signal Process.* **82** 791–801. https://doi.org/10.1016/S0165-1684(02)00158-5.

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrsch. Verw. Gebiete* **34** 33–58. MR0402883 https://doi.org/10.1007/BF00532688

KÖNIG, C., MUNK, A. and WERNER, F. (2020). Supplement to "Multidimensional multiscale scanning in exponential families: Limit theory and statistical consequences." https://doi.org/10.1214/18-AOS1806SUPP.

KOU, J. (2017). Identifying the support of rectangular signals in Gaussian noise. Preprint. Available at arXiv:1703.06226.

KULLDORFF, M., HEFFERNAN, R., HARTMAN, J., ASSUNÇÃO, R. and MOSTASHARI, F. (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2**. https://doi.org/10.1371/journal.pmed.0020059.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015 https://doi.org/10.1007/978-3-642-20212-4

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

LI, H., MUNK, A. and SIELING, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Stat.* **10** 918–959. MR3486421 https://doi.org/10.1214/16-EJS1131

MASSART, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT constructions. *Ann. Probab.* **17** 266–291. MR0972785

NAUS, J. I. and WALLENSTEIN, S. (2004). Multiple window and cluster size scan procedures. *Methodol. Comput. Appl. Probab.* **6** 389–400. MR2108559 https://doi.org/10.1023/B:MCAP.0000045087.33227.8c

POZDNYAKOV, V., GLAZ, J., KULLDORFF, M. and STEELE, J. M. (2005). A martingale approach to scan statistics. *Ann. Inst. Statist. Math.* **57** 21–37. MR2165605 https://doi.org/10.1007/BF02506876

PROKSCH, K., WERNER, F. and MUNK, A. (2018). Multiscale scanning in inverse problems. *Ann. Statist.* **46** 3569–3602. MR3852662 https://doi.org/10.1214/17-AOS1669

RIO, E. (1993). Strong approximation for set-indexed partial-sum processes, via KMT constructions. II. *Ann. Probab.* **21** 1706–1727. MR1235436

RIVERA, C. and WALTHER, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scand. J. Stat.* **40** 752–769. MR3145116 https://doi.org/10.1111/sjos.12027

SCHMIDT-HIEBER, J., MUNK, A. and DÜMBGEN, L. (2013). Multiscale methods for shape constraints in deconvolution: Confidence statements for qualitative features. *Ann. Statist.* **41** 1299–1328. MR3113812 https://doi.org/10.1214/13-AOS1089

SCHWARTZMAN, A., GAVRILOV, Y. and ADLER, R. J. (2011). Multiple testing of local maxima for detection of peaks in 1D. *Ann. Statist.* **39** 3290–3319. MR3012409 https://doi.org/10.1214/11-AOS943

SHARPNACK, J. and ARIAS-CASTRO, E. (2016). Exact asymptotics for the scan statistic and fast alternatives. *Electron. J. Stat.* **10** 2641–2684. MR3546971 https://doi.org/10.1214/16-EJS1188

ŠIDÁK, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62** 626–633. MR0216666

SIEGMUND, D. and VENKATRAMAN, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.* **23** 255–271. MR1331667 https://doi.org/10.1214/aos/1176324466

SIEGMUND, D. and YAKIR, B. (2000). Tail probabilities for the null distribution of scanning statistics. *Bernoulli* **6** 191–213. MR1748719 https://doi.org/10.2307/3318574

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90. MR0790201 https://doi.org/10.1093/biomet/72.1.67

TAYLOR, J. E. and WORSLEY, K. J. (2007). Detecting sparse signals in random fields, with an application to brain mapping. *J. Amer. Statist. Assoc.* **102** 913–928. MR2354405 https://doi.org/10.1198/016214507000000815

TU, I. (2013). The maximum of a ratchet scanning process over a Poisson random field. *Statist. Sinica* **23** 1541–1551. MR3222809

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

WALTHER, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.* **38** 1010–1033. MR2604703 https://doi.org/10.1214/09-AOS732

ZHANG, N. R., YAKIR, B., XIA, L. C. and SIEGMUND, D. (2016). Scan statistics on Poisson random fields with applications in genomics. *Ann. Appl. Stat.* **10** 726–755. MR3528358 https://doi.org/10.1214/15-AOAS892