Submitted to the Annals of Statistics
arXiv: 1611.04537

MULTISCALE SCANNING IN INVERSE PROBLEMS

By Katharina Proksch[*,‡], Frank Werner[*,‡] and Axel Munk[*,†,‡]

Institute for Mathematical Stochastics, University of Göttingen and Max Planck Institute for Biophysical Chemistry

In this paper we propose a multiscale scanning method to determine active components of a quantity f w.r.t. a dictionary $\mathcal{U}$ from observations Y in an inverse regression model $Y = Tf + \xi$ with linear operator T and general random error ξ. To this end, we provide uniform confidence statements for the coefficients $\langle \varphi, f \rangle$, $\varphi \in \mathcal{U}$, under the assumption that $(T^*)^{-1}(\mathcal{U})$ is of wavelet-type. Based on this we obtain a multiple test that allows to identify the active components of $\mathcal{U}$, i.e. $\langle f, \varphi \rangle \neq 0$, $\varphi \in \mathcal{U}$, at controlled, family-wise error rate. Our results rely on a Gaussian approximation of the underlying multiscale statistic with a novel scale penalty adapted to the ill-posedness of the problem. The scale penalty furthermore ensures convergence of the statistic's distribution towards a Gumbel limit under reasonable assumptions. The important special cases of tomography and deconvolution are discussed in detail. Further, the regression case, when $T = \mathrm{id}$ and the dictionary consists of moving windows of various sizes (scales), is included, generalizing previous results for this setting. We show that our method obeys an oracle optimality, i.e. it attains the same asymptotic power as a single-scale testing procedure at the correct scale. Simulations support our theory and we illustrate the potential of the method as an inferential tool for imaging. As a particular application we discuss super-resolution microscopy and analyze experimental STED data to locate single DNA origami.

1. Introduction. Suppose we have access to observations $Y_{\mathbf{j}}$ which are linked to an unknown quantity $f \in \mathbb{H}_1$ via the inverse regression model

$$Y_{\mathbf{j}} = Tf(\mathbf{x}_{\mathbf{j}}) + \xi_{\mathbf{j}}, \quad \mathbf{j} \in I_n^d := \{1, \dots, n\}^d,\ d \in \mathbb{N}. \tag{1}$$

[*]The authors gratefully acknowledge financial support by the German Research Foundation DFG through subproject A07 of CRC 755.

[†]Funding through the VW foundation is also gratefully acknowledged.

[‡]We thank Haisen Ta and Jan Keller from the Department of NanoBiophotonics, Max Planck Institute for Biophysical Chemistry, for providing the experimental data and expertise. We also thank Cristina Butucea for helpful comments and several anonymous referees and the editors for a number of insightful questions and constructive suggestions which helped us to improve the quality of the paper substantially.

MSC 2010 subject classifications: Primary 62G10; secondary 62G15, 62G20, 62G32

Keywords and phrases: multiscale analysis, scan statistic, ill-posed problem, deconvolution, super-resolution, Gumbel extreme value limit

Here, $T : \mathbb{H}_1 \to \mathbb{H}_2 \subset C[0,1]^d$ is a bounded linear operator acting between proper Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$. In model (1), n stands for the level of discretization such that, more rigorously, the model reads $Y_{\mathbf{j},n} = Tf(\mathbf{x}_{\mathbf{j},n}) + \xi_{\mathbf{j},n}$ with triangular schemes of sampling points $\mathbf{x}_{\mathbf{j}} = \mathbf{x}_{\mathbf{j},n}$ in the d-cube $[0,1]^d$ and independent, centered but not necessarily identically distributed random variables $\xi_{\mathbf{j}} = \xi_{\mathbf{j},n}$, $\mathbf{j} \in I_n^d$. For ease of notation, this dependence on n is suppressed whenever it is not relevant. Here and throughout the paper, bold print letters and numbers denote vectors and multi-indices, whereas scalars are printed in regular type face.

Models of the kind (1) underly a plenitude of applied problems varying from astrophysics and tomography to cell biology (see e. g. O'Sullivan, 1986; Bertero et al., 2009) and have received considerable interest in the statistical literature. Most of research targets (regularized) estimation of f and associated theory. An early approach for estimation is based on a singular value decomposition (SVD) of the operator, where f is expanded in a series of eigenfunctions of T^*T (see e. g. Mair and Ruymgaart, 1996; Johnstone et al., 2004; Cavalier and Golubev, 2006; Bissantz et al., 2007; Kerkyacharian et al., 2010; Johnstone and Paul, 2014; Albani et al., 2016). Given a proper choice of the regularization parameter, SVD-based estimators are well-known to be minimax optimal (Johnstone and Silverman, 1991). Adaptive estimation in this context was studied, e. g. by Goldenshluger (1999); Tsybakov (2000); Cavalier et al. (2003); Chernousova and Golubev (2014). Since in SVD-based estimation the basis for the expansion is entirely defined by the operator, as an alternative, wavelet-based methods which incorporate the properties of the function of interest have also been frequently employed. Examples are wavelet-vaguelette (Donoho, 1995) and vaguelette-wavelet methods (Abramovich and Silverman, 1998), where f and Kf are expanded in a wavelet and vaguelette basis or vice versa, and the coefficients are estimated by proper thresholding. This allows for a natural adaptation to the local smoothness of the unknown function (see e. g. Cavalier and Tsybakov, 2002). Related to this, Cohen et al. (2004) proposed an adaptive estimator based on a combination of linear Galerkin projection methods and adaptive wavelet thresholding. Besides of these selective references a vast amount of work has been devoted to recovery of f during the last decades and the common ground of all these works is that the ill-posedness of an inverse problem usually only gives poor (minimax) rates for estimation and makes full recovery of f a very difficult problem in general (in the setup of (1) see, e. g. Willer (2009) or for deconvolution, see, e. g., the monograph by Meister (2009) and the references given there).

A possibility to deal with this intrinsic difficulty is to relax the ambitious

goal of recovering the entire function f. Indeed, in many applications, only certain properties or aspects of f are of primary interest and a full, precise reconstruction is not necessary any more. Examples of practical relevance are the detection and localization of "hot spots" in astrophysical image analysis (Friedenberg and Genovese, 2013), functional magnetic resonance imaging (Schwartzman et al., 2008), non-destructive testing (Kazantsev et al., 2002), and image deformation in microscopy (Bissantz et al., 2009), to mention a few. For a theoretical account in deconvolution see (Butucea and Comte, 2009). In a similar spirit, the detection of certain geometric shapes in image analysis has been studied by Genovese et al. (2012), but the authors do not take into account the underlying inverse problem. All these issues can be treated by means of statistical testing, presumably a simpler task than estimation.

In contrast to estimation, hypothesis testing in inverse problems has been investigated much less, early references are Butucea (2007); Holzmann et al. (2007). Ingster et al. (2012) treat the problem of testing $f = 0$ against $f \in \Theta_q(r)$ where $\Theta_q(r)$ is a suitable smoothness class restricted to $\|f\| \geqslant r$ by means of the classical minimax testing approach (see e.g. the series of papers by Ingster (1993)). Also Laurent et al. (2011, 2012) follow this path and investigate the differences and commonalities of testing in the image space ($Tf = 0$) and the preimage space ($f = 0$). The authors prove that in several situations it does not matter if first f is approximately reconstructed using an SVD-based regularization method and then tested to be 0, or if Tf is directly tested to be 0, see also Holzmann et al. (2007) for a similar observation. More precisely, minimax testing procedures for one of these problems are also minimax for the rephrased problem and the asymptotic detection boundary for both testing problems coincides. For related results in the multivariate setting or for more general regularization schemes see Ingster et al. (2014); Marteau and Mathé (2014). In contrast to the problem treated here, in all these studies only "global" features of the full signal are investigated, such as testing that the full signal is zero, and no simultaneous inference on sub-structures of the signal is targeted. In fact, this is a much more challenging task in an inverse problems setup and it turns out also to be substantially different to the corresponding direct testing problem of "hot spot" detection. This will be the topic of this paper.

In direct problems ($T = \mathrm{id}$ in (1)), finding relevant sub-structures, such as the detection of regions of activity, is of "scanning-type", which means that it can be reformulated as a (multiple) testing problem for structures on the grid I_n^d in (1) and scanning-type procedures can be employed. These have received much attention in the literature over the past decades. Walther

(2010) considers the two dimensional problem of detecting spatial clusters in the Bernoulli model by scanning with rectangular windows of varying sizes, see also Kabluchko (2011), Butucea and Ingster (2013) and Sharpnack and Arias-Castro (2016) for results in a Gaussian setting. In a similar spirit, scan statistics have been employed in the context of multiscale inference about higher order qualitative characteristics such as modes of a density (see Dümbgen and Walther, 2008; Rufibach and Walther, 2010; Li et al., 2016; Eckle et al., 2017).

However, in an inverse problem as in (1), it is not obvious how to perform statistically efficient "scanning" because local properties of f may propagate in a non-local manner into Tf. If, e. g., f is a function on $[0,1]^d$ and we want to infer on the support of f, we find that despite the fact that *globally* testing $f \equiv 0$ is equivalent to testing $Tf \equiv 0$, this is not true for localized tests on regions $B \subset [0,1]^d$ we are interested in here. This is due to the fact that $(Tf)_{|_B}$ is not necessarily related to $f_{|_B}$ only. Indeed, we will see that reducing this problem to the image domain $\mathbb{H}_2$, i. e., simultaneously testing $H_B : (Tf)_{|_B} \equiv 0$ against $K_B : (Tf)_{|_B} > 0$ cannot lead to a competitive procedure as it does not take into account the propagation of (multiscale) features of f by T (cf. Figure 2(f)). Instead, it becomes necessary to employ probe functionals $\varphi_i = \varphi_{i,n}$ (again dependent on the discretization level n, but this dependence will be suppressed whenever not relevant below), which are compatible with the operator T and hence allow for transportation of "local" information from Tf back to $\langle f, \varphi_i \rangle$. If the probe functionals φ_i are chosen properly, the values $\langle f, \varphi_i \rangle$ hold information about "local" features of f, e. g. in form of a wavelet-type analysis, see also Schmidt-Hieber et al. (2013); Eckle et al. (2016), who infer on shape characteristics in i.i.d. density deconvolution. Arias-Castro et al. (2005) propose a scanning procedure based on a multiscale dictionary of beamlets that allows to detect line segments hidden in a noisy image, however, not in an inverse problems context.

The problem we consider in our paper is as follows: Given model (1) and an associated sequence of dictionaries

$$\mathcal{U} = \mathcal{U}_n = \left\{\varphi_{1,n}, ..., \varphi_{N(n),n}\right\} \subset R\left(T^*\right), \tag{2}$$

of cardinality $N = N(n) \to \infty$ as $n \to \infty$, we provide a sequence of multiple tests ("scanning") for the associated sequence of multiple testing problems

$$(H_{J,n}) \qquad \langle f, \varphi_{i,n} \rangle = 0 \qquad \text{for all} \qquad i \in J$$

vs.

$$(K_{J,n}) \qquad \exists\, i \in J \qquad \text{such that} \qquad \langle \varphi_{i,n}, f \rangle > 0,$$

simultaneously over all subsets $J \subset I_{N(n)} =: \{1, \dots, N(n)\}$. It is clear that the structure of the testing problem stays the same if $\cdot > 0$ in $(K_{J,n})$ is replaced by $\cdot < 0$ or $|\cdot| \neq 0$, hence we restrict ourselves to $(K_{J,n})$ in the following. Moreover, it is also clear that as $n \to \infty$, there is a detection boundary, given by a sequence $(\mu_{i,n})_{i\in\mathbb{N}}$, dividing the space of all signals into the asymptotically detectable region and the non-detectable region such that $\cdot > 0$ will be replaced by $\cdot > \mu_{i,n}$ later on.

The underlying idea of the present paper is to provide for each local testing problem a local test which detects those coefficients $\langle f, \varphi_{i,n}\rangle$, $i \in J$, which are strong enough, and hence by performing all these tests simultaneously, we expect to (asymptotically) detect all positive coefficients above the detection boundary. If f admits a sparse representation w.r.t. $\mathcal{U}$, this is $f \approx \sum c_{i,n}\varphi_{i,n}$ with $\|c\|_0$ small, then the simultaneous testing problem $H_{J,n}$ against $K_{J,n}$, $J \subset I_N(n)$ allows to detect exactly those i with $c_{i,n} > 0$. However, we emphasize that sparsity of f w.r.t. $\mathcal{U}$ is not required or assumed here.

With this choice of a sequence of multiple tests we will not simply control the error of a wrong rejection of $f \equiv 0$, rather we control the *family-wise error rate (FWER)* of making any wrong decision, cf. Dickhaus (2014, Def. 1.2). Mathematically, our test is a level-α-test for the simultaneous testing problem $H_{J,n}$ against $K_{J,n}$, $J \subset I_N(n)$, i. e., it guarantees that

$$\sup_{J:J\subset I_{N(n)}} \mathbb{P}_{H_{J,n}}[\text{"at least one (wrong) rejection in } J\text{"}] \leqslant \alpha + o(1), \tag{3}$$

as n and hence $N(n) \to \infty$. Consequently, all rejections (i. e. decisions for signal strength > 0) will be made at a *uniform* error control, no matter what the underlying configuration of $\langle f, \varphi_{i,n}\rangle$'s is.

Fundamental to our simultaneous scanning procedure are uniform confidence statements for the coefficients $\langle f, \varphi_{i,n}\rangle$, $i \in I_{N(n)}$ in the inverse regression model (1). Conceptually related, Nickl and Reiß (2012) and Söhl and Trabs (2012) provide uniform Donsker-type results in the context of i.i.d. deconvolution for single-scale contrasts $\langle f, \varphi\rangle$. As one particular example the results of the latter authors can be used to derive uniform statements with respect to both the regularization parameter h (which plays the role of a scale parameter) and variable location t via the functionals $\langle I_{(-\infty,0]}(\cdot - t), \widehat{f}_h\rangle =: \widehat{F}_h(t)$ as estimators of the distribution function F, where $\widehat{f}_h$ is a deconvolution estimator of the density f. We consider dictionaries which are different in that they are closely related to estimating the regression function f in (1) (which would correspond to estimating f, not F, in their model), where uniform control with respect to the scale parameter requires the use of very different techniques.

Multiscale approaches have also been discussed in the Bayesian literature, see e. g. Castillo and Nickl (2014); Ray (2017), but not in the inverse problems setup. Even though it seems promising to exploit Gaussian approximations based on posterior distributions as in Castillo and Nickl (2014), this leads to additional difficulties in our general setup as typically conjugacy is lost in inverse regression problems if the likelihood and/or the prior are non-Gaussian. Consequently, sampling from the posterior becomes a computationally involved large-scale problem. Nevertheless, we stress that in principle recent developments for nonparametric Bayesian credible sets (see e. g. Knapik et al., 2011; Ray, 2013) can offer an alternative route to the present methodology.

1.1. ***M**ultiscale **I**nverse **SCA**nning **T**est: MISCAT.* As we have assumed that $\varphi_{i,n} \in \mathcal{R}(T^*)$ for all $i \in I_{N(n)}$, there exists a sequence of dictionaries $\mathcal{W} = \mathcal{W}_n = \{\Phi_{i,n} \,|\, i \in I_{N(n)}\} \subset \mathbb{H}_2$ such that $\varphi_{i,n} = T^*\Phi_{i,n}$. In the following we will assume that $\mathcal{W}$ obeys a certain wavelet-type structure, i.e. for each $i \in I_{N(n)}$ there is an associated *scale* $\mathbf{h}_{i,n} = (h_{i,n,1}, \dots, h_{i,n,d})^T \in (0,1]^d$ and an associated *translation* $\mathbf{t}_{i,n} \in [\mathbf{h}_{i,n}, \mathbf{1}]$. The products $\mathbf{h}_{i,n}^{\mathbf{1}} := h_{i,n,1} \cdot \ldots \cdot h_{i,n,d}$ will be referred to as *sizes of scales*. In contrast to the direct problem ($T = \mathrm{id}$), in an inverse problem the condition $\varphi_i = T^*\Phi_i$ implies a non-standard scaling of the Φ_i's which can be chosen to depend only on $\mathbf{h}_i$ and not on $\mathbf{t}_i$ in many cases. To highlight this scaling property, with a slight abuse of notation, we will also introduce a sequence of dictionary functions $\Phi_{\mathbf{h}_{i,n}}$ and assume that $\mathcal{W}_n$ is as follows:

$$\mathcal{W}_n = \left\{ \Phi_{i,n}(\mathbf{z}) := \Phi_{\mathbf{h}_{i,n}}\left(\tfrac{\mathbf{t}_{i,n}-\mathbf{z}}{\mathbf{h}_{i,n}}\right) \,\middle|\, \mathrm{supp}(\Phi_{\mathbf{h}_{i,n}}) \subset [0,1]^d,\ i \in I_{N(n)} \right\}. \tag{4}$$

Here and in the following, division of a vector by a vector is meant componentwise. All quantities depend on n, and this dependence is suppressed in the following, e.g. we write Φ_i instead of $\Phi_{i,n}$. Note that if $\Phi_{\mathbf{h}_i} \equiv \Phi$ for all $i \in I_N$, then the dictionary (4) is a wavelet dictionary in the classical sense, which is appropriate for direct regression problems, i.e. $T = \mathrm{id}$ in (1) (see e. g. Arias-Castro et al., 2005). For our asymptotic results we will further assume that the normed functions $\Phi_{\mathbf{h}_i}/\|\Phi_{\mathbf{h}_i}\|$ satisfy an average Hölder condition, see (AHC) or (15) below. Such conditions are satisfied for many important operators T such as the Radon transform (see Section 3.1) and convolution operators (see Section 3.2).

To construct a level-α-test for simultaneously testing $H_{J,n}$ against $K_{J,n}$, $J \subset I_N$ we can now employ

$$\langle f, \varphi_i \rangle_{\mathbb{H}_1} = \langle Tf, \Phi_i \rangle_{\mathbb{H}_2} \tag{5}$$

to estimate the local coefficients $\langle f, \varphi_i \rangle$ by their empirical counterparts

$$\langle Y, \Phi_i \rangle_n := \frac{1}{n^d} \sum_{\mathbf{j} \in I_n^d} Y_{\mathbf{j}} \Phi_i(\mathbf{x}_{\mathbf{j}}), \tag{6}$$

cf. (4) for the definition of Φ_i and see Section 3 for details. MISCAT combines these local statistics to a global level-α-test in the sense of (3) no matter what the (local) dependency structure is. To this end, we take the maximum of the local test statistics, yielding a multiple "dictionary scanning" test statistic of the form

$$\mathcal{S}(Y) := \max_{i \in I_N} S(Y, i), \quad \text{with} \quad S(Y, i) := \omega_i \left(\frac{\langle Y, \Phi_i \rangle_n}{\sigma_i} - \omega_i \right), \tag{7}$$

where $\sigma_i^2 := \mathrm{Var}[\langle Y, \Phi_i \rangle_n]$ depend on the variances $\sigma^2(\mathbf{j})$ of the errors $\xi_{\mathbf{j}}$, which are unknown in general. For simplicity, all results will be stated with known σ_i^2, as all results remain valid if the unknown ones are replaced by appropriate estimates (see Remark 3). The weights

$$\omega_i = \omega_{\mathbf{h}_i}(K, C_d) = \sqrt{2 \log(K / \mathbf{h}_i^{\mathbf{1}})} + C_d \frac{\log(\sqrt{2 \log(K / \mathbf{h}_i^{\mathbf{1}})})}{\sqrt{2 \log(K / \mathbf{h}_i^{\mathbf{1}})}} \tag{8}$$

provide a proper scale calibration (see Section 2) if $K / \mathbf{h}_i \geqslant \sqrt{e}$ for all $i \in I_N$. Since for all results $\max_{i \in I_N} \mathbf{h}_i \to \mathbf{0}$, this is satisfied for any fixed $K > 0$ if n is large enough and we may assume throughout this paper, without loss of generality, that $\min_{i \in I_N} K / \mathbf{h}_i^{\mathbf{1}} \geqslant \sqrt{e}$. In this sense, our results hold for any constant $K > 0$, however, in many situations K can be chosen such that the weak limit of $\mathcal{S}(Y)$ in (7) is a standard Gumbel distribution (see Remark 2(c) and Theorems 3 and 5). C_d is an explicit constant only depending on the dimension, the system of scales considered and the degree of L^2-smoothness of $\Phi_{\mathbf{h}_i}$ (see Theorem 1 and Remark 2(b)). Our scale balancing (8) is in line with Dümbgen and Spokoiny (2001) and others (but notably different as explained in detail below), who pointed out that, in a multiscale setting, some elements of the dictionary may dominate the behavior of the maximum of a scanning statistic and it is most important to balance all local tests on the different scales in order to obtain good overall power, i.e. a scale dependent correction is necessary.

MISCAT now selects all probe functionals $\Phi_{i,n}$ as "active", where $\mathcal{S}(Y, i)$ is above a certain (universal) threshold, which guarantees (3), to be specified now. To this end, notice that in (3) we have

$$\sup_{J : J \subset I_N} \mathbb{P}_{H_{J,n}} \left[\text{"}\exists \text{ rejection in } J\text{"} \right] \leqslant \mathbb{P}_0 \left[\text{"}\exists \text{ rejection in } I_{N(n)}\text{"} \right], \tag{9}$$

where $\mathbb{P}_0 = \mathbb{P}_{0,n} = \mathbb{P}_{H_{I_{N(n)},n}}$, corresponding to $f \perp \mathcal{U}_n$. The reason for this is that the chance of a false positive among a selection of possible false positives is highest if this selection is as large as possible and all positives are false. Therefore, in order to control the FWER, we only need a universal global threshold $q_{1-\alpha}$ such that $\mathbb{P}_0\left[\mathcal{S}\left(Y\right) > q_{1-\alpha}\right] \leqslant \alpha$. To obtain this universal threshold $q_{1-\alpha}$ we will determine the $\mathbb{P}_0$-limiting distribution of $\mathcal{S}(Y)$ under a general moment condition including many practically relevant models. Theorem 1(a) in Section 2 provides a distribution free (i. e. independent of any unknown quantities such as f) limit, which is obtained as an almost surely bounded Gaussian approximation for the scan statistic (7) by replacing the errors by a standard Brownian sheet W, i. e.

$$\mathcal{S}(W) := \max_{i \in I_N} S(W,i), \quad \text{with} \quad S(W,i) := \omega_i \left(\frac{|\int \Phi_i(\mathbf{z}) \, \mathrm{d}W_{\mathbf{z}}|}{\|\Phi_i\|_2} - \omega_i \right). \tag{10}$$

Since $\mathcal{S}(W)$ does not depend on any unknown quantities, it can be used to simulate $q_{1-\alpha}$. Exploiting the specific and new choice of calibration in (8) we will furthermore show in Theorem 1(b) that $\mathcal{S}(Y)$ convergences in distribution towards a Gumbel limit for a wide-range of dictionary functions Φ_i. As $\mathcal{S}(Y)$ can be seen as a maximum over extreme value statistics of different scales, it follows that the contributions of the different scales are balanced in an ideal way. This result is remarkable, as it provides a general recipe how to calibrate multiscale statistics depending on the degree of smoothness of the probe functionals Φ_i and the system of scales considered. To best of our knowledge, this is new even in $d = 1$, and in addition, it generalizes results by Sharpnack and Arias-Castro (2016) to other systems than rectangular scanning (see Remark 2), and to inverse problems and non-Gaussian errors. Note that the calibration proposed by Dümbgen and Spokoiny (2001) for direct regression problems (which is frequently employed in multiscale procedures, see e.g. Rohde (2008); Walther (2010); Schmidt-Hieber et al. (2013); Eckle et al. (2016)) is tailored to a continuous observation setting in which all scales within a range $(0, a]$, $a \in \mathbb{R}^+$ are considered. If this calibration is used in a discrete setting like (1), the overall test-statistic converges to a degenerate limit, since the largest scale $h_{\max}$ has to satisfy $h_{\max} \to 0$ as $n \to \infty$, otherwise the finite sample approximations do not converge to their continuous counterparts. Therefore, we propose a different scale calibration which also takes into account the ill-posedness and yields a proper weak limit in many of such cases.

The approximation in (10) requires a coupling technique to replace the observation errors by i.i.d. Gaussian random variables. To this end we do not make use of strong approximations by KMT-like constructions (see Komlós

et al. (1975) for the classical KMT results and, e. g. Rio (1993) or Dedecker et al. (2014) for generalizations) as, for instance, Schmidt-Hieber et al. (2013) in the univariate case, $d = 1$, but we take a different route and employ a coupling of the supremum based on recent results by Chernozhukov et al. (2014). Doing so, we can prove the approximation in (10) to hold for a much larger range of scales.

A major benefit of MISCAT is its wide range of applicability and its multiscale detection power. Given the operator T, one chooses a dictionary $\mathcal{U}$ of probe functionals as in (2) such that $\mathcal{W}$ is of the form (4). We will demonstrate this for the case of T being the Radon transform in Section 3.1 and for T being a convolution operator in Section 3.2. For the latter situation we will also discuss an optimal choice of the probe functionals φ_i. Once the dictionaries $\mathcal{U}$ and $\mathcal{W}$ have been obtained, the quantiles $q_{1-\alpha}$ from the Gaussian approximation (10) or its finite sample analogues can be simulated. As it is well-known that convergence towards the Gumbel limit is extremely slow, it is beneficial that for deconvolution we find that the limit only depends on the degree of smoothness (see Theorem 4), and hence the finite sample distribution can be pre-simulated in a universal manner.

We will show in Section 2.4 that the power of MISCAT asymptotically coincides with the power of a single-scale oracle test which knows the correct size of the unknown object beforehand. More generally, if prior scale information is available, our method can be adapted immediately to this situation by restricting (7) to this subset, which mat lead to different calibration constants in (8) (see Remark 2(b)). This will further increase detection power in finite sample situations.

1.2. *MISCAT in action: Locating fluorescent markers in STED super-resolution microscopy.* In Section 3.2, we specify and refine our results to deconvolution which is applied to a data example from nanobiophotonics in Section 4.2 which we briefly review in the following. Suppose that the operator T is a convolution operator having a kernel k such that

$$(Tf)(\mathbf{y}) = (k * f)(\mathbf{y}) = \int_{\mathbb{R}^d} k(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) \, \mathrm{d}\mathbf{y}. \tag{11}$$

In our subsequent application the convolution kernel k corresponds to the point spread function of a microscope and the object of interest, f, is an image such that $d = 2$. We assume that k is finitely smooth, which is equivalent to a polynomial decay of its Fourier coefficients. Futhermore, if we choose $\mathcal{U}$ to be of wavelet-type, then the specific structure of the convolution ensures that $\mathcal{W}$ is as in (4) (cf. (29) and (30) below). Consequently, in this situation

we may choose the dictionaries $\mathcal{U}$ and $\mathcal{W}$ such that each $\varphi_i \geqslant 0$ has compact support $\operatorname{supp}(\varphi_i) \subset [\mathbf{t}_i - \mathbf{h}_i, \mathbf{t}_i]$. Consequently, if $f \geqslant 0$, we find

$$\langle f, \varphi_i \rangle > 0 \qquad \Rightarrow \qquad \exists \mathbf{x} \in [\mathbf{t}_i - \mathbf{h}_i, \mathbf{t}_i] \quad \text{s.t.} \quad f(\mathbf{x}) > 0, \tag{12}$$

i. e., there must be a point $\mathbf{x} \in [\mathbf{t} - \mathbf{h}, \mathbf{t}]$ belonging to the support of f. Employing this, we can use MISCAT to segment f into active and (most likely) inactive parts, which is of particular interest in many imaging modalities.
With this setup, MISCAT will be used to infer on the location of fluorescent markers in DNA origami imaged by a super-resolution STED microscope (cf. Hell, 2007). In STED microscopy, the specimen is illuminated by a laser beam along a grid with a diffraction-limited spot centered at the current grid point and the entire specimen is scanned this way, pixel by pixel, leading to observations as in (1) with a convolution T as in (11). The error distribution and the kernel k in (11) are well-known experimentally, see Supplement A for a detailed description of the mathematical model.
The investigated specimen consists of DNA origami, which have been designed in a way such that each of the clusters contains up to 24 fluorescent markers, arrayed in two strands of up to 12 having a distance of 71 nanometers (nm) (cf. the sketch in the upper left of Figure 1). As the ground truth is basically known, this serves as a real world phantom. Data were provided by the lab of Stefan Hell of the Department of NanoBiophotonics of the Max Planck Institute for Biophysical Chemistry, cf. Figure 1.

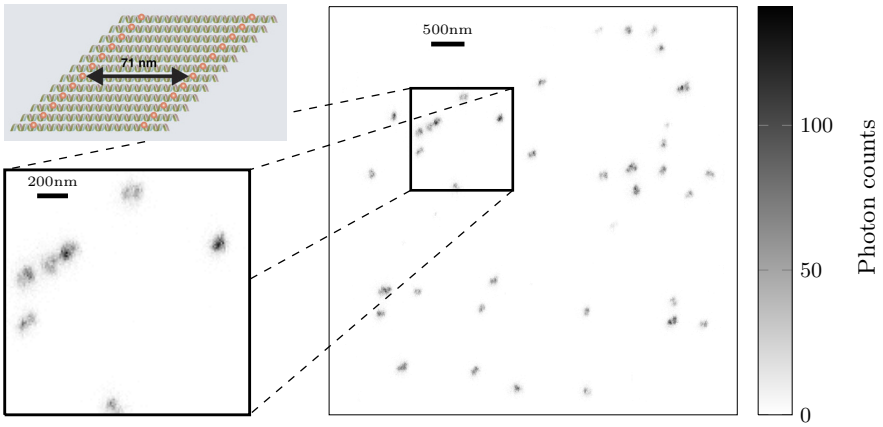
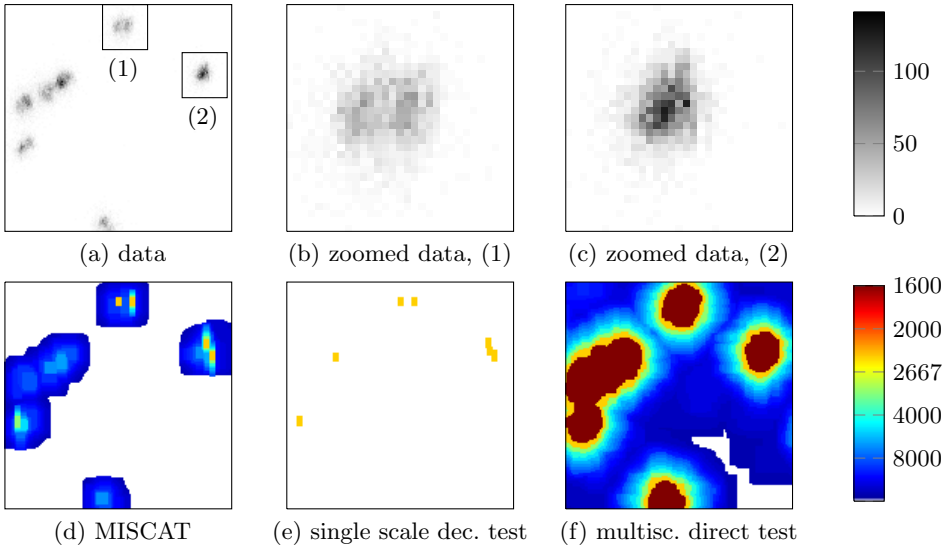


FIG 1. *Experimental data of the DNA origami sample and zoomed region (150×150 pixels). The sketch in the upper left shows the structure of the investigated DNA origami sample (red dots represent possible positions for fluorophores) see (Ta et al., 2015)).*

To infer on the positions of the fluorescent markers, we apply MISCAT with

a set of scales defined by boxes of size $k_x \times k_y$ pixels, $k_x, k_y = 4, 6, ..., 20$. One pixel in the measurements in Figure 1 is of size 10 nm $\times$ 10 nm. To highlight our multiscale approach we also display results of a single scale version of MISCAT (see Remark 2(b) and Section 4.2) using only boxes of size 4×6 pixels (these are the smallest boxes found by MISCAT), and to highlight the deconvolution effect, we apply a direct multiscale scanning test not designed for deconvolution (i.e. $T = \mathrm{id}$ in the model (1) and $\Phi_i = \varphi_i$ in (7)) based on indicator functions as probe functionals using the scale calibration suggested by Dümbgen and Spokoiny (2001). For details see Section 4.2.
In Figure 2 the zoomed region of Figure 1 is shown together with *significance maps* for all three tests. The significance map color-codes for each pixel the smallest scale (volume of the box in nm^2) on which it is significant. In case that a pixel belongs to significant boxes of different scales, only the smallest one is displayed for ease of visualization by the color coding. For



(a) data (b) zoomed data, (1) (c) zoomed data, (2)

(d) MISCAT (e) single scale dec. test (f) multisc. direct test

FIG 2. *Experimental data and corresponding* 90% *significance maps computed by different tests. The color-coding of the significance maps always show the size of smallest significance in* nm^2*, cf. the main text. (a)–(c) data and zoomed regions, (d) MISCAT, (e) a single scale test with deconvolution, (f) a multiscale scanning test without deconvolution. We emphasize that MISCAT performs* 1.587.600 *tests on the data in (a), and out of those* 13.973 *local hypotheses are rejected. The FWER control ensures that with (asymptotic) probability at least* 90% *among the selected regions there is no wrong detection.*

instance, in Figure 2(d) MISCAT marked several boxes as significant, and the smallest scale on which significant boxes were found is of size 2400 nm^2 (yellow). These results show that MISCAT is able (at least for some of the

single DNA origamis) to distinguish both strands. In view of the zoomed data in Figure 2(b) and Figure 2(c) this is quite remarkable as not visible from the data. The latter is due to the fact that the distance between the two strands of 71nm is slightly smaller than the full width at half maximum (FWHM, see Supplement A for details) of the convolution kernel k ($\approx$ 76nm), and there is a common understanding that objects which are closer to each other than a distance of approximately the FWHM cannot be identified as separate objects. Hence, MISCAT allows to discern objects below the resolution level of the STED microscope. The single scale variant of MISCAT (for explanation see Section 4) in Figure 2(e) has clearly more power in detecting small features on this single scale. While the multiscale test detects 4 boxes of 4×6 pixels, the single scale test detects several more, however, at the price of overseeing many DNA origamis at different scales. Note that the investigated specimen consists only of structures, which are present on a few (known) scales. For illustrative purposes, MISCAT, as employed here, does not use this information, as in general, these scales are not a priori known in living cell imaging. It is also clearly visible in Figure 2(f) that ignoring the deconvolution does not lead to a competitive test: distinguishing between different DNA origamis fails completely, as the support of the DNA-origami has been severely blurred by the STED microscope. We emphasize that the FWER control in (3) with $\alpha = 0.1$ implies that with (asymptotic) probability $\geqslant 90\%$, each of the 13.973 detections out of 1.587.600 local tests in Figure 2(d) is correct.

2. General Theory.

2.1. *Framework and Notation.* Recall the general framework introduced in Section 1 and model (1) and that all quantities may depend on the sample size n. Throughout this paper, $\{Tf(\mathbf{x}_{\mathbf{j},n}) \,|\, \mathbf{j} \in I_n^d\}$ is the discretization of the function Tf on the grid $\{(j_1/n, \ldots, j_d/n) \,|\, 1 \leqslant j_k \leqslant n,\ 1 \leqslant k \leqslant d\}$. This discretization model is a prototype for many inverse problems and in particular matches the application to imaging considered in Section 4 below. For different applications alternative discretization schemes may be of interest as well but, for the sake of a clearer display, we consider uniform sampling on a complete grid since most of the results presented below do not crucially depend on the specific discretization. We make the following assumption on the dictionaries $\mathcal{U}$ and $\mathcal{W}$ in (2) and (4).

ASSUMPTION 1. *Let $\mathcal{U}$ as in* (2) *and $\mathcal{W}$ as in* (4).

(a) Dictionary source condition *Let*

$$\text{(DSC)} \qquad \varphi_i \in \mathcal{R}(T^*), \quad \textit{i. e.,} \quad \varphi_i = T^*\Phi_i.$$

(b) Growth of the dictionary *For some* $\kappa > 0$

$$\text{(G)} \qquad |\mathcal{U}| = |\mathcal{W}| = N = O(n^\kappa).$$

(c) Scale restrictions *For the smallest and the largest scale in* (4)*, i. e.,* $\mathbf{h}_{\min} = (h_{\min}, \dots, h_{\min})^T$ *and* $\mathbf{h}_{\max} = (h_{\max}, \dots, h_{\max})^T$*, respectively,*

$$\text{(SR)} \qquad h_{\min} \gtrsim n^{-1} \log(n)^{15/d \vee 3} \log\log(n)^2 \quad \textit{and} \quad h_{\max} = o\big(\log(n)^{-2}\big).$$

(d) Average Hölder condition *Suppose that the functions* $\Phi_{\mathbf{h}_i}$ *in* (4) *are uniformly bounded, supported on* $[0,1]^d$*, vanishing at the boundary and*

$$\text{(AHC)} \qquad \int |\Phi_{\mathbf{h}_i}(\mathbf{t}-\mathbf{z}) - \Phi_{\mathbf{h}_i}(\mathbf{s}-\mathbf{z})|^2 \,\mathrm{d}\mathbf{z} \leqslant L\|\mathbf{t}-\mathbf{s}\|_2^{2\gamma} \|\Phi_{\mathbf{h}_i}\|_2^2,$$

for some $\gamma \in (0,1]$ *and all* $i \in I_N$ *uniformly as* n *and hence* $N \to \infty$*.*

REMARK 1. *(a) Assumption* (DSC) *is a smoothness condition on the functions of the dictionary* $\mathcal{U}$ *related to* T*. Instead of posing such an assumption on the dictionary, it is common to pose such an assumption on* f*, e. g. the so-called benchmark source condition* $f \in \mathcal{R}(T^*)$*, which requires the unknown solution* f *to be at least as smooth as any function in the range of* T^**. For deconvolution problems with real-valued kernel this means that* f *is at least as smooth as the kernel itself. In this paper, as we want to reconstruct pairings* $\langle f, \varphi_i \rangle$ *instead of* f*, we may relax this and pose conditions on the functions* φ_i *instead of* f*, see also (Burger et al., 2013). Note, that if additionally* f *admits a sparse representation w.r.t. the dictionary* $\mathcal{U}$*, then* (DSC) *implies* $f \in \mathcal{R}(T^*)$*. We emphasize that our approach strongly relies on the condition* (DSC)*, see also Donoho (1995); Anderssen (1986). For a strategy how to estimate a linear functional* $\langle f, \varphi \rangle$ *for* $\varphi \notin R(T^*)$ *we refer to Mathé and Pereverzev (2002).*

(b) Assumption (G) *is rather mild. In particular it implies that positions and scales* $(\mathbf{t}_i, \mathbf{h}_i)$ *from any grid of polynomial size can be used. In the example of imaging this is naturally satisfied as the* $\mathbf{t}_i$ *are grid points of the pixel grid and the sizes of the scales* $\mathbf{h}_i$ *are given by rectangular groups of pixels and are hence also only of polynomial order in* n*. Furthermore, to serve as an approximation for a continuous version, the grid can be chosen sufficiently fine and still* (G) *is satisfied. The constant* κ *only enters into our results via some constants.*

(c) *As already discussed in the introduction, the scale restrictions* (SR) *are also rather mild. The lower bound on* $h_{\min}$ *is up to a poly-log factor of the same order as the sampling error, and the upper bound on* $h_{\max}$ *is required to ensure asymptotic unbiasedness of our local test statistics. For some of the results presented below, a slightly stricter bound on* $h_{\max}$ *will be necessary, and this is emphasized in the corresponding theorems.*

(d) *Assumption* (AHC) *is a smoothness condition on the dictionary* $\mathcal{W}$. *In the case* $\gamma < 1$, *the class of functions satisfying Assumption* (AHC) *corresponds to the class* $H_2^{(\gamma,\dots,\gamma)}$, *defined by Nikol'skiĭ (1951, pp. 256-257), see also Tsybakov (2009, p. 13). It holds, for instance, if all* $\Phi_{\mathbf{h}_i}$ *are Hölder-continuous of order* γ. *In case* $T = id$, *the 'classical' scanning function* $\Phi_{\mathbf{h}_i} \equiv I_{(0,1)^d}$ *satisfies condition* (AHC) *with* $\gamma = 1/2$ *and* $L = d$. *In Section 3 we discuss this condition in more detail and show its validity if* T *is the Radon transform and if* T *is a convolution operator (see Section 3.1 and Section 3.2, respectively).*

The following assumptions concern the noise $\xi_{\mathbf{j}}$, $\mathbf{j} \in I_n^d$ in model (1).

ASSUMPTION 2. *Let* $\xi_{\mathbf{j}}$, $\mathbf{j} \in I_n^d$ *in* (1) *be independent and centered random variables. Assume that there exists a function* $\sigma \in C^1[0,1]^d$ *such that* $\mathrm{Var}[\xi_{\mathbf{j}}] = \sigma^2(\mathbf{x}_{\mathbf{j}})$ *and*

$$\text{(M1)} \qquad \mathbb{E}|\xi_{\mathbf{j}}|^{2J} \leqslant \frac{1}{2} J! \, \mathbb{E}\xi_{\mathbf{j}}^4 \qquad \textit{for all} \qquad J \geqslant 2.$$

Assume further that

$$\text{(M2)} \qquad 0 < \liminf_{n\to\infty} \inf_{\mathbf{j}\in I_n^d} \mathbb{E}[|\xi_{\mathbf{j}}|^2] \quad \textit{and} \quad \limsup_{n\to\infty} \sup_{\mathbf{j}\in I_n^d} \mathbb{E}[|\xi_{\mathbf{j}}|^4] < \infty.$$

Note that (M1) is in fact equivalent to the well-known Cramér condition that the moment generating function exists in a small neighborhood of 0 (cf. Lin, 2017, Thm. 1) and is satisfied by many distributions, including Gaussian and Poisson. The latter is most relevant for our subsequent application.

2.2. *Asymptotic Theory.* We are now in the position to provide some general asymptotic properties of MISCAT such as a uniform Gaussian approximation of the test statistic, a.s. boundedness of the simulated quantiles, and weak convergence under further specification of assumptions towards an explicit Gumbel-type distribution. The latter is for ease of presentation only shown when using the full set of possible scales. If MISCAT is restricted to smaller subsets of scales (e.g. resulting from prior information), this may change the limit distribution, see Remark 2 below.

THEOREM 1. *Suppose we are given observations from model* (1) *with random noise satisfying Assumption 2 and dictionaries* $\mathcal{U}$ *and* $\mathcal{W}$ *as specified in Assumption 1. Let* $h_{\max} \leqslant n^{-\delta}$ *for some (small)* $\delta > 0$ *in* (SR) *and suppose that the approximation error of* $\langle \mathbb{E}[Y], \Phi_i \rangle_n := \frac{1}{n^d} \sum_{\mathbf{j} \in I_n^d} Tf(\mathbf{x_j}) \Phi_i(\mathbf{x_j})$ *is asymptotically negligible, i. e.,*

$$n^{\frac{d}{2}} \max_{i \in I_N} \frac{\langle \mathbb{E}[Y], \Phi_i \rangle_n - \langle Tf, \Phi_i \rangle}{\|\Phi_i\|_2} = o\bigg(\frac{1}{\log(n)^2 \log\log(n)^2} \bigg). \tag{13}$$

For any constant $K > 0$ *and* $C_d = 2d + d/\gamma - 1$ *consider the calibration values* $\omega_i = \omega_i(K, C_d)$ *as in* (8).

(a) Then, for a standard Brownian sheet W *on* $[0,1]^d$*, it holds*

$$\lim_{n \to \infty} \big| \mathbb{P}_0\big(\mathcal{S}(Y) \leqslant q\big) - \mathbb{P}_0\big(\mathcal{S}(W) \leqslant q\big) \big| = 0, \qquad q \in \mathbb{R}$$

where $\mathcal{S}(Y)$ *and* $\mathcal{S}(W)$ *are defined in* (7) *and* (10)*, respectively. Consequently, under* H_0*,* $\mathcal{S}(Y)$ *and* $\mathcal{S}(W)$ *converge weakly towards the same limit. Furthermore, the approximating statistic* $\mathcal{S}(W)$ *is almost surely bounded and does not depend on any unknown quantity.*

(b) Instead of (AHC) *assume the stronger condition that there exists a function* Ξ *supported on* $[0,1]^d$ *with* $\|\Xi\|_2 = 1$ *such that*

$$\max_{i \in I_N} \left| \int \left(\frac{\Phi_{\mathbf{h}_i}(\mathbf{t}_i - \mathbf{z})}{\|\Phi_{\mathbf{h}_i}\|_2} - \Xi(\mathbf{t}_i - \mathbf{z}) \right) \mathrm{d}W_{\mathbf{z}} \right| = o_{\mathbb{P}}\left(\frac{1}{\sqrt{\log(n)}} \right) \tag{14}$$

and

$$\int |\Xi(D_\Xi(\mathbf{t} - \mathbf{z})) - \Xi(D_\Xi(\mathbf{s} - \mathbf{z}))|^2 \, \mathrm{d}\mathbf{z} = \sum_{j=1}^{d} |t_j - s_j|^{2\gamma} (1 + o(1)) \tag{15}$$

with $\gamma \in (0,1]$ *and a symmetric, positive definite matrix* $D_\Xi \in \mathbb{R}^{d \times d}$*. Suppose that the set of scales* $\mathcal{H} := \big\{ \mathbf{h}_i \;\big|\; i \in I_N \big\}$ *is complete, i.e.* $\mathcal{H} = \{h_{\min}, ..., h_{\max}\}^d$*, where*

$$-\log(h_{\max}) = \delta \log(n) + o(\log(n)), \; -\log(h_{\min}) = \Delta \log(n) + o(\log(n)) \tag{16}$$

with $0 < \delta < \Delta \leqslant 1$*. If the grids of positions* $\mathbf{t}$ *and scales* $\mathbf{h}$ *are furthermore sufficiently fine, i.e.*

$$\max_{i \in I_N} \min_{j \in I_N : \mathbf{t}_i \neq \mathbf{t}_j} \|\mathbf{t}_i - \mathbf{t}_j\|_\infty = O\big(n^{-1}\big) \tag{17}$$

and

(18)
$$\max_{i\in I_N}\min_{j\in I_N;\, h_{i,l}\neq h_{j,l}} \left|(h_{j,l}-h_{i,l})/\sqrt{h_{i,l}h_{j,l}}\right| \to 0 \qquad \textit{for all} \qquad 1\leqslant l\leqslant d$$

then it holds

$$\lim_{n\to\infty} \mathbb{P}_0\big(\mathcal{S}(Y)\leqslant\lambda\big) = \exp\left(-\exp\left(-\lambda\right)\cdot \tfrac{H_{2\gamma}\det\left(D_{\bar{\Xi}}^{-1}\right) I_d(\delta,\Delta)}{\sqrt{2\pi K}}\right), \tag{19}$$

with

$$I_d(\delta,\Delta) := \frac{(-1)^{d-1}}{(d-1)!}\sum_{k=0}^{d}(-1)^k\binom{d}{k}\log\big(k\delta+(d-k)\Delta\big) > 0 \tag{20}$$

and Pickands' constant $H_{2\gamma}$ (cf. Pickands, 1969).

Detailed proofs are deferred to Supplement A, and the main ideas are described in Section 6.

Remark 2.

(a) *Assumption* (13) *is a mild assumption on the integral approximation as the required rate is very slow. It is satisfied, in particular, if Tf and Φ in* (4) *are Hölder-continuous of some order, or if Tf is Hölder-continuous and Φ is an indicator function. Note that due to the ill-posedness of the problem, Tf being Hölder-continuous does typically not require f to be continuous.*

(b) *Although it might seem marginal, a proper choice of the constant C_d is crucial for the boundedness of $\mathcal{S}(W)$. The choice $C_d = 2d + d/\gamma - 1$ used in the formulation of the theorem is adjusted to the case where a dense grid of scales in the sense of* (18) *is considered. In particular, this includes the case where* ***all*** *scales in Assumption 1* (SR) *ranging from $\mathbf{h}_{\min}$ to $\mathbf{h}_{\max}$ are used. If now, for instance, $T = \mathrm{id}$ and Φ in* (4) *is chosen to be the indicator function of $[0,1]^d$, we have $\gamma = 1/2$ and consequently $C_d = 4d - 1$, which coincides with the constant of Sharpnack and Arias-Castro (2016) for the Gaussian case.*
However, in many situations a less dense grid of scales might be of interest, e.g. under prior scale information on the object of interest f. Then for the choice $C_d = 2d + d/\gamma - 1$ the statistic $\mathcal{S}(W)$ is still a.s. bounded from above, but (19) *might not be valid anymore. To avoid this, C_d has to be adjusted. Suppose in what follows that the grid of positions still satisfies* (17)*. In the least dense regime, when $\mathcal{S}(W)$*

behaves as in a single scale scenario, the proper choice is $C_d = d/\gamma - 1$. *Another interesting special case is when only squares in a dense range are considered (this is* $\mathbf{h}_i = (h_i, ..., h_i)$ *and* (18) *is satisfied), where one should choose* $C_d = 1 + d/\gamma$.
All these choices of C_d *are specified in more detail in Corollary 1 in Section 5 and follow from our general result in Theorem 7.*

(c) *As specified in the theorem,* $\mathcal{S}(W)$ *is bounded for any choice of the constant* $K > 0$. *In fact,* K *does not affect the asymptotic power of MISCAT as it only determines the location of the limiting distribution. For* $\gamma \in \{1/2, 1\}$, $H_{2\gamma}$ *can be computed explicitly (see Pickands, 1969), i.e.* $H_1 = 1$ *and* $H_2 = \pi^{-\frac{d}{2}}$. *In these cases the explicit choice* $K = |\det D_\Xi^{-1}| I_d(\delta, \Delta) H_{2\gamma}/\sqrt{2\pi}$ *yields standard Gumbel limit distributions. If* $\gamma = 1$ *and if the correlation function* r_Ξ *of the Gaussian field* $Z_\mathbf{t} = \int \Xi\big(\mathbf{t} - \mathbf{z}\big)\, \mathrm{d}W_\mathbf{z}$ *is twice differentiable in* $\mathbf{0}$, *the matrix* D_Ξ *can be computed via* $D_\Xi^* D_\Xi = \mathrm{Hess}_{r_\Xi}(\mathbf{0})^{-1}$. *For* T *being the Radon transform or a convolution operator, this allows us to give explicit constants* K *in* (27) *and* (37), *respectively, ensuring standard Gumbel limit distributions.*

(d) *In the situation of Theorem 1 (b) under a weaker assumption than* (14) *and* (15) *it can be shown that the limiting distribution is stochastically bounded by Gumbel distributions and is hence non-degenerate in the limit. This will be done in Theorem 4 in the situation of deconvolution.*

2.3. *Statistical Inference.* In the following, let $q_{1-\alpha}$ denote the $1-\alpha$-quantile of the approximating process $\mathcal{S}(W)$. To compare the local test statistics $\mathcal{S}(Y, i)$ in (7) with $q_{1-\alpha}$, we have assumed so far to know the local variances $\sigma_i^2 = \mathrm{Var}[\langle Y, \Phi_i \rangle_n]$. The next Remark shows that they can easily be estimated without changing the limiting distribution of $\mathcal{S}(W)$.

REMARK 3. *As mentioned before, the local variances* σ_i^2, $i \in I_N$, *depend on* $Var[\xi_\mathbf{j}] = \sigma^2(\mathbf{x_j})$ *(cf. Assumption 2),* $\mathbf{j} \in I_n^d$, *which are typically unknown in applications. Nevertheless, all results remain valid if the* C^1*-function* σ^2 *(see Assumption 2) can be estimated from the data by* $\hat{\sigma}^2$ *such that*

$$\max_{i \in I_N} \big|\hat{\sigma}^2(\mathbf{t}_i) - \sigma^2(\mathbf{t}_i)\big| = o_\mathbb{P}\big(\log(n)^{-\frac{1}{2}}\big). \tag{V}$$

The local variances σ_i^2 *can then be estimated by* $\hat{\sigma}_i^2 := \big\langle \hat{\sigma}^2, \Phi_i^2 \big\rangle_n$. *Condition* (V) *is e.g. satisfied for (suitable) kernel-type estimators or point-wise maximum likelihood estimators as used in Section 4.2.*

We conclude by Theorem 1 that $\lim_{n\to\infty} \mathbb{P}_0\big(\mathcal{S}(Y, i) \leqslant q_{1-\alpha}\, \forall\, i \in I_N\big) \geqslant 1-\alpha$, and hence (3) is valid, i.e. all rejections are significant findings. Conversely, it

can be shown that, with overall confidence of approximately $(1-\alpha)\cdot 100\%$, all relevant components are found, provided that the signal is sufficiently strong.

LEMMA 1. *Suppose we are given observations from model* (1) *with random noise satisfying Assumption 2 and dictionaries* $\mathcal{U}$ *and* $\mathcal{W}$ *as specified in Assumption 1. Let* $\mathcal{I}_\alpha$ *denote the set of all large components, i. e.*

$$\mathcal{I}_\alpha := \big\{ i \;\big|\; \langle \varphi_i \,,\, f\rangle > 2\big(\tfrac{q_{1-\alpha}}{\omega_i} + \omega_i\big)\sigma_i \big\}.$$

Then, under the assumptions of Theorem 1

$$\lim_{n\to\infty} \mathbb{P}\big(\mathcal{S}\,(Y,i) > q_{1-\alpha} \quad \textit{for all} \quad i \in \mathcal{I}_\alpha\big) \geqslant 1-\alpha$$

For general T it is not clear if the detection guarantee in Lemma 1 is optimal in the sense that weaker signals cannot be detected by any procedure. However, in the next subsection we will show that in special situations MISCAT obeys an oracle optimality property.

2.4. *Asymptotic Optimality.* For signals built from block signals, the asymptotic power of MISCAT can be computed explicitly which reveals an oracle optimality property of MISCAT in the following sense. Suppose that $f = \mu_{n,\mathbf{h}_\star} I_{[\mathbf{t}_\star - \mathbf{h}_\star, \mathbf{t}_\star]}$. If one knew the correct scale $\mathbf{h}_\star$, one would perform a single-scale test in order to find the location $\mathbf{t}_\star$. Hence, in this idealized situation, the "oracle scan statistic" $\mathcal{S}^\star(Y)$ given by

$$\mathcal{S}^\star(Y) = \sup_{i\in I_N} \omega_{\mathbf{h}_\star}\big(K, \tfrac{d}{\gamma} - 1\big)\Big(\sigma_i^{-1}\Big\langle Y, \Phi_{\mathbf{h}_\star}\big(\tfrac{\mathbf{t}_i - \cdot}{\mathbf{h}_\star}\big)\Big\rangle_n - \omega_{\mathbf{h}_\star}\big(K, \tfrac{d}{\gamma} - 1\big)\Big)$$

would be used. Note the different adjustment of weights due to Remark 2(b). It turns out that MISCAT performs as well in terms of its asymptotic power as the oracle test corresponding to $\mathcal{S}^\star\,(Y)$. Moreover, the following theorem guarantees that signals will be detected asymptotically with probability 1, if $\mu_{n,\mathbf{h}} \geqslant \max_{\mathbf{t}} \sigma(\mathbf{t})(\sqrt{2\log(1/\mathbf{h}_\star)} + \beta_n) n^{-\frac{d}{2}} \|\Phi_{i_\star}\|_2$, where $i_\star$ is such that $(\mathbf{t}_i, \mathbf{h}_i) = (\mathbf{t}_\star, \mathbf{h}_\star)$ and $\beta_n \to \infty$. In this setting, if the errors are i.i.d. standard normal and $T = \mathrm{id}$, the single scale test is minimax optimal if $\|\Phi_i\|_2 = \sqrt{\mathbf{h}_i^{\mathbf{1}}}$, which follows from the arguments in Kou (2017) (see also Arias-Castro et al., 2005; Chan and Walther, 2013, for related results). A rigorous proof for the case $d = 2$, $\Phi = I_{[0,1]^2}$ and $h_1 = h_2$ can be found in Butucea and Ingster (2013). Thus, also the multiscale procedure MISCAT is minimax optimal in this case. If $T \neq \mathrm{id}$, optimality depends on both dictionaries $\mathcal{W}$ and $\mathcal{U}$ and special care has to be put into the choice of dictionary functions. This

is discussed in more detail in Section 3.2.1 below. Under general noise, the following can be said:

THEOREM 2 (Asymptotic Power of MISCAT). *Suppose we are given observations from model* (1) *with random noise satisfying Assumption 2 and dictionaries* $\mathcal{U} = \{\varphi_i \,|\, \varphi_i(\mathbf{z}) = \varphi((\mathbf{t}_i - \mathbf{z})/\mathbf{h}_i),\, \varphi(\mathbf{z}) > 0, \mathbf{z} \in (0,1)^d\}$ *and* $\mathcal{W}$ *as specified in Assumption 1. Suppose* (16) *with* $0 < \delta < \Delta \leqslant 1$ *and fix a scale* $\mathbf{h}_\star = \mathbf{h}_\star(n) \in [\mathbf{h}_{\min}, \mathbf{h}_{\max}]$ *and a subset* $\mathcal{T}_\star \subset I_N$ *such that* $\mathbf{h}_i = \mathbf{h}_\star$ *for all* $i \in \mathcal{T}_\star$*. Now consider the set of functions* f *with support given by the union of all corresponding boxes which are sufficiently strong, i.e.*

$$\mathscr{S}_{\mathcal{T}_\star}(\mathbf{h}_\star, \mu_n) := \left\{ f \,\middle|\, (13) \text{ holds, } \operatorname{supp}(f) = \bigcup_{i \in \mathcal{T}_\star} I_{\mathbf{t}_i, \mathbf{h}_\star},\, \frac{\langle \varphi_i, f \rangle}{\|\Phi_i\|_2} \geqslant \frac{\mu_n}{n^{d/2}}, i \in \mathcal{T}_\star \right\},$$

where $I_{\mathbf{t}_i, \mathbf{h}_\star} := [\mathbf{t}_i - \mathbf{h}_\star, \mathbf{t}_i]$*. Assume that* $\sigma \in C^1([0,1]^d)$ *and* $\mathbf{t}_\star \in (0,1)^d$ *where* $\mathbf{t}_\star \in \operatorname{argmax}\{\sigma(\mathbf{t}) \,|\, \mathbf{t} \in [0,1]^d\}$ *and let* $K > 0$*.*

(a) If $\{\mathbf{h}_i \,|\, i \in I_N\} = \{\mathbf{h}_\star\}$*, i. e. for each* $\mathbf{t}$ *we consider scanning windows of (correct) size* $\mathbf{h}_\star$*, then MISCAT with the single-scale-calibration* $\omega_i(K, d/\gamma - 1)$ *as in* (8) *(cf. Remark 2(b)) attains power*

$$\begin{aligned} \inf_{f \in \mathscr{S}_{\mathcal{T}_\star}(\mathbf{h}_\star, \mu_n)} \mathbb{P}_f\big(\mathcal{S}^\star(Y) > q_{1-\alpha}\big) &= \inf_{f \in \mathscr{S}_{\{\mathbf{t}_\star\}}(\mathbf{h}_\star, \mu_n)} \mathbb{P}_f\big(\mathcal{S}^\star(Y) > q_{1-\alpha}\big) \\ &= \alpha + (1-\alpha) \cdot \overline{\psi}\left(\sqrt{2 \log\left(\tfrac{1}{\mathbf{h}_\star^1}\right)} - \frac{\mu_n}{\sigma(\mathbf{t}_\star)} \right) + o(1). \end{aligned}$$

Here and in the following, $\overline{\psi}(x) := \int_x^\infty (2\pi)^{-1/2} \exp\left(-y^2/2\right) \mathrm{d}y$ *is the tail function of the standard normal distribution.*

(b) In general, MISCAT with the multiscale-calibration $\omega_i(K, 2d + d/\gamma - 1)$ *as in* (8) *satisfies*

$$\inf_{f \in \mathscr{S}_{\mathcal{T}_\star}(\mathbf{h}_\star, \mu_n)} \mathbb{P}_f\big(\mathcal{S}(Y) > q_{1-\alpha}\big) + o(1) \geqslant \inf_{f \in \mathscr{S}_{\mathcal{T}_\star}(\mathbf{h}_\star, \mu_n)} \mathbb{P}_f\big(\mathcal{S}^\star(Y) > q_{1-\alpha}\big), \tag{21}$$

i. e. the multiscale procedure performs at least as well as the oracle procedure.

This complements the results from Theorems 4 and 6 in Sharpnack and Arias-Castro (2016), where a similar expansion of the power is provided for the case $T = \mathrm{id}$ and $\Phi = I_{[0,1]^d}$.

3. Examples.

3.1. *The d-dimensional Radon transform.* Assume one observes a discretized and noisy sample of the Radon transform of f,

$$Y_{\mathbf{k},l} = Tf(\boldsymbol{\vartheta}_{\mathbf{k}}, u_l) + \xi_{\mathbf{k},l}; \quad u_l = \tfrac{l-1/2}{n}, \quad l = 1, \dots, n \tag{22}$$

and $\boldsymbol{\vartheta}_{\mathbf{k}} \in \mathbb{S}^{d-1}, \mathbf{k} \in I_n^{d-1}$ are design points which are uniformly distributed w.r.t. the angles in a parametrization using polar coordinates, where

$$Tf(u, \boldsymbol{\vartheta}) = \int_{\langle \mathbf{v}, \boldsymbol{\vartheta} \rangle = u} f(\mathbf{v}) \, \mathrm{d}\mu_{d-1}(\mathbf{v})$$

denotes Radon transformation (cf. Natterer, 1986), $\mathrm{d}\mu_{d-1}$ denotes the $(d-1)$-dimensional Lebesgue measure on the hyperplane $\{\mathbf{v} \,|\, \langle \mathbf{v}, \boldsymbol{\vartheta} \rangle = u\}$ and $\xi_{\mathbf{k},l}$ are i.i.d., $\mathrm{Var}[\xi_{(\mathbf{1},1)}] = \sigma^2$. In this case fix $\widetilde{\varphi} : \mathbb{R}^+ \to \mathbb{R}$, set $\varphi(\mathbf{x}) := \widetilde{\varphi}(\|x\|_2)$, $\mathrm{supp}(\widetilde{\varphi}) \subset [0,1]$ and define

$$\mathcal{U} = \left\{ \varphi_i = h_i^{-d/2} \varphi\left(\tfrac{\cdot - \mathbf{t}_i}{h_i} \right) \,\middle|\, i \in I_N \right\}, \tag{23}$$

i.e. we consider a dictionary $\mathcal{U}$ of rotationally invariant functions. We now construct the corresponding dictionary $\mathcal{W}$. To this end we need to fix some more notation. Let $\mathrm{d}\boldsymbol{\vartheta}$ denote the common surface measure on $\mathbf{S}^{d-1}$ such that for measurable $S \subset \mathbb{S}^{d-1}$ we have $|S| = \int_S \mathrm{d}\boldsymbol{\vartheta}$. Let further $\mathcal{F}_d f$ denote the d-dimensional Fourier transform of f, defined by

$$\mathcal{F}_d f(\boldsymbol{\xi}) = \int f(\mathbf{x}) \, \exp\left(\mathrm{i}\langle \mathbf{x}, \boldsymbol{\xi} \rangle\right) \mathrm{d}\mathbf{x}, \;\; f(\mathbf{x}) = \frac{1}{(2\pi)^d} \int \mathcal{F}_d f(\boldsymbol{\xi}) \, \exp\left(-\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x} \rangle\right) \mathrm{d}\boldsymbol{\xi}.$$

Lemma 2. *Let $\mathcal{U}$ be as in (23), $\varphi \in \mathcal{R}(T^*)$. Then*

$$\mathcal{W} = \left\{ \Phi_i \,\middle|\, \Phi_i(u, \boldsymbol{\vartheta}) = h_i^{-\frac{d}{2}} \Phi\left(\tfrac{u - \langle \boldsymbol{\vartheta}, \mathbf{t}_i \rangle}{h_i} \right) \right\}, \tag{24}$$

where, due to the rotational invariance of φ, the function Φ, defined by

$$\Phi(x) := \frac{1}{2(2\pi)^d} \mathcal{F}_1\left((\mathcal{F}_d \varphi)(\cdot \boldsymbol{\vartheta}) \,|\, \cdot \,|^{d-1} \right)(x), \qquad x \in \mathbb{R}, \tag{25}$$

is independent of $\boldsymbol{\vartheta}$.

Consequently, the functions Φ_{h_i} as in (4) are in $\mathcal{W}$ as in (24), i.e. we have the special structure $\Phi_{h_i} = C_{h_i} \Phi$ and hence we can define $\Xi := \Phi_{h_i} / \|\Phi_{h_i}\|_{L^2(\mathbb{R} \times \mathbb{S}^{d-1})}$. It turns out that (AHC) and (15) are satisfied if φ is sufficiently smooth. This is made more precise in the following Lemma.

LEMMA 3. *Let* $4\pi\|\mathcal{F}_1\big((\mathcal{F}_d\varphi)(\cdot\boldsymbol{\vartheta})|\cdot|^{d-1}\big)(u-\langle\mathbf{t}_i,\boldsymbol{\vartheta}\rangle)\|^{-1}_{L^2(\mathbb{R}\times\mathbb{S}^{d-1})} =: C_{\varphi,d}$. *If* $\varphi\in H^{\frac{d+1}{2}}(\mathbb{R}^d)$, (15) *holds with*

$$D_\Xi^{-2} := \operatorname{diag}\left(C_{\varphi,d}\int_{\mathbb{R}^d}\omega_1^2\|\boldsymbol{\omega}\|^{d-1}\big|(\mathcal{F}_d\varphi)(\boldsymbol{\omega})\big|^2\,\mathrm{d}\boldsymbol{\omega}\right). \tag{26}$$

In general, the dictionary functions Φ may be of unbounded support. In this case the results from Theorem 1 b) remain valid if we exclude a small boundary region from our analysis. Here, we only consider positions $\mathbf{t}_i\in[\mathbf{0},\mathbf{1}-\boldsymbol{\rho}]$, where $\boldsymbol{\rho}=(\rho,\dots,\rho)^T$, $\rho>0$ and we obtain the following extreme value theorem for MISCAT in the case of the Radon transform.

THEOREM 3 (MISCAT for the Radon Transform). *Suppose that we have access to observations following model* (22). *Let* $\mathbf{t}_i\in[\mathbf{0},\mathbf{1}-\boldsymbol{\rho}]$, *where* $\boldsymbol{\rho}=(\rho,\dots,\rho)^T$, $\rho>0$. *Assume also that the approximation error of* $\langle\mathbb{E}[Y],\Phi_{h_i}\rangle_n$ *is asymptotically negligible, i. e.,* (13) *holds and* $\varphi\in H^{\frac{d+1}{2}}(\mathbb{R}^d)$, *such that the integral in* (26) *is finite. If furthermore* (16) *holds true with* $0<\delta<\Delta\leqslant 1$ *and the grids of positions* $\mathbf{t}$ *and scales* h *are sufficiently fine, i.e. satisfy* (17) *and* (18) *and if the calibration*

$$\omega(K,1+d)\quad\textit{with}\quad K=(1-\rho)^d(2\pi)^{-\frac{d+1}{2}}\det(D_\Xi^{-2})^{\frac12}\log(\Delta/\delta) \tag{27}$$

is used (see (8) *and Remark 2(b)), where* D_Ξ^{-2} *is defined in* (26), *then one has* $\lim_{n\to\infty}\mathbb{P}_0\big[\mathcal{S}(Y)\leqslant\lambda\big]=e^{-e^{-\lambda}}$. *Furthermore the statements of Lemma 1 and Theorem 2 also hold.*

3.2. *Deconvolution.* We discuss now in detail the case of deconvolution, i. e. (1) specializes to

$$Y_{\mathbf{j}} = (k*f)(\mathbf{x}_{\mathbf{j}})+\xi_{\mathbf{j}},\quad \mathbf{j}\in\{1,\dots,n\}^d, \tag{28}$$

where the function k is a convolution kernel and the operation "$*$" denotes convolution as defined in (11). In our subsequent data example k corresponds to the point-spread function (PSF) of a microscope (see e. g. Bertero et al., 2009; Aspelmeier et al., 2015; Hohage and Werner, 2016) .
Assume that there exist positive constants $\underline{c},\overline{C}$ and a such that

$$\underline{c}\big(1+\|\boldsymbol{\xi}\|_2^2\big)^{-a}\leqslant|\mathcal{F}_dk(\boldsymbol{\xi})|\leqslant\overline{C}\big(1+\|\boldsymbol{\xi}\|_2^2\big)^{-a}. \tag{D1}$$

Assumption (D1) is a standard assumption characterizing mildly ill-posed deconvolution problems (see e. g. Fan, 1991; Meister, 2009). For any fixed function φ, $\|\varphi\|_2>0$, generating a dictionary

$$\mathcal{U}=\big\{\varphi_i\,\big|\,\varphi_i(\mathbf{z})=\varphi\big(\tfrac{\mathbf{t}_i-\mathbf{z}}{\mathbf{h}_i}\big),\ i\in I_N\big\}, \tag{29}$$

the corresponding dictionary $\mathcal{W}$ inherits the required wavelet-type structure:

$$\mathcal{W} = \big\{\Phi_i \,\big|\, \Phi_i(\mathbf{z}) = \Phi_{\mathbf{h}_i}\big(\tfrac{\mathbf{t}_i - \mathbf{z}}{\mathbf{h}_i}\big),\ \Phi_{\mathbf{h}_i} := \mathcal{F}_d^{-1}\Big(\tfrac{\mathcal{F}_d\varphi}{\overline{\mathcal{F}_d k(\cdot/\mathbf{h}_i)}}\Big),\ i \in I_N\big\}, \tag{30}$$

and the results from the previous section transfer to deconvolution as follows.

THEOREM 4 (MISCAT for deconvolution). *Suppose model* (28) *with convolution kernel k satisfying Assumption* (D1) *and $\xi_{\mathbf{j}}$ satisfying Assumption 2. Let $\mathbf{t}_i \in [\boldsymbol{\rho} + \mathbf{h}_i, \mathbf{1} - \boldsymbol{\rho}]$, where $\boldsymbol{\rho} = (\rho, \dots, \rho)^T$, $\rho > 0$. Consider the dictionary $\mathcal{W}$, given by* (30) *such that Assumption 1 is satisfied and, in addition, φ belongs to a Sobolev space $H^{2a+\gamma\vee 1/2}(\mathbb{R}^d)$ Assume further that the approximation error of $\langle \mathbb{E}[Y], \Phi_i\rangle_n$ is asymptotically negligible, i. e.,* (13) *holds.*

(a) The results of Theorem 1a) carry over to this general setting.

(b) Furthermore, let the grids of positions $\mathbf{t}$ and scales $\mathbf{h}$ sufficiently fine, i.e. satisfy (17) *and* (18)*. Then there exist positive constants $\underline{D}_\gamma$ and $\overline{D}_\gamma$ such that for any fixed $\lambda \in \mathbb{R}$*

$$e^{-\underline{D}_\gamma e^{-\lambda}} \leqslant \lim_{n\to\infty} \mathbb{P}_0\big[\mathcal{S}(Y) \leqslant \lambda\big] \leqslant e^{-\overline{D}_\gamma e^{-\lambda}}.$$

Hence, under H_0, $\mathcal{S}(Y)$ is asymptotically non-degenerate.

(c) In the situation of (b), let $\mathbf{h}_i = (h_i, \dots, h_i)$ for all $i \in I_N$ and assume that (16) *holds true with $0 < \delta < \Delta \leqslant 1$. If the stronger condition* (14) *holds, then with the calibration $w(K, 1+d)$ we obtain*

$$\lim_{n\to\infty} \mathbb{P}_0\big(\mathcal{S}(Y) \leqslant \lambda\big) = \exp\left(-\exp(-\lambda) \cdot \frac{H_{2\gamma}\det\big(D_\Xi^{-1}\big)\log(\Delta/\delta)}{\sqrt{2\pi}K}\right),$$

REMARK 4. *(a) In Theorem 4 we need to exclude a small boundary region of the observations from the analysis since, in general, the functions $\Phi_{\mathbf{h}_i}$ in $\mathcal{W}$ might be of unbounded support. Then the results of Theorem 1 transfer to this setting.*

(b) The results from Theorem 4 (c) require assumption (14) *which basically means that the convolution kernel k should decay exactly like a polynomial if $\|\boldsymbol{\xi}\|_2 \to \infty$ in contrast to the weaker assumption* (D1) *which only requires upper and lower polynomial bounds and can hence only ensure upper and lower Gumbel bounds. In Section 4 we provide a specific example for which both* (D1) *and* (14) *are satisfied.*

3.2.1. *Optimal detection in deconvolution.* In this section we discuss and specify the results from Sections 2.3 and 2.4 for deconvolution. The results given in Lemma 1 also hold in the general deconvolution setting. The following lemma contains a related result in the situation of (32) concerning the support inference about the signal f itself.

LEMMA 4. *Given observations from model* (28) *with random noise satisfying Assumption 2 and k as in* (32) *and given a non-negative function $\varphi \in \mathcal{R}(T^*)$, define the dictionary $\mathcal{W}$ as in* (30)*. Suppose that the signal f is non-negative as well. Let further $\mathcal{I}_\alpha(f)$ denote the set*

$$\mathcal{I}_\alpha(f) := \big\{ i \,\big|\, f\big|_{\mathrm{supp}(\varphi_i)} > 2q_{i,1-\alpha}\|\sigma\Phi_i\|_2/(h_{i,1}h_{i,2}n^{d/2}) \big\}.$$

Then, under the assumptions of Theorem 1,

$$\lim_{n\to\infty} \mathbb{P}\big(\langle \Phi_i \,,\, Y\rangle_n > q_{i,1-\alpha}\|\sigma\Phi_i\|_2/n \quad \textit{for all} \quad i \in \mathcal{I}_\alpha(f) \big) \geqslant 1-\alpha.$$

The result above immediately shows that the choice of φ in (29) has a high influence on the detection properties of the corresponding test via the variances $\|\sigma\Phi_i\|_2^2$. Extending an argument from Schmidt-Hieber et al. (2013) for $d = 1$ to general d, we can provide a mother wavelet φ which minimizes the asymptotic variance of the test statistic over all tensor-type probe functions. It only depends on the polynomial order of decay of the convolution kernel in Fourier space ($\hat{=}$ degree of ill-posedness) and is (for $d = 2$) given by

$$\varphi\,(x,y) = x^{\beta_1+1}\,(1-x)^{\beta_1+1}\,y^{\beta_2+1}\,(1-y)^{\beta_2+1}\,\mathbf{1}_{(0,1)}\,(x)\,\mathbf{1}_{(0,1)}\,(y)\,, \tag{31}$$

where the two parameters $\beta_1, \beta_2 \in \mathbb{N}$ equal the polynomial order of decay of the convolution kernel in x and y direction. This choice will be considered in the following.

The previous lemma implies the consistency of the testing procedure for the signal itself, i. e., testing $f = 0$ versus $f > 0$, if the minimal scale satisfies $h_{\min} \gtrsim (\log(n)/n)^{1/(4a+1)}$. Moreover, in the situation of Theorem 5 (c) the optimality results of Section 2.4 carry over to the deconvolution setting. For a comparison consider the rate of estimation of the $2a$-th derivative of a Hölder β function w.r.t L^∞ risk in $d = 1$. We restrict to this case as otherwise the deconvolution is no longer equivalent to estimating derivatives, cf. (33). This is possible with minimax rate $(\log n/n)^{\beta/(2\beta+4a+1)}$, which is attained for $h \sim (\log n/n)^{1/(2\beta+4a+1)}$ (see e. g. Johnstone et al., 2004), i.e. such a function can be distinguished from 0 by means of estimation on a box $[t-h, t]$ as long as it is asymptotically larger than h^β. Posing the same

question to MISCAT, the above result show that for $f\big|_{[t-h,t]} \sim h^\beta$ and $h \sim (\log n/n)^{1/(2\beta+4a+1)}$ it recognizes $[t-h,t]$ as active with (asymptotic) probability $\geqslant 1-\alpha$. Consequently any support points found by estimation will also be found by MISCAT.

4. Simulations and real data applications. In this section we investigate the finite sample properties of the proposed multiscale test. To this end, we apply MISCAT in a 2-dimensional mildly ill-posed deconvolution problem. In Section 4.2 we then analyze experimental STED data to locate single DNA origami in a sample.

Specifying the setting described in Section 3.2 to this situation, the data is given by (28). The convolution kernel k is chosen from the parametric family $\big\{k_{a,b} \;\big|\; a \in \mathbb{N}, b > 0\big\}$ defined in Fourier space via

$$(\mathcal{F}_2 k_{a,b})(\boldsymbol{\xi}) = (1 + b^2 \|\boldsymbol{\xi}\|_2^2)^{-a}, \qquad \boldsymbol{\xi} \in \mathbb{R}^2. \tag{32}$$

Model (32) is a 2-dimensional generalization of the one-dimensional family of auto-convolutions of a scaled version of the density of the Laplace distribution with itself with radially symmetric PSF. For any convolution kernel $k_{a,b}$ Assumption (D1) is obviously satisfied and we obtain

$$\Phi_{\mathbf{h}_i} = \sum_{j=0}^{a} \sum_{k=0}^{j} \binom{a}{j} \binom{j}{k} \Big(\frac{b}{h_{i,1}}\Big)^{2k} \Big(\frac{b}{h_{i,2}}\Big)^{2(j-k)} \partial^{(2k\,,\,2(j-k))} \varphi. \tag{33}$$

Alternatively, the functions $\Phi_{\mathbf{h}_i}$ can be computed by means of the Fourier transform as in (30). However, (33) shows that a compactly supported function φ results in a dictionary $\mathcal{W}$ which consists of compactly supported functions as well. Consequently, the results from Theorem 4 can be obtained even without excluding a small boundary region, and furthermore a Gumbel limit theorem can be obtained as follows. Let

$$\Xi = \frac{\widetilde{\Xi}}{\|\widetilde{\Xi}\|_2}, \quad \text{where} \quad \widetilde{\Xi} = b^{2a} \sum_{k=0}^{a} \binom{a}{k} \partial^{2k,2(a-k)} \varphi. \tag{34}$$

and consider the case $\mathbf{h}_i = (h_i, h_i)$ for all $i \in I_N$. Then

$$\|\Phi_{\mathbf{h}_i}\|_2 = \left(\frac{1}{h_i}\right)^{2a} \|\Xi + h_i^2 \Xi_{n,i}\|_2 \quad \text{and} \quad \frac{\Phi_{\mathbf{h}_i}}{\|\Phi_{\mathbf{h}_i}\|_2} = \frac{\Xi + h_i^2 \Xi_{n,i}}{\|\Xi + h_i^2 \Xi_{n,i}\|_2}, \tag{35}$$

where

$$\Xi_{n,i} := \sum_{j=0}^{a-1} h_i^{2(a-1-j)} \sum_{k=0}^{j} \binom{j}{k} \binom{a}{j} \partial^{2k,2(j-k)} \varphi. \tag{36}$$

In this setting, it is easy to verify that condition (14) holds.

THEOREM 5 (MISCAT for our application).
Suppose that we have access to observations following model (28) *with convolution kernel* $k_{a,b}$ *satisfying Assumption* (32), $d = 2$ *and random noise satisfying Assumption 2. Assume that the dictionary is given by* (30) *with dictionary functions* $\Phi_{\mathbf{h}_i}$ *defined in* (33) *such that Assumption 1 is satisfied, and that* (13) *holds.*

(a) The results of Theorem 1(a) carry over to this particular convolution setting.
(b) If $\varphi \in H^{2a+\gamma\wedge 1/2}\left(\mathbb{R}^2\right)$ *and if the grids of positions* $\mathbf{t}$ *and scales* $\mathbf{h}$ *are sufficiently fine, i.e. satisfy* (17) *and* (18)*, then the results of Theorem 4(b) carry over to this particular convolution setting.*
(c) Suppose furthermore that $\mathbf{h}_i = (h_i, h_i)$ *for all* $i \in I_N$*, that* (16) *holds true with* $0 < \delta < \Delta \leqslant 1$ *and that the grids of positions* $\mathbf{t}$ *and scales* $\mathbf{h}$ *are sufficiently fine, i.e. satisfy* (17) *and* (18)*. If in addition,* φ *is* $(2a+1)$*-times differentiable in* $L^2(\mathbb{R}^d)$*, let* $\varphi_{\boldsymbol{\alpha}} = \sum_{k=0}^{a}\binom{a}{k}\partial^{2k+\alpha_1,2(a-k)+\alpha_2}\varphi$*,* $\boldsymbol{\alpha} \in \{0,1\}^2$*,* $|\boldsymbol{\alpha}| = 1$*. Then, for* $\omega(K, 1+d)$ *with*

$$K = b^{4a}\log(\Delta/\delta)(2\pi)^{-\frac{3}{2}}\|\widetilde{\Xi}\|_2^{-1}\sqrt{\|\varphi_{0,1}\|_2^2\|\varphi_{1,0}\|_2^2 - \langle\varphi_{0,1}\varphi_{1,0}\rangle} \tag{37}$$

(see (8) *and Remark 2(b)), we obtain* $\lim_{n\to\infty}\mathbb{P}_0\big[\mathcal{S}(Y) \leqslant \lambda\big] = e^{-e^{-\lambda}}$.

4.1. 2*-dimensional support inference.* In Supplement A we present detailed simulations to infer on the support of a testfunction of size 512×512, i.e. $n = 512$ with a kernel $k_{a,b}$ as in (32). This setting is close to our subsequent data example. We apply MISCAT using 196 different scales defined by boxes consisting of $k_x \times k_y$ pixels, $k_x, k_y = 4, 6, ..., 30$. Concerning the positions $\mathbf{t}$ we use again all possible upper left points of boxes fitting in the image, which results in 48.219.136 local tests in total. To implement MISCAT, we first fix φ as in (31), and then compute the 196 functions $\Phi_{\mathbf{h}_i}$ as in (30) by using the Fourier convolution theorem, i.e. $\Phi_{\mathbf{h}_i} = \mathcal{F}_d^{-1}\left(\frac{\mathcal{F}_d\varphi}{\mathcal{F}_d k_{a,b}(\cdot/\mathbf{h}_i)}\right)$ as in (30). This can be done explicitly exploiting the structure of $k_{a,b}$, and is efficiently implemented using FFT, which results in $\mathcal{O}\left(196\cdot 512^2\log\left(512^2\right)\right)$ flops. Similarly, all local statistics $\langle Y, \Phi_{i,n}\rangle_n$ with fixed scale $\mathbf{h} \equiv \mathbf{h}_i$ can also be computed by two FFT's using the Fourier convolution theorem. Consequently, MISCAT for deconvolution problems can be performed in general in $\mathcal{O}\left(\#\text{scales}\cdot\#\text{pixels}\log\left(\#\text{pixels}\right)\right)$, and in the setting here the evaluation of the roughly 50 million local test statistics takes less than one minute on a standard laptop. Finally, to perform MISCAT, the (asymptotic) quantiles of the approximating Gaussian test statistic $\mathcal{S}(W)$ can be pre-

computed, which corresponds to many evaluations of the maximum in (10) and is costly.

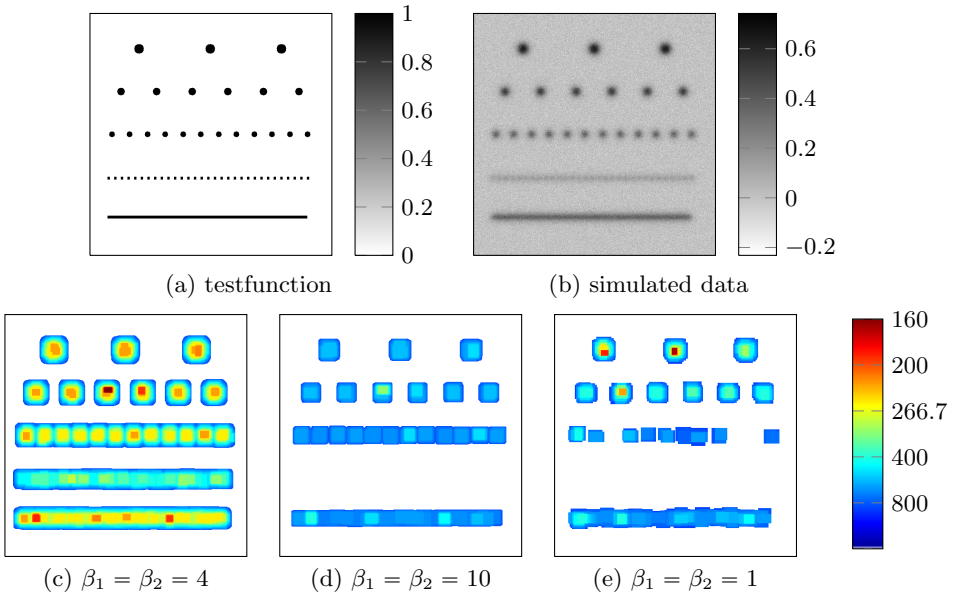
Let us now briefly conclude the findings in Supplement A. First of all, our simulations suggest a non-degenerated behavior of the distribution of the penalized maximum statistic.

Concerning support inference we observe that MISCAT with correctly specified degree of ill-posedness (this is $\beta_1 = \beta_2 = 2a$ in (31) with a as in (32)) is able to detect large objects even in a large noise regime, and for sufficiently small amount of noise it is even able to separate objects which have a distance of 9 pixels, which is even less than the FWHM (see Supplement A). We furthermore find that misspecification of ill-posedness does not provide false detections, but loss in detection power, where underspecification of the ill-posedness ($\beta_1 = \beta_2 = 1$ in (31)) has less severe effects to MISCAT than overspecification ($\beta_1 = \beta_2 = 10$ in (31)). This can also be seen from Figure 3 where a synthetic testfunction, simulated data from a homogeneous Gaussian model with noise level $\sigma = 0.05$ and significance maps of the three corresponding tests (correctly specified, overspecified and underspecified ill-posedness) is shown. The significance map color-codes for each pixel the smallest scale on which it was significant, for details see Supplement A.

Finally we also investigate robustness of MISCAT to the noise distribution in Supplement A. We investigate empirical levels under the student's t distribution $t(\nu)$, which has $\nu - 1$ moments and hence does not satisfy (M1) for any ν. Nevertheless, for sufficiently large parameter ν we still obtain an empirical level close to our theoretical value α. Furthermore we investigate a model mimicking data coming from CCD sensors, which consists of Poissonian observations which are additionally corrupted by additive Gaussian noise. This model satisfies (M1), and our simulations suggest that MISCAT keeps its level quite stable over a large range of parameters. Only in the situation of a low Poisson intensity, the heavier tail behavior of the Poisson distribution dominates and the empirical level deteriorates.

4.2. *Locating fluorescent markers in STED super-resolution microscopy.* Based on the results from Section 4.1, we are now able to rigorously treat the real world application from Section 1.2 from 2-dimensional STED (stimulated emission depletion) super-resolution microscopy (Hell and Wichmann, 1994; Klar and Hell, 1999; Hell, 2007). A brief overview over the experimental setup is already given in the introduction, and for a detailed mathematical model we refer to Supplement A, where we argue there that our measurements are described reasonably by

$$Y_{\mathbf{j}} \overset{\text{independent}}{\sim} \mathrm{Bin}\left(t, \left(k_{2,0.016} * f\right)(\mathbf{x}_{\mathbf{j}})\right), \qquad \mathbf{j} \in \{1, ..., 600\}^2.$$



(a) testfunction (b) simulated data

(c) $\beta_1 = \beta_2 = 4$ (d) $\beta_1 = \beta_2 = 10$ (e) $\beta_1 = \beta_2 = 1$

FIG 3. *(a) Synthetic testfunction, (b) simulated data from a homogeneous Gaussian model with noise level* $\sigma = 0.05$*, (c) 90%-significance map of MISCAT with correctly specified ill-posedness (*$\beta_1 = \beta_2 = 4$ *in* (31)*), (d) 90%-significance map of MISCAT with overspecified ill-posedness (*$\beta_1 = \beta_2 = 10$ *in* (31)*), (e) 90%-significance map of MISCAT with underspecified ill-posedness (*$\beta_1 = \beta_2 = 1$ *in* (31)*). The color-coding shows the smallest scale (in pixels) on which the corresponding pixel was significant.*

Here $\mathrm{Bin}(t, p)$ denotes the Binomial distribution with parameters $t \in \mathbb{N}$ and $p \in [0, 1]$, observations are obtained on the grid $\{\mathbf{x_j} \mid \mathbf{j} \in \{1, ..., 600\}^2\}$ and $f(\mathbf{x})$ is the probability that a photon emitted at grid point $\mathbf{x}$ is recorded at the detector in a single excitation pulse. The kernel $k_{2,0.016}$ is as in (32), and in actual experiments t is roughly 10^3.

With this kernel we design a test using the optimal probe function φ in (31) (i.e. $\beta_1 = \beta_2 = 4$) and a set of scales defined by boxes of size $k_x \times k_y$ pixels, $k_x, k_y = 4, 6, ..., 20$, resulting in 28.100.601 local tests. The variances σ_i^2 in (7) used in the test statistic are estimated from the data point-wise using a maximum likelihood estimator. Furthermore we ease the problem by neglecting all boxes in which no photons where observed, i.e. we drop all pairs $(\mathbf{t}_i, \mathbf{h}_i)$ such that $Y_\mathbf{j} = 0$ for all $\mathbf{x_j} \in [\mathbf{t}_i - \mathbf{h}_i, \mathbf{t}_i]$. Even though this choice is data dependent and hence random, the uniformity over all pairs $(\mathbf{t}_i, \mathbf{h}_i)$ of our confidence statements ensures that those stay valid.

With this test we analyze the data shown in Figure 1, cf. Section 1.2 for details. In total, MISCAT marks 94.141 out of 28.100.601 boxes as signifi-

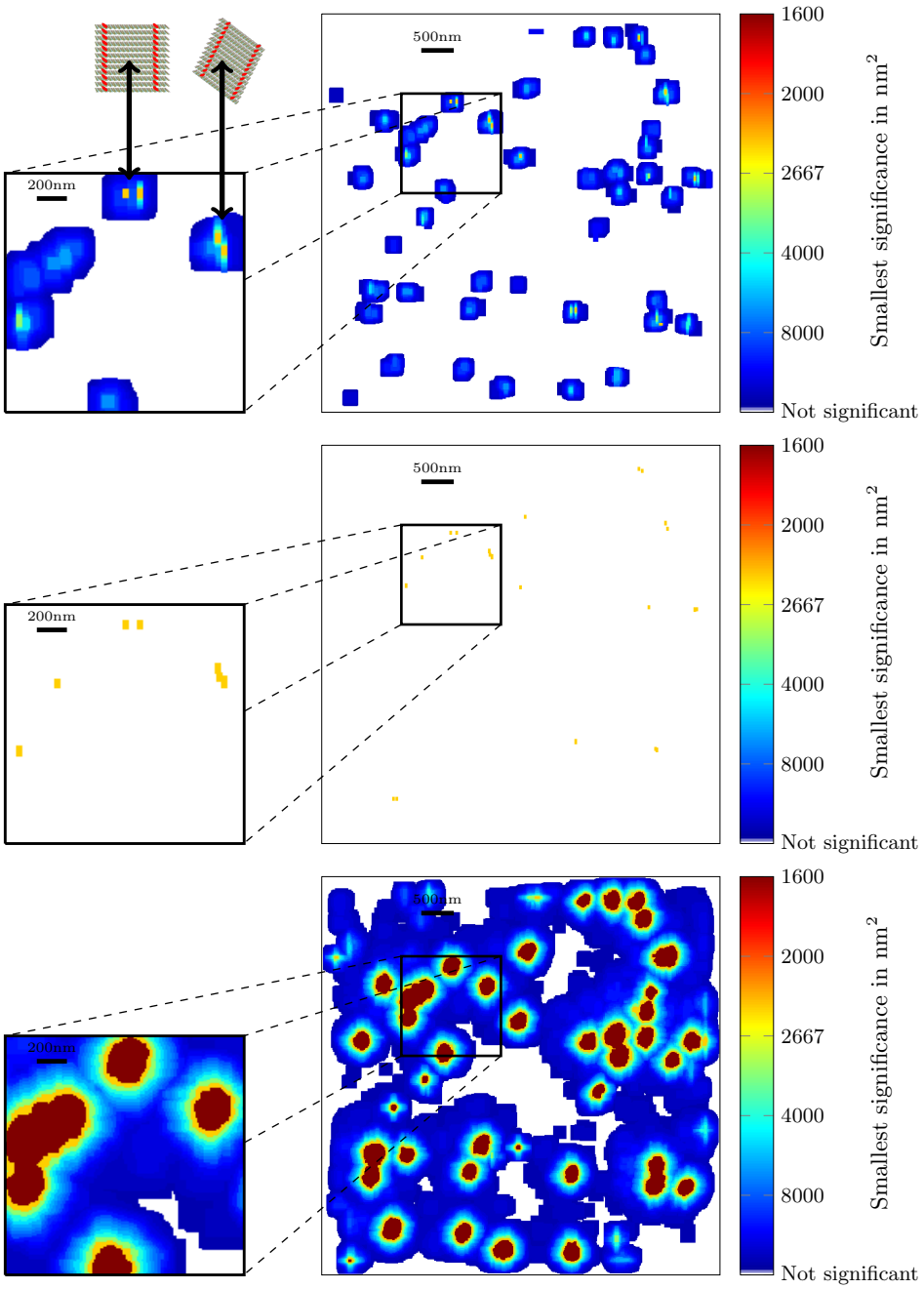


FIG 4. 90% *significance maps and excerpts for different tests computed from the data in Figure 1. From top to bottom: MISCAT, a single scale test with deconvolution, and the standard multiscale test without deconvolution.*

cant. For a comparison, we also use two different tests, namely an analog of MISCAT using only one single scale of size 4×6 pixels (these are the smallest boxes found by MISCAT, see Remark 2(b)), and the multiscale scanning test ignoring the deconvolution ($T = \mathrm{id}$), boiling down to the test statistic of Dümbgen and Spokoiny (2001):

$$\max_i \frac{\sqrt{\log(3/\mathbf{h}_i^{\mathbf{1}})}}{\log\big(\log(3/\mathbf{h}_i^{\mathbf{1}})\big)} \left[\frac{1}{\sqrt{\mathbf{h}_i^{\mathbf{1}}}} \sum_{\mathbf{x_j} \in [\mathbf{t}_i - \mathbf{h}_i, \mathbf{t}_i]} Y_{\mathbf{j}} - \sqrt{2\log(3/\mathbf{h}_i^{\mathbf{1}})} \right].$$

For all tests we again use empirical quantiles computed in 10^4 runs of the test statistics applied to Gaussian white noise.
The full result is depicted in Figure 4. As mentioned in the introduction, MISCAT is able (at least for some of the single DNA origamis) to infer on position and rotation as indicated in the first panel in Figure 4. Remarkably, this information is not visible by eye, cf. Figure 2.

5. Multiscale Extreme Value Theory. In this section we state the results that are the core of the proofs of our theorems from the previous sections. The following theorem guarantees that the Gaussian approximation $\mathcal{S}(W)$ is asymptotically bounded from above almost surely.
Let $Z_{\mathbf{t},\mathbf{h}} := 1/\sqrt{\mathbf{h}^{\mathbf{1}}} \int \Xi\big(\frac{\mathbf{t}-\mathbf{z}}{\mathbf{h}}\big)\, \mathrm{d}W_{\mathbf{z}}$.

THEOREM 6 (MISCAT: a.s. boundedness). *Let Ξ be a normed function, i.e. $\|\Xi\|_2 = 1$, supported on $[0,1]^d$ such that* (AHC) *holds. Let $(\mathbf{t},\mathbf{h}) \in \mathcal{H} \times \mathcal{T}_{\mathbf{h}} \subset [\mathbf{h}_{\min}, \mathbf{h}_{\max}] \times [\mathbf{h}, \mathbf{1}]$, where $h_{\max} \leqslant n^{-\delta}$. There exists a function F which is independent of n, such that $\lim_{\lambda\to\infty} F(\lambda) = 0$ and for $\lambda > 0$*

$$\mathbb{P}\bigg(\sup_{\mathbf{h}\in\mathcal{H}} \sup_{\mathbf{t}\in\mathcal{T}_{\mathbf{h}}} \omega_{\mathbf{h}}\big(Z_{\mathbf{t},\mathbf{h}} - \omega_{\mathbf{h}}\big) > \lambda \big) \leqslant F(\lambda).$$

This implies in particular that $\mathcal{S}(W)$ is almost surely bounded. Furthermore, there exists a positive constant $\underline{D}_\gamma$ such that for any fixed $\lambda \in \mathbb{R}$

$$e^{-\underline{D}_\gamma} e^{-\lambda} \leqslant \lim_{n\to\infty} \mathbb{P}\bigg(\sup_{\mathbf{h}\in\mathcal{H}} \sup_{\mathbf{t}\in\mathcal{T}_{\mathbf{h}}} \omega_{\mathbf{h}}\big(Z_{\mathbf{t},\mathbf{h}} - \omega_{\mathbf{h}}\big) \leqslant \lambda \big).$$

The following theorem yields a weak limit for multiscale statistics of the type $\mathcal{S}(W)$.

THEOREM 7 (A general multiscale Gumbel limit theorem). *Let Ξ be a normed function, i.e. $\|\Xi\|_2 = 1$, supported on $[0,1]^d$ such that* (15) *holds.*

Let $K > 0$ be a fixed, positive constant. If furthermore (16) *holds true with* $0 < \delta < \Delta \leqslant 1$,

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\sup_{\mathbf{h}\in[h_{\min},h_{\max}]^d} \sup_{\mathbf{t}\in[\mathbf{h},\mathbf{1}]} \omega_{\mathbf{h}}\big(Z_{\mathbf{t},\mathbf{h}} - \omega_{\mathbf{h}}\big) \leqslant \lambda \Bigg) = e^{-e^{-\lambda}\cdot \frac{H_{2\gamma}|\det D_{\Xi}^{-1}| I_d(\delta,\Delta)}{\sqrt{2\pi}K}},$$

where $\omega_{\mathbf{h}}$ and $I_d(\delta,\Delta)$ are defined as in (8) *and* (20)*, respectively.*

Corollary 1 below follows immediately from the proofs of the previous Theorems. Two special cases are discussed in Remark 2.

Corollary 1. *Suppose that the assumptions of Theorem 6 hold. Assume that $\mathbf{h}_i \in \mathcal{H}_1 \times \ldots \times \mathcal{H}_d$, where possibly $\mathcal{H}_i \neq \mathcal{H}_j$ for $i \neq j$. Let for $\mathcal{P} := \big\{\lfloor \log(1/h_{\max})\rfloor - 2, \lfloor \log(1/h_{\max})\rfloor - 1, \ldots, \lceil \log(1/h_{\min})\rceil\big\}$, and $j \in \{1,\ldots,d\}$*

$$\mathcal{P}_j := \big\{ p \in \mathcal{P} \,\big|\, \exists\ h_{i,j} \in \mathcal{H}_j :\ h_{i,j} \in \big[e^{-(p+1)}, e^{-p}\big)\big\}.$$

Choose the constant C_d in (8) *such that there exist positive constants $\underline{d}$ and $\overline{D}$ such that*

$$\underline{d} \leqslant \log(n)^{-\frac{C_d - d/\gamma + 1}{2}} \big|\mathcal{P}_1 \times \ldots \times \mathcal{P}_d\big| \leqslant \overline{D}.$$

(a) The results of Theorem 6 remain valid.

(b) If, in addition, the grid of positions $\mathbf{t}$ is sufficiently fine, i.e. (17) *holds and for each j, the sets $\mathcal{P}_j = \mathcal{P}_{j,n}$ are increasing sets with respect to $n \in \mathbb{N}$, i. e., $|\mathcal{P}_{j,n}| \leqslant |\mathcal{P}_{j,n+1}|$ and $\sum_{p_{j,n}\in\mathcal{P}_{j,n}} p_{j,n}$ is increasing, there exists a constant $C_{\mathcal{P}} > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\sup_{\mathbf{h}\in\mathcal{H}_1\times\ldots\times\mathcal{H}_d} \sup_{\mathbf{t}\in\mathcal{T}_{\mathbf{h}}} \omega_{\mathbf{h}}\big(Z_{\mathbf{t},\mathbf{h}} - \omega_{\mathbf{h}}\big) \leqslant \lambda \Bigg) = e^{-e^{-\lambda}\cdot \frac{H_{2\gamma}|\det D_{\Xi}^{-1}| C_{\mathcal{P}}}{\sqrt{2\pi}K}}.$$

6. Proofs. This section contains only the key ideas and main steps of the proof of our key result, Theorem 7. A complete version of this proof can be found in Supplement A.

Proof of Theorem 7. Recall that with a slight abuse of notation, we denote $\mathbf{h}^{\boldsymbol{\alpha}} = h_1^{\alpha_1} \cdot \ldots \cdot h_d^{\alpha_d}$, $\quad \mathbf{h_p} = (h_1,\ldots,h_d)^T$, $\quad \mathbf{1/h_p} = \big(1/h_1,\ldots,1/h_d\big)^T$ and inequalities between vectors or multi-indices are meant component-wise.

Step I: Proof for scales on a dyadic grid.

We show later in Step II.3 that the supremum over $[h_{\min}, h_{\max}]$ and the

supremum over $[h_{\min}/\log(n)^{\frac{1}{\gamma}}\log\log(n), h_{\max}]$ are asymptotically equivalent and consider first the supremum over the slightly enlarged set. Define a dyadic grid $\mathcal{H}_{\text{dyad}} \subset [h_{\min}/\log(n)^{\frac{1}{\gamma}}\log\log(n), h_{\max}]$ as follows:

$$\mathcal{H}_{\text{dyad}} := \{2^{-p} \,|\, p \in \mathcal{P}\}, \quad \mathcal{P} = \big\{\big\lceil b_\gamma(h_{\max})\big\rceil, \dots, \big\lfloor b_\gamma(h_{\min})\big\rfloor\big\},$$

where $b_\gamma(h) := \log\big(\frac{\log(n)^{\frac{1}{\gamma}}\log\log(n)}{h}\big)/\log(2)$. Here and below log denotes the natural logarithm. Define $p_{\min} := \min\mathcal{P}$ and $p_{\max} := \max\mathcal{P}$. Let

$$\begin{aligned} &\max_{\mathbf{h}\in\mathcal{H}^d_{\text{dyad}}} \sup_{\mathbf{t}\in[\mathbf{h},\mathbf{1}]} \omega_{\mathbf{h}}\bigg(\frac{1}{\sqrt{\mathbf{h^1}}}\int \Xi\Big(\frac{\mathbf{t}-\mathbf{z}}{\mathbf{h}}\Big)\,\mathrm{d}W_{\mathbf{z}} - \omega_{\mathbf{h}}\bigg) \\ &\overset{\mathcal{D}}{=} \max_{\mathbf{h}\in\mathcal{H}^d_{\text{dyad}}} \sup_{\mathbf{t}\in[\mathbf{1},\mathbf{1}/\mathbf{h}]} \omega_{\mathbf{h}}\bigg(\int \Xi(\mathbf{t}-\mathbf{z})\,\mathrm{d}W_{\mathbf{z}} - \omega_{\mathbf{h}}\bigg) =: M_n, \end{aligned} \tag{38}$$

by stationarity of $Z_{\mathbf{t},\mathbf{h}}$ for fixed $\mathbf{h}$. We now consider the term M_n.

Step I.1: Partition of the parameter set

The form of M_n in (38) reveals a redundancy pattern that we will exploit later on. Observe that the suprema with respect to $\mathbf{t}$ of the rescaled version M_n are taken over subsets of the rectangle $[\mathbf{1}, \mathbf{1}/\mathbf{h}_{\min}]$. For smaller scales, the supremum with respect to $\mathbf{t}$ is taken over larger sets. Obviously, for $\mathbf{p} \in \mathcal{P}^d$,

$$[\mathbf{1}/\mathbf{h_p}, \mathbf{1}/\mathbf{h_{p+1}}] \subset [\mathbf{1}, \mathbf{1}/\mathbf{h_s}] \quad \forall \quad \mathbf{s} > \mathbf{p}+\mathbf{1}.$$

In order to exploit this fact we partition the parameter set $[\mathbf{1}, \mathbf{1}/\mathbf{h}_{\min}]$ into suitable blocks, i.e. into blocks $B_{\mathbf{p+1},\mathbf{q+1}}$ that are approximately equal to $[\mathbf{1}/\mathbf{h_p}, \mathbf{1}/\mathbf{h_{p+1}}]$ in order to split the suprema with respect to $\mathbf{t}$ into suitable sub-suprema. To achieve that those sub-suprema are independent, we separate the blocks by small bands of width 1. This ensures independence, since $\text{supp}(\Xi) \subset [0,1]^d$. The bands only yield a contribution which is asymptotically negligible, which we will show in Step I.3 below.

To be precise, we define subsets of $[\mathbf{1}, \mathbf{1}/\mathbf{h}_{\min}]$ as follows

$$B_{\mathbf{p}} := \big[\tfrac{\mathbf{1}}{\mathbf{h_{p-1}}}, \tfrac{\mathbf{1}}{\mathbf{h_p}} - \mathbf{1}\big], \text{ and } R_{\mathbf{p}} = \big[\tfrac{\mathbf{1}}{\mathbf{h_{p-1}}}, \tfrac{\mathbf{1}}{\mathbf{h_p}}\big] \backslash B_{\mathbf{p}}, \tag{39}$$

where $\mathbf{h}_{\mathbf{p}_{\min}-1} := \mathbf{1}$. The large blocks $B_{\mathbf{p}}$ yield the main contributions. The sets $R_{\mathbf{p}}$ are asymptotically negligible (see Step I.3 below). Define further for $\mathbf{q} \in \mathcal{P}^d$

$$\mathcal{B}_{\mathbf{q}} := \bigcup_{\mathbf{p}\in\mathcal{P}^d,\, \mathbf{p}\leqslant\mathbf{q}} B_{\mathbf{p}} \quad \text{and} \quad M_{\mathcal{B}} := \max_{\mathbf{p}\in\mathcal{P}^d} \omega_{\mathbf{h_p}}\Big(M_{\mathcal{B}_{\mathbf{p}}} - \omega_{\mathbf{h_p}}\Big).$$

Write

$$M_{\mathcal{B}} = \max_{\mathbf{q}\in\mathcal{P}^d} \max_{\mathbf{p}\leqslant\mathbf{q}} \sup_{\mathbf{t}\in B_{\mathbf{p}}} \omega_{\mathbf{h_q}}\bigg(\int \Xi(\mathbf{t}-\mathbf{z})\,\mathrm{d}W_{\mathbf{z}} - \omega_{\mathbf{h_q}}\bigg)$$
$$= \max_{\mathbf{p}\in\mathcal{P}^d} \max_{\mathbf{q}\geqslant\mathbf{p}} \sup_{\mathbf{t}\in B_{\mathbf{p}}} \omega_{\mathbf{h_q}}\bigg(\int \Xi(\mathbf{t}-\mathbf{z})\,\mathrm{d}W_{\mathbf{z}} - \omega_{\mathbf{h_q}}\bigg).$$

Fix $\lambda \in \mathbb{R}$. Since the blocks $B_{\mathbf{p}}$ are constructed such that the sub-maxima over different blocks are independent, we find

$$\mathbb{P}\big(M_{\mathcal{B}} \leqslant \lambda\big) = \prod_{\mathbf{p}\in\mathcal{P}^d} \mathbb{P}\bigg(\max_{\mathbf{p}\leqslant\mathbf{q}} \sup_{\mathbf{t}\in B_{\mathbf{p}}} \omega_{\mathbf{h_q}}\bigg(\int \Xi(\mathbf{t}-\mathbf{z})\,\mathrm{d}W_{\mathbf{z}} - \omega_{\mathbf{h_q}}\bigg) \leqslant \lambda\bigg)$$
$$= \prod_{\mathbf{p}\in\mathcal{P}^d} \mathbb{P}\bigg(\sup_{\mathbf{t}\in B_{\mathbf{p}}} \int \Xi(\mathbf{t}-\mathbf{z})\,\mathrm{d}W_{\mathbf{z}} \leqslant \Lambda_{\min,\mathbf{p}}\bigg),$$

where $\Lambda_{\min,\mathbf{p}} := \min_{\mathbf{p}\leqslant\mathbf{q}}\big(\frac{\lambda}{\omega_{\mathbf{h_q}}} + \omega_{\mathbf{h_q}}\big)$. Now we have broken the proof down into $|\mathcal{P}|^d$ "one-scale" extreme value problems and use results for those. Let $\mathrm{Leb}(B_{\mathbf{p}})$ denote the Lebesgue-measure of $B_{\mathbf{p}}$ and let $\Lambda_{\mathbf{p}}$ denote

$$\Lambda_{\mathbf{p}} := \lambda/\omega_{\mathbf{h_p}} + \omega_{\mathbf{h_p}}. \tag{40}$$

For any fixed $\lambda \in \mathbb{R}$ we have that $\Lambda_{\min,\mathbf{p}} = \Lambda_{\mathbf{p}}$, for sufficiently large n. Thus,

$$\mathbb{P}\big(M_{\mathcal{B}} \leqslant \lambda\big) = \prod_{\mathbf{p}\in\mathcal{P}^d} \bigg(1 - \mathbb{P}\bigg(\sup_{\mathbf{t}\in B_{\mathbf{p}}} \int \Xi(\mathbf{t}-\mathbf{z})\,\mathrm{d}W_{\mathbf{z}} > \Lambda_{\mathbf{p}}\bigg)\bigg).$$

Step I.2: Derivation of the weak limit on the dyadic grid.
Next, we estimate

$$P_{n,\mathbf{p}}(\lambda) := \mathbb{P}\bigg(\sup_{\mathbf{t}\in B_{\mathbf{p}}} \int \Xi(\mathbf{t}-\mathbf{z})\,\mathrm{d}W_{\mathbf{z}} \leqslant \Lambda_{\mathbf{p}}\bigg)$$

using Theorem 7.2 in Piterbarg (1996). In the supplementary material we give the explicit calculations which show that the following holds:

$$P_n(\lambda) = \exp\bigg(-e^{-\lambda}\frac{H_{2\gamma}|\det D_{\Xi}^{-1}|}{K\sqrt{2\pi}} \sum_{\mathbf{p}\in\mathcal{P}^d} \Big(\log\Big(\frac{K}{\mathbf{h_p^1}}\Big)\Big)^{-d}\bigg)(1+o(1)),$$

for some $\delta_e > 0$. Notice that

$$\sum_{\mathbf{p}\in\mathcal{P}^d} \Big(\log\Big(\frac{K}{\mathbf{h_p^1}}\Big)\Big)^{-d} \sim \int_{[\delta\log(n),\Delta\log(n)]^d} \bigg(\frac{1}{\log(K)+z_1+\cdots+z_d}\bigg)^d \mathrm{d}\mathbf{z} =: I_{n,d}.$$

By induction with respect to $d \in \mathbb{N}$ we show in the supplementary material that

$$\tfrac{(d-1)!}{(-1)^d} I_{n,d} = \log\Big(\prod_{j \text{ even}} (k\delta + (d-k)\Delta)^{\binom{d}{j}} \Big) - \log\Big(\prod_{j \text{ odd}} (k\delta + (d-k)\Delta)^{\binom{d}{j}} \Big).$$

Hence, the statement of the theorem holds true for scales on the dyadic grid.
Step I.3: Negligibility of the remainder terms In the supplementary material, we first show that, asymptotically, the slight enlargement of the domain of the scales from the beginning of Step I does not have an impact. Then, we show that the contribution of the separating regions, $\mathcal{R}_{\mathbf{p}}$, $\mathbf{p} \in \mathcal{P}^d$ are asymptotically negligible.
Step II: Show that the dyadic grid is sufficiently dense
We now show

$$\Delta_{\gamma,n} = \Big| \max_{\mathbf{h} \in \mathcal{H}^d_{\text{dyad}}} \sup_{\mathbf{t} \in \mathcal{T}} \omega_{\mathbf{h}} \big(Z_{\mathbf{t},\mathbf{h}} - \omega_{\mathbf{h}} \big) - \sup_{\mathbf{h} \in [h_{\min}, h_{\max}]^d} \sup_{\mathbf{t} \in \mathcal{T}} \omega_{\mathbf{h}} \big(Z_{\mathbf{t},\mathbf{h}} - \omega_{\mathbf{h}} \big) \Big| = o_{\mathbb{P}}(1).$$

Let $\varepsilon > 0$.

$$\mathbb{P}\big(\Delta_{n,\gamma} > \varepsilon \big) \leqslant \mathbb{P}\Big(\max_{\mathbf{p} \in \mathcal{P}^d} \Big| \omega_{\mathbf{h_p}} \Big(\sup_{\mathbf{t} \in \mathcal{T}} \big| Z_{\mathbf{t},\mathbf{h_p}} - Z_{\mathbf{t},\mathbf{h}} \big| + \max_{\mathbf{h} \in [\mathbf{h_p}, \mathbf{h_{p+1}}]} \big| \omega_{\mathbf{h}} - \omega_{\mathbf{h_p}} \big| \Big) \Big| > \varepsilon \Big).$$

Step II.1: Fineness of the dyadic grid I
Let $h \in [h_{\min}, h_{\max}]$. Set $p = \lfloor \log(\log(n)^{\frac{1}{\gamma}} \log\log(n)/h) \rfloor$ and assign the element h_{dyad} of the dyadic grid to h:

$$h_{\text{dyad}} = \operatorname{argmin}\big\{ |g - h| \,|\, g \in \{2^{-p}, \dots, 2^{-p_{\min}}\} \big\}. \tag{41}$$

We show in the supplementary material that $|\omega_{\mathbf{h}} - \omega_{\mathbf{h}_{\text{dyad}}}| = o(1/\sqrt{\log(n)})$.
Step II.2: Estimation of the covering numbers.
We show in detail in the supplementary material that there exists a constant C_{cov}, depending only on the dimension d and the function Ξ via the constants L_Ξ and γ from condition (AHC) such that for $\varepsilon \in (0, d)$,

$$\mathcal{N}\big(\mathcal{T} \times \mathcal{H}, \rho, \varepsilon \big) \leqslant C_{\text{cov}} \left(\frac{1}{\varepsilon} \right)^{\frac{2d}{\gamma}} \left(\frac{1}{h_{\min}} - \frac{1}{h_{\max}} \right)^d, \tag{42}$$

where $\rho^2\big((\mathbf{t},\mathbf{h}), (\mathbf{s},\mathbf{l}) \big) = \mathbb{E}|Z_{\mathbf{t},\mathbf{h}} - Z_{\mathbf{s},\mathbf{l}}|^2$ and $\mathcal{N}\big(\mathcal{T} \times \mathcal{H}, \rho, \varepsilon \big)$ denotes the covering numbers of $\mathcal{T} \times \mathcal{H}$ with respect to ρ.
Step II.3: Fineness of the dyadic grid II.
By an application of Dudley's theorem, using the estimates from Step II.2, we show in the supplementary material that

$$\max_{\mathbf{p} \in \mathcal{P}^d} \omega_{\mathbf{h_p}} \Big| \sup_{\mathbf{t} \in \mathcal{T}} Z_{\mathbf{t},\mathbf{h_p}} - \sup_{\mathbf{h} \in [\mathbf{h_p}, \mathbf{h_{p+1}}]} \sup_{\mathbf{t} \in \mathcal{T}} Z_{\mathbf{t},\mathbf{h}} \Big| = o(1).$$

Hence, the supremum over the dyadic grid and the supremum over the full range $[h_{\min}, h_{\max}]^d$ have the same limit. □

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to 'Multiscale scanning in inverse problems'

(doi: TBA; .pdf). This supplementary material contains an explanation of the full width at half maximum (FWHM), a detailed mathematical model for superresolution STED microscopy, a detailed simulation study for 2-dimensional support inference, and detailed proofs of all theoretical results provided in the main document.

References.

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85:115–129.

Albani, V., Elbau, P., de Hoop, M. V., and Scherzer, O. (2016). Optimal convergence rates results for linear inverse problems in Hilbert spaces. *Numer. Funct. Anal. Optim.*, 37(5):521–540.

Anderssen, R. S. (1986). The linear functional strategy for improperly posed problems. In *Inverse Problems*, pages 11–30. Springer.

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(7):2402–2425.

Aspelmeier, T., Egner, A., and Munk, A. (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annu. Rev. Stat. Appl.*, 2:163–202.

Bertero, M., Boccacci, P., Desiderà, G., and Vicidomini, G. (2009). Image deblurring with Poisson data: from cells to galaxies. *Inverse Probl.*, 25(12):025004, 18.

Bissantz, N., Claeskens, G., Holzmann, H., and Munk, A. (2009). Testing for lack of fit in inverse regression—with applications to biophotonic imaging. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(1):25–48.

Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6):2610–2636.

Burger, M., Flemming, J., and Hofmann, B. (2013). Convergence rates in regularization if the sparsity assumption fails. *Inverse Probl.*, 29(2):025013.

Butucea, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations. *Ann. Statist.*, 35(5):1907–1930.

Butucea, C. and Comte, F. (2009). Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli*, 15(1):69–98.

Butucea, C. and Ingster, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688.

Castillo, I. and Nickl, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.*, 42(5):1941–1969.

Cavalier, L. and Golubev, Y. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.*, 34(4):1653–1677.

Cavalier, L., Golubev, Y., Lepski, O., and Tsybakov, A. (2003). Block thresholding and sharp adaptive estimation in severely ill-posed inverse problems. *Teor. Veroyatnost. i Primenen.*, 48(3):534–556.

Cavalier, L. and Tsybakov, A. (2002). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statist. Sinica*, 23(1):409–428.

Chernousova, E. and Golubev, Y. (2014). Spectral cut-off regularizations for ill-posed linear models. *Math. Methods Statist.*, 23(2):116–131.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42:1564–1597.

Cohen, A., Hoffmann, M., and Reiß, M. (2004). Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, 42(4):1479–1501.

Dedecker, J., Merlevède, F., and Rio, E. (2014). Strong approximation of the empirical distribution function for absolutely regular sequences in $\mathbb{R}^d$. *Electron. J. Probab.*, 19(9):1–56.

Dickhaus, T. (2014). *Simultaneous statistical inference*. Springer, Heidelberg. With applications in the life sciences.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2(2):101–126.

Dümbgen, L. and Spokoiny, V. (2001). Multiscale testing of qualitative hypotheses. *The Annals of Statistics*, 29(1):124–152.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Ann. Statist.*, 36(4):1758–1785.

Eckle, K., Bissantz, N., and Dette, H. (2016). Multiscale inference for multivariate deconvolution. arXiv:1611.05201.

Eckle, K., Bissantz, N., Dette, H., Proksch, K., and Einecke, S. (2017+). Multiscale inference for a multivariate density with applications to x-ray astronomy. To appear in: Annals of the Institute of Statistical Mathematics; DOI 10.1007/s10463-017-0605-1.

Fan, J. (1991). Asymptotic normality for deconvolution kernel density estimators. *Sankhyā Ser. A*, 53(1):97–110.

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(502):456–468.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012). The geometry of nonparametric filament estimation. *J. Amer. Statist. Assoc.*, 107(498):788–799.

Goldenshluger, A. (1999). On pointwise adaptive nonparametric deconvolution. *Bernoulli*, 5(5):907–925.

Hell, S. (2007). Far-field optical nanoscopy. *Science*, 316:1153 – 1158.

Hell, S. W. and Wichmann, J. (1994). Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Opt. Lett.*, 19(11):780–782.

Hohage, T. and Werner, F. (2016). Inverse problems with poisson data: statistical regularization theory, applications and algorithms. *Inverse Probl.*, 32:093001, 56.

Holzmann, H., Bissantz, N., and Munk, A. (2007). Density testing in a contaminated sample. *J. Multivariate Anal.*, 98(1):57–75.

Ingster, Y., Laurent, B., and Marteau, C. (2014). Signal detection for inverse problems in a multidimensional framework. *Math. Methods Statist.*, 23(4):279–305.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I-III. *Math. Methods Statist.*, 2:85–114, 171–189, 249–268.

Ingster, Y. I., Sapatinas, T., and Suslina, I. A. (2012). Minimax signal detection in ill-posed inverse problems. *Ann. Statist.*, 40(3):1524–1549.

Johnstone, I. M., Kerkyacharian, G., Picard, D., and Raimondo, M. (2004). Wavelet

deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(3):547–573.

Johnstone, I. M. and Paul, D. (2014). Adaptation in some linear inverse problems. *Stat*, 3(1):187–199.

Johnstone, I. M. and Silverman, B. W. (1991). Discretization effects in statistical inverse problems. *J. Complexity*, 7:1–34.

Kabluchko, Z. (2011). Extremes of the standardized Gaussian noise. *Stochastic Process. Appl.*, 121(3):515–533.

Kazantsev, I., Lemahieu, I., Salov, G., and Denys, R. (2002). Statistical detection of defects in radiographic images in nondestructive testing. *Signal Processing*, 82(5):791 – 801.

Kerkyacharian, G., Kyriazis, G., Le Pennec, E., Petrushev, P., and Picard, D. (2010). Inversion of noisy Radon transform by SVD based needlets. *Appl. Comput. Harmon. Anal.*, 28(1):24–45.

Klar, T. A. and Hell, S. W. (1999). Subdiffraction resolution in far-field fluorescence microscopy. *Opt. Lett.*, 24(14):954–956.

Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, 39(5):2626–2657.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 32:111–131.

Kou, J. (2017). Identifying the support of rectangular signals in gaussian noise. arXiv preprint 1703.06226.

Laurent, B., Loubes, J.-M., and Marteau, C. (2011). Testing inverse problems: a direct or an indirect problem? *J. Statist. Plann. Inference*, 141(5):1849–1861.

Laurent, B., Loubes, J.-M., and Marteau, C. (2012). Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electron. J. Stat.*, 6:91–122.

Li, H., Munk, A., Sieling, H., and Walther, G. (2016). The essential histogram. arXiv:1612.07216.

Lin, G. D. (2017). Recent developments on the moment problem. *arXiv 1703:01027*.

Mair, B. A. and Ruymgaart, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56(5):1424–1444.

Marteau, C. and Mathé, P. (2014). General regularization schemes for signal detection in inverse problems. *Math. Methods Statist.*, 23(3):176–200.

Mathé, P. and Pereverzev, S. V. (2002). Direct estimation of linear functionals from indirect noisy observations. *J. Complexity*, 18(2):500–516.

Meister, A. (2009). *Deconvolution problems in nonparametric statistics*, volume 193 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin.

Natterer, F. (1986). *The mathematics of computerized tomography*. B. G. Teubner, Stuttgart; John Wiley & Sons, Ltd., Chichester.

Nickl, R. and Reiß, M. (2012). A Donsker theorem for Lévy measures. *J. Funct. Anal.*, 263(10):3306–3332.

Nikol'skiĭ, S. M. (1951). Inequalities for entire functions of finite degree and their application in the theory of differentiable functions of several variables. In *Trudy Mat. Inst. Steklov., v. 38*, pages 244–278. Izdat. Akad. Nauk SSSR, Moscow.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1(4):502–527.

Pickands, III, J. (1969). Upcrossing probabilities for stationary Gaussian processes. *Trans. Amer. Math. Soc.*, 145:51–73.

Piterbarg, V. I. (1996). *Asymptotic methods in the theory of Gaussian processes and fields*,

volume 148 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI.

Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.*, 7:2516–2549.

Ray, K. (2017+). Adaptive bernstein-von mises theorems in gaussian white noise. *Ann. Statist.*

Rio, E. (1993). Strong approximation for set-indexed partial-sum processes, via KMT constructions. II. *Ann. Probab.*, 21(3):1706–1727.

Rohde, A. (2008). Adaptive goodness-of-fit tests based on signed ranks. *Ann. Statist.*, 36(3):1346–1374.

Rufibach, K. and Walther, G. (2010). The block criterion for multiscale inference about a density, with applications to other multiscale problems. *J. Comput. Graph. Statist.*, 19(1):175–190.

Schmidt-Hieber, J., Munk, A., and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: Confidence statements for qualitative features. *Ann. Statist.*, 41(3):1299–1328.

Schwartzman, A., Dougherty, R. F., and Taylor, J. E. (2008). False discovery rate analysis of brain diffusion direction maps. *Ann. Appl. Stat.*, 2(1):153–175.

Sharpnack, J. and Arias-Castro, E. (2016). Exact asymptotics for the scan statistic and fast alternatives. *Electron. J. Stat.*, 10(2):2641–2684.

Söhl, J. and Trabs, M. (2012). A uniform central limit theorem and efficiency for deconvolution estimators. *Electron. J. Stat.*, 6:2486–2518.

Ta, H., Keller, J., Haltmeier, M., Saka, S. K., Schmied, J., Opazo, F., Tinnefeld, P., Munk, A., and Hell, S. W. (2015). Mapping molecules in scanning far-field fluorescence nanoscopy. *Nat. Commun.*, 6:7977.

Tsybakov, A. (2000). On the best rate of adaptive estimation in some inverse problems. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(9):835–840.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York.

Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.*, 38(2):1010–1033.

Willer, T. (2009). Optimal bounds for inverse problems with Jacobi-type eigenfunctions. *Statist. Sinica*, 19(2):785–800.

Am Fassberg 11
37077 Göttingen
E-mail: Frank.Werner@mpibpc.mpg.de

Goldschmidtstrasse 7
37077 Göttingen
E-mail: kproksc@uni-goettingen.de; munk@math.uni-goettingen.de