

Chapter 11

Photonic Imaging with Statistical Guarantees: From Multiscale Testing to Multiscale Estimation



Axel Munk, Katharina Proksch, Housen Li and Frank Werner

Abstract In this chapter we discuss how to obtain statistical guarantees in photonic imaging. We start with an introduction to hypothesis testing in the context of imaging, more precisely we describe how to test if there is signal in a specific region of interest (RoI) or just noise. Afterwards we extend this approach to a family of RoIs and examine the occurring problems such as inflation of type I error and dependency issues. We discuss how to control the family-wise error rate by different modifications, and provide a connection to extreme value theory. Afterwards we present possible extension to inverse problems. Moving from testing to estimation, we finally introduce a method which constructs an estimator of the desired quantity of interest with automatic smoothness guarantees.

2010 Mathematics Subject Classification: Primary 62-01, 62G10 · Secondary 62G20, 62F17

A. Munk (✉) · K. Proksch · H. Li · F. Werner
Institute for Mathematical Stochastics, Universität Göttingen,
Goldschmidtstr. 7, 37077 Göttingen, Germany
e-mail: munk@math.uni-goettingen.de

K. Proksch
e-mail: kproksc@uni-goettingen.de

H. Li
e-mail: housen.li@mathematik.uni-goettingen.de

F. Werner
e-mail: f.werner@math.uni-goettingen.de

A. Munk · F. Werner
Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen, Germany

11.1 Introduction

The analysis of a photonic image typically involves a reconstruction of the measured object of interest which becomes the subject of further evaluation. This approach is frequently employed in photonic image analysis, though it can be quite problematic for several reasons.

1. As the image is noisy and often inherently random, a full reconstruction relies on the choice of a regularisation functional and corresponding a priori assumptions on the image, often implicitly hidden in a reconstruction algorithm. Related to this, the reconstruction relies on the choice of one or several tuning parameters. A proper choice is a sensible task, in particular when the noise-level is high and/or inhomogeneous.
2. The sizes of the objects might be below the resolution of the optical device which further hinders a full reconstruction.
3. As the resolution increases, the object to be recovered becomes random in itself as its fine structure then depends on, e.g., the conformational states of a protein and the interpretation of the recovered object might be an issue.

It is the aim of this chapter to provide a careful discussion of such issues and to address the analysis of photonic images with statistical guarantees. This will be done in two steps. In Sect. 11.2 we survey some recent methodology, which circumvents a full recovery of the image, to extract certain relevant information in such difficult situations mentioned above. Based on this (see Sect. 11.3), we will extend such methods also to situations in which a full reconstruction is reasonable, but still a difficult task, e.g., when the multiscale nature of the object has to be recovered. In both scenarios we will put a particular emphasis on statistical guarantees for the provided methods.

An example where a full recovery of the object of interest is typically not a valid task is depicted in the centre of Fig. 11.1 where a detail of a much larger image is shown (see Fig. 1 in [1] for the full image). The investigated specimen consists of DNA origami which have been designed in such a way that each of the signal clusters contains up to 24 fluorescent markers, arrayed in two strands of up to 12, having a distance of 71 nanometers (nm) (see left panel of Fig. 11.1 for a sketch of such a DNA origami). As the ground truth is basically known, this serves as a real world phantom.

Data were recorded with a STED (STimulated Emission Depletion) microscope at the lab of Stefan Hell of the Department of NanoBiophotonics of the Max Planck Institute for Biophysical Chemistry. In contrast to classical fluorescence microscopy, the resolution in STED microscopy is in theory not limited and can be enhanced by increasing the intensity of the depletion laser [2]. However, this increase comes at the price of a decrease in intensity of the focal spot, which bounds the resolution in practice. Therefore a convolution of the underlying signal with the PSF of the STED microscope is unavoidable and a full reconstruction of the DNA origami (or the shape of the markers) appears to be difficult. However, for most purposes this is also

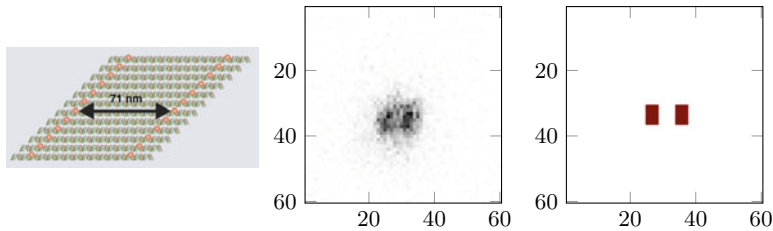


Fig. 11.1 (Detail of Fig. 1 in [1]) Left: Sketch of single DNA origami, middle: detail of image of randomly distributed DNA origami, right: detected strands of markers

not relevant. Instead, less ambitious tasks will provide still important information, e.g., the location of these fluorescent markers. This can be done via a statistical test, which is presumably a much simpler task than reconstruction (estimation in statistical terminology) and it can be tailored towards answering particular questions “How many strands of markers are there?” and “Where are the DNA origamis located?”. The right panel of Fig. 11.1 shows the locations of markers as found by such a statistical test (from the data in the middle panel in Fig. 11.1) which will be introduced later on.

11.2 Statistical Hypothesis Testing

11.2.1 Introduction

We will see that proper testing in the above example (Fig. 11.1) is already a complex task. Therefore, in this section, we first introduce the concept of statistical testing in a basic setting. The first step in statistical hypothesis testing is to define the so-called *null hypothesis*, H , and the *alternative hypothesis*, K :

$$\begin{aligned} H &: \text{“Hypothesis to be disproven”} \\ K &: \text{“Hypothesis to be substantiated”}. \end{aligned}$$

For example, H might correspond to the hypothesis that no marker is contained in a certain given region of the image, K corresponds to the contrary that there is at least one marker in this region. A *statistical significance test* is a decision rule which, based on given data, allows to discriminate between H and K . If a certain criterion is met, H is rejected and K is assumed. If not, H cannot be rejected. For instance, the photon count in a certain given region of a noisy image gives rise to the belief that at least one marker is contained therein. This could be tested, for example by checking whether the total number of photons detected in this region is larger than a certain threshold. However, due to the involved randomness of photon

emissions and background noise such a finding is associated with a (certain) risk of being incorrect. A statistical test aims to control this risk. Hence, prior to performing a statistical test, a tolerable risk α is specified, typically in the range of 0.01 up to 0.1, corresponding to accepting the error rate that, on average, in at most $\alpha \cdot 100\%$ of the cases the null-hypothesis H is falsely rejected. Such an α is called *significance level*. This is written as

$$\mathbb{P} ("H \text{ is rejected although } H \text{ is true}") \leq \alpha. \quad (11.1)$$

Here, \mathbb{P} stands generically for all possible distributions under H and $\mathbb{P}(A)$ denotes the probability¹ of an event A . If the test criterion is chosen such that (11.1) holds, the corresponding test is called a *level- α -test*. The ability of a test to correctly reject H is called *detection power*. If H corresponds to the hypothesis that no marker is located in a certain given region, the test (i.e., the data based decision procedure) is then constructed in such a way that the probability α to falsely detect a marker in an empty region is controlled. H and K are chosen in such a way that the false rejection of H is to be considered the more serious error and controlled in advance. In our scenario, this means that we consider wrong detection of a fluorophore as the more serious error than missing a fluorophore.

11.2.2 A Simple Example

To demonstrate this concept more rigorously, we now consider a very simple Gaussian model, which can be seen as a proxy for more complicated models. Assume that one observes data

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (11.2)$$

where $\mu_i \geq 0$ denote possible “signals” hidden in observations Y_i , and $\varepsilon_i \sim \mathcal{N}(0, 1)$ are independent normal random variables with variances $\sigma^2 = 1$ (for simplicity). Assume for the moment that all signals have the same strength, $\mu_i \equiv \mu \geq 0$. The interest lies in establishing that $\mu > 0$, i.e., presence of such signal in the data. Hence, we set

$$H : \mu = 0 \text{ (to be disproven)} \quad \text{vs.} \quad K : \mu > 0 \text{ (to be substantiated)}. \quad (11.3)$$

The goal is now to find a suitable criterion which, given Y_1, \dots, Y_n , allows to decide in favour or against H in such a way that the error to wrongly reject H is controlled by α . From a statistical perspective the aim is to infer about the mean of

¹More formally, (11.1) is meant as $\mathbb{P} ("H \text{ is rejected although it holds"}) \leq \alpha$ under all possible configurations under H . Only where necessary this will be made explicit in the following by an additional subscript.

the Y_i which should be close to the empirical mean $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ of the data. An intuitive decision rule would be to check whether \bar{Y} is “clearly” larger than zero, $\bar{Y} > \gamma_\alpha$, say, for a suitable threshold $\gamma_\alpha > 0$. We consider the *normalized* (i.e. with unit variance) sum

$$T(Y) := \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$$

and choose, for prescribed $\alpha > 0$, the threshold γ_α such that we have equality in (11.1). As under the assumption H we have that $\mu = 0$, this gives²

$$\begin{aligned} \mathbb{P}(H \text{ is falsely rejected}) &= \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \geq \gamma_\alpha\right) = \mathbb{P}(\mathcal{N}(0, 1) \geq \gamma_\alpha) \\ &= 1 - \Phi(\gamma_\alpha) \stackrel{!}{=} \alpha, \end{aligned} \quad (11.4)$$

since $\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \sim \mathcal{N}(0, 1)$. Here Φ denotes the cumulative distribution function of a standard normal random variable: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy$. If H holds true, i.e., $\mu = 0$, (11.4) holds if we choose $\gamma_\alpha = z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of the standard normal distribution, e.g., $z_{1-\alpha} = 1.6449$, when $\alpha = 0.05$. The statistical test that rejects H whenever $T(Y) > z_{1-\alpha}$ is called *Z-test* and is a level- α test. Furthermore, if a signal is present, i.e., $\mu_i \equiv \mu > 0$ we have that

$$\begin{aligned} \mathbb{P}_{\mu_i \equiv \mu}(H \text{ is correctly rejected}) &= \mathbb{P}_{\mu_i \equiv \mu}\left(\frac{1}{n} \sum_{i=1}^n (\mu + \varepsilon_i) \geq \gamma_\alpha\right) \\ &= 1 - \mathbb{P}(\mathcal{N}(0, 1) > \mu\sqrt{n} - z_{1-\alpha}) = 1 - \Phi(\mu\sqrt{n} - z_{1-\alpha}). \end{aligned}$$

Since $1 - \Phi(x) \leq \exp(-\frac{1}{2}x^2)$ for $x \geq 1/\sqrt{2\pi}$ (see, e.g., [3], inequality (1.8)), we obtain

$$\begin{aligned} \mathbb{P}_{\mu_i \equiv \mu}(H \text{ is correctly rejected}) &\geq 1 - \exp\left(-\frac{1}{2}n\left(\mu - \frac{z_{1-\alpha}}{\sqrt{n}}\right)^2\right) \\ &\geq 1 - \exp\left(-\frac{1}{4}n\mu^2\right), \end{aligned}$$

for sufficiently large n . This means that, if the number n of data points grows, the *detection power* (the case when $\mu > 0$) of the Z-test converges to 1 exponentially fast. This test has been derived in an intuitive way but it can be proven that it is a *uniformly most powerful* (UMP) test (see [4], Chap. 3.4). This means that for all $\mu > 0$ (i.e. the alternative K holds) the detection power is maximized among all

²Here \mathbb{P} corresponds to only one configuration of distributions when all $\mu_i \equiv \mu = 0$.

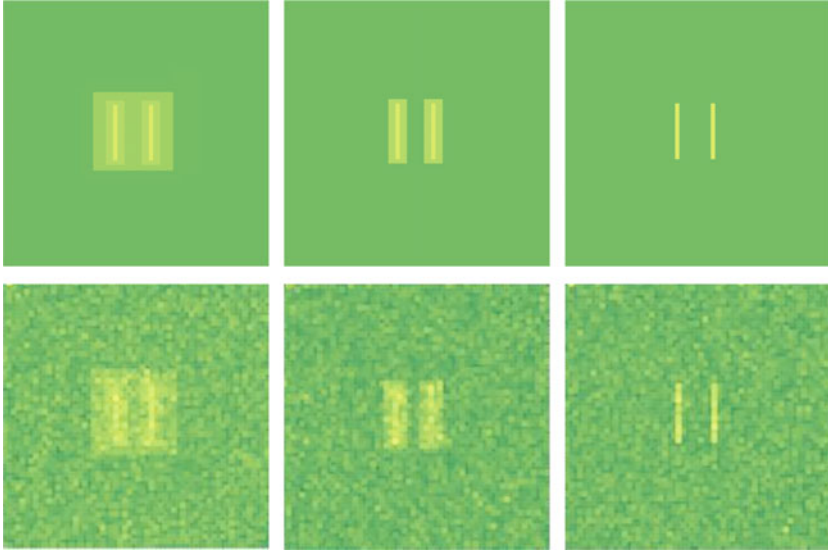


Fig. 11.2 Three different signals (upper row) and noisy signals (lower row)

level- α -tests, i.e., all possible decision rules one might think of which satisfy (11.1) in our set up based on the data Y_1, \dots, Y_n .

Z-test

Comparison of the normalized empirical mean of the set of measurements to a given threshold to assess difference in location to a given constant μ_0 . When $\mu_0 = 0$ the Z-test rejects $H : \mu = \mu_0 = 0$ in favor of $K : \mu > 0$ if

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i > z_{1-\alpha}.$$

This is the best possible test at level α if the data Y_1, \dots, Y_n are independent and $\mathcal{N}(\mu, 1)$ distributed.

11.2.3 Testing on an Image

Subsequently, we consider three illustrative synthetic images of size 60×60 , shown in Fig. 11.2 (see the upper panel for a noise-less version and the lower panel for

a noisy image). These serve the purpose of explaining how to extend the above simple Z-test to detect a signal in an image, which is a more complex task. To illustrate, we assume for the moment that in these images the intensity on each pixel $Y_i, i = 1, \dots, n$, follows a $\mathcal{N}(\mu_i, 1)$ distribution, where each μ_i takes one of the four values 0, 2, 3.5 and 5 (see Fig. 11.2). Now, our goal is to segment the image into regions with signal and empty regions while maintaining statistical error guarantees. Note that we do not aim to recover the exact value of each μ_i , only whether it is positive or not (no signal). To this end we will perform many “local” statistical Z-tests on different (and possibly overlapping) regions of this image. We will discuss several approaches (Scenarios 1–5) which provide a step-by-step derivation of our final solution (Scenario 5). As it turns out, the crucial issue will be to control the statistical error of wrong decisions of all these tests simultaneously (overall error).

Scenario 1 (Known position, one test for central 20×20 square) *Assume for now that we are only interested whether there is some signal in the central 20×20 square (framed in blue in the upper row of Fig. 11.3), i.e. we fix the location to be investigated. For this task, we now perform a Z-test at level $\alpha = 0.05$ for the central square with $n = 20 \times 20 = 400$ pixels, i.e., the test statistic*

$$T_{\text{central } 20 \times 20 \text{ square}}(Y) := \frac{1}{20} \sum_{\text{central } 20 \times 20 \text{ square}} Y_i \quad (11.5)$$

is compared to $z_{1-\alpha} = 1.6449$. The test allows for exactly two outcomes: rejection (of the hypothesis H : no signal in the 20×20 square) or no rejection. In the second row of Fig. 11.3 the results are depicted. In each of the three test images, the Z-test correctly recognizes that there is signal in the central square, and to visualize this, the square is marked in green. The test decision is correct, however, we cannot draw more (localized) information from this test. Nevertheless, this gives us a first guide how to obtain a segmentation into regions, our final task. Note, that the Z-test, as we derived it in Sect. 11.2.2, is still applicable although we did not assume the alternative that all signals have the same strengths (recall Sect. 11.2.2). This will only affect the power. Crucial is that the test controls the error at level α correctly under the assumption that all signals $\mu_i = 0$.

Given a region of interest (RoI), performing one test on the whole region, as done in the previous scenario, only allows to infer on the entire RoI, i.e., the largest scale there is, finer details cannot be discerned. In the following step we consider the finest possible scales, i.e., tests on single pixels, hoping that we can extract more detailed information on different parts of the image, simultaneously.

Scenario 2 (Known position, pixel-wise tests in 20×20 square) *Assume again that we are only interested in testing within the central 20×20 square. We now perform*

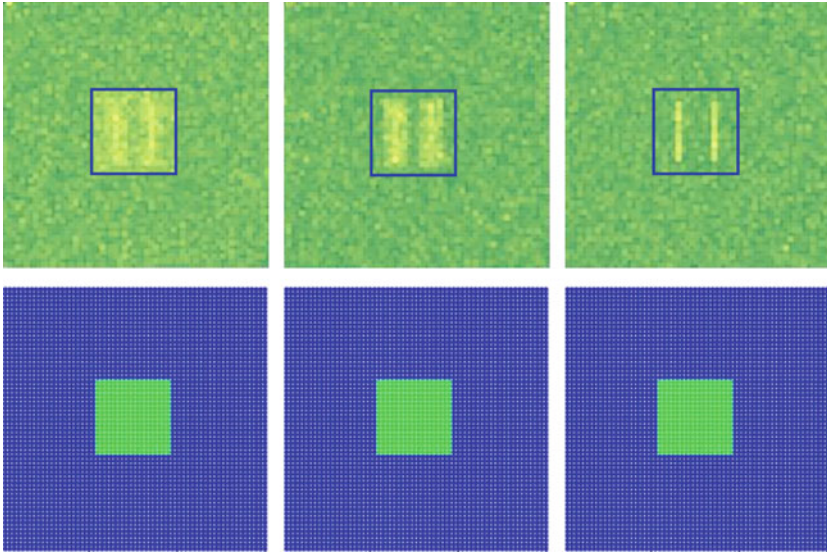


Fig. 11.3 Noisy signals (upper row) and test results from Scenario 1 (lower row). The square is marked in green to show that the test was significant for all three images

a test for each entry in the 20×20 central region separately, in total 400 tests. The test statistics $T_{i-\text{th pixel}}(Y)$ are given by the pixel values. For simplicity, we consider tests for the presence of a signal at pixel “ i ” which are only based on the observation Y_i at pixel i , i.e.

$$T_{i-\text{th pixel}}(Y) = Y_i, \quad (11.6)$$

and are compared to $z_{1-\alpha} = 1.6449$. Again, each test allows for two outcomes: rejection or no rejection. In the second row of Fig. 11.4, exemplary results are depicted (all pixels, for which positive test decisions have been made are marked green).

It is obvious that Scenario 2 gives more detailed information on the signal, but at the expense of several false detections. This is an important issue and will be discussed in more detail in the following section. It is also obvious that parts of the weak signal are missed (see Fig. 11.4: Only 71.25% of the active pixels are detected in the left test image and 85% in the second one). This is due to the fact that the local tests do not take into account neighboring information (surrounding data) from which they could borrow detection strength. This will also be refined in the subsequent sections.

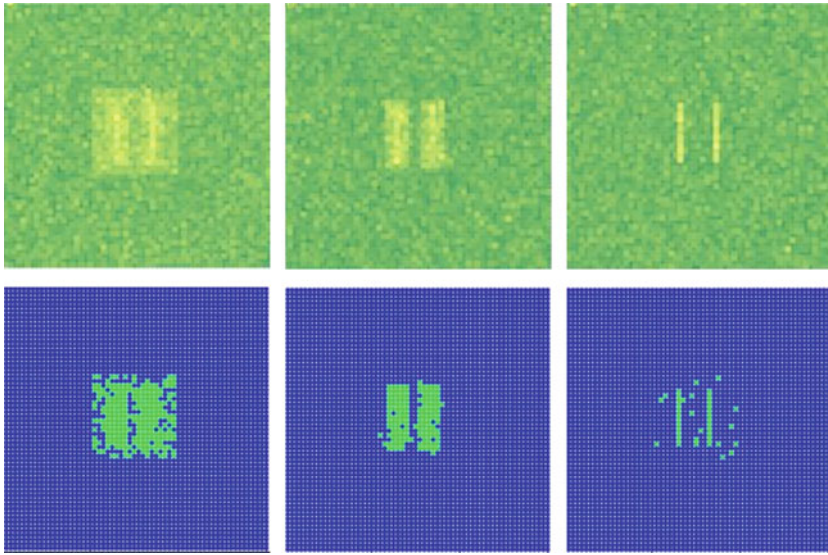


Fig. 11.4 Noisy signals (upper row) and the corresponding test results from Scenario 2 (lower row). The significant pixels are marked green, insignificant pixels are blue

False test decisions

There are two kinds of possible false test decisions:

1. Type I error (probability of its occurrence is controlled by α).
Here: Selection of a RoI although it does not contain any signal (see lower right panel of Fig. 11.4).
2. Type II error (a missed rejection, not controlled).
Here: Missing to select a RoI that contains signal (see lower left panel of Fig. 11.4).

11.2.4 Testing Multiple Hypotheses

In Scenario 2 in the previous section we applied 400 single Z-tests in the central square of the synthetic image. It is obvious from Fig. 11.4 that this approach suffers from many false detections, in particular when the signal gets sparser (see lower right plot in Fig. 11.4). This issue becomes even more severe if the number of tests increases, as the following test scenario illustrates.

Scenario 3 (Unknown position, pixel-wise tests, whole image) *If we do not have prior information on the particular region which we should investigate, we need*

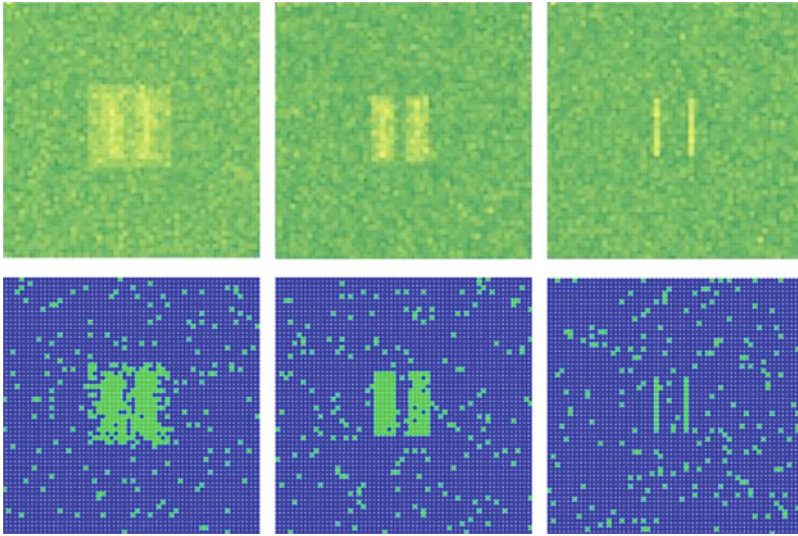


Fig. 11.5 Noisy signal (upper row) and the corresponding test results from Scenario 3 (lower row). The significant pixels are marked green

to scan the entire image. In generalization of Scenario 2 (the RoI is now the full image) to the case of unknown signal position, all single pixels of the entire image are tested. This results in 3600 tests. The results are shown in the second row of Fig. 11.5. Obviously, the number of false rejections increases with the number of tests. In fact, this did not just randomly happen, it is a systematic flaw which we encounter when we naively perform many tests on the same image, simultaneously.

11.2.4.1 Number of False Rejections

The statistical control of false rejections is a general problem one encounters in multiple testing (i.e., testing many hypotheses simultaneously on the same data). The increase of false rejections with increasing number of tests is denoted as *multiplicity effect*.

Figure 11.6 shows the probabilities that out of n independent Z-tests, at least 1 (solid line), 10 (dashed line), 75 (dotted line) and 150 (dash-dotted line) false rejections occur. The curves suggest that in the situation of Scenario 3 we need to expect at least 150 false detections. In fact, the probability that many wrong rejections are made within N tests, each at level α , performed on a data set converges to 1 exponentially fast.

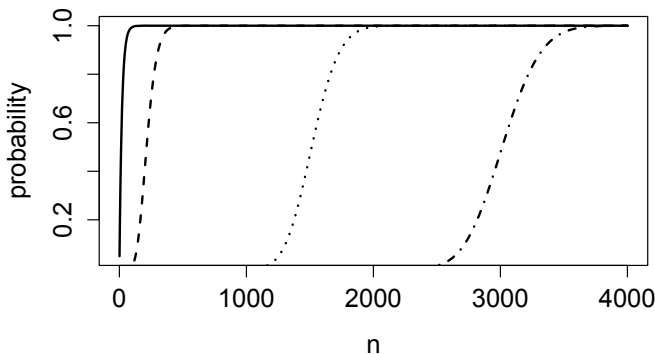


Fig. 11.6 Exact probabilities (y-axis) that out of n (x-axis) independent Z-tests with $\alpha = 0.05$, at least 1 (solid line), 10 (dashed line), 75 (dotted line) and 150 (dash-dotted line) false rejections occur. Here, $n = 1, 2, \dots, 4000$, where the respective probabilities are zero as long as n is smaller than 10, 75 or 150, respectively

Lemma 11.1 *If $0 < \alpha \leq 1/2$, $N \geq 2$ and $k \leq N \log(1 + \alpha)/\log(N)$, we have that*

$$\mathbb{P}(\text{at least } k \text{ out of } N \text{ false rejections}) \geq 1 - (1 - \alpha^2)^N.$$

Proof The random variables $I\{i\text{-th test rejects}\}$, where I denotes the indicator function, follow a Bernoulli distribution with parameter α . Therefore, if $\alpha \leq 1/2$, we can estimate the probability that out of $N \geq 2$ tests k false rejections are made, as

$$\begin{aligned} \mathbb{P}(\text{at least } k \text{ out of } N \text{ false rejections}) &= 1 - \sum_{j=0}^{k-1} \mathbb{P}(\text{exactly } j \text{ false rejections}) \\ &= 1 - \sum_{j=0}^{k-1} \binom{N}{j} (1 - \alpha)^{N-j} \alpha^j \geq 1 - (1 - \alpha)^N \sum_{j=0}^{k-1} \binom{N}{j}. \end{aligned}$$

It follows, e.g., by induction over k for any $N \geq 2$, that $\sum_{j=0}^{k-1} \binom{N}{j} \leq N^k$, which implies

$$\mathbb{P}(\text{at least } k \text{ out of } N \text{ false rejections}) \geq 1 - (1 - \alpha)^N N^k.$$

For $k \leq N \log(1 + \alpha)/\log(N)$ we thus find

$$\mathbb{P}(\text{at least } k \text{ out of } N \text{ false rejections}) \geq 1 - (1 - \alpha^2)^N.$$

Hence, the probability of making at least k out of N false rejections converges to 1 exponentially fast, as $N \rightarrow \infty$. \square

To reduce the number of false detections, so-called *multiplicity adjustments* have to be made. Two general approaches in this regard are the control of the *family wise error rate* (FWER) and of the *false discovery rate* (FDR). Here, we will mainly focus on the FWER but will briefly discuss FDR control in Sect. 11.2.9. For further reading we refer to the monograph by [5] and the references given there.

Multiplicity effect

If multiple tests are performed without accounting for multiplicity, the chances of making many type I errors are quite large if the false null hypotheses are sparse (see Fig. 11.5).

11.2.4.2 Control of FWER

One possible way to deal with multiplicity is to control the family wise error rate (FWER), that is, controlling the probability of making *any* wrong decision in all tests that are performed. Assume model (11.2) and denote by $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ the vector of all true means and by $\mathbb{P}_{\boldsymbol{\mu}}$ the probability under configuration $\boldsymbol{\mu}$. In the previous example of imaging, the sample size n corresponded to the number of pixels. Scenarios 2 and 3 were based on many single tests (on many single pixels). Such single tests will be referred to as *local tests* in the sequel. Each of the N (say) local tests corresponds to its own (local) hypotheses H_i versus K_i . For example, in the setup of Scenario 3, a local hypothesis is $H_i : \mu_i = 0$ versus the local alternative $K_i : \mu_i > 0$, for some $i = 1, \dots, n$. In this case $n = N$ when all local hypotheses are tested. If only a few are tested, then $N \ll n$. If in addition all 2×2 RoIs are tested a total of $N \approx 2n$ tests are performed.

Assume now that all local tests H_i vs. K_i are performed, each at error level α/N . Then the risk of making *any* wrong rejection is controlled at level α , that is, the FWER is controlled.

Theorem 11.1 (Bonferroni correction) *Given N testing problems H_i vs. K_i , $i = 1, \dots, N$ and local tests at level α/N , we have for any configuration $\boldsymbol{\mu}$*

$$\mathbb{P}_{\boldsymbol{\mu}} (\text{"at least one wrong rejection"}) \leq \alpha.$$

Proof

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}} (\text{"at least one wrong rejection"}) &\leq \sum_{i=1}^N \mathbb{P}_{\boldsymbol{\mu}} (\text{" i - th test falsely rejects"}) \\ &\leq \sum_{i=1}^N \mathbb{P}_{H_i} (\text{" i - th test falsely rejects"}) \leq \sum_{i=1}^N \frac{\alpha}{N} = \alpha. \end{aligned} \quad (11.7)$$

Since the right hand side is independent of μ we say that the FWER is controlled *in the strong sense*. As a consequence, each finding can be considered α -significant and hence can be used as a segment for the final segmentation. Performing tests at an adjusted level such as α/N instead of α is called *level adjusted testing* and the multiple test “reject those H_i which are significant at the adjusted level α/N ” is called *Bonferroni procedure*. We stress that although Theorem 11.1 was formulated for the special case of signal detection in independent Gaussian noise, the Bonferroni procedure strongly controls the FWER in much more generality and in particular without any assumptions on the dependency structure between different tests [5, see, e.g., Chap. 3.1.1, for a more detailed discussion].

Scenario 4 (Unknown position, pixel-wise, Bonferroni adjustment) *In the situation of Scenario 3, we now perform a Bonferroni procedure for the entire image, i.e., for all $60 \times 60 = 3600$ entries (see Fig. 11.7). The local testing problems are*

$$H_i : \text{“No signal in } i\text{-th pixel”}, \text{ i.e., } \mu_i = 0 \quad \text{vs.} \quad K_i : \mu_i > 0.$$

Now $n = N = 3600$ and $\alpha/N \approx 1.3889 \times 10^{-5}$ for $\alpha = 0.05$. In this scenario all single entries are compared to $z_{1-\frac{0.05}{3600}} \approx 4.19096$. (Recall that in Scenarios 2 and 3 we compared each entry to the much smaller threshold 1.6449 and note that any level adjustment corresponds to an increase of the threshold for testing.) The result is shown in the second row of Fig. 11.7. While no false findings were provided by any of these tests, too few detections have been made at all as only parts of the signal have been detected.

Bonferroni multiplicity adjustment

Adjustment (increase) of the thresholds when multiple tests are performed simultaneously to *control the overall type I error*, i.e., the FWER. This is a very general but also a conservative method (in particular if the signal is not sparse).

11.2.5 Connection to Extreme Value Theory

There is a close connection between the control of the FWER in the situation of Scenario 3 and extreme value theory. Recall that the aim is to control \mathbb{P}_μ (“at least one wrong rejection”) for any configuration μ . By monotonicity, we have that

$$\mathbb{P}_\mu (\text{“at least one wrong rejection”}) \leq \mathbb{P}_{\mu=0} (\text{“at least one wrong rejection”}),$$

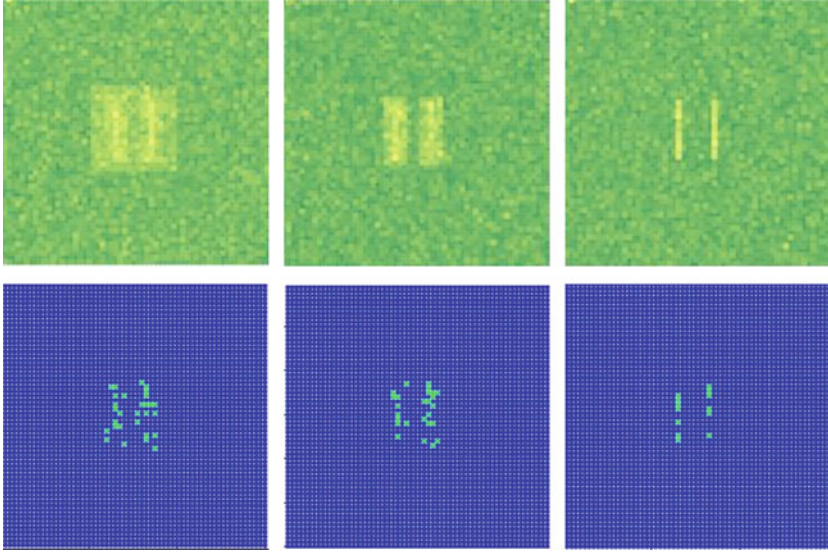


Fig. 11.7 Noisy signals (upper row) and the corresponding test results from Scenario 4 (FWER-controlled, lower row). The significant pixels are marked green, insignificant pixels are blue

which implies that the FWER is controlled if we choose the threshold q for our multiple tests such that

$$\mathbb{P}_{\mu=0}(\text{“at least one wrong rejection”}) = \mathbb{P}(\exists i \in \{1, \dots, N\} : \varepsilon_i > q) \leq \alpha. \quad (11.8)$$

Now, since $\mathbb{P}(\exists i \in \{1, \dots, N\} : \varepsilon_i > q) = \mathbb{P}(\max\{\varepsilon_1, \dots, \varepsilon_N\} > q)$, q can be chosen as the $(1 - \alpha)$ -quantile of $\max\{\varepsilon_1, \dots, \varepsilon_N\}$ under the global null hypothesis, H ,

$$H = \bigcap_{i=1}^{3600} H_i : \text{“No signal at all”}, \quad (11.9)$$

i.e., $\mu_i = 0$ for all $i = 1, \dots, 3600$. In this case we have equality in (11.8). Note that

$$\begin{aligned} \mathbb{P}(\max\{\varepsilon_1, \dots, \varepsilon_N\} > q) &= 1 - \mathbb{P}(\max\{\varepsilon_1, \dots, \varepsilon_N\} \leq q) \\ &= 1 - \mathbb{P}(\varepsilon_1 \leq q \text{ and } \varepsilon_2 \leq q \text{ and } \dots \text{ and } \varepsilon_N \leq q) \\ &= 1 - (\mathbb{P}(\varepsilon_1 \leq q))^N = 1 - (\Phi(q))^N. \end{aligned}$$

Therefore,

$$\mathbb{P}(\max\{\varepsilon_1, \dots, \varepsilon_N\} > q) = \alpha \Leftrightarrow (\Phi(q))^N = 1 - \alpha \Leftrightarrow \Phi(q) = (1 - \alpha)^{\frac{1}{N}},$$

which yields $q = z_{(1-\alpha)^{1/N}}$. Since

$$(1 - \alpha)^{1/N} < 1 - \alpha/N,$$

by Bernoulli's inequality, strict monotonicity of Φ^{-1} implies the same ordering of the thresholds, i.e., $z_{(1-\alpha)^{1/N}} < z_{1-\alpha/N}$. However, it is easy to show that for $N \geq 1$ and $\alpha \leq 1/2$

$$1 - \frac{\alpha + \alpha^2}{N} < (1 - \alpha)^{1/N}$$

and therefore the difference between $z_{(1-\alpha)^{1/N}}$ and $z_{1-\alpha/N}$ is quite small, e.g. $z_{(1-0.05)^{1/3600}} \approx 4.18516$ and $z_{1-0.05/3600} \approx 4.19096$.

The following lemma shows that $z_{1-\alpha/N} \approx \sqrt{2 \log(N)}$ (and therefore also $z_{(1-\alpha)^{1/N}} \approx \sqrt{2 \log(N)}$).

Lemma 11.2 *There exists $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$*

$$\sqrt{2 \log(N)} - \frac{\log \log(N)}{\sqrt{2 \log(N)}} \leq z_{1-\alpha/N} \leq \sqrt{2 \log(N)}$$

Proof To bound the normal quantiles from above and below, we use

$$\frac{\varphi(x)}{x + \frac{1}{x}} < 1 - \Phi(x) < \frac{\varphi(x)}{x},$$

[6, see inequality (10)], where φ and Φ denote the density and cdf of the standard normal distribution, respectively. Since, for sufficiently large N ,

$$1 - \Phi(\sqrt{2 \log(N)}) \leq \frac{\varphi(\sqrt{2 \log(N)})}{\sqrt{2 \log(N)}} = \frac{1}{N \sqrt{4\pi \log(N)}} \leq \frac{\alpha}{N},$$

and therefore

$$1 - \frac{\alpha}{N} \leq \Phi(\sqrt{2 \log(N)}) \Leftrightarrow z_{1-\frac{\alpha}{N}} = \Phi^{-1}(1 - \frac{\alpha}{N}) \leq \sqrt{2 \log(N)},$$

the right hand side follows. We further have that

$$\begin{aligned} 1 - \Phi\left(\sqrt{2 \log(N)} - \frac{\log \log(N)}{\sqrt{2 \log(N)}}\right) &\geq \frac{\varphi\left(\sqrt{2 \log(N)} - \frac{\log \log(N)}{\sqrt{2 \log(N)}}\right)}{\sqrt{2 \log(N)} - \frac{\log \log(N)}{\sqrt{2 \log(N)}} + \frac{1}{\sqrt{2 \log(N)} - \frac{\log \log(N)}{\sqrt{2 \log(N)}}}} \\ &\geq \frac{\log(N)}{N \sqrt{2\pi}} \exp\left(-\frac{\log \log(N)}{4 \log(N)}\right) \geq \frac{\alpha}{N}, \end{aligned}$$

for sufficiently large N , and the left hand side follows.

11.2.5.1 Towards Better Detection Properties

The Bonferroni approach is valid in most generality. Nevertheless, as we have seen in Fig. 11.7, if applied pixel-wise the level adjustment (and the resulting increase of the threshold) is (much) too strict for our purposes. This is not caused by the Bonferroni-adjustment per se, as it can be shown that the detection power of the Bonferroni approach cannot be considerably improved in general [7, Sect. 1.4.1]. The issue is that we have only considered each single pixel as input for our local tests. Therefore, we will extend this from single pixels to larger systems of RoIs, which allow to “borrow strength from neighbouring pixels”. This makes sense as soon as the signal has some structure, e.g., whenever signal appears in (small) clusters or filament-like structures. To see this, suppose that for $k > 1$ we have $\mu_1 = \mu_2 = \dots = \mu_k = \mu$. An uncorrected pixel-wise Z-test would compare each Y_i to the threshold $z_{1-\alpha}$, i.e., signal in a pixel would be detected if

$$Y_i = \underbrace{(Y_i - \mu)}_{\mathcal{N}(0,1)} + \mu > z_{1-\alpha}.$$

This is almost impossible if μ is too small or the noise takes a negative value and becomes even worse if a multiplicity adjustment is performed. If we instead group the first k pixels together and perform a grouped Z-test, i.e., compare $\frac{1}{\sqrt{k}} \sum_{j=1}^k Y_j$ to $z_{1-\alpha}$, a signal would be detected if

$$\sqrt{k}\mu + \mathcal{N}(0, 1) > z_{1-\alpha}.$$

This way, the signal is “magnified” by a factor \sqrt{k} . Unfortunately, performing, for any k , every test that groups k pixels together and thereby incorporating the fact that positions i and numbers k of relevant pixels are in general not known in advance, is infeasible.³ However, if the data is clustered spatially we can construct a reasonable test procedure that follows a similar path. Instead of performing all tests that group any configurations of k pixels, we perform *all tests* that merge all pixels in a $k \times k$ square, for many different values of k and “scan” the image for signal in such regions in a computationally and statistically feasible way. Now the local tests become (locally highly) correlated (see Sect. 11.2.6) and a simple Bonferroni adjustment does not provide the best detection power any more, although (11.7) is still valid. This will be the topic of Sects. 11.2.6 and 11.2.7.

³One issue is computational limitation. Additionally, this has a systematic statistical burden as then tests have to be performed over all possible subsets of the image. For n pixels, these are of size 2^n , which is a collection of sets such that the resulting error probabilities can no longer be controlled in a reasonable way.

Amplification of the signal strength by aggregation
If a signal is spatially grouped in clusters, cluster-wise tests can increase its detectability. The average of all signal strengths inside the test region is magnified by a factor of $\sqrt{\text{size of cluster}}$.

11.2.6 Scanning

In a way, the two approaches of aggregating data over the entire image (Scenario 1) and performing pixel-wise tests (as done in Scenarios 2–4) are the most extreme scenarios. As a rule of thumb, aggregation makes detection easier at the cost of losing spatial precision whereas pixel-wise testing provides the highest possible spatial precision but makes detection more difficult (after Bonferroni level adjustment as we have seen in Scenario 4. Recall that since the tests are independent we know that there is no substantially better way to control the FWER). In a next step we will combine both ideas. We test on various squares of different sizes to achieve accuracy (small regions) where possible and gain detection power (larger regions) where the signal is not strong enough to be detected pixel-wise, i.e., on small spatial scales. As the system of all subsquares of an image consists of many overlapping squares, we have to deal with locally highly dependent test statistics. Table 11.1 illustrates this effect presenting simulated values of the family wise error rate, based on 1000 simulation runs each, with preassigned value $\alpha = 0.05$. Squares of size $h \times h$, $h \in \{1, 2, 3, 4, 5\}$ in an image of 60×60 are considered. The parameter h is denoted as a spatial scale. The results of this small simulation study demonstrate that the Bonferroni correction is much too strict if we aggregate data in larger squares. The following scenario is tailored towards dealing with this specific type of dependency structure and is called *multiscale scanning*. Here, the level adjustment is made in an optimal spatially adaptive way, i.e., such that the thresholds are both, large enough so that the FWER is controlled but on the other hand so small that smaller thresholds can no longer guarantee the control of the FWER. The key is now to exploit that the system of all $h \times h$ squares fitting into the $n \times n$ image is highly redundant. For instance, if a square is shifted one pixel to the right, say, both squares share most of their pixels and their contents should not be treated as independent. We discussed in

Table 11.1 Simulated values of FWER at nominal level $\alpha = 0.05$ for a matrix of local averages of $h \times h$ pixels

| $h \times h$ | 1×1 | 2×2 | 3×3 | 4×4 | 5×5 | 10×10 |
|---------------------|--------------|--------------|--------------|--------------|--------------|----------------|
| Observed error rate | 0.049 | 0.046 | 0.043 | 0.028 | 0.025 | 0.016 |

Sect. 11.2.5 that instead of the Bonferroni threshold $z_{1-\frac{\alpha}{N}}$ the $(1 - \alpha)$ -quantile of the distribution of the maximum of N independent standard normal random variables under the global null hypothesis, H , could be used as a threshold as well. This idea can be transferred to this setting by using the $(1 - \alpha)$ -quantile of

$$\max_h w(h)(T_{h \times h \text{ square}}(Y) - w(h)),$$

where $w(h)$ is a size-dependent correction term, given by

$$w(h) := \sqrt{2 \ln \frac{N}{h^2}} + 7 \ln \left(\sqrt{2 \ln \frac{N}{h^2}} \right) / \sqrt{2 \ln \frac{N}{h^2}}. \quad (11.10)$$

Under H , the quantiles can be simulated as described in Algorithm 1. Recall that in Lemma 11.2 it was shown that the quantile $z_{1-\frac{\alpha}{N}}$ and therefore also the quantile of the maximum, $z_{(1-\alpha)^N}$, are approximately of size $\sqrt{2 \log(N)}$. When pixels are aggregated over $h \times h$ squares, the corresponding quantiles can be shown to be of first asymptotic order $\sqrt{2 \log(N/h^2)}$ (the leading term of $w(h)$ in (11.10), see Theorem 11.2 for details), which corresponds to the case of N/h^2 independent tests. This is incorporated into the construction of the thresholds as described in Algorithm 1.

Algorithm 1: Simulation of the thresholds

Parameters : Number of Monte-Carlo runs $M \in \mathbb{N}$, largest size $h_{\max} \in \mathbb{N}$, significance level $\alpha \in (0, 1)$

```

1 for  $n = 1, 2, \dots, M$  do
2   Draw i.i.d. data  $Y_i \sim \mathcal{N}(0, 1)$  for  $1 \leq i \leq n$ ;
3   for  $1 \leq h \leq h_{\max}$  do
4     Compute all test statistics  $T_{h \times h \text{ square}}(Y)$ ;
5     Compute all  $w(h)(T_{h \times h \text{ square}}(Y) - w(h))$ ;
6     Save their maximal value in  $q_h$ ;
7   Set  $t_i := \max_{1 \leq h \leq h_{\max}} q_h$ ;
8 Sort the values  $t_i$  such that  $t_1 \leq \dots \leq t_M$ ;
9 Choose  $j \in \{1, \dots, M\}$  such that  $j/M \leq \alpha < (j+1)/M$ ;
10 Set  $q_{1-\alpha}^h = t_j/w(h) + w(h)$ ;
```

In line 12 of Algorithm 1, the size-dependent thresholds $q_{1-\alpha}^h = t_j/w(h) + w(h)$ are defined. Comparing each $T_{h \times h \text{ square}}(Y)$ to $q_{1-\alpha}^h$ yields a multiplicity adjusted multiple test procedure. Note that in Algorithm 1 the quantile of the maximum over all, locally correlated, test statistics under the global null hypothesis is approximated. This way, the dependence structure is taken into account precisely.

Scenario 5 (Unknown position, multiscale scanning) *We now aggregate test results for several different scanning tests. We consider testing each pixel, as well as testing each 2×2 , 3×3 , 4×4 and 5×5 square. In total these are 16.830 tests. We now*

Table 11.2 Scale dependent quantiles for the scanning test with windows of variable sizes

| α | $q_{1-\alpha}^1$ | $q_{1-\alpha}^2$ | $q_{1-\alpha}^3$ | $q_{1-\alpha}^4$ | $q_{1-\alpha}^5$ | Bonferroni for 16.830 tests |
|----------|------------------|------------------|------------------|------------------|------------------|--------------------------------|
| 0.1 | 5.115 | 4.760 | 4.531 | 4.345 | 4.208 | 4.380 |
| 0.05 | 5.267 | 4.921 | 4.698 | 4.527 | 4.385 | 4.528 |
| 0.01 | 5.581 | 5.2538 | 5.043 | 4.883 | 4.750 | 4.875 |

adjust the level in a way that accounts for local correlations. We fix $\alpha = 0.05$ and calculate all test statistics $T_{h \times h \text{ square}}(Y)$ (see (11.5)). The local hypotheses $H_{h \times h \text{ square}}$ are

$$H_{h \times h \text{ square}} : “\mu_i \equiv 0 \text{ in } h \times h \text{ square.}” \quad (11.11)$$

Each $T_{h \times h \text{ square}}(Y)$ is compared to the size-dependent thresholds $q_{1-\alpha}^h$, which have been generated according to Algorithm 1 and are listed in Table 11.2. We reject the local hypotheses that there is no signal in a particular $h \times h$ square if the corresponding test statistic is larger than the threshold, that is, if

$$T_{h \times h \text{ square}}(Y) = \frac{1}{h} \sum_{h \times h \text{ square}} Y_i > q_{1-\alpha}^h. \quad (11.12)$$

All significant squares are stored and finally, after all square-wise comparisons have been made, for each pixel, the smallest square that was significant is plotted. Findings for the different sizes are color-coded and for each pixel the color corresponding to the smallest square in which signal was detected is plotted. The results are shown in Fig. 11.8. One big advantage of this approach is that also the weak signal is now completely included in the segmentation in contrast to even the unadjusted approach of Scenario 2 (compare the lower left plots of Figs. 11.4 and 11.8). Also, the color-coding visualizes regions of strong signal and therefore contains “structural information” on the data.

The procedure in Scenario 5 is such that the FWER is still controlled in a strong sense, although the thresholds can be chosen smaller than in a Bonferroni approach. This is much more so if N and h get larger, but is visible starting from $h = 4$, which matches the values given in Table 11.1. This was possible due to the strong local correlations between tests. Roughly speaking, for each size of the moving window a Bonferroni-type adjustment is made for the (maximum) number of non-overlapping squares of that size which is a considerable relaxation. Remarkably, the prize for including many different sizes is extremely small. More theoretical details can be found in Sect. 11.2.7.

To conclude this section, it should be stressed that in many situations, we do not encounter rectangular signals, however, small rectangles can be considered as building blocks for more complex structures. If specific shape information is available,

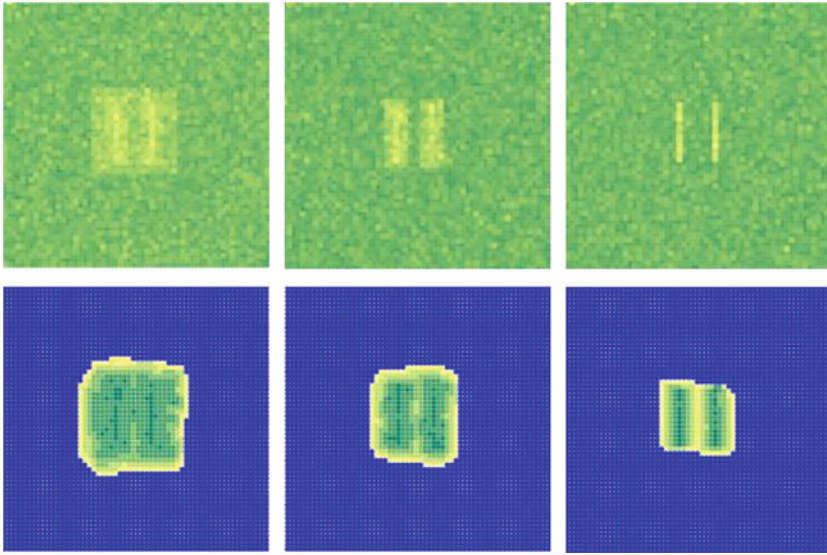


Fig. 11.8 Noisy signals (upper row) and the corresponding test results from Scenario 5 (lower row). Significant 5×5 squares are plotted in yellow. Significant $1 \times 1 - 4 \times 4$ squares are plotted in green with increasing brightness. For each pixel, the smallest square which was found significant was plotted. Insignificant regions are coloured in blue

this can be incorporated into the testing procedure as long as the regions are not too irregular and the set of regions satisfies a Vapnik-Cervonenkis-type complexity condition (see [8] for more details). The literature on multiscale scanning methods is vast. In the particular context of imaging, the reader may also consult [9–12] for related ideas.

Multiscale Scanning

With probability guarantee of $1 - \alpha$ all of the RoIs chosen in the multiscale scanning procedure described in Scenario 5, are valid. Hence, we obtain localized RoIs where the signal is sufficiently strong and profit from aggregation, as described in Sect. 11.2.5.1, where the signal is weak and point-wise detection is too difficult.

11.2.7 Theory for the Multiscale Scanning Test

The following theorem is the theoretical foundation for Scenario 5.

Theorem 11.2 Assume that an $n \times n$ array of independent $\mathcal{N}(\mu_i, 1)$ variables is observed and $\mathcal{H} \subset \{1, \dots, n\}$ is a set of side lengths of squares. Denote for $h \in \mathcal{H}$

by $\mathcal{S}(h)$ the set of all $h \times h$ -squares. Let $N = n^2$, $w(h)$ as defined in (11.10) and let further $\tilde{q}_{1-\alpha}$ denote the $1 - \alpha$ -quantile of

$$\max_{h \in \mathcal{H}} \max_{S \in \mathcal{S}(h)} w(h) (T_S(Y) - w(h)) \quad (11.13)$$

under the global hypothesis H : “no signal in any of the squares”. Reject each hypothesis $H_{h \times h \text{ square}}$ (see (11.11)) for which

$$T(R) \geq \frac{\tilde{q}_{1-\alpha}}{w(h)} + w(h). \quad (11.14)$$

- (i) This yields a multiple test for which the FWER at level α is controlled asymptotically (as $|\mathcal{H}|/n \rightarrow 0$, $n \rightarrow \infty$) in the strong sense.
- (ii) This test is minimax optimal in detecting sparse rectangular regions of the signal.

Claims (i) and (ii) follow from Theorems 7 and 2 in [1]. Roughly speaking, the essence of the previous theorem is that we only need multiplicity control for approximately n^2/h^2 (corresponding to the number of independent) tests instead of $(n - h + 1)^2$ (corresponding to the actual number of all) tests. Control of the FWER in the strong sense means that all significant squares can be used in the final segmentation (lower row of Fig. 11.8).

In this chapter we mainly focused on control of the FWER, however weaker means of error control are of interest as well. A very prominent one is the *false discovery rate* (FDR, [13]), which we briefly discuss in Sect. 11.2.9.

11.2.8 Deconvolution and Scanning

In photonic imaging additional difficulties arise. Firstly, we have to deal with non-Gaussian and non i.i.d. data (see Chap. 4), e.g., following a Poisson distribution with inhomogeneous intensities λ_i . Then, as long as the intensity is not too small, a Gaussian approximation validates model (11.2) as a reasonable proxy for such situations. A formal justification for the corresponding multiscale tests is based on recent results by [14], for details see [1]. The price to pay for such an approximation is a lower bound on the sizes of testing regions that can be used, due to the fact that several data points (of logarithmic order in n) need to be aggregated so that a Gaussian approximation is valid. For ease of notation, we only discussed the Gaussian case in Sect. 11.2.7, generalizations to other distributions can be found in [8].

Secondly, convolution with the PSF of the imaging device induces blur. The first row of Fig. 11.9 shows the convolved synthetic images that were shown in the upper row of Fig. 11.2, where the images in the central row are noisy versions of these convolved images. Note, that some structures are no longer identifiable by eye after convolution. When applying the multiscale scanning approach in Scenario 5 naively

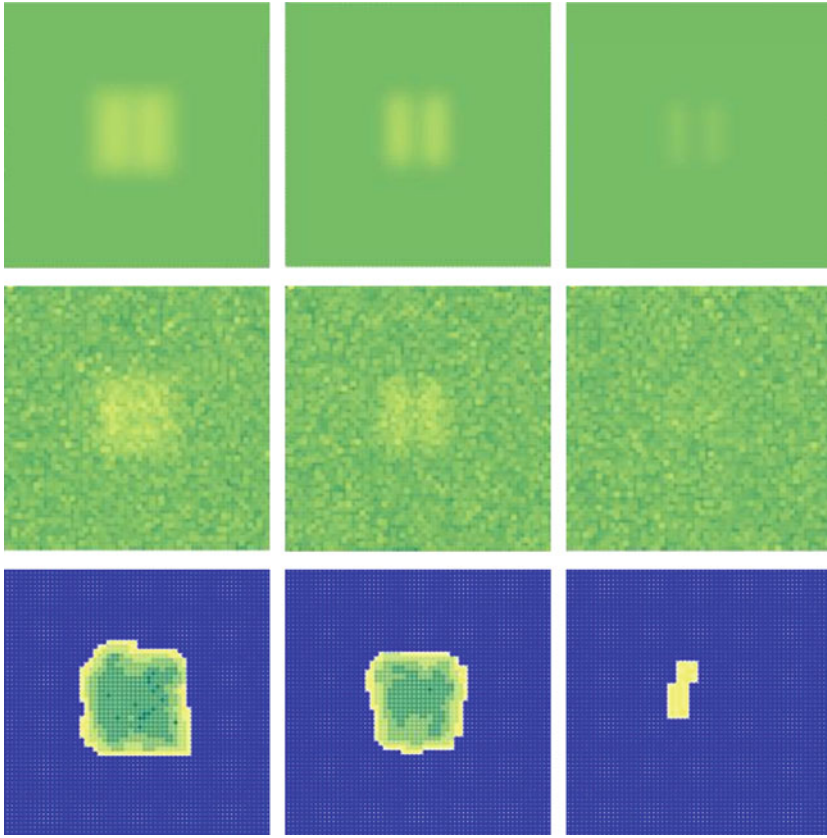


Fig. 11.9 Signals after convolution (upper row), noisy version (central row) and the corresponding test results from Scenario 5, naively applied to the convolved data (lower row). Significant 5×5 squares are plotted in yellow. Significant $1 \times 1 - 4 \times 4$ squares are plotted in green with increasing brightness. For each pixel, the smallest square which was found significant was plotted. Insignificant regions are coloured in blue

to the convolved data (central row of Fig. 11.9). The result (lower row of Fig. 11.11) demonstrates that this is indeed not a competitive strategy and it strongly suggests to take the convolution into account.

We now briefly sketch how to adapt the multiscale scanning procedure (Scenario 5) to the convolution setting. Notice that in the case of data (11.2), we can write the test statistic (11.5) for a particular square S as

$$T_S(Y) = \langle \mathcal{I}_S, Y \rangle,$$

where $Y = (Y_1, \dots, Y_n)$ denotes the data vector and \mathcal{I}_S denotes the scaled indicator function on S , i.e., $\mathcal{I}_S(j) = 1/\sqrt{|S|}$ if $j \in S$ and 0 else. Now, the indicator functions

are considered as a system of probe functions, which are tested on the data Y . In case of convolution with the PSF k (e.g. of the microscope), model (11.2) turns into

$$Y_i^* = (\boldsymbol{\mu} * k)_i + \varepsilon_i, \quad i = 1, \dots, n \quad (11.15)$$

where “ $*$ ” denotes convolution. The goal is to find a probe function, acting on the convolved data, denoted as \mathcal{I}_S^* such that

$$\langle \mathcal{I}_S^*, Y^* \rangle \approx \langle \mathcal{I}_S, Y \rangle,$$

that is, \mathcal{I}_S^* should locally deconvolve. Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$. Then, if \mathcal{F} denotes the discrete Fourier transform, by Plancherel isometry and the convolution theorem

$$\langle \mathcal{I}_S, \boldsymbol{\mu} \rangle = \left\langle \mathcal{F}^{-1} \left(\frac{\mathcal{F} \mathcal{I}_S}{\mathcal{F} k} \right), \boldsymbol{\mu} * k \right\rangle.$$

This means that (provided $\mathcal{F} k \neq 0$)

$$\mathcal{I}_S^* = \mathcal{F}^{-1} \left(\frac{\mathcal{F} \mathcal{I}_S}{\mathcal{F} k} \right) \quad (11.16)$$

is a reasonable choice of a probe system for the data (11.15) and a statistic that adapts to the convolution is given by

$$T_S^*(Y^*) = \langle \mathcal{I}_S^*, Y^* \rangle.$$

Scenario 5 can now be performed, following Algorithm 1 to derive suitable thresholds, replacing \mathcal{I}_S by \mathcal{I}_S^* and the FWER is controlled. More precisely, it can be shown that Theorem 11.2 also applies in this scenario (see [1] for details). Figure 11.10 d shows the result of this adapted test procedure (MISCAT) applied to our original data (Fig. 11.10 a). As a comparison, we also applied Scenario 5 naively to the data set (Fig. 11.10f). Analogously to [15], \mathcal{I}_S can be chosen such that MISCAT with \mathcal{I}_S^* performs optimally in terms of detection power.

Deconvolution and scanning

In convolution problems sums of pixel values over spatial regions (e.g. squares) will be replaced by probe functionals over the pixels (weighted sums) which can be designed in an optimal way for a given convolution K . The resulting multiscale test scans over all probe functionals which results in substantially more precise segmentation results (for a direct comparison see lower left and lower right panel of Fig. 11.10). It still controls the FWER.

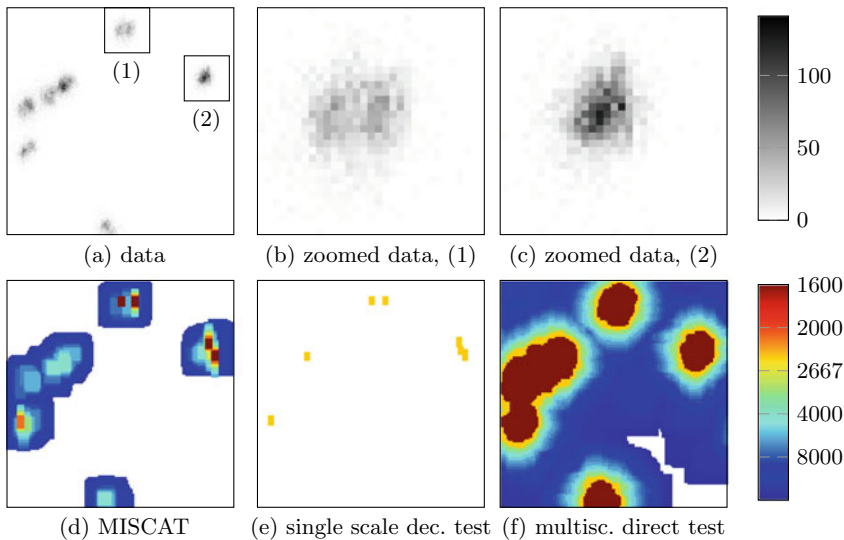


Fig. 11.10 (Figure 2 in [1]) Experimental data and corresponding 90% significance maps computed by different tests. The color-coding of the significance maps always show the size of smallest significance in nm^2 , cf. the main text. **a–c** data and zoomed regions, **d** MISCAT, **e** a single scale test with deconvolution, **f** a multiscale scanning test without deconvolution

11.2.9 FDR Control

As discussed in the previous sections of this chapter, as the sample size increases (and therefore the number of tests), the control of the FWER becomes more difficult and thus this may result in low detection power, e.g., in three dimensional imaging. Therefore, a strategy to obtain less conservative procedures of error control is to relax the FWER. The most prominent relaxation is the *false discovery rate* (FDR [13]), defined as

$$\text{FDR} = \mathbb{E} \left[\frac{\#\text{false rejections}}{\max\{\#\text{all rejections}, 1\}} \right],$$

that is, the average proportion of false rejections among all rejections. Hence, in contrast to the FWER this criterion scales with the number of rejections. The control of the FDR is a weaker requirement than the control of the FWER in general. Procedures that control the FDR are often written in terms of *p-values*. In the situation of the Z-test with test statistics $T_{h \times h \text{ square}}(Y)$ as in (11.12) the p-values are given as

$$p_{h \times h \text{ square}} = 1 - \Phi \left(T_{h \times h \text{ square}}(Y) \right), \quad (11.17)$$

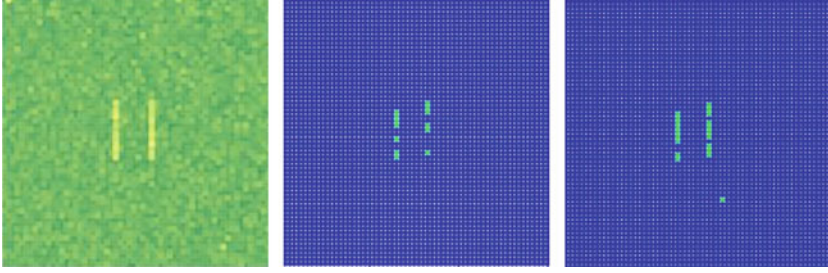


Fig. 11.11 Noisy signal (left) test result of pixel-wise tests after Bonferroni adjustment (middle) and test results from Scenario 6 (right) with FDR control. In both multiple testing procedures $\alpha = 0.05$. Significant pixels are marked green, insignificant regions are coloured in blue

where Φ denotes the cumulative distribution function of the standard normal distribution. The smaller the p-value, the stronger the evidence that the null hypothesis should be rejected.

Benjamini-Hochberg Procedure ([13]) *Consider a multiple testing procedure consisting of independent tests with p-values p_1, \dots, p_N . Sort the p-values increasingly, $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(N)}$, and reject all null hypotheses for which $p_i \leq \alpha \frac{\hat{k}}{N}$, where $\hat{k} = \max\{k \mid p_{(k)} \leq \alpha k/N\}$.*

Reference [16] already proposed the above procedure but pointed out that this approach lacks a theoretical justification, which has been given by [13], who showed that $\text{FDR} \leq \frac{N_0}{N} \alpha$, where N_0 denotes the number of true null hypotheses.

Scenario 6 (Benjamini-Hochberg (BH) Procedure) *In the situation of Scenario 3, we also performed a BH procedure for all $60 \times 60 = 3600$ entries of the third test image (see left panel of Fig. 11.11). The result is displayed in the right most panel of Fig. 11.11, while in the centre, for a comparison, the result of the Bonferroni procedure on the same data set is displayed. Obviously, more parts of the signal have been found, however, still several positives are missed and a false discovery is included.*

There is a vast literature on FDR control and many generalizations have been proposed. For instance, if $\frac{N_0}{N}$ is much smaller than 1, corresponding to the case of a non-sparse signal, the procedure controls the FDR at much smaller level than α and refined versions of the BH procedure in which N_0/N is estimated from the data have been proposed (see, e.g., [17, 18] and the references given therein).

While the BH procedure grants control of the FDR in test Scenarios 2 and 3 due to independence between pixels, the situation in Scenario 5 is more delicate due to the strong local correlations, in particular in the presence of convolution, where a suitable FDR-procedure is still an open problem and currently investigated by the authors. We stress that while FDR-control under specific dependency structures has been investigated by many authors, e.g., [19, 20] and the references given therein.

Non of the existing methods provide a procedure tailored to deconvolution problems as they occur in photonic imaging. The construction of such adjusted methods is a worthwhile focus for future research.

11.3 Statistical Multiscale Estimation

If one is further interested in the recovery (estimation) of the unknown signal, the multiscale testing procedure developed in Sects. 11.2.6 and 11.2.8 actually provides a collection of feasible candidates for this task in the sense that all signals which fall in the acceptance region of one of the afore-mentioned tests can be considered as “likely” as they cannot be rejected by such a scanning test. More precisely, if we assume model (11.15), any signal $\tilde{\mu}$ which satisfies

$$\max_{h \in \mathcal{H}} \max_{S \in \mathcal{S}(h)} w(h) \left(\langle \mathcal{I}_S^*, Y^* - \tilde{\mu} * k \rangle - w(h) \right) \leq q_{1-\alpha}^*, \quad (11.18)$$

cannot be rejected. Here \mathcal{H} and $\mathcal{S}(h)$ are defined in Theorem 11.2, and $q_{1-\alpha}^*$ is the $(1 - \alpha)$ -quantile of the left hand side of (11.18) with $(Y^* - \tilde{\mu} * k)$ being replaced by noise ε , $w(h)$ is the scale correction term given in (11.10), and \mathcal{I}_S^* is as in (11.16). Among all the candidates $\tilde{\mu}$ lying in (11.18), we will pick the most regular estimate. This is done by means of a (convex) functional $\mathcal{S}(\cdot)$, defined on a domain \mathcal{D} for μ , which encodes prior information about the unknown signal, e.g. sparsity or smoothness. Thus, the final estimator $\hat{\mu}$ is defined as

$$\hat{\mu} \in \arg \min_{\tilde{\mu}} \mathcal{S}(\tilde{\mu}) \quad \text{subject to } \tilde{\mu} \text{ satisfies (11.18)}. \quad (11.19)$$

Because of the choice of $q_{1-\alpha}^*$ we readily obtain the *regularity guarantee*

$$\mathbb{P}_{\mu} \left(\mathcal{S}(\hat{\mu}) \leq \mathcal{S}(\mu) \right) \geq 1 - \alpha \quad \text{uniformly over all } \mu \in \mathcal{D},$$

i.e., the resulting estimator is at least as regular as the truth with probability $1 - \alpha$, whatever the configuration μ of the truth is. Furthermore, the remaining residuum $Y^* - \hat{\mu} * k$ is accepted as pure noise by the multiscale procedure described in Sect. 11.2.8.

Before we discuss possible ways how to solve the minimization problem (11.19), note that $\langle \mathcal{I}_S^*, Y - \tilde{\mu} * k \rangle = \langle \mathcal{I}_S^*, Y \rangle - \langle \mathcal{I}_S, \tilde{\mu} \rangle$ in (11.18), and hence the computation can be sped up by avoiding convolutions between $\tilde{\mu}$ and k . Next we emphasize that the discretization of (11.19) has the form

$$\arg \min_{\tilde{\mu}} \mathcal{S}(\tilde{\mu}) \quad \text{subject to } \underline{\lambda} \leq K \tilde{\mu} \leq \bar{\lambda}, \quad (11.20)$$

where $\underline{\lambda}, \bar{\lambda}$ are vectors, K a matrix, and “ \leq ” acts element-wise. Thus, whenever \mathcal{S} is convex, the whole problem is convex (but, however, non-smooth) and can be solved by many popular methods. In Algorithm 2 we give one possibility which arises from applying the primal-dual hybrid gradient method [21] to an equivalent reformulation of the first order optimality conditions of (11.20) (which are necessary and sufficient by convexity).

Algorithm 2: Primal dual hybrid gradient method for (11.20)

Parameters : Set $\sigma, \tau > 0$ s.t. $\sigma\tau\|K\|^2 < 1$, and $\theta \in [0, 1]$
Initialization: Set $\tilde{\mu}_0 = \mu_0 \in \mathcal{D}(K)$ and $\nu_0 \in \mathcal{R}(K)$
1 for $n = 1, 2, \dots$ **do**
2 $\nu_n = \nu_{n-1} + \sigma K \tilde{\mu}_{n-1}$;
3 $\nu_n = \max\{\nu_n - \sigma \bar{\lambda}, 0\} + \min\{\nu_n - \sigma \underline{\lambda}, 0\}$;
4 $\mu_n = \arg \min_{\tilde{\mu}} \frac{1}{2\tau} \|\tilde{\mu} - (\mu_{n-1} - \tau K^* \nu_n)\|^2 + \mathcal{S}(\tilde{\mu})$;
5 $\tilde{\mu}_n = \mu_n + \theta(\mu_n - \mu_{n-1})$;

Algorithm 2 relies on efficient computations of the so-called proximal operator of \mathcal{S} , see line 4. In most cases, it has either an analytic form if \mathcal{S} is ℓ^p -norm ($1 \leq p \leq \infty$), or an efficient solver if \mathcal{S} is the total variation semi-norm [22].

One alternative to Algorithm 2 is the alternating direction method of multipliers (ADMM), which can be applied directly to (11.20) and is compatible with any convex functional \mathcal{S} [23]. However, Algorithm 2 avoids the projection onto the intersection of convex sets, and turns out to be much faster in practice if step 4 in Algorithm 2 can be efficiently computed. For further algorithms relevant for this problem, see Chaps. 6 and 12.

We stress that a crucial part of the estimator $\hat{\mu}$ in (11.18)–(11.19) is the choice of probe functionals \mathcal{I}_S^* from Sect. 11.2.8. In Fig. 11.12, this estimator $\hat{\mu}$ is referred to as MiScan (short for multiscale image scanning), whereas MrScan (short for multiscale residual scanning) denotes the estimator of a similar form as $\hat{\mu}$ but with \mathcal{I}_S^* being replaced by \mathcal{I}_S see [23–26] i.e., the convolution is not explicitly taken into account in the probe functional. MiScan recovers significantly more features over a range of scales (i.e., various sizes) compared to MrScan.

There is good theoretical understanding on the estimator $\hat{\mu}$ by (11.18)–(11.19) for the regression model (11.2), that is, $k = \delta_0$, the Dirac delta function, in model (11.15). In case of \mathcal{S} being Sobolev norms, [27] shows the minimax optimality of $\hat{\mu}$ for Sobolev functions for fixed smoothness, and [28] further show the optimality over Sobolev functions with varying smoothness (adaptation). In case of \mathcal{S} being the total variation semi-norm, [29] show the minimax optimality of such an estimator for functions with bounded variation. All the results above are established for L^p -risks ($1 \leq p \leq \infty$). For the more general model (11.15), [30] provide some asymptotic analysis

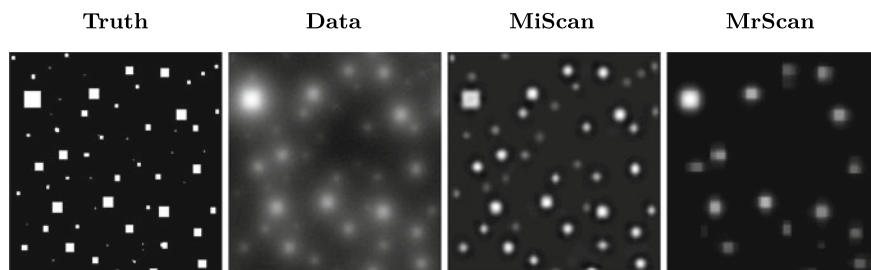


Fig. 11.12 Comparison on a deconvolution problem ($\text{SNR} = 100$, and the convolution kernel k satisfying $\mathcal{F}k = 1/(1 + 0.09\|\cdot\|^2)$). MiScan is defined by (11.18)–(11.19); MrScan is similar to MiScan but with \mathcal{I}_S^* replaced by \mathcal{I}_S ; For both methods, the regularization functional \mathcal{S} is chosen as the total variation semi-norm

with respect to a relatively weak error measure, the Bregman divergence. A detailed analysis of MiScan exploring the probe functionals in (11.16) in a convolution model is still open and currently investigated by the authors.

References

1. Proksch, K., Werner, F., Munk, A.: Multiscale scanning in inverse problems. *Ann. Statist.* **46**(6B), 3569–3602 (2018). <https://doi.org/10.1214/17-AOS1669>
2. Hell, S.W.: Far-field optical nanoscopy. *Science* **316**, 1153–1158 (2007)
3. Feller, W.: An introduction to probability theory and its applications. Vol. I, 2nd ed, John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London (1957)
4. Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses. Springer Texts in Statistics, 3rd edn. Springer, New York (2005)
5. Dickhaus, T.: Simultaneous statistical inference. Springer, Heidelberg (2014). <https://doi.org/10.1007/978-3-642-45182-9>. With applications in the life sciences
6. Gordon, R.D.: Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Stat.* **12**, 364–366 (1941)
7. Donoho, D., Jin, J.: Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**(3), 962–994 (2004). <https://doi.org/10.1214/009053604000000265>
8. König, C., Munk, A., Werner, F.: Multidimensional multiscale scanning in exponential families: limit theory and statistical consequences (2018+). *Ann. Statist.*, to appear
9. Arias-Castro, E., Donoho, D.L., Huo, X.: Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* **51**(7), 2402–2425 (2005). <https://doi.org/10.1109/TIT.2005.850056>
10. Donoho, D.L., Huo, X.: Beamlets and multiscale image analysis. In: *Multiscale and Multiresolution Methods, Lect. Notes Comput. Sci. Eng.*, vol. 20, pp. 149–196. Springer, Berlin (2002). https://doi.org/10.1007/978-3-642-56205-1_3
11. Sharpnack, J., Arias-Castro, E.: Exact asymptotics for the scan statistic and fast alternatives. *Electron. J. Stat.* **10**(2), 2641–2684 (2016). <https://doi.org/10.1214/16-EJS1188>
12. Siegmund, D., Yakir, B.: Tail probabilities for the null distribution of scanning statistics. *Bernoulli* **6**(2), 191–213 (2000). <https://doi.org/10.2307/3318574>
13. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**(1), 289–300 (1995). [http://links.jstor.org/sici?sici=0035-9246\(1995\)57:1<289:CTFDRA>2.0.CO;2-E&origin=MSN](http://links.jstor.org/sici?sici=0035-9246(1995)57:1<289:CTFDRA>2.0.CO;2-E&origin=MSN)

14. Chernozhukov, V., Chetverikov, D., Kato, K.: Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42**(4), 1564–1597 (2014). <https://doi.org/10.1214/14-AOS1230>
15. Schmidt-Hieber, J., Munk, A., Dümbgen, L.: Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Ann. Statist.* **41**(3), 1299–1328 (2013). <https://doi.org/10.1214/13-AOS1089>
16. Simes, R.J.: An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**(3), 751–754 (1986). <https://doi.org/10.1093/biomet/73.3.751>
17. Kumar Patra, R., Sen, B.: Estimation of a two-component mixture model with applications to multiple testing. *J. Roy. Statist. Soc. Ser. B* **78**(4), 869–893 (2016). <https://doi.org/10.1111/rssb.12148>
18. Storey, J.D., Taylor, J.E., Siegmund, D.: Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. Ser. B* **66**(1), 187–205 (2004). <https://doi.org/10.1111/j.1467-9868.2004.00439.x>
19. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**(4), 1165–1188 (2001). <https://doi.org/10.1214/aos/1013699998>
20. Finner, H., Dickhaus, T., Roters, M.: Dependency and false discovery rate: asymptotics. *Ann. Statist.* **35**(4), 1432–1455 (2007). <https://doi.org/10.1214/0090536070000000046>
21. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* **40**(1), 120–145 (2011). <https://doi.org/10.1007/s10851-010-0251-1>
22. Chambolle, A.: An algorithm for total variation minimization and applications. *J. Math. Imaging Vision* **20**(1-2), 89–97 (2004). <https://doi.org/10.1023/B:JMIV.0000011320.81911.38>. Special issue on mathematics and image analysis
23. Frick, K., Marnitz, P., Munk, A.: Statistical multiresolution Dantzig estimation in imaging: fundamental concepts and algorithmic framework. *Electron. J. Stat.* **6**, 231–268 (2012). <https://doi.org/10.1214/12-EJS671>
24. Aspelmeier, T., Egner, A., Munk, A.: Modern statistical challenges in high-resolution fluorescence microscopy. *Annu. Rev. Stat. Appl.* **2**, 163–202 (2015)
25. Frick, K., Marnitz, P., Munk, A.: Statistical multiresolution estimation for variational imaging: with an application in Poisson-biophotonics. *J. Math. Imaging Vision* **46**(3), 370–387 (2013). <https://doi.org/10.1007/s10851-012-0368-5>
26. Li, H.: Variational estimators in statistical multiscale analysis. Ph.D. thesis, Georg-August-Universität Göttingen (2016)
27. Nemirovski, A.: Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Tekhn. Kibernet.* (in Russian) **3**, 50–60 (1985). *J. Comput. System Sci.*, 23:1–11, 1986 (in English)
28. Grasmair, M., Li, H., Munk, A.: Variational multiscale nonparametric regression: smooth functions. *Ann. Inst. Henri Poincaré Probab. Stat.* **54**(2), 1058–1097 (2018). <https://doi.org/10.1214/17-AIHP832>
29. del Álamo, M., Li, H., Munk, A.: Frame-constrained total variation regularization for white noise regression (2018). arXiv preprint [arXiv:1807.02038](https://arxiv.org/abs/1807.02038)
30. Frick, K., Marnitz, P., Munk, A.: Shape-constrained regularization by statistical multiresolution for inverse problems: asymptotic analysis. *Inverse Probl.* **28**(6), 065,006, 31 (2012). <https://doi.org/10.1088/0266-5611/28/6/065006>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

