The Block Criterion for Multiscale Inference About a Density, With Applications to Other Multiscale Problems

Kaspar RUFIBACH and Guenther WALTHER

The use of multiscale statistics, that is, the simultaneous inference about various stretches of data via multiple localized statistics, is a natural and popular method for inference about, for example, local qualitative characteristics of a regression function, a density, or its hazard rate. We focus on the problem of providing simultaneous confidence statements for the existence of local increases and decreases of a density and address several statistical and computational issues concerning such multiscale statistics. We first review the benefits of employing scale-dependent critical values for multiscale statistics and then derive an approximation scheme that results in a fast algorithm while preserving statistical optimality properties. The main contribution is a methodology for calibrating multiscale statistics that does not require a case-by-case derivation of its specific form. We show that in the above density context the methodology possesses statistical optimality properties and allows for a fast algorithm. We illustrate the methodology with two further examples: a multiscale statistic introduced by Gijbels and Heckman for inference about a hazard rate and local rank tests introduced by Dümbgen for inference in nonparametric regression.

Code for the density application is available as the R package modehunt on CRAN. Additional code to compute critical values, reproduce the hazard rate and local rank example and the plots in the paper as well as datasets containing simulation results and an appendix with all the proofs of the theorems are available online as supplemental material.

Key Words: Fast algorithm; Local increase in a density; Multiscale test.

1. INTRODUCTION

There has been considerable recent interest in the inference about qualitative characteristics of a regression function, a density, or its hazard rate, such as the number or location of monotone or convex regions, local extrema, or inflection points. As the location and extent of these local characteristics are not known in advance, it is natural to em-

Kaspar Rufibach is Lecturer in Biostatistics, Institute of Social- and Preventive Medicine, University of Zurich, Switzerland (E-mail: *kaspar.rufibach@ifspm.uzh.ch*). Guenther Walther is Associate Professor, Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305 (E-mail: *gwalther@stanford.edu*).

*Journal of Computational and Graphical Statistics, Volume 19, Number 1, Pages 175–190
DOI: 10.1198/jcgs.2009.07071*

ploy multiscale methods for such an inference; that is, one simultaneously examines local regions of various sizes and locations. This approach was used by Chaudhuri and Marron (1999, 2000) and Dümbgen and Spokoiny (2001) in conjunction with kernel estimates with varying bandwidths, by Dümbgen (2002) and Rohde (2008) with local rank tests, by Hall and Heckman (2000) with local linear smoothers, by Ganguli and Wand (2004) with local splines, and by Gijbels and Heckman (2004) and Dümbgen and Walther (2008) with local spacings.

For a concise exposition we will focus our theoretical investigations on the problem of inference about a density, as did Dümbgen and Walther (2008). But it will become clear that the methodology introduced in this article can be adapted to the other contexts cited above, and it may be relevant for multiscale methods beyond the area of shape-restricted inference. We illustrate this by applying our methodology in two other contexts: Section 5 considers a multiscale statistic that was recently introduced by Gijbels and Heckman (2004) for inference about a hazard rate, and Section 6 considers local rank tests that were introduced by Dümbgen (2002) for inference in nonparametric regression.

Our main focus is thus to detect and localize local increases (or decreases) of a univariate density f based on a vector $\mathbf{X}$ of iid observations $X_1, \ldots, X_n$. The nature of the problem suggests considering local test statistics on multiple intervals and then performing a simultaneous inference with these multiple tests. This general program of multiscale inference was implemented in this context by Dümbgen and Walther (2008) as follows:

Consider all intervals $\mathcal{I}_{jk} := (X_{(j)}, X_{(k)})$, $1 \le j < k-1 \le n-1$, and on each such interval $\mathcal{I}_{jk}$ compute the local test statistic $T_{jk}(\mathbf{X}) := \sqrt{\frac{3}{k-j-1}} \sum_{i=j+1}^{k-1} (2 \frac{X_{(i)}-X_{(j)}}{X_{(k)}-X_{(j)}} - 1)$. If f is nonincreasing on $\mathcal{I}_{jk}$, then one obtains the deterministic inequality $T_{jk}(\mathbf{X}) \le T_{jk}(\mathbf{U})$, where $\mathbf{U}$ is the vector of $U[0,1]$ random variables $U_i := F(X_i)$, $1 \le i \le n$. Thus we can conclude with confidence $1-\alpha$ that f must have an increase on every interval $\mathcal{I}_{jk}$ for which $T_{jk}(\mathbf{X})$ exceeds a critical value $c_{jk}(\alpha)$ that can be simulated with $U[0,1]$ random variables. More precisely: With confidence $1-\alpha$ one can claim that f must have an increase on every $\mathcal{I}_{jk} \in \mathcal{D}^+(\alpha)$, where

$$\mathcal{D}^+(\alpha) := \{\mathcal{I}_{jk} : T_{jk}(\mathbf{X}) > c_{jk}(\alpha)\},$$

provided that

$$\mathbb{P}\Big\{|T_{jk}(\mathbf{U})| \le c_{jk}(\alpha) \text{ for all } 1 \le j < k-1 \le n-1\Big\} \ge 1-\alpha,$$

and furthermore f must have a decrease on every $\mathcal{I}_{jk}$ for which $T_{jk}(\mathbf{X}) < -c_{jk}(\alpha)$; see Remark 1 in Section 8. Note that this approach yields a guaranteed finite sample simultaneous confidence level $1-\alpha$ for the above statements.

A central problem of the multiscale inference is the choice of the local critical values $c_{jk}(\alpha)$. The traditional approach to this problem is to treat all of the local test statistics as equal, that is, one sets $c_{jk}(\alpha) := \tilde{\kappa}_n(\alpha)$, where $\tilde{\kappa}_n(\alpha)$ is chosen, for example, via Monte Carlo, such that

$$\mathbb{P}\Big\{\max_{1 \le j < k-1 \le n-1} |T_{jk}(\mathbf{U})| \le \tilde{\kappa}_n(\alpha)\Big\} \ge 1-\alpha. \tag{1.1}$$

(Of course, such an equal treatment requires that all local statistics are first standardized to the same mean and variance; $T_{jk}(\mathbf{U})$ has mean 0 and variance unity for all (j,k).) It can be shown that for these critical values $\tilde{\kappa}_n(\alpha) \sim \sqrt{2\log n}$.

Dümbgen and Spokoiny (2001) pioneered an approach that assigns different critical values to different scales $\frac{k-j}{n}$. In the present context their method amounts to setting $c_{jk}(\alpha) := \sqrt{2 \log \frac{en}{k-j}} + \kappa_n(\alpha)$, that is, $\kappa_n(\alpha)$ is chosen via Monte Carlo such that

$$\mathbb{P}\left\{ \max_{1 \le j < k-1 \le n-1} \left(|T_{jk}(\mathbf{U})| - \sqrt{2 \log \frac{en}{k-j}} \right) \le \kappa_n(\alpha) \right\} \ge 1 - \alpha. \tag{1.2}$$

A motivation for this choice is as follows: There are $\sim \frac{1}{h}$ disjoint intervals $\mathcal{I}_{jk}$ of 'length' $h = \frac{k-j}{n}$. As the distribution of $T_{jk}(\mathbf{U})$ is roughly standard normal, $\max_{j,k} T_{jk}(\mathbf{U})$ over these intervals will be of size $\sqrt{2 \log \frac{1}{h}}$. Intervals of this length that overlap with those disjoint intervals will result in local statistics $T_{jk}(\mathbf{U})$ that are correlated and will not affect the overall maximum in a relevant way. Thus $\max_{j,k} T_{jk}(\mathbf{U})$ over small intervals $\mathcal{I}_{jk}$ with $k - j \le const$ will be of the size $\sim \sqrt{2 \log n}$, whereas $\max_{j,k} T_{jk}(\mathbf{U})$ over large intervals $\mathcal{I}_{jk}$ with $k - j \ge const \cdot n$ will stay bounded. Consequently, in the traditional approach (1.1) the overall critical value $\tilde{\kappa}_n(\alpha)$ will essentially be determined by the stochastically larger null distribution at the small scales, with a corresponding loss of power at large scales. Method (1.2) counters this by first subtracting off $\sqrt{2 \log \frac{en}{k-j}}$, the putative size of $\max_{j,k} T_{jk}(\mathbf{U})$ on scale $\frac{k-j}{n}$, thus putting the various scales on a more equal footing.

This approach has strong theoretical support: As detailed in Section 2, it can be shown that this calibration leads to optimal large sample power properties for detecting increases and decreases on small scales *and* on large scales. In contrast, the traditional approach (1.1) will necessarily lead to a suboptimal performance except for signals on the smallest scale. One disadvantage of the calibration (1.2) is the fact that its particular form depends on the particular setup at hand. For example, likelihood ratio type statistics will require a different calibration, whose particular form must be derived from theoretical considerations that are nontrivial. The main contribution of this article is a method of calibration that is generally applicable without the need for such case-by-case specifications, which is simple to implement and which is shown to share the large sample optimality properties of the calibration (1.2) in the statistical context under consideration here.

We start with a small simulation study in Section 2 to investigate the effects of different calibrations in a finite sample context. Section 3 addresses a computational problem inherent in multiscale inference: There are of the order n^2 intervals $\mathcal{I}_{jk}$, and on each such interval a local test statistic needs to be computed. We will introduce a methodology to choose a particular subset of intervals that results in a total computational cost of $O(n \log n)$ while essentially retaining the optimal power properties. This efficient computational strategy provides the main ideas for the general method of calibration, which is introduced in Section 4 and is shown to combine computational efficiency with statistical optimality. In Section 5 we apply this methodology to a multiscale statistic introduced by Gijbels and Heckman (2004) for inference about a hazard rate, and in Section 6 we apply it to the local rank statistics introduced by Dümbgen (2002) for inference about a regression function. We summarize our findings in Section 7. Some computational details, remarks, and further illustrations are in Section 8. Proofs are deferred to a supplementary file that is available online.

2. CALIBRATING THE MULTISCALE STATISTIC

In this section we will investigate the effects of the different calibrations (1.1) and (1.2). Dümbgen and Walther (2008) showed that the relevant quantity for detecting an increase of the density f on an interval I is $H(f, I) := \inf_I f'|I|^2/\sqrt{F(I)}$, where $|I|$ denotes the length of I, and they established the following theorem for the calibration (1.2):

Theorem 1. *Let f_n be a density with distribution function F_n that satisfies $H(f_n, I_n) \geq C_n\sqrt{\log(e/F_n(I_n))/n}$ for a bounded interval I_n. Then*

$$\mathbb{P}_{f_n}(\mathcal{D}^+(\alpha) \text{ contains an interval } \mathcal{J} \subset I_n) \to 1,$$

provided that $C_n = \sqrt{24} + \frac{b_n}{\sqrt{\log(e/F_n(I_n))}}$ with $b_n \to \infty$.

Note that I_n and f_n may vary with n. This theorem allows us to deduce large sample optimality on small scales (i.e., intervals I_n with $F_n(I_n) \to 0$) as well as on large scales (intervals I_n with $\liminf F_n(I_n) > 0$):

Optimality for small scales. In this case we can take $C_n = \sqrt{24} + \epsilon_n$ for certain $\epsilon_n \to 0$ and there is a threshold effect for $H(f_n, I_n)$ at $\sqrt{24\log(e/F_n(I_n))/n}$: If the factor $\sqrt{24}$ is replaced by $\sqrt{24} + \epsilon_n$ for certain $\epsilon_n \to 0$, then the multiscale statistic will detect and localize the increase with power converging to 1. On the other hand, it can be shown that in the case $\sqrt{24} - \epsilon_n$ no procedure can detect such an increase with nontrivial asymptotic power.

Optimality for large scales. If $C_n \to \infty$, then the multiscale procedure will detect the increase with power converging to 1. It was shown by Dümbgen and Walther (2008) that $C_n \to \infty$ is also a necessary condition for any test to have asymptotic power 1.

This optimality result for small scales as well as for large scales supports the strategy (1.2), which employs larger critical values for smaller scales than for larger scales. In a finite sample context, this arrangement of critical values will simply shift power from the small scales to the large scales when compared to the traditional calibration (1.1). The above results show that as the sample size gets large, this disadvantage at small scales disappears, whereas the advantage at large scales persists, so then strategy (1.2) will dominate strategy (1.1). It is of interest to see from what sample size on this effect sets in, and what the trade-off in power looks like for smaller sample sizes.

We performed a simulation study for samples with $n = 200$, 1000, and 5,000 observations from a density that equals 1 on [0, 1] apart from a linear increase with slope s on an interval $[a, b]$: $f(x) = 1\{x \in [0, 1]\} + s(x - (a + b)/2)1\{x \in [a, b]\}$. To examine the power on a large scale we set $b - a = 1/2$ and as a small scale we took $b - a$ just large enough to get meaningful power, viz. 0.15, 0.07, and 0.03, respectively. In each simulation run, the interval $[a, b]$ was located randomly in [0, 1]. The finite sample critical values $\tilde{\kappa}_n(0.95)$ and $\kappa_n(0.95)$ were simulated with 10^5 Monte Carlo samples. Figure 1 shows the power of each method as a function of the slope parameter s. The relevant graphs are the dashed blue curve for the traditional method (1.1) and the dashed black curve for the method with the additive correction term (1.2).

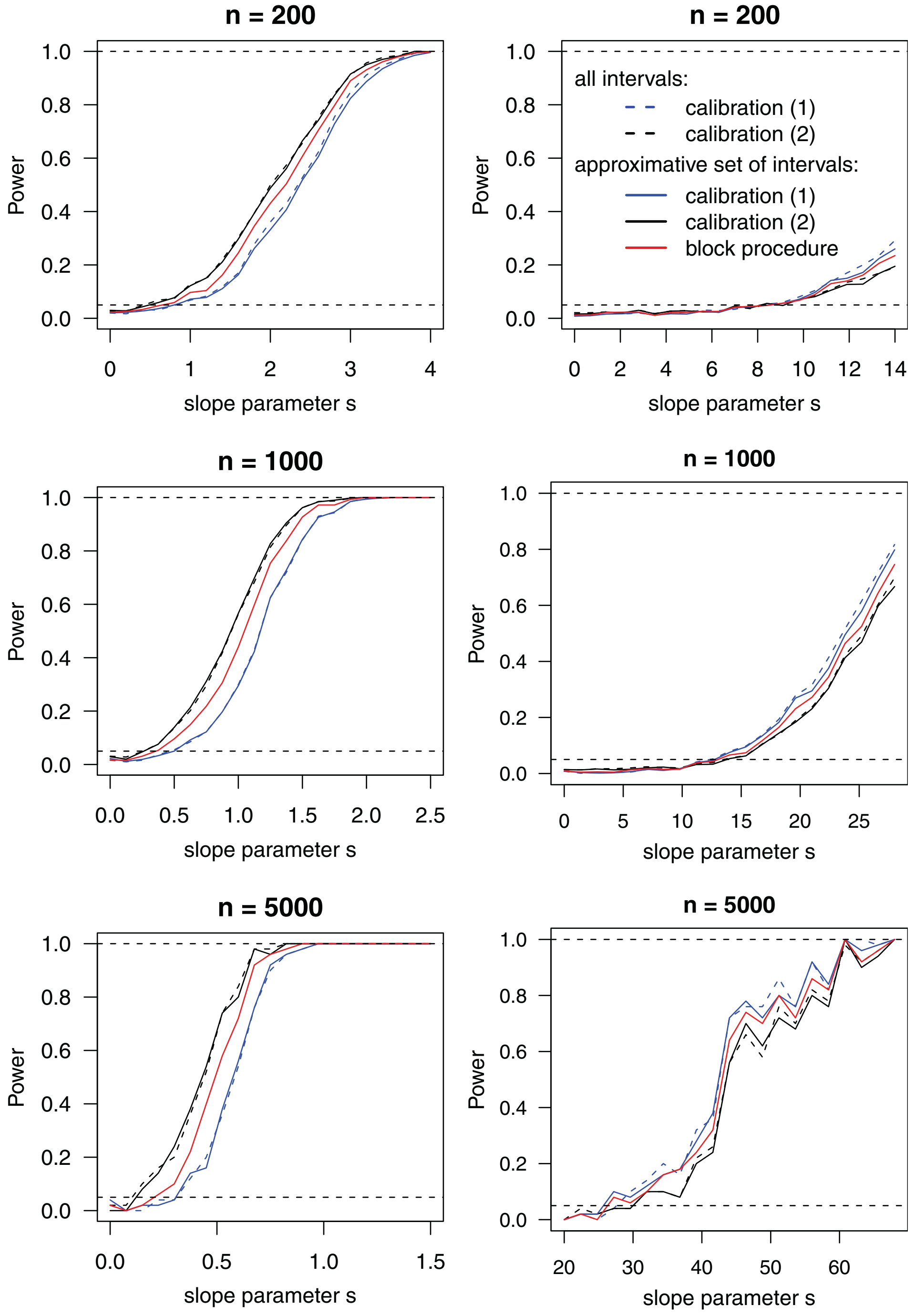

Figure 1. Power curves for sample sizes $n = 200$, 1000, and 5,000 for increases on a large scale (left) and on a small scale (right). Each curve is based on 1000 simulations at each of 20 lattice points and gives the proportion of simulations that produce an interval $\mathcal{I}_{jk} \in \mathcal{D}^{+}(0.05)$ with $\mathcal{I}_{jk} \cap [a, b] \neq \emptyset$.

The plots in Figure 1 show that the method with the additive correction term (1.2) has a clear advantage on the large scale, whereas the traditional method (1.1) has more power on the small scale under consideration. However, it turns out that this advantage extends only over a small part of the scale range: The scale $b - a$ above which method (1.2) has more power than (1.1) was found for the three sample sizes to be 0.25, 0.13, and 0.06, respectively. The plot with $n = 5{,}000$ shows the onset of the threshold effect described above.

Thus we conclude that for sample sizes in the hundreds, there is a trade-off in power between the two methods, with method (1.2) having more power on a large part of the scale range. For sample sizes in the thousands, this advantage extends to all but the smallest scales.

3. A FAST APPROXIMATION

Computing a local test statistic on intervals at various locations and sizes is computationally expensive: There are of the order n^2 intervals $\mathcal{I}_{jk}$, and on each such interval one has to compute a local test statistic. The idea for a fast but accurate approximation is based on the following observation: For large intervals, there is not much lost by considering only endpoints with indices on an appropriate grid, because the distance between potential endpoints will be small compared to the length of the interval (where distance and length are in terms of empirical measure). We will show how this idea can be finessed in a way that reduces the computational complexity in the above density case to $O(n \log n)$, while at the same time essentially retaining the optimality results with respect to power.

The algorithm can be described as follows: We start out by first considering as potential endpoints only every Dth observation, and we consider only intervals that contain between M and $2M - 1$ observations. Then we increase M to $2M$ and D to $\sqrt{2}D$ and iterate while $M \leq n/2$. This algorithm produces a sparse collection of intervals that approximates the collection of all intervals. The indices of the endpoints of these intervals lie on a grid that is finer for small intervals and coarser for larger intervals. Incrementing D only by a factor of $\sqrt{2}$ while the interval size is doubled results in an approximation loss that becomes negligible relative to the size of the interval and yields the optimal computational and statistical properties as detailed below; see also Remark 3.

Table 1 gives a more formal description of the algorithm.

Thus we set the notation such that $\mathcal{I}_{app}(1)$ contains the largest intervals, and $\mathcal{I}_{app}(l_{max})$ contains the smallest intervals. $\mathcal{I}_{app} := \bigcup_{l=1}^{l_{max}} \mathcal{I}_{app}(l)$ is then the collection of all intervals that we are considering for our approximation. We define $\mathcal{D}^{+}_{app}(\alpha)$ analogously to $\mathcal{D}^{+}(\alpha)$ with $\mathcal{I}_{app}$ in place of all intervals $\{\mathcal{I}_{jk}, 1 \leq j < k \leq n\}$.

Theorem 2. *$\mathcal{I}_{app}$ contains $O(n \log n)$ intervals. Moreover, Theorem 1 continues to hold when $\mathcal{D}^{+}(\alpha)$ is replaced by $\mathcal{D}^{+}_{app}(\alpha)$ provided $C_n = \sqrt{24} + \frac{b_n}{(\log(e/F_n(I_n)))^{1/4}}$ with $b_n \to \infty$.*

Table 1. Pseudo-code to enumerate the sets of intervals $\mathcal{I}_{app}(l), l = 1, \ldots, l_{max}$.

```
Set $D, M > 1$
$l_{max} \leftarrow \lfloor \log_2(n/M) \rfloor$
for $l = 1, \ldots, l_{max}$ do
    $\mathcal{I}_{app}(l) \leftarrow \{\}$
    $d_l \leftarrow \text{round}(D2^{(l_{max}-l)/2})$
    $m_l \leftarrow \text{round}(M2^{l_{max}-l})$
    Add all intervals $\mathcal{I}_{jk}$ to $\mathcal{I}_{app}(l)$ for which
    (a) $j, k \in \{1 + id_l,\ i = 0, 1, \ldots\}$ (we consider only every $d_l$th observation)
    and
    (b) $m_l \leq k - j - 1 \leq 2m_l - 1$ ($\mathcal{I}_{jk}$ contains between $m_l$ and $2m_l - 1$ observations)
end %for
```

Thus in the above density case the multiscale statistics on $\mathcal{I}_{app}$ can be computed in $O(n \log n)$ steps; see Remark 4. At the same time this procedure retains the statistical optimality properties on both large and small scales. The slightly different result of Theorem 2 compared to Theorem 1 affects only the secondary structure of the threshold effect at small scales, that is, the rate at which $C_n = \sqrt{24} + \epsilon_n \to \sqrt{24}$.

$\mathcal{I}_{app}$ will be a closer approximation to the collection of all intervals if the initial value of D is chosen smaller and the initial value of M is chosen larger (as this results in fewer iterations of the algorithm that increase D). We found that $D = 2$ and $M = 10$ yields a very good approximation for a range of sample sizes, as illustrated in Figure 1: The relevant power curves are the solid blue and black lines, which have to be compared to the respective dashed lines. Thus we use $D = 2$ and $M = 10$ in the following. Section 8 provides further simulation results that illustrate how different choices of D and M affect the approximation.

4. THE BLOCK CRITERION

Section 2 has shown that employing different critical values for different scales can result in advantageous statistical properties in the above density context. Therefore, it is worthwhile to explore such a calibration in other settings. One disadvantage of the method (1.2) is that the form of the correction term depends on the particular situation at hand, namely on the tail behavior of the local statistics as well as on a certain entropy and the behavior of the increments of a certain stochastic process; see theorem 7.1 in the article by Dümbgen and Walther (2008). Deriving these properties is typically a nontrivial task. It is thus of interest to develop methodology that does not require these case-by-case specifications.

The motivation for our methodology derives from the above computational considerations that group intervals into blocks: As each block contains intervals of about the same length (scale), we will assign to each such interval the same critical value. Then we set these critical values such that the significance level of the lth block decreases as $\sim l^{-2}$.

More formally, in the above density setting let $\alpha \in (0, 1)$ and define $q_l(\alpha)$ to be the $(1 - \alpha)$-quantile of $\max_{\mathcal{I}_{jk} \in \mathcal{I}_{app}(l)} |T_{jk}(\mathbf{U})|$. We suppress the dependence of $q_l(\alpha)$ on the sample size n for notational simplicity. Let $\tilde{\alpha}$ be the largest number such that

$$\mathbb{P}\left(\bigcup_{l=1}^{l_{max}}\left\{\max_{\mathcal{I}_{jk}\in\boldsymbol{\mathcal{I}}_{app}(l)}|T_{jk}(\mathbf{U})| > q_l\left(\frac{\tilde{\alpha}}{(A+l)^2}\right)\right\}\right) \le \alpha, \tag{4.1}$$

where $A \ge 0$; we found that $A := 10$ works well in practice and we use this choice in the following. Section 8 shows that the critical values $q_l(\frac{\tilde{\alpha}}{(A+l)^2})$ can be simulated with a simple extension of the algorithm used for methods (1.1) and (1.2).

Now we can define $\mathcal{D}^+_{block}(\alpha)$ analogously to $\mathcal{D}^+(\alpha)$ by taking as critical value in the lth block $q_l(\frac{\tilde{\alpha}}{(A+l)^2})$. By construction, we can again claim with guaranteed finite sample simultaneous confidence $1-\alpha$ that f must have an increase on every $\mathcal{I}_{jk} \in \mathcal{D}^+_{block}(\alpha)$. The next theorem shows that in the setup under consideration here, we obtain the same statistical optimality properties as for the method (1.2).

Theorem 3. *Theorem 1 continues to hold when $\mathcal{D}^+(\alpha)$ is replaced by $\mathcal{D}^+_{block}(\alpha)$ provided $C_n = \sqrt{24} + \frac{b_n}{(\log(e/F_n(I_n)))^{1/4}}$ with $b_n \to \infty$.*

The proof of Theorem 3 shows that if we apply the block procedure to all intervals $\mathcal{I}_{jk}$, that is, we do not enforce (a) in Table 1, then we recover the stronger assertion of Theorem 1. Of course, in that case we would lose the computational efficiency that $\boldsymbol{\mathcal{I}}_{app}$ affords.

In Figure 1, the power curves of the block method are depicted by a solid red line, which has to be compared to the solid blue line of the traditional method (1.1) and the solid black line of the method (1.2) that uses an additive correction term. Thus one sees that in a finite sample context, the block method is intermediate between the other two methods. In particular, it gives more power to small scales than method (1.2), and we found this to be a desirable feature in many examples that we investigated. The increased power at small scales arises by construction: Whereas the significance level for the lth block can be shown to decrease exponentially as $\sim\exp(-c\sqrt{l})$ for method (1.2) (see Proposition 1 in the Supplemental materials), the block method employs the slower polynomial decrease $\sim l^{-2}$. Another reason for the better power at small scales of the block method is the fact that the critical values in each block automatically adapt to the exact finite sample distribution of the local test statistics.

The block calibration described in this section can be readily adapted to other settings. The next two sections explore how this calibration performs when applied to multiscale statistics that have recently been introduced for inference on hazard rates and for regression functions. A theoretical treatment of these cases is beyond the scope of this article, so we evaluate the performance with simulation studies.

5. INFERENCE ABOUT A HAZARD RATE

Gijbels and Heckman (2004) considered the problem of detecting a local increase in a hazard rate. They constructed a multiscale statistic by localizing a statistic introduced by Proschan and Pyke (1967): Let $X_1, \ldots, X_n$ be an iid sample from a distribution F whose left endpoint of support is 0. Consider the normalized spacings $D_i := (n-i+1)(X_{(i)} -$

Table 2. Proportion of rejections of the null hypothesis at the 5% significance level in 10,000 simulations for the hazard rate example.

	$n = 50$		$n = 250$	
	Method (1.1)	Block method	Method (1.1)	Block method
$a_1 = -0.2, \sigma = 0.1$	0.128	0.147	0.556	0.622
$a_1 = 0, \sigma = 0.2$	0.122	0.146	0.317	0.384

$X_{(i-1)})$, $i = 1, \ldots, n$, where $X_{(0)} := 0$. Gijbels and Heckman (2004) considered the local sign test

$$\max_{1 \le s \le n-1} \max_{1 \le k \le n-s} \mathcal{V}^*_{sk}, \tag{5.1}$$

where $\mathcal{V}^*_{sk} = v_k^{-1/2}(\sum_{s \le i < j \le s+k} V_{ij} - \mu_k)$, $V_{ij} = 1(D_i > D_j)$, $\mu_k = (k+1)k/4$, and $v_k = (2k+7)(k+1)k/72$. Thus s indexes the starting point and k indexes the bandwidth (scale) of the local statistic. Alternatively one can reparameterize the local statistic by start- and endpoint. Then (5.1) is seen to be equivalent to

$$\max_{1 \le j < k \le n} W_{jk}, \tag{5.2}$$

where $W_{jk} = v_{k-j}^{-1/2}(\sum_{j \le i < i' \le k} V_{ii'} - \mu_{k-j})$.

We can now apply the algorithm given in Table 1 and the block methodology described in Section 4 with W_{jk} in place of T_{jk}. Gijbels and Heckman (2004) showed that guaranteed finite sample significance levels can be obtained by simulating critical values using the standard exponential distribution for the X_i as null distribution for the null hypothesis of a constant failure rate.

We illustrate the methodology by repeating the simulation study of Gijbels and Heckman (2004). $n = 50$ observations were drawn from a distribution whose hazard rate h is modeled via $\log h(t) = a_1 \log t + \beta(2\pi\sigma^2)^{-1/2} \exp\{-(t-\mu)^2/(2\sigma^2)\}$, $t > 0$. Parameter values $a_1 \le 0, \beta = 0$ pertain to the null hypothesis of a nonincreasing failure rate, whereas $\beta > 0$ will result in a local increase for certain values of a_1, μ, σ. Gijbels and Heckman (2004) considered alternatives with $\beta = 0.3$, $\mu = 1$ and various values of a_1, σ. The values $a_1 = -0.2, \sigma = 0.1$ result in a local increase on a small scale, whereas the values $a_1 = 0, \sigma = 0.2$ result in a local increase on a large scale. Table 2 shows the power of the multiscale test (5.2) against these alternatives with the calibration (1.1) used by Gijbels and Heckman (2004) and the block method of Section 4.

6. LOCAL RANK TEST FOR NONPARAMETRIC REGRESSION

As a further example we apply our methodology to the local rank tests introduced by Dümbgen (2002) in the context of regression. Consider the standard nonparametric regression model $Y_i = f(x_i) + \varepsilon_i$, for $i = 1, \ldots, n$, with an unknown regression function f and independent random errors ε_i having continuous distribution function and mean zero. Denoting the ranks of Y_i among the numbers $Y_{j+1}, \ldots, Y_k$ by $R_{jk}(i)$, a local monotone trend

Table 3. Proportion of rejections of the null hypothesis at the 5% significance level for sample size $n = 800$ in 10,000 simulations for the regression example.

	Method (1.1)	Method (1.2)	Block method
$b - a = 0.02, c = 0.2$	0.353	0.169	0.355
$b - a = 0.5, c = 0.05$	0.916	0.976	0.943

of the observations $Y_{j+1}, \ldots, Y_k$ can be detected via the linear rank statistics

$$Z_{jk}(\mathbf{Y}) = \sum_{i=j+1}^{k} \beta\left(\frac{i-j}{k-j+1}\right) q\left(\frac{R_{jk}(i)}{k-j+1}\right)$$

for appropriate functions β and q on $(0, 1)$. For the case of the Wilcoxon Score function $\beta(x) = q(x) = 2x - 1$ it was shown by Dümbgen (2002) that the appropriately standardized local test statistic $|Z_{jk}(\mathbf{Y})|$ can be written as

$$\frac{6 \sum_{j<a<b\leq k} (b-a)\,\mathrm{sign}(Y_a - Y_b)}{(k-j)(k-j+1)\sqrt{k-j-1}}.$$

Dümbgen (2002) calibrated these local test statistics using calibration (1.2) with the additive correction factor $-\sqrt{2 \log \frac{n}{k-j}}$. The null distribution for constant f is obtained by simulation with uniform random variables in place of the Y_i.

We can now apply the algorithm given in Table 1 and the block methodology described in Section 4 with Z_{jk} in place of T_{jk}. We compared the block calibration with the calibrations (1.1) and (1.2) in a simulation study. We used the regression function $f_{a,b,c}(x) = c(x-a)/(b-a)1\{x \in [a, b]\}$ for $x \in [0, 1] \supseteq [a, b]$ with 800 equispaced design points on $[0, 1]$ and errors from a logistic distribution with $\mu = 0$ and $\sigma = 0.05$. For a given interval length $b - a$ the interval $[a, b]$ was randomized in $[0, 1]$. The results of the simulation are summarized in Table 3.

7. CONCLUSIONS

Employing a calibration for multiscale statistics that varies with scale can result in important improvements in terms of power. In the context of certain inferences about a density we described an approximation scheme that allows for an $O(n \log n)$ algorithm to compute an appropriate multiscale statistic while preserving statistical optimality properties. We introduced a general block method for calibrating multiscale statistics. This method has the advantage that its specification does not depend on the particular problem at hand. We investigated the performance of this block method in several settings. It was shown that the block method is computationally efficient and possesses statistical optimality properties for certain inferences about local increases and decreases of a density. We also applied the block method to two multiscale statistics that have recently been introduced for detecting local increases in a hazard rate and in a regression function. Simulation studies show that the block method produces favorable results in these settings.

All the methods described in this article are implemented in the `R` package `modehunt`, available from CRAN. The package also provides tables of critical values for some combinations of α and n as well as functions to simulate finite sample critical values.

8. COMPUTATIONAL DETAILS, REMARKS, AND FURTHER ILLUSTRATIONS

8.1 SIMULATING THE NULL DISTRIBUTION OF THE MULTISCALE STATISTIC

It was explained in Section 1 that the joint finite sample null distribution of the local statistics T_{jk} for the null hypothesis of a constant density can be obtained by Monte Carlo simulation using n iid $U[0,1]$ random variables $\mathbf{U} = (U_1, \ldots, U_n)$; see also Remark 1 below. After S Monte Carlo simulation runs the critical values $\tilde{\kappa}_n(\alpha)$ and $\kappa_n(\alpha)$ for methods (1.1) and (1.2) are taken as the $100(1-\alpha)$th percentile of the S simulated values of the respective $\max_{1 \le j < k-1 \le n-1}(\cdots)$ given in (1.1) and (1.2).

For method (4.1) we need to find l_{max} critical values $q_l(\tilde{\alpha}/(A+l)^2)$, $l = 1, \ldots, l_{max}$, for a given $\alpha \in (0,1)$. To this end, for each simulation run we record the max of each block in an $S \times l_{max}$ array $\mathcal{A}$ whose (s,l)th entry is $\max_{\mathcal{I}_{jk} \in \boldsymbol{\mathcal{I}}_{app}(l)} |T_{jk}(\mathbf{U})|$ for the sth simulation run. Next we sort the columns of $\mathcal{A}$ and store these in the array $\mathcal{B}$. This is done so that we can efficiently compute various percentiles of the different columns. Now the desired critical values are given by $q_l := \mathcal{B}(S - [(S-i)(A+1)^2/(A+l)^2], l)$, $l = 1, \ldots, l_{max}$, where i is the smallest integer $i \in \{1, \ldots, S\}$ such that the proportion of rows r of $\mathcal{A}$ for which $\sum_{l=1}^{l_{max}} 1(\mathcal{A}(r,l) > q_l) > 0$ is not larger than α. This index i can be quickly found via bisection.

We were initially concerned about the required number S of Monte Carlo simulation runs, for two reasons: first, we have to estimate several critical values q_l simultaneously; second, those critical values are further out in the tails. However, we found that over multiple sets of $S = 5 \cdot 10^5$ Monte Carlo simulations the standard error of these critical values was not larger than that for κ_n over multiple sets of $S = 10^5$ Monte Carlo simulations. We thus recommend $S = 5 \cdot 10^5$ Monte Carlo runs. The computing time is in the order of minutes for the examples in this article.

8.2 REMARKS

Remark 1: Finding local increases or decreases is a multiple testing problem, so an important issue is to justify the validity of the resulting confidence statement. Key to this are the *deterministic* inequalities $T_{jk}(\mathbf{X}) \le T_{jk}(\mathbf{U})$ if f is nonincreasing on $\mathcal{I}_{jk}$ ('$\ge$' if f is nondecreasing on $\mathcal{I}_{jk}$), where $U_i := F(X_i)$, $1 \le i \le n$. These inequalities yield

$$\begin{aligned}
&\mathbb{P}_f\big(T_{jk}(\mathbf{X}) \ge \tilde{c}_{jk}(\alpha) \text{ for some } 1 \le j < k-1 \text{ where } f \text{ is nonincreasing on } \mathcal{I}_{jk}\big)\\
&\quad \le \mathbb{P}\big(T_{jk}(\mathbf{U}) \ge \tilde{c}_{jk}(\alpha) \text{ for some } 1 \le j < k-1\big)\\
&\quad \le \alpha
\end{aligned}$$

provided the $\tilde{c}_{jk}(\alpha)$ are chosen such that $\mathbb{P}(T_{jk}(\mathbf{U}) \le \tilde{c}_{jk}(\alpha) \text{ for all } 1 \le j < k-1) \ge 1-\alpha$. Hence we can claim with finite sample confidence $1-\alpha$ that f must have an increase on every $\mathcal{I}_{jk}$ with $T_{jk}(\mathbf{X}) \geq \tilde{c}_{jk}(\alpha)$. Statements about increases *and* decreases require the control of $|T_{jk}(\mathbf{U})|$ with cutoffs $c_{jk}(\alpha)$ in place of $T_{jk}(\mathbf{U})$ and $\tilde{c}_{jk}(\alpha)$, as detailed in Section 1. Note that for the three forms of calibrations c_{jk} discussed in this article, $1-\alpha$ confidence statements about increases only (using the $\tilde{c}_{jk}(\alpha)$) remain valid at level $1-\alpha'$ with some $\alpha' \in (\alpha, 2\alpha)$ if the analysis concerns both increases and decreases (using $c_{jk}(\alpha)$).

Remark 2: A key point in establishing Theorem 1 is to show that $\kappa_n(\alpha)$ stays bounded in n, that is, after subtracting off $\sqrt{2\log(\cdots)}$ to adjust for multiple testing over different intervals on a given scale, there is no further adjustment necessary for combining the multiple scales.

Remark 3: It is shown below that consistent detection of an increase is possible only if the corresponding interval contains at least $\log n$ observations, that is, $m_l \geq \log n$, where the notation m_l is from Table 1. For these scales l one finds $d_l/m_l = C\sqrt{2^l/n} \leq C(\log n)^{-1/2}$, that is, the approximation error at the endpoints relative to the size of the interval becomes negligible as n increases.

As an alternative approximation scheme one can consider the univariate version of the multivariate algorithm given by Walther (2009). That algorithm also uses $m_l = \text{round}(n2^{-l})$, but $d_l = \text{round}(n2^{-l}l^{-1/2}/6)$. In that case $d_l/m_l = Cl^{-1/2}$, hence the approximation error relative to the size of the interval decreases with the size of the interval. The rate of decrease of d_l/m_l may be sufficient to establish statistical optimality as in Theorem 2. Furthermore, if one considers only intervals with $m_l \geq \log n$, then that approximation scheme can be shown to result in only $O(n)$ intervals. Of course, the computational complexity of $O(n\log n)$ cannot be improved because the data need to be sorted.

Remark 4: Theorem 2 establishes that $\mathcal{I}_{app}$ contains $O(n\log n)$ intervals. A naive computation of the local test statistics will add another factor n to the computational complexity for computing the local statistics on these $O(n\log n)$ intervals. But the particular statistic used here can be computed in constant time after computing the vector of cumulative sums of the observations once in the beginning, thus resulting in overall complexity of $O(n\log n)$.

8.3 FURTHER SIMULATIONS AND ILLUSTRATIONS

Figure 1 in Section 3 illustrated that for the choice $D=2$ and $M=10$, $\mathcal{I}_{app}$ provides a good approximation to the collection of all intervals in the context of the calibration (1.1), which does not use an additive correction term, and the calibration (1.2) with additive correction term. Figure 2 illustrates how the quality of the approximation changes for different values of D and M. The power curves in Figure 2 are for the same model as in Figure 1, but with a different sample size $n=500$ to avoid displaying redundant information. The plots on the top row show the performance of $\mathcal{I}_{app}$ for various choices of D and M with the calibration (1.1), so we are trying to approximate the dashed power curve. The plots

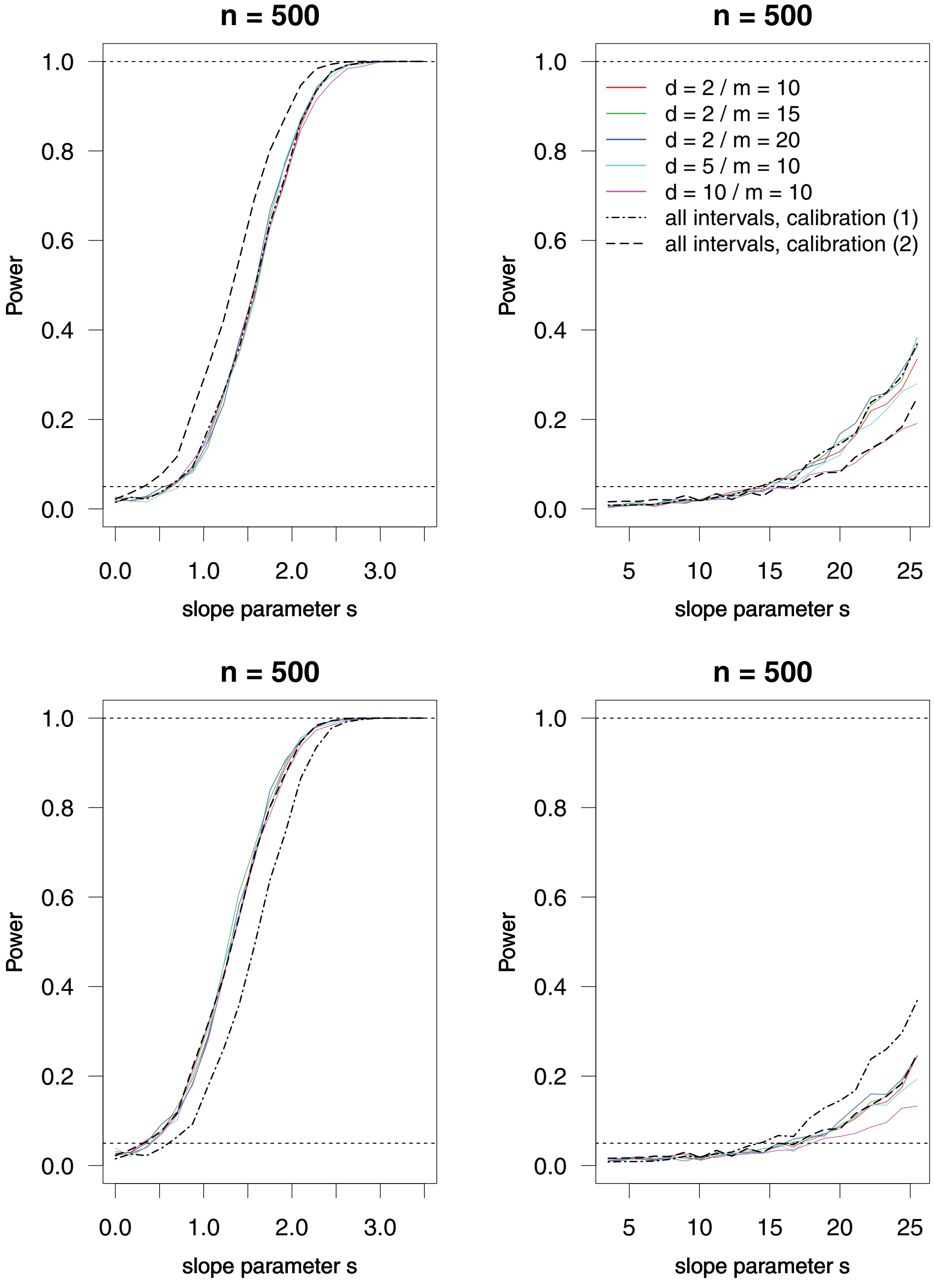

Figure 2. Power curves for the same model as in Figure 1 with various choices of D and M for $\mathcal{I}_{app}$. The left (right) plots show power for increases on a large (small) scale.

in the bottom row use the calibration (1.2), so the approximation is for the dashed-dotted curve. These plots show small gains if one would use $D = 2$ and $M = 20$ instead of $D = 2$ and $M = 10$, but this would come at the cost of a longer running time for the algorithm.

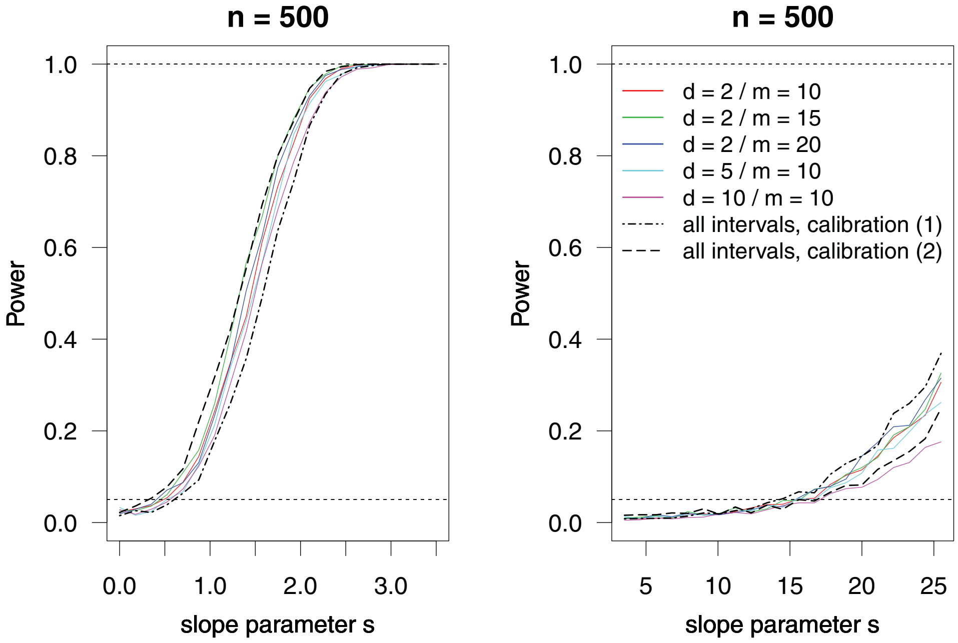

Figure 3. Power curves for the same model as in Figure 1 with various choices of D and M for the block procedure. Also shown are power curves for the calibration (1.1) and the calibration (1.2), both using all intervals, that is, no approximation. The left (right) plot shows power versus alternatives on a large (small) scale.

On the other hand, one sees that $D = 10$ and $M = 10$ is just too coarse an approximation as it leads to a significant loss of power in some cases. These simulations confirm $D = 2$ and $M = 10$ as an appropriate choice.

Figure 3 illustrates how various choices of D and M affect the block procedure. As in Figure 1 one sees that the power of the block procedure is typically intermediate between the calibration (1.1), which does not use an additive correction term, and the calibration (1.2) with additive correction term. An exception is the choice $D = 10$ and $M = 10$ which results in a large power loss due to the coarseness of the approximation, just as above. For alternatives on large scales the choice $D = 2$ and $M = 15$ or $M = 20$ results in more power than with $D = 2$ and $M = 10$. This is due to a finer approximation that comes at the cost of a longer running time of the algorithm.

To demonstrate the performance of the block procedure on a smooth density we considered simulations from $f(x) = pN(\mu, \sigma) + (1 - p)U[0, 1]$. Figure 4 gives the power curves for $n = 200$ and our default choice $D = 2$ and $M = 10$ when the increase is on a large scale ($\sigma = 0.05$, $p \in [0.01, 0.4]$) and on a small scale ($\sigma = 0.001$, $p \in [0.01, 0.075]$). We obtain the same qualitative behavior of the power curves as for the discontinuous density considered before: The power of the block procedure is intermediate between the calibration (1.1), which does not use an additive correction term, and the calibration (1.2) with additive correction term.

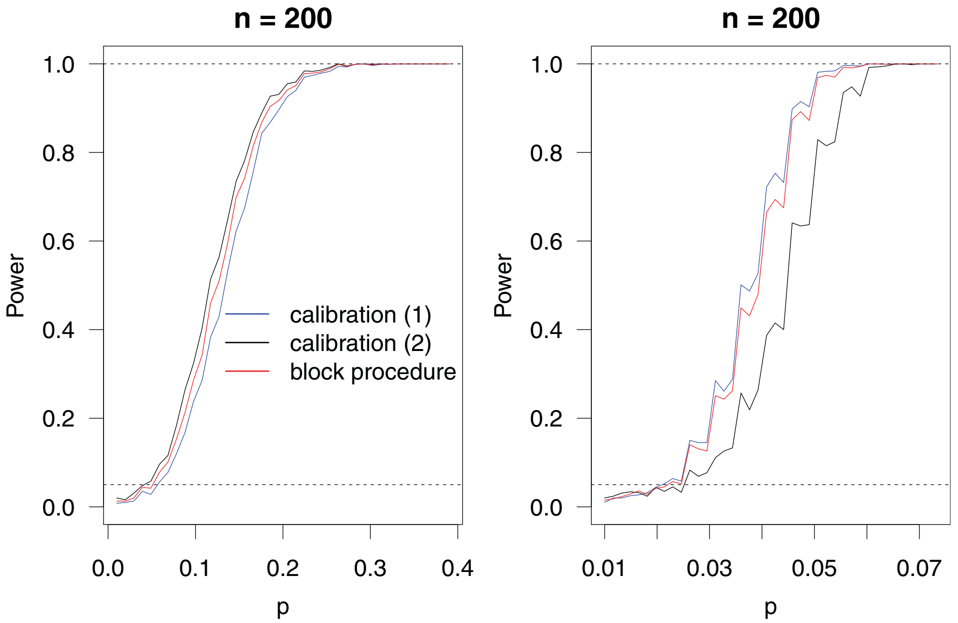

Figure 4. Power curves for the calibrations (1.1), (1.2), and the block procedure versus increases on a large scale (left) and on a small scale (right) in the case of a uniform density with a Gaussian bump.

SUPPLEMENTAL MATERIALS

Proofs and computer code: The supplement materials contain the proofs of all the theorems in the article (proof.pdf) and an R package, which implement all proposed methods. The package also contains tables and functions to simulate critical values. The file block_method_code.zip contains additional code to reproduce the hazard and local rank test examples and to generate the plots in the article. In addition, datasets containing the power curve data to reproduce the figures in the article are provided. (Supplemental Materials.zip)

ACKNOWLEDGMENTS

The work of Kaspar Rufibach was supported by the Swiss National Science Foundation. The work of Guenther Walther was supported by NSF grant DMS-0505682 and NIH grant 1R21AI069980.

[Received June 2007. Revised June 2009.]

REFERENCES

Chaudhuri, P., and Marron, J. S. (1999), "SiZer for the Exploration of Structures in Curves," *Journal of the American Statistical Association*, 94, 807–823. [176]

——— (2000), "Scale Space View of Curve Estimation," *The Annals of Statistics*, 28, 408–428. [176]

Dümbgen, L. (2002), "Application of Local Rank Tests to Nonparametric Regression," *Journal of Nonparametric Statistics*, 14, 511–537. [176,177,183,184]

Dümbgen, L., and Spokoiny, V. G. (2001), "Multiscale Testing of Qualitative Hypotheses," *The Annals of Statistics*, 29, 124–152. [176,177]

Dümbgen, L., and Walther, G. (2008), "Multiscale Inference About a Density," *The Annals of Statistics*, 36, 1758–1785. [176,178,181]

Ganguli, B., and Wand, M. P. (2004), "Feature Significance in Geostatistics," *Journal of Computational and Graphical Statistics*, 13, 954–973. [176]

Gijbels, I., and Heckman, N. (2004), "Nonparametric Testing for a Monotone Hazard Function via Normalized Spacings," *Journal of Nonparametric Statistics*, 16, 463–477. [176,177,182,183]

Hall, P., and Heckman, N. E. (2000), "Testing for Monotonicity of a Regression Mean by Calibrating for Linear Functions," *The Annals of Statistics*, 28, 20–39. [176]

Proschan, F., and Pyke, R. (1967), "Tests for Monotone Failure Rate," in *Proceedings of the Fifth Berkeley Symposium*, Vol. 3, Berkeley and Los Angeles: University of California Press, pp. 293–313. [182]

Rohde, A. (2008), "Adaptive Goodness-of-Fit Tests Based on Signed Ranks," *The Annals of Statistics*, 36, 1346–1374. [176]

Walther, G. (2009), "Optimal and Fast Detection of Spatial Clusters With Scan Statistics," *The Annals of Statistics*, to appear. [186]