

A general criterion for multiscale inference

Kaspar Rufibach
Department of Statistics
Stanford University (until August 31, 2007)
Supported by Swiss National Science Foundation
Joint work with Prof. Günther Walther, Stanford

IMSV, University of Bern
September 28, 2007

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

Goal: Inference about qualitative characteristics of a regression, density, or hazard function.

Number and **location** of monotone or convex regions, local extrema, etc.

Inference is **local** in nature.

Multiscale statistics: examine stretches of data at multiple locations and of multiple lengths.

General scheme in density estimation:

- X_1, \dots, X_n i.i.d. f , f a PDF ($\mathbf{X} = (X_{(1)}, \dots, X_{(n)})$).
- $\mathcal{I}_{jk} := (X_{(j)}, X_{(k)})$ for $1 \leq j < k \leq n$.
- Compute a suitable **local** test statistic $T_{jk}(\mathbf{X})$ on each \mathcal{I}_{jk} .

Crucial question: How to do simultaneous inference with the T_{jk} ?

Dümbgen and Walther (2006): Find **local de-/increases** of f .

Test statistic:

$$T_{jk}(\mathbf{X}) := \sum_{i=j+1}^{k-1} (2X_{(i;j,k)} - 1) 1\{X_{(i;j,k)} \in (0, 1)\}$$

for

$$X_{(i;j,k)} := \frac{X_{(i)} - X_{(j)}}{X_{(k)} - X_{(j)}}, j \leq i \leq k.$$

$\Rightarrow T_{jk}$ locally most powerful in a suitable parametric model.

Theorem (Worst case scenario)

If f non-increasing then $T_{jk}(\mathbf{X}) \leq T_{jk}(\mathbf{U})$, \mathbf{U} vector of uniform rv's.

Suppose we know “local critical values” c_{jk} such that

$$P\left\{|T_{jk}(\mathbf{U})|/\sigma_{jk} \leq c_{jk}(\alpha) \text{ for all } j, k\right\} \geq 1 - \alpha, \quad (1)$$

where $\sigma_{jk}^2 = \text{Var}(T_{jk}(\mathbf{U})) = \sqrt{(k-j-1)/3}$.

$$\mathbb{E}(T_{jk}(\mathbf{U})) = 0.$$

We can then conclude that f must have an increase on every \mathcal{I}_{jk} with

$$T_{jk}(\mathbf{X})/\sigma_{jk} > c_{jk}.$$

Holds with finite sample confidence $1 - \alpha$!

Again in troubles: How to find the c_{jk} 's such that (1) holds ?

Demonstrate example!

Traditional (straightforward) calibration

Treat all local test statistics equally, i.e. $c_{jk}(\alpha) = c(\alpha) =: \kappa_n^*(\alpha)$.

Critical value $\kappa_n^*(\alpha)$ found via Monte Carlo such that

$$P\left(\max_{j,k} |T_{jk}(\mathbf{U})|/\sigma_{jk} \leq \kappa_n^*(\alpha)\right) \geq 1 - \alpha.$$

Fancy calibration (Dümbgen and Spokoiny, 2001)

Set

$$c_{jk}(\alpha) = \kappa_n(\alpha) + \sqrt{2 \log \frac{en}{k-j}},$$

choose $\kappa_n(\alpha)$ via Monte Carlo such that

$$P\left(\max_{j,k} \left(|T_{jk}(\mathbf{U})|/\sigma_{jk} - \sqrt{2 \log \frac{en}{k-j}}\right) \leq \kappa_n(\alpha)\right) \geq 1 - \alpha.$$

Heuristic:

- Consider a fix scale (=interval length):
 \mathcal{I}_{jk} with length $\sim h = (k - j)/n$.
- We have $\sim 1/h$ such disjoint intervals.
- $T_{jk}^{\#}(\mathbf{U}) := T_{jk}(\mathbf{U})/\sigma_{jk} \sim \mathcal{N}(0, 1)$.
- $\max_{j,k} T_{jk}^{\#}(\mathbf{U}) \sim \sqrt{2 \log \frac{1}{h}}$ (max of $\mathcal{N}(0, 1)$'s!).
Overlapping \mathcal{I}_{jk} 's do not affect max.
- **Small scale:** $k - j \sim \text{const} \Rightarrow T_{jk}^{\#}(\mathbf{U}) \sim \sqrt{2 \log n}$.
- **Large scale:** $k - j \sim \text{const} \cdot n \Rightarrow T_{jk}^{\#}(\mathbf{U}) \sim \text{const}$.

Null distribution on small scales **stochastically larger** than that on large scales!

$$P\left(\max_{j,k} |T_{jk}^\#(\mathbf{U})| \leq \kappa_n^*(\alpha)\right) \geq 1 - \alpha.$$

$\kappa_n^*(\alpha)$ dominated by small scales \Rightarrow power loss for large scales!

Suboptimal on all but the very smallest scales.

$$P\left(\max_{j,k}\left(|T_{jk}^{\#}(\mathbf{U})| - \sqrt{2 \log \frac{en}{k-j}}\right) \leq \kappa_n(\alpha)\right) \geq 1 - \alpha.$$

$\kappa_n^*(\alpha)$ dominated by small scales \Rightarrow power loss for large scales!
Suboptimal on all but the very smallest scales.

By subtracting off

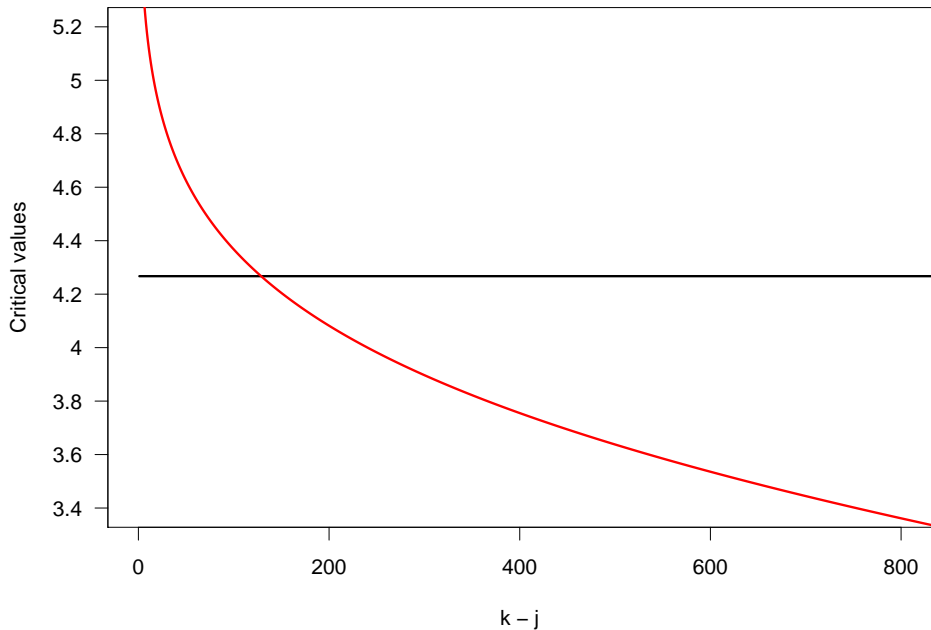
$$\sqrt{2 \log \frac{en}{k-j}} = \text{putative size of } \max_{j,k} T_{jk}^{\#}(\mathbf{U}) \text{ on scale } \frac{k-j}{n}$$

all scales are put on an “equal” footing

\Rightarrow larger critical values for smaller scales than for large scales

(“more difficult for small intervals to become significant”).

Critical values for $n = 1000$ and $\alpha = 0.05$



Theorem (Dümbgen and Walther (2006))

Let f_n be a PDF that has a (precisely specified) at-least-slope on a bounded interval I_n . Then the probability that the Dümbgen-Spokoiny calibrated procedure detects at least one interval $\mathcal{I}_{jk} \subset I_n$ with $T_{jk}(\mathbf{X}) > c_{jk}(\alpha)$ converges to 1.

⇒ **Minimax optimal** for detecting an increase of f_n on \mathcal{I}_{jk} both if the increase is on a

small: $\frac{k-j}{n} \rightarrow 0$ and on a

large: $\liminf \frac{k-j}{n} > 0$ scale.

⇒ Searching over all (i.e. large and small) scales does not incur drawback (at least asymptotically)!

Drawbacks of Dümbgen-Spokoiny calibration:

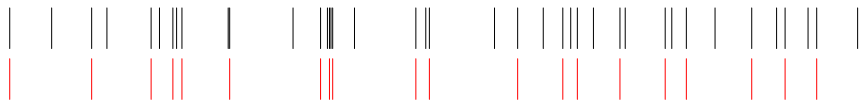
- Test statistic $T_{jk}(\mathbf{X})$ needs to be computed on $\#\mathcal{I}_{jk} \sim n^2$ intervals.
- Calibration term $\sqrt{2 \log \frac{en}{k-j}}$:
 - ⇒ depends on T_{jk} , i.e. the problem at hand.
 - ⇒ Highly non-trivial to derive: tail behavior of T_{jk} , behavior of increments $T_{jk} - T_{j'k'}$, certain entropy computations (Theorem 8 in Dümbgen and Walther, 2006).
- Simulations show: this calibration, being minimax optimal however, gives a lot of weight to large scales (“Überkompensation”).

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

\mathcal{I}_{app} : Approximation to $\mathcal{I}_{\text{all}} := \{\mathcal{I}_{jk} : 1 \leq j < k \leq n\}$

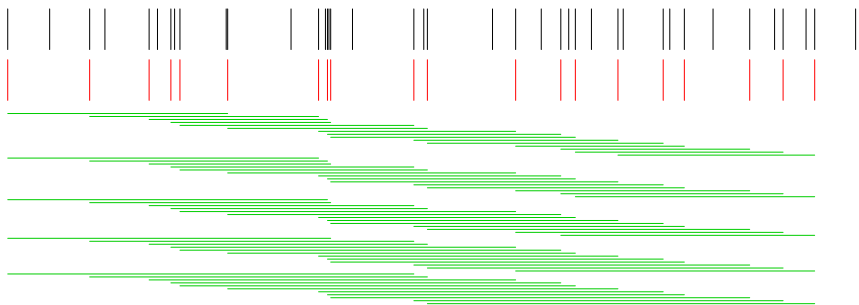


\mathcal{I}_{app} : Approximation to $\mathcal{I}_{\text{all}} := \{\mathcal{I}_{jk} : 1 \leq j < k \leq n\}$



\Rightarrow Take as interval endpoint only every d -th observation ($d_0 = 2$).

\mathcal{I}_{app} : Approximation to $\mathcal{I}_{\text{all}} := \{\mathcal{I}_{jk} : 1 \leq j < k \leq n\}$

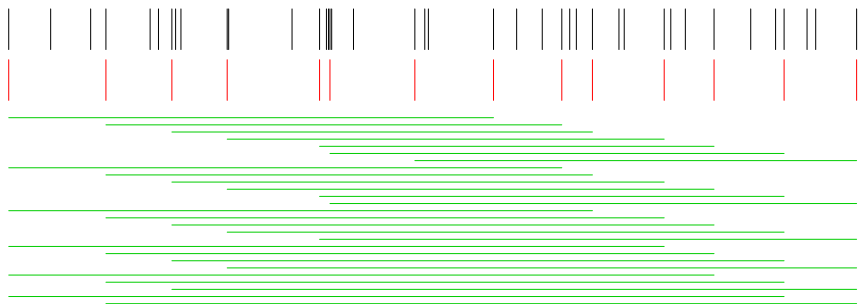


⇒ Take as interval endpoint only **every d -th** observation ($d_0 = 2$).

⇒ Only consider intervals on that grid that contain

#observations $\in \{m, \dots, 2m - 1\}$ ($m_0 = 10$).

\mathcal{I}_{app} : Approximation to $\mathcal{I}_{\text{all}} := \{\mathcal{I}_{jk} : 1 \leq j < k \leq n\}$



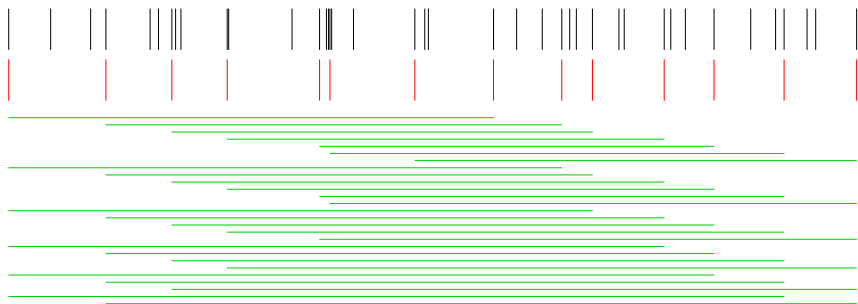
\Rightarrow Take as interval endpoint only every d -th observation ($d_0 = 2$).

\Rightarrow Only consider intervals on that grid that contain

#observations $\in \{m, \dots, 2m - 1\}$ ($m_0 = 10$).

\Rightarrow Set $d_1 = \sqrt{2}d_0 = 3$, $m_1 = 2m_0 = 20$.

\mathcal{I}_{app} : Approximation to $\mathcal{I}_{\text{all}} := \{\mathcal{I}_{jk} : 1 \leq j < k \leq n\}$



\Rightarrow Take as interval endpoint only every d -th observation ($d_0 = 2$).

\Rightarrow Only consider intervals on that grid that contain

$$\#\text{observations} \in \{m, \dots, 2m - 1\} \quad (m_0 = 10).$$

\Rightarrow Set $d_1 = \sqrt{2}d_0 = 3$, $m_1 = 2m_0 = 20$.

\Rightarrow Iterate while $m_k \leq n/2$.

Idea of the block algorithm

- Relative to interval length not much lost by considering interval endpoints on a fixed grid.
- Iteration increments: d by $\sqrt{2}$ and m by 2 \Rightarrow yields a negligible approximation loss!

Properties:

- $\log(n)$ iterations or **blocks**.
- Totally $O(n \log(n))$ intervals – compare to $O(n^2)$!

Theorem

Theorem of Dümbgen and Walther (2006) continues to hold when \mathcal{I}_{all} is replaced by \mathcal{I}_{app} (up to some peanuts). Yeah!

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion**
- ④ Power
- ⑤ Take home messages

Recall: Dümbgen-Spokoiny calibration employs **different critical values on different scales**.

Group intervals into blocks: intervals **within one block** roughly contain the **same number** of observations (up to factor of 2)!

Idea: assign the same critical value to intervals in a block.

Significance level in the i -th block: $(10 + i)^{-2}$,

$i = 1$: largest and $i = \log(n)$: smallest intervals.

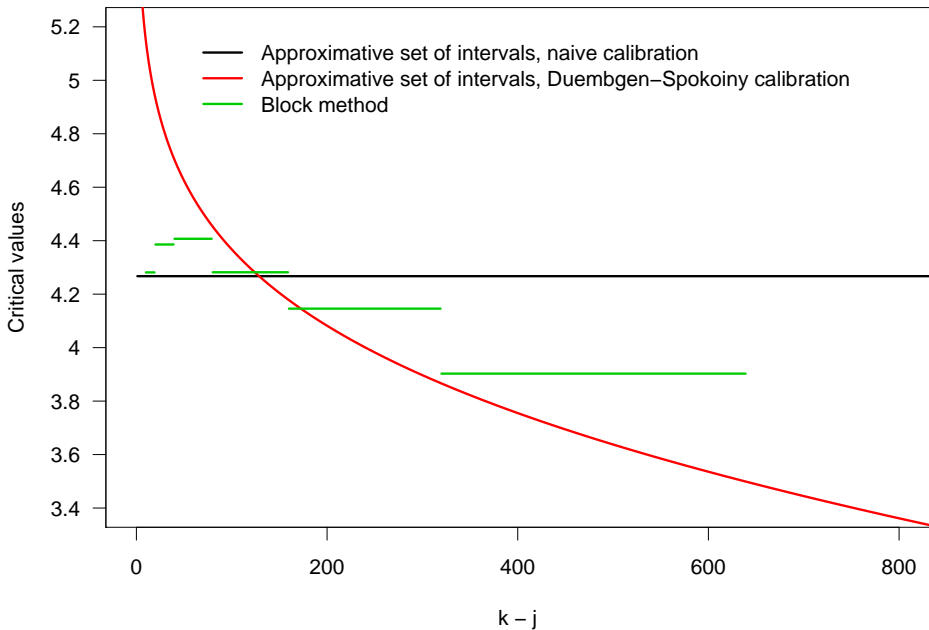
Formally: Let $q_i(\alpha)$ be the $(1 - \alpha)$ -quantile of $\max_{\mathcal{I}_{jk} \in \text{block } i} |T_{jk}(\mathbf{U})|$.

Then, let $\tilde{\alpha}$ be the largest number s.t.

$$P\left(\bigcup_{i=1}^{\#\text{blocks}} \left\{ \max_{\mathcal{I}_{jk} \in \text{block } i} |T_{jk}(\mathbf{U})| > q_i\left(\frac{\tilde{\alpha}}{(10 + i)^2}\right) \right\}\right) \leq \alpha.$$

$\tilde{\alpha}$ found via Monte Carlo.

Critical values for $n = 1000$ and $\alpha = 0.05$



Properties of this new calibration

Theorem

Theorem of Dümbgen and Walther (2006) continues to hold when the block procedure is applied either on \mathcal{I}_{all} or \mathcal{I}_{app} (up to some peanuts).

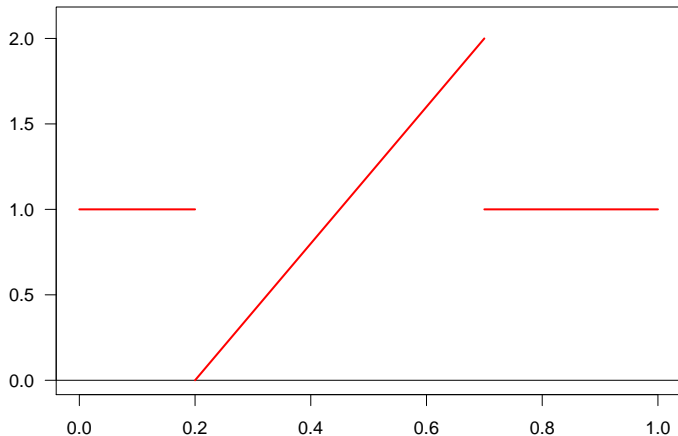
- No derivation of penalty term necessary.
- Meant to be applied on \mathcal{I}_{app} \Rightarrow **efficient algorithm**.
- Gives **more weight to smaller scales**: polynomial decrease in significance level vs. almost exponential decrease.

- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

Consider toy model

$$f_{a,b,s}(x) = 1\{x \in [0, 1]\} + s\left(x - \frac{a+b}{2}\right)1\{x \in [a, b]\}$$

where $0 \leq a < b \leq 1$, $|s| \leq 2/(b-a)$.



Power comparisons:

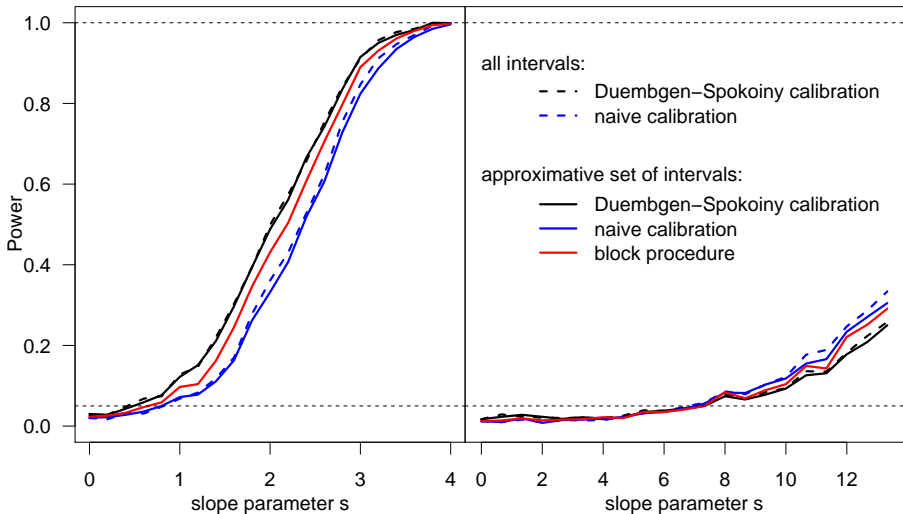
- Fix n and scale $b - a$: large scale = 0.5, small scale s.t. power curves meaningful.
- Set $s = 2/(b - a)$ (so biggest possible).
- In each simulation run: Randomly select $a \sim U[0, 1 - b]$, draw a sample \mathbf{X} from $f_{a,b,2/(b-a)}$.
- Do this 1'000 times, report proportion of simulations where a given method produced at least one interval \mathcal{I}_{jk} such that

$$\mathcal{I}_{jk} \cap [a, b] \neq \emptyset.$$

Power curves for $n = 200$

Large scale: $b-a = 0.5$

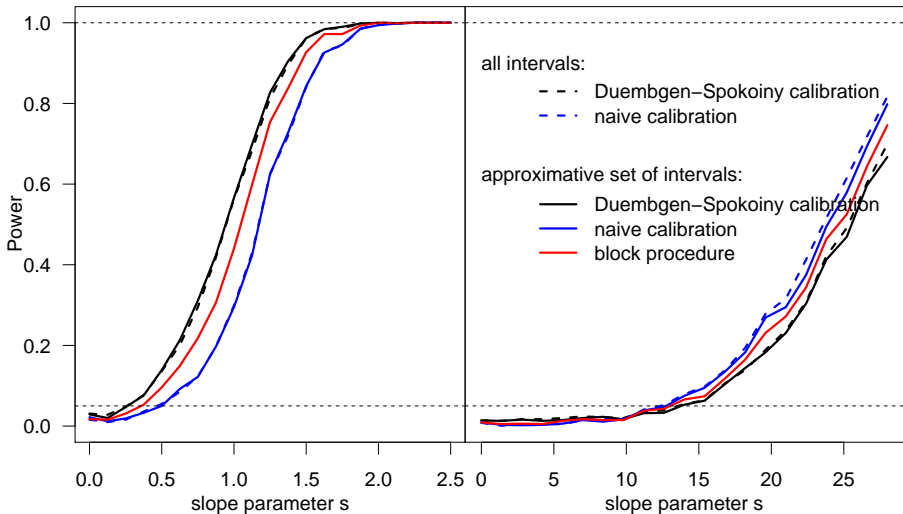
Small scale: $b-a = 0.15$



Power curves for $n = 1000$

Large scale: $b-a = 0.5$

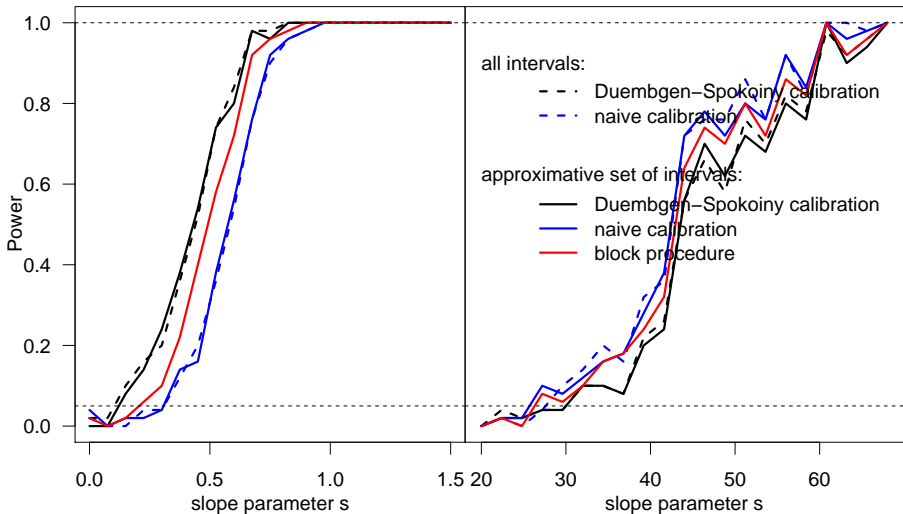
Small scale: $b-a = 0.07$



Power curves for $n = 5000$

Large scale: $b-a = 0.5$

Small scale: $b-a = 0.03$



- ① Multiscale inference
- ② Block method: approximate total set of intervals
- ③ Finally: the general criterion
- ④ Power
- ⑤ Take home messages

- Employing **different critical values on different scales** results in advantageous (asymptotical) statistical properties in mode hunting.
- These properties are **preserved** when replacing \mathcal{I}_{all} by \mathcal{I}_{app} .
- The block procedure is **computationally efficient**.
- No derivation of penalty term necessary.
- Block procedure trades off power on small **and** large scales.
- Same procedure applies to detection of hazard rate increase.

All the methods (incl. tables of critical values) are available in R-package `modehunt`, available from CRAN.