OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

MATH FACULTY OF MATHEMATICS

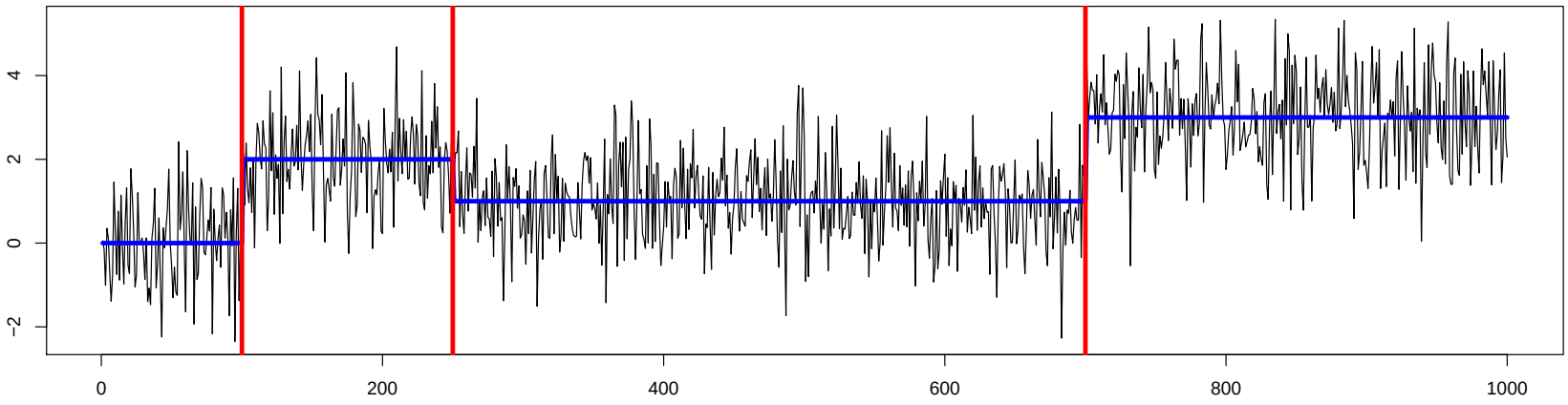
Data segmentation based on moving sum statistics[1]

Claudia Kirch
joint with
H. Cho (University of Bristol)
B. Eichinger (Karlsruhe Institute of Technology)
A. Meier (Otto-von-Guericke University Mageburg)

Bernoulli-IMS One World Symposium 2020

[1] implemented in the R-package mosum on CRAN

Multiple mean change model



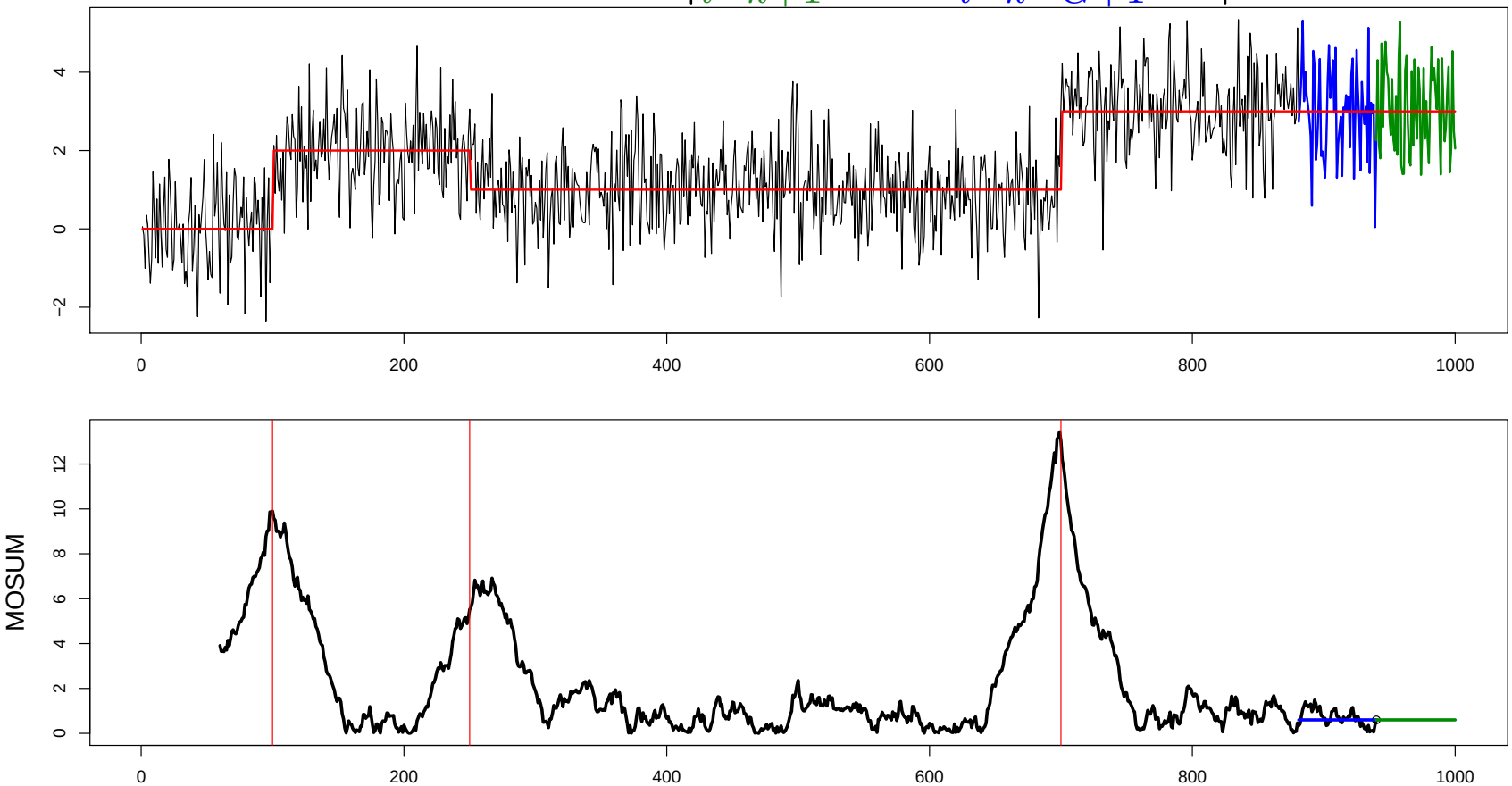
$$X_i = \begin{cases} \mu_{1,n} + \varepsilon_i, & i \leqslant k_{1,n} \\ \mu_{2,n} + \varepsilon_i, & k_{1,n} < i \leqslant k_{2,n} \\ \dots \\ \mu_{q_n+1,n} + \varepsilon_i & k_{q_n,n} < i \leqslant n \end{cases}$$

- $q_n \in \mathbb{N}$ - unknown number of change points
- $k_{1,n}, \dots, k_{q_n,n}$ - change points
- $\mu_{1,n}, \dots, \mu_{q_n+1,n} \in \mathbb{R}$ - expected values
- $\{\varepsilon_i : 1 \leqslant i \leqslant n\}$ -centered residual sequence

Moving sum statistics

For a given bandwidth G:

$$T_{k,n}(G) = \frac{1}{\sqrt{2G}} \left| \sum_{i=k+1}^{k+G} X_i - \sum_{i=k-G+1}^{k} X_i \right| .$$

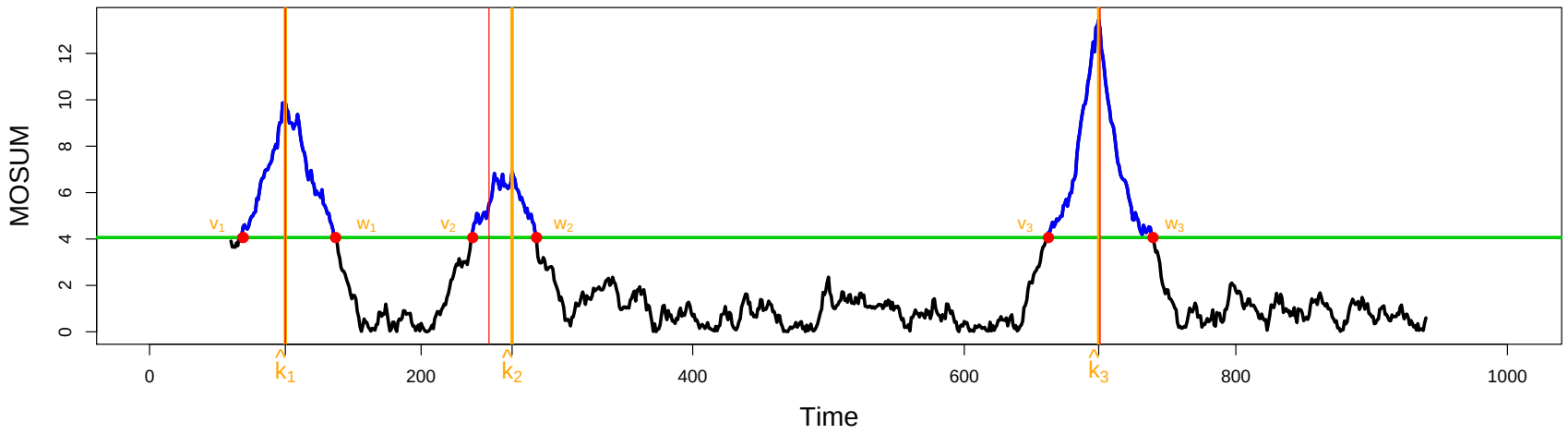


Linear computational complexity!

Change point estimation via moving sums

- Consider a threshold $D_n(G, \alpha)$ such that – if no change is present – $P(\max_{G \leqslant k \leqslant n-G} T_{k,n}(G) > \sigma\, D_n(G, \alpha)) \to \alpha$.
- Each point k^* with
  - $T_{k^*,n}(G) \geqslant \sigma\, D_n(G, \alpha)$
  - $T_{k^*,n}(G) \geqslant T_{k,n}(G)$ for all $|k - k^*| \leq \eta G$

  is an estimator for a change point.
- The unknown variance σ^2 can be replaced by a local estimator.



Homogeneous change points

$$\min_{1\le j\le q_n} (\mu_{j+1}-\mu_j)^2 \cdot \min_{1\le j\le q_n} (k_{j+1}-k_j) \to \infty$$

(minimal jump size)^2 · (minimal distance between CPs)

Then (under certain assumptions, e.g. $\alpha = \alpha_n \to 0$):

Single-scale mosum with appropriate bandwidth

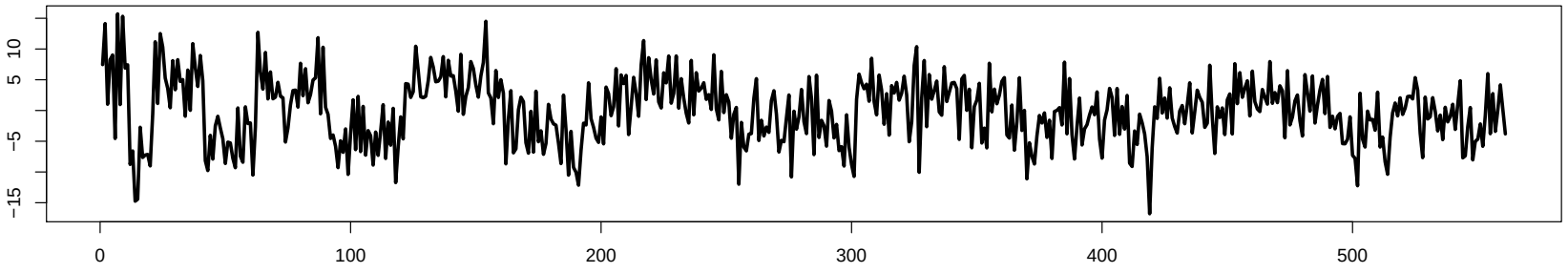
- yields consistent estimators for number and location of change points.
- achieves minimax optimal separation rate.
- achieves minimax optimal localisation even for an unbounded number of change points, in situations where such minimax-results are available.
- Generalizations beyond mean changes have been obtained (Reckrühm, 2019) including e.g. changes in count time series or (non-)linear regression.

Multiscale change points

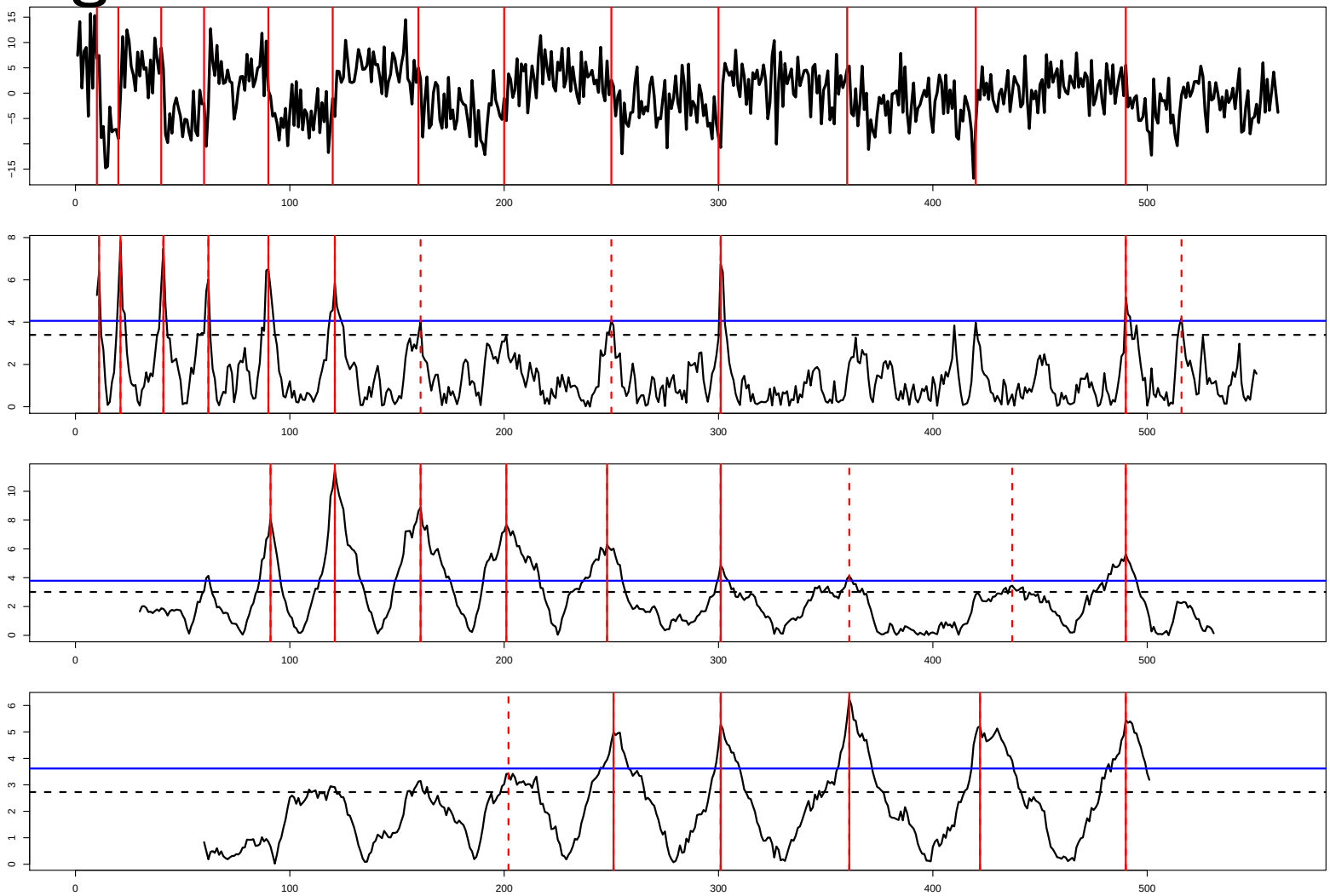
$$\min_{1\le j\le q_n}\left[(\mu_{j+1}-\mu_j)^2\cdot\min(k_{j+1}-k_j,k_j-k_{j-1})\right]\to\infty$$

(compare with homogeneous change points:
$\min_{1\le j\le q_n}(\mu_{j+1}-\mu_j)^2\cdot\min_{1\le j\le q_n}(k_{j+1}-k_j)\to\infty$)

Example: Mix-Signal:

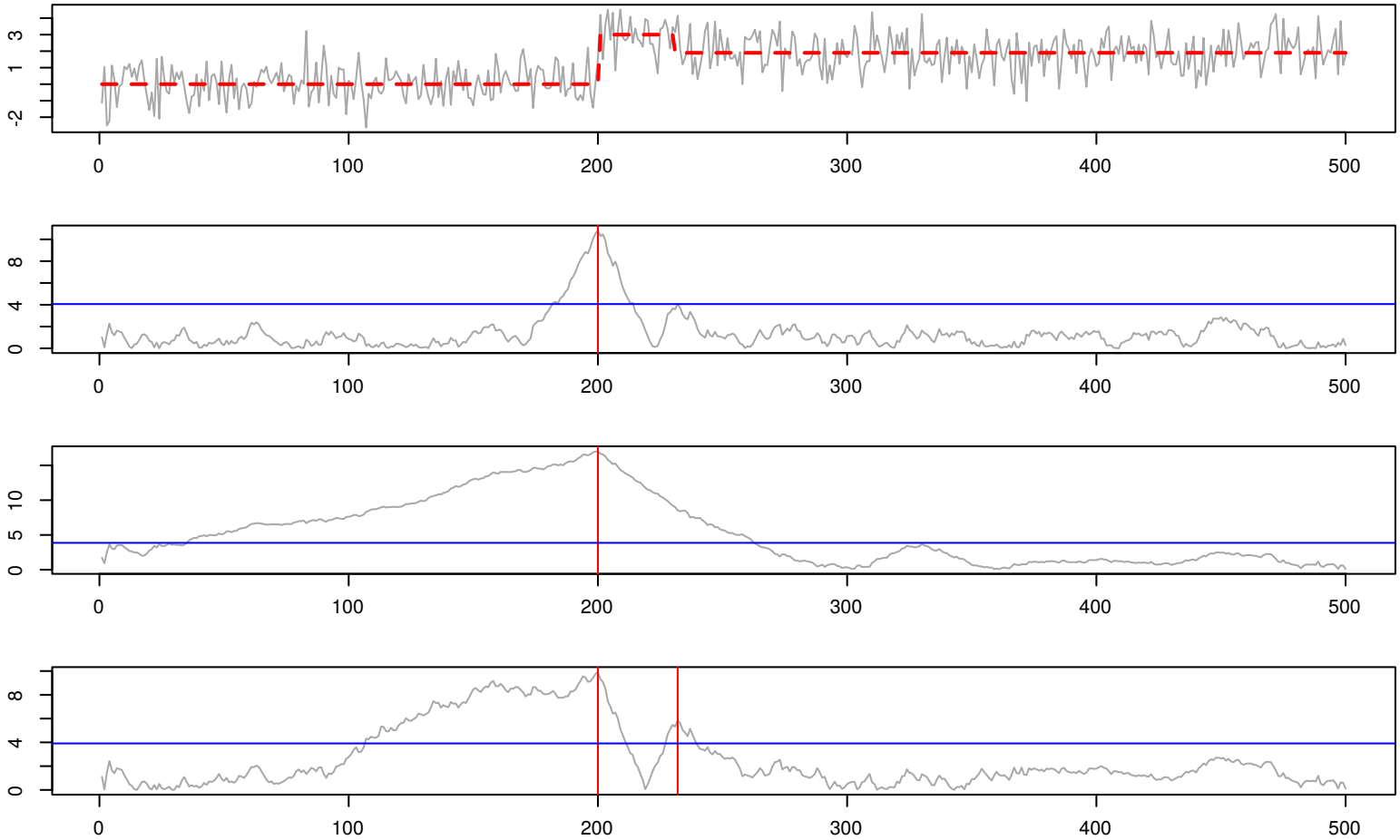


Mix-Signal



Bandwidths: $G = 10, 30, 60,$ Solid: $\alpha = 0.1$, Dashed: $\alpha = 0.5$

Asymmetric bandwidths



$(G = 30, 120, (30, 120))$

Second change point is only detected with asymmetric bandwidths.

Localized Pruning

Removing candidates obtained from multiple bandwidths by a *top down search*:

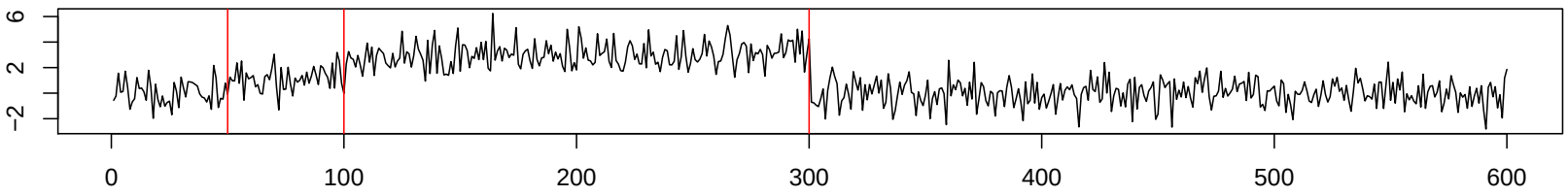
- Find sets: Adding further candidates, monotonically increases $\mathcal{SC}$ (Schwarz-like information criterion).
- Global procedure: Pick the one with smallest $\mathcal{SC}$ among those with smallest numbers of elements.
- Local procedure: Adaptation due to possible boundary effects.

Computational Speed:

- Usually truncation of the search space.
- Above properties are needed to prove consistency.

Localized pruning can be combined with other candidate-generating methods.

Usage example in R-package mosum




```
x <- testData(lengths = c(50, 50, 200, 300), means = c(0,
1, 3, 0),sds = rep(1, 4), seed = 123)

mlp <- multiscale.localPrune(x)

print(mlp$cpts); print(mlp$pooled.cpts)
```

Output:

```
 [1] 50 100 300
 [1] 29 43 47 48 50 51 53 89 94 96 100 101 300
```

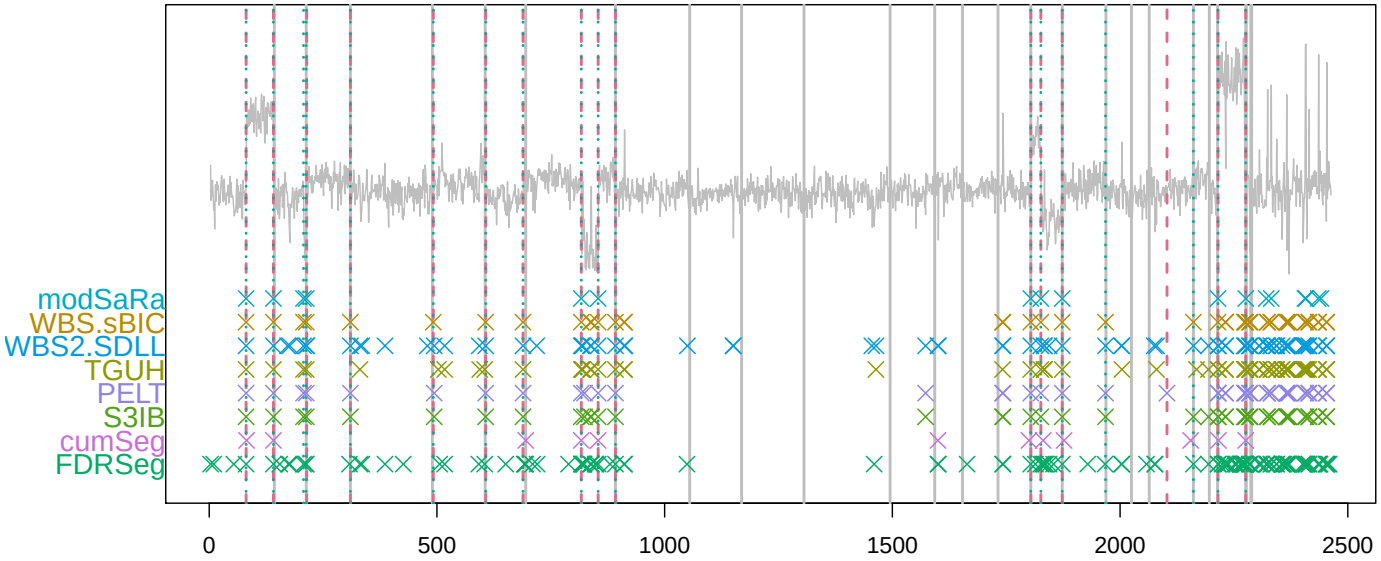
Localized pruning

- No bandwidth choice necessary.
- Extends consistency and optimality properties of single-scale mosum to the multiscale change point problem.
- Theoretic properties obtained under general assumptions permitting heavy tails and dependence.
- Competitive in terms of performance and run time.

Comparison for sub-Gaussian and sub-linear changes:

	Detection lower bound		Localisation		Computational	Beyond
Methodology	Multiscale	Rate	Multiscale	Rate	complexity	sub-Gaussianity
MoLP	✓	$\boldsymbol{\log(n)}$	✓	$\boldsymbol{\log(q_n)}$	$\boldsymbol{O(n\log(n))}$	✓
Chan and Chen (2017)	✗	$\log(n)$	✗	$\log(n)$	$O(n\log(n))$	✗
Single-scale MOSUM	✗	$\log(n)$	✓	$\log(q_n)$	$O(n)$	✓
Fromont et al. (2020)	✓	$\log(n)$	✓	$\log(q_n)$	$O(n^2)$	✗
Wang et al. (2018) (LSE)	✗	$\log(n)$	✓	$\log(n)$	$O(n^2)$	✗
Wang et al. (2018) (mWBS)	✗	$\log(n)$	✓	$\log(n)$	$O(nR_n)$	✗
Baranowski et al. (2019)	✗	$\log(n)$	✓	$\log(n)$	$O(nR_n)$	✗
Frick et al (2014)	✗	$\log(n)$	✗	$\log(n)$	$O(n^2)$	✓
Li et al. (2019)	✗	$q_n\log(n)$	✗	$q_n\log(n)$	–	✗
Fryzlewicz (2018)	✗	$\log^2(n)$	✗	$\log^2(n)$	$O(n\log^2(n))$	✗

DNA Data: Normalized copy number ratios:



Vertical solid lines:

Boundaries between chromosomes

Vertical broken lines:

Change point estimators from our pruning procedure,
dashed: MOSUM+locPrun, dotted: WBS+locPrun

Crosses: Different competing procedures.

Right end: Many procedures struggle with this variance change!

Literature

Eichinger, Kirch
A MOSUM procedure for the estimation of multiple random change points.
Bernoulli, 24:526-564, 2018.

Meier, Cho, Kirch
`mosum`: A package for moving sums in change point analysis.
To appear in *J. Stat. Soft.*[2]

Cho, Kirch
Two-stage data segmentation permitting multiscale change points, heavy tails and dependence.
Preprint arXiv:1910.12486v3, 2020.

Thank you very much for your attention!

[2] Preprint available at https://drive.google.com/file/d/1KSbGfx-sg6B1CcjN4i2RJSaXVejOA7bp.