Shape Constraints and Multiscale Methods for Density Estimation

Lutz Dümbgen,   Kaspar Rufibach

(University of Berne)

Günther Walther

(Stanford University)

I.   Estimating a Log–Concave Density

II.   Inference via Multiscale Methods

I. Estimating a Log–Concave Density

Consider order statistics

$$X_1 < X_2 < \cdots < X_n$$

of a random sample from unknown density f.

Shape constraint: f assumed log–concave,

$$f = \exp(\psi) \quad \text{with} \quad \psi : \mathbb{R} \to [-\infty, \infty) \text{ concave}.$$

Goal: Compute and analyze the NPMLE

$$\widehat{f} = \exp(\widehat{\psi}).$$

Why log–concavity?

- Many standard models satisfy this constraint, e.g.

$$\begin{aligned} \mathcal{N}(\mu, \sigma^2) & \\ \mathrm{Gamma}(a,b) &\quad (a \geq \mathbf{1}, b > 0) \\ \mathrm{Beta}(a,b) &\quad (a \geq \mathbf{1}, b \geq \mathbf{1}) \\ \mathrm{Weibull}(a,b) &\quad (a \geq \mathbf{1}, b > 0) \\ \mathrm{Gumbel} & \\ \ldots & \end{aligned}$$

- Log–concave densities are unimodal.

  As opposed to NPMLE of a unimodal density,

  - no trying out of many potential modes,
  - no “spiking” near the estimated mode.

Estimation of $\psi = \log f$

$$\widehat{F}_{\mathrm{emp}} \;:=\; n^{-1} \sum_{i=1}^{n} 1\{X_i \le \cdot\} \qquad \text{(empirical c.d.f.)}$$

$$\widehat{\psi} \;:=\; \mathop{\arg\max}_{\psi \text{ concave}} \left(\underbrace{\int \psi \, d\widehat{F}_{\mathrm{emp}}}_{\text{log-likelihood}} \;-\; \underbrace{\int \exp(\psi(x))\, dx}_{\text{Lagrange term}} \right)$$

Theorem 1 (Existence and uniqueness)

- $\widehat{\psi}$ exists and is unique,
- $\widehat{\psi}$ is piecewise linear and continuous on $[X_1, X_n]$ with knots only in $\{X_1, X_2, \ldots, X_n\}$,
- $\widehat{f} \equiv 0$ on $\mathbb{R} \setminus [X_1, X_n]$.

Characterisation and properties of $\widehat{\psi}$, $\widehat{f}$

By definition,

$$\frac{d}{dt}\bigg|_{t=0} \left(\int (\widehat{\psi} + t\Delta)\, d\widehat{F}_{\text{emp}} - \int \exp(\widehat{\psi}(x) + t\Delta(x))\, dx \right) \;\leq\; 0$$

whenever $\psi + t\Delta$ is concave for some $t > 0$.

Lemma

$$\int \Delta\, d\widehat{F}_{\text{emp}} \;\leq\; \int \Delta(x)\widehat{f}(x)\, dx$$

whenever $\psi + t\Delta$ is concave for some $t > 0$.

In addition to $\widehat{F}_{\text{emp}}$, $\widehat{\psi}$ and $\widehat{f}$ define

$$\widehat{F}(r) \ := \ \int_{-\infty}^{r} \widehat{f}(x)\, dx \, .$$

Setting $\Delta(x) := x$ or $\Delta(x) := -x^2$ in the previous Lemma yields:

Corollary 1

$$\mathbf{Mean}(\widehat{\boldsymbol{F}}) \ = \ \mathbf{Mean}(\widehat{\boldsymbol{F}}_{\mathbf{emp}}) \, ,$$

$$\mathbf{Var}(\widehat{\boldsymbol{F}}) \ \leq \ \mathbf{Var}(\widehat{\boldsymbol{F}}_{\mathbf{emp}}) \, .$$

Let

$$\widehat{\boldsymbol{\mathcal{S}}} \;:=\; \Big\{\text{knots of } \widehat{\boldsymbol{\psi}}\Big\} \quad \supset \; \{X_1, X_n\}\,.$$

Corollary 2

$$\int \boldsymbol{\Delta(x)}\, \widehat{\boldsymbol{F}}\boldsymbol{(dx)} \;=\; \int \boldsymbol{\Delta(x)}\, \widehat{\boldsymbol{F}}_{\mathbf{emp}}\boldsymbol{(dx)}$$

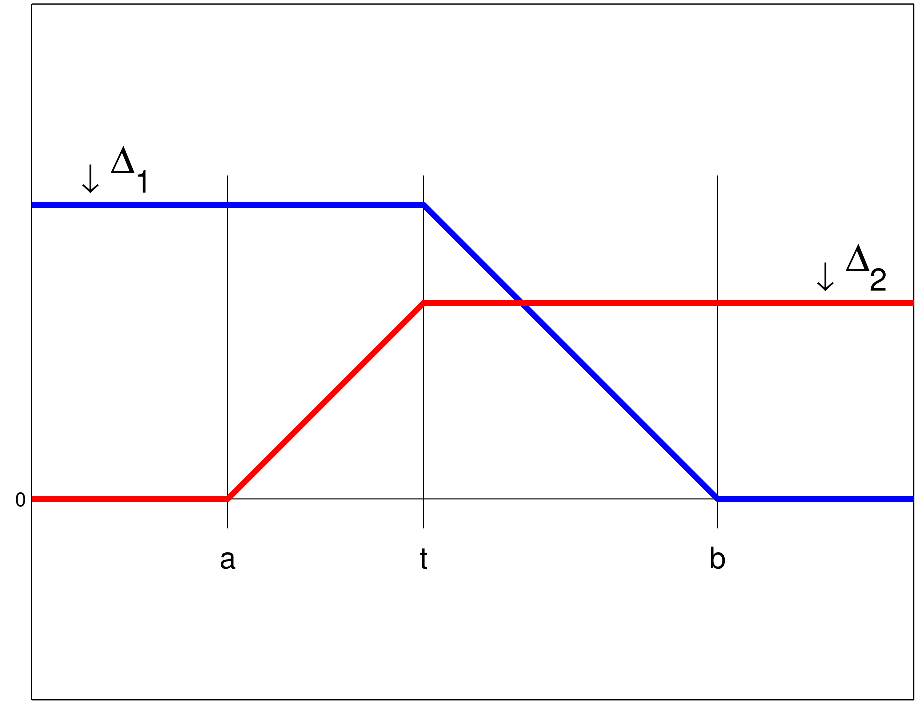
whenever $\Delta : \mathbb{R} \to \mathbb{R}$ is continuous and piecewise linear with knots only in $\widehat{\mathcal{S}}$.

Corollary 3 For $a < t < b$ with $a, b \in \widehat{\mathcal{S}}$,

$$\int_a^t \widehat{F}_{\mathrm{emp}}(x)\,dx \;\geq\; \int_a^t \widehat{F}(x)\,dx\,,$$

$$\int_t^b \widehat{F}_{\mathrm{emp}}(x)\,dx \;\leq\; \int_t^b \widehat{F}(x)\,dx\,,$$

$$\int_a^b \widehat{F}_{\mathrm{emp}}(x)\,dx \;=\; \int_a^b \widehat{F}(x)\,dx\,.$$



Corollary 4

$$\widehat{F}(X_1) = 0\,, \quad \widehat{F}(X_n) = 1$$

and

$$\widehat{\boldsymbol{F}} \in \left[\widehat{\boldsymbol{F}}_{\mathbf{emp}} - \boldsymbol{n}^{-1},\, \widehat{\boldsymbol{F}}_{\mathbf{emp}}\right] \quad \text{on } \widehat{\boldsymbol{\mathcal{S}}}\,.$$

Corollary 5

$$\left\|\widehat{\boldsymbol{F}} - \boldsymbol{F}\right\|_\infty \le \mathbf{3}\left\|\widehat{\boldsymbol{F}}_{\mathbf{emp}} - \boldsymbol{F}\right\|_\infty + \boldsymbol{n}^{-1}$$

whence

$$\left\|\widehat{F} - F\right\|_\infty = O_p\left(n^{-1/2}\right)\,.$$

Conjecture

$$\sup_{\mathbb{R}} \left|\widehat{\boldsymbol{F}}_{\mathbf{emp}} - \widehat{\boldsymbol{F}}\right| = \boldsymbol{o}_{\boldsymbol{p}}\left(\boldsymbol{n}^{-1/2}\right).$$

Theorem 2 (Consistency of $\widehat{\psi}$)

Suppose that ψ is Hölder–continuous with exponent $\beta \in [1, 2]$ on $[a, b] \subset \{f > 0\}$, i.e. for some constant L,

$$\left|\psi'(x) - \psi'(y)\right| \le L|x - y|^{\beta-1} \quad \text{for all } x, y \in [a, b]\,.$$

Then

$$\boldsymbol{\sup_{[a+\delta_n, b-\delta_n]} \left|\widehat{\psi} - \psi\right| = O_p\left(\left(\frac{\log n}{n}\right)^{\beta/(2\beta+1)}\right)}$$

where $\delta_n \downarrow 0$.

Theorem 3 (Consistency of $\widehat{F}$)

Suppose that ψ is twice continuously differentiable on $[a, b] \subset \{f > 0\}$ with $\psi'' < 0$. Then

$$\sup_{[a+\delta_n, b-\delta_n]} \left| \widehat{F} - \widehat{F}_{\text{emp}} \right| = o_p\left(n^{-1/2}\right).$$

Remark 1

Rate $O_p\left((\log(n)/n)^{\beta/(2\beta+1)}\right)$ is optimal under the given conditions.

Remark 2

Integrating the log–concave density estimator $\widehat{f}$ yields an estimator $\widehat{F}$ which is essentially equivalent to $\widehat{F}_{\mathrm{emp}}$.

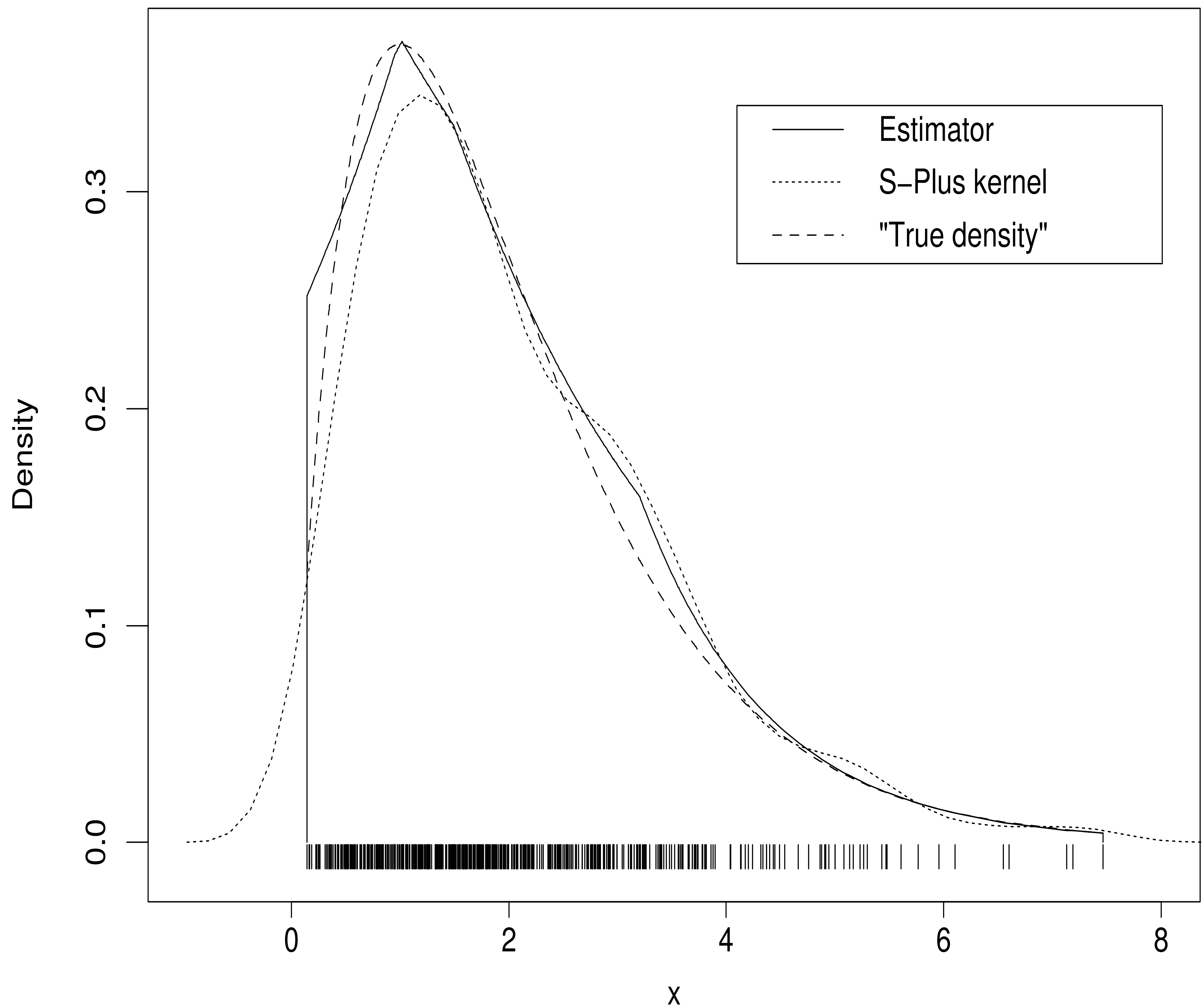
This is not true in case of a kernel density estimator $\widehat{f}$ with nonnegative kernel and optimal bandwidth of order $O(n^{-1/2})$.

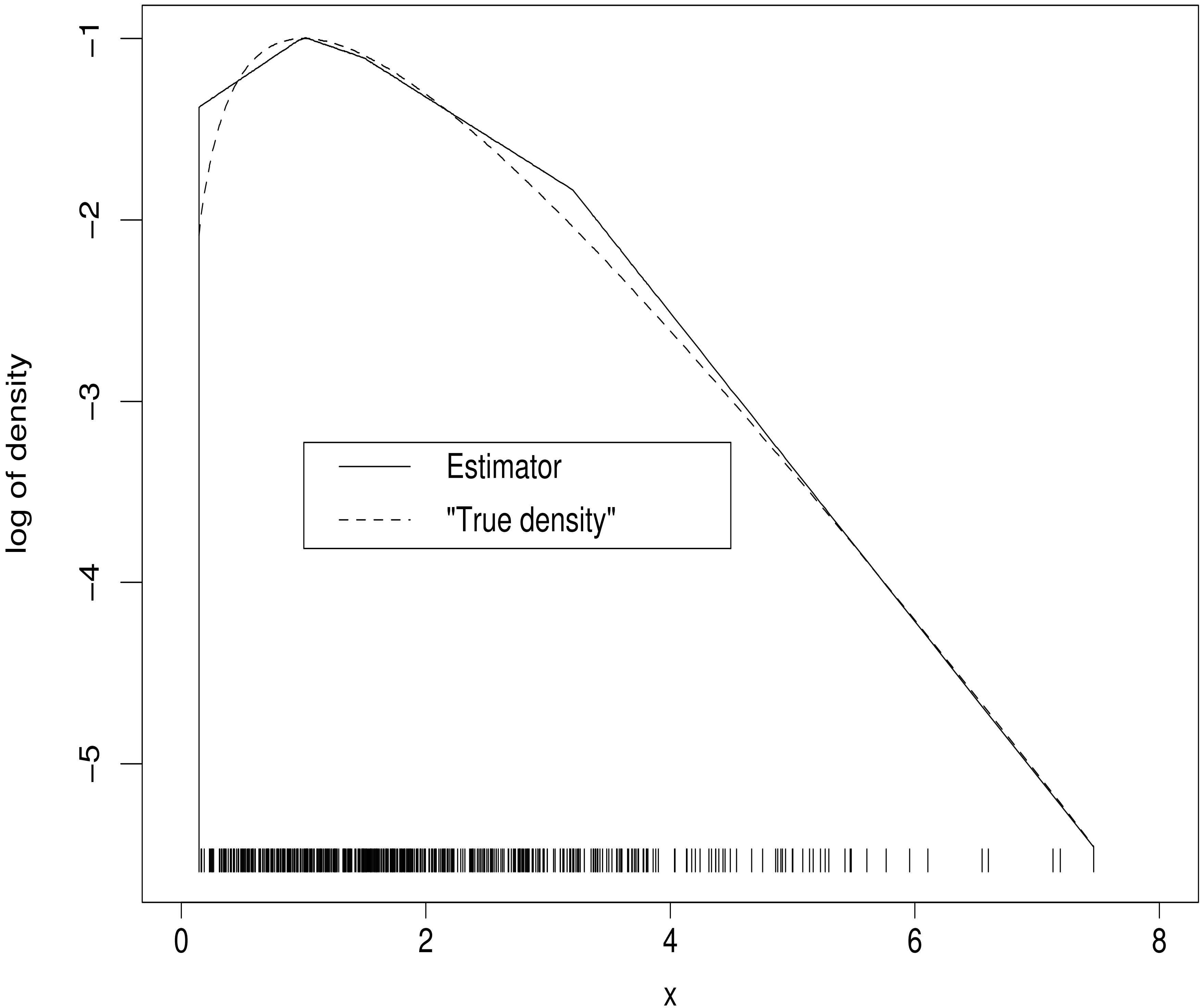
References

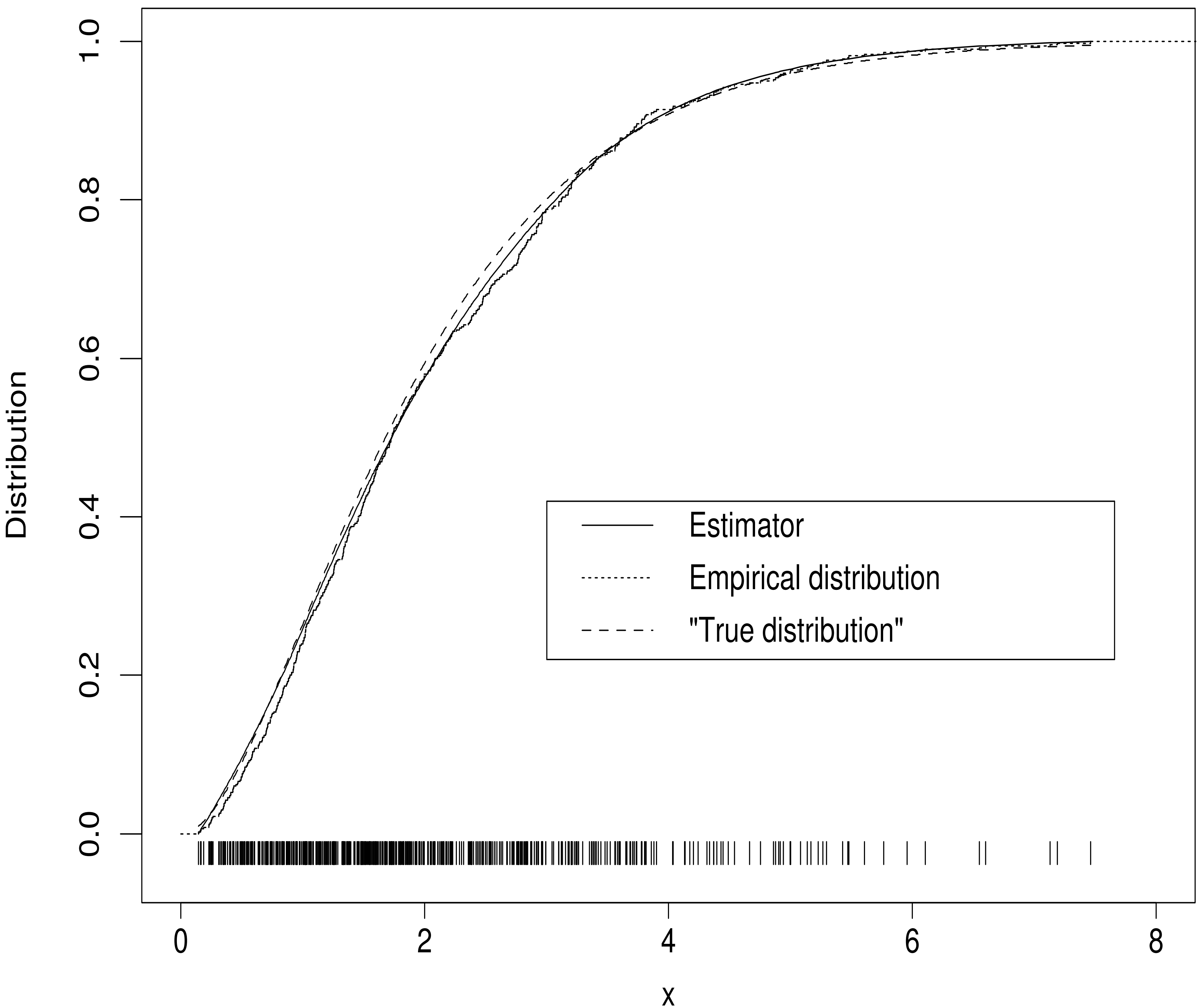
- K. Rufibach and LD (2004). *Maximum likelihood estimation of a log–concave density: basic properties and uniform consistency.* Preprint
- K. Rufibach (2004). *Computing maximum likelihood estimators of log–concave density functions.* Preprint

Numerical example

Sample of size $n = 500$ from $\mathrm{Gamma}(2, 1)\ldots$


Estimator
S−Plus kernel
"True density"
Density
x
0.0
0.1
0.2
0.3
0
2
4
6
8


Estimator
"True density"
log of density
x
−1
−2
−3
−4
−5
0
2
4
6
8



Distribution
x
Estimator
Empirical distribution
"True distribution"
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8

II. Inference via Multiscale Methods

Goal: Identify intervals $[s, t]$ on which unknown curve f has a certain property, for instance,

- increases ($f' \not\leq 0$ on $[s, t]$)
- decreases ($f' \not\geq 0$ on $[s, t]$)
- has a local extremum (maximum or minimum)
- bends upward ($f'' \not\leq 0$ on $[s, t]$)
- bends downward ($f'' \not\geq 0$ on $[s, t]$)

II.1 The multiscale approach in general

For (almost) any interval $[s, t]$ consider a test statistic

$$T_{s,t} = T_{s,t}(\mathrm{data})$$

for the null hypothesis

$$H_{s,t} \; : \; f \text{ hasn't specified property on } [s, t]$$

Multiple test: For $\alpha \in (0, 1)$ let $c_{s,t}(\alpha)$ be a critical value such that

$$\mathbb{P}\left(T_{s,t} > c_{s,t}(\alpha) \text{ and } H_{s,t} \text{ true for some } [s, t]\right) \;\le\; \alpha\,.$$

Claim with confidence $1 - \alpha$ that f has specified property on **any** interval $[s, t]$ such that

$$T_{s,t} > c_{s,t}(\alpha)\,.$$

References

- P. Chaudhuri and J.S. Marron (1999). SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.* ***94***, 807–823
- LD and V.G. Spokoiny (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* ***29***, 124–152
- LD (2002). Application of local rank tests to nonparametric regression. *J. Nonpar. Statist.* ***14**, 511–537*
- ...
- LD and G. Walther (2005). Multiscale Inference about a Density. Preprint in preparation
- LD, K. Rufibach and G. Walther (2005). Bump hunting via multiscale testing. Preprint in preparation

II.2 Conditional densities and spacings

If the support of f is known to be (contained in)

$$[a, \infty) \quad \text{or} \quad (-\infty, b] \quad \text{or} \quad [a, b]$$

with real endpoints a, b, then add

$$X_0 := a \quad \text{or} \quad X_{n+1} := b$$

(or both) to the ordered sample. After adjusting and renumbering the observations we end up with

$$\boldsymbol{X} = (\boldsymbol{X_i})_{\boldsymbol{i=0}}^{\boldsymbol{n+1}}, \quad \boldsymbol{X_0} < \boldsymbol{X_1} < \cdots < \boldsymbol{X_{n+1}}.$$

Proposition

For $0 \leq j < k \leq n+1$ with $k - j > 1$ define (random) interval

$$\mathcal{I}_{jk} \;:=\; [X_j, X_k]$$

and (random) density

$$f_{jk}(x) \;:=\; \frac{1\{x \in \mathcal{I}_{jk}\}\, f(x)}{F(X_k) - F(X_j)}$$

Then **conditional on X_j and X_k**,

$$(X_i)_{i=j+1}^{k-1} \;=_{\mathcal{L}}\; (Y_s)_{s=1}^{k-j-1} \;:\; \text{ordered sample from } f_{jk}$$

Thus use

$$\left(\boldsymbol{X}_i\right)_{i=j+1}^{k-1} \quad \text{or} \quad \left(\frac{\boldsymbol{X}_i - \boldsymbol{X}_j}{\boldsymbol{X}_k - \boldsymbol{X}_j}\right)_{i=j+1}^{k-1}$$

for inference about

shape of $\boldsymbol{f}$ on $\boldsymbol{\mathcal{I}_{jk}}$.

II.3 Local monotonicity properties (mode hunting)

Subsequent “distribution–free” method relies on the following fact:

$$F_o \; := \; \text{c.d.f. of } f_{0,n+1} \, ,$$

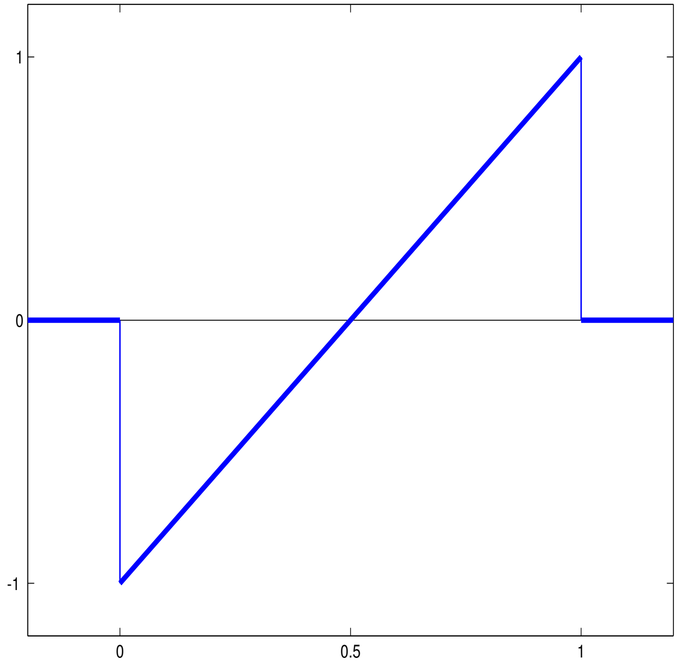
$$\boldsymbol{U} = (U_i)_{i=0}^{n+1} \; := \; (F_o(X_i))_{i=0}^{n+1} \; .$$

Then

$$(\boldsymbol{U}_{\boldsymbol{i}})_{\boldsymbol{i}=1}^{\boldsymbol{n}} \; =_{\boldsymbol{\mathcal{L}}} \; \text{ordered sample from } \boldsymbol{\mathcal{U}}[\mathbf{0}, \mathbf{1}]$$

Test statistic for an increase of f

$$\beta(x) := 1\{0 < x < 1\}(2x - 1)$$

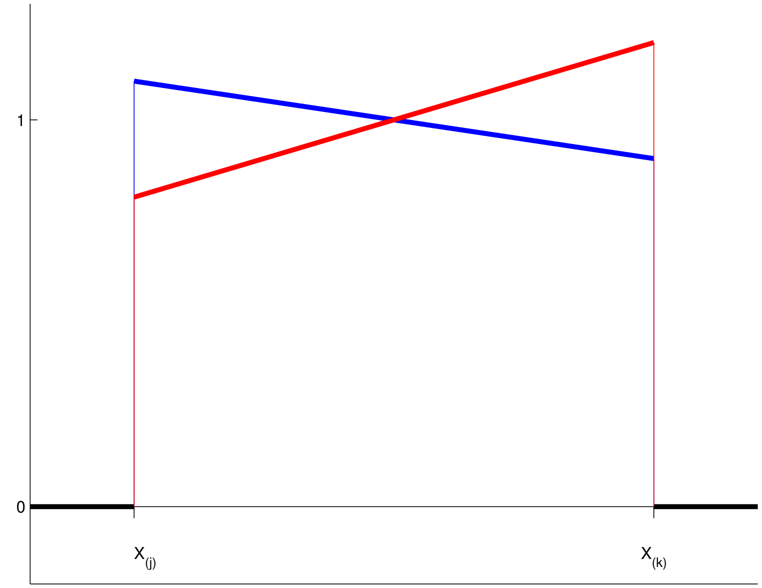


$$\boldsymbol{T_{jk}(X) := \sqrt{\frac{3}{k-j-1}} \sum_{i=j+1}^{k-1} \beta\left(\frac{X_i - X_j}{X_k - X_j}\right)}$$

Possible interpretation of T_{jk}

Locally most powerful test of

" $\boldsymbol{\lambda} \leq \mathbf{0}$ " versus " $\boldsymbol{\lambda} > \mathbf{0}$ "



within parametric model where

$$f_{jk}(x) \;=\; 1\{x \in \mathcal{I}_{jk}\}\left(1 + \boldsymbol{\lambda}\,\beta\left(\frac{x - X_j}{X_k - X_j}\right)\right)$$

Proposition

With $\boldsymbol{U} = \left(F_o(X_i)\right)_{i=0}^{n+1}$,

$$\boldsymbol{T_{jk}(X)} \begin{cases} \geq \ \boldsymbol{T_{jk}(U)} & \text{if} \quad \boldsymbol{f'} \geq \boldsymbol{0} \text{ on } \boldsymbol{\mathcal{I}_{jk}}\,, \\ \leq \ \boldsymbol{T_{jk}(U)} & \text{if} \quad \boldsymbol{f'} \leq \boldsymbol{0} \text{ on } \boldsymbol{\mathcal{I}_{jk}}\,. \end{cases}$$

Application

Let $c_{jk}(\alpha)$ be critical values such that

$$\mathbb{P}\left(|T_{jk}(U)| > c_{jk}(\alpha) \text{ for some } (j,k)\right) \;\leq\; \alpha$$

Then claim with confidence $1-\alpha$ that for arbitrary intervals $\mathcal{I}_{jk}$:

- $f' \not\leq 0$ on $\mathcal{I}_{jk}$ whenever $T_{jk}(X) > c_{jk}(\alpha)$
- $f' \not\geq 0$ on $\mathcal{I}_{jk}$ whenever $-T_{jk}(X) > c_{jk}(\alpha)$

Moreover, for arbitrary intervals $\mathcal{I}_{jk}, \mathcal{I}_{\ell m}$, the density f has a proper

- **local minimum** on $\mathcal{I}_{jm}$ whenever

$$-T_{jk}(X) > c_{jk}(\alpha)\,, \quad T_{\ell m}(X) > c_{\ell m}(\alpha)\,, \quad k \leq \ell$$

- **local maximum** on $\mathcal{I}_{jm}$ whenever

$$T_{jk}(X) > c_{jk}(\alpha)\,, \quad -T_{\ell m}(X) > c_{\ell m}(\alpha)\,, \quad k \leq \ell$$

- **local extremum** on $\mathcal{I}_{\min(j,\ell),\max(k,m)}$ whenever

$$\pm T_{jk}(X) > c_{jk}(\alpha)\,, \quad \mp T_{\ell m}(X) > c_{\ell m}(\alpha)$$

Finding the critical values $c_{jk}(\alpha)$

$$T(\boldsymbol{U}) \;:=\; \max_{k-j>1}\left(\left|T_{jk}(\boldsymbol{U})\right| \boldsymbol{-G}\left(\frac{\boldsymbol{k-j}}{\boldsymbol{n+1}}\right)\right)$$

$$G(u) \;:=\; \sqrt{2\log\left(\frac{e}{u}\right)}$$

$$\boldsymbol{\kappa(\alpha) \;:=\; (1-\alpha)-}\text{ quantile of } \boldsymbol{T(U)}$$

$$\boldsymbol{c_{jk}(\alpha) \;:=\; \kappa(\alpha)+G\left(\frac{k-j}{n+1}\right)}$$

Theorem 1

$$T(\boldsymbol{U}) \ \to_{\mathcal{L}} \ \boldsymbol{T} \in [0, \infty) \quad \text{as } n \to \infty$$

Theorem 2

Suppose that $\pm \boldsymbol{f}' \geq \boldsymbol{c} > \mathbf{0}$ in some neighborhood of $x \in \mathbb{R}$. Then x is localized with asymptotic probability one and precision

$$\boldsymbol{O_p}\left(\left(\frac{\log \boldsymbol{n}}{\boldsymbol{n}}\right)^{1/3}\right).$$

Suppose that $\boldsymbol{f}(\boldsymbol{x}) > \mathbf{0}$, $\boldsymbol{f}'(\boldsymbol{x}) = \mathbf{0}$ and $\pm \boldsymbol{f}'' \geq \boldsymbol{c} > \mathbf{0}$ in some neighborhood of $x \in \mathbb{R}$. Then this local extremum is localized with asymptotic probability one and precision

$$\boldsymbol{O_p}\left(\left(\frac{\log \boldsymbol{n}}{\boldsymbol{n}}\right)^{1/5}\right).$$

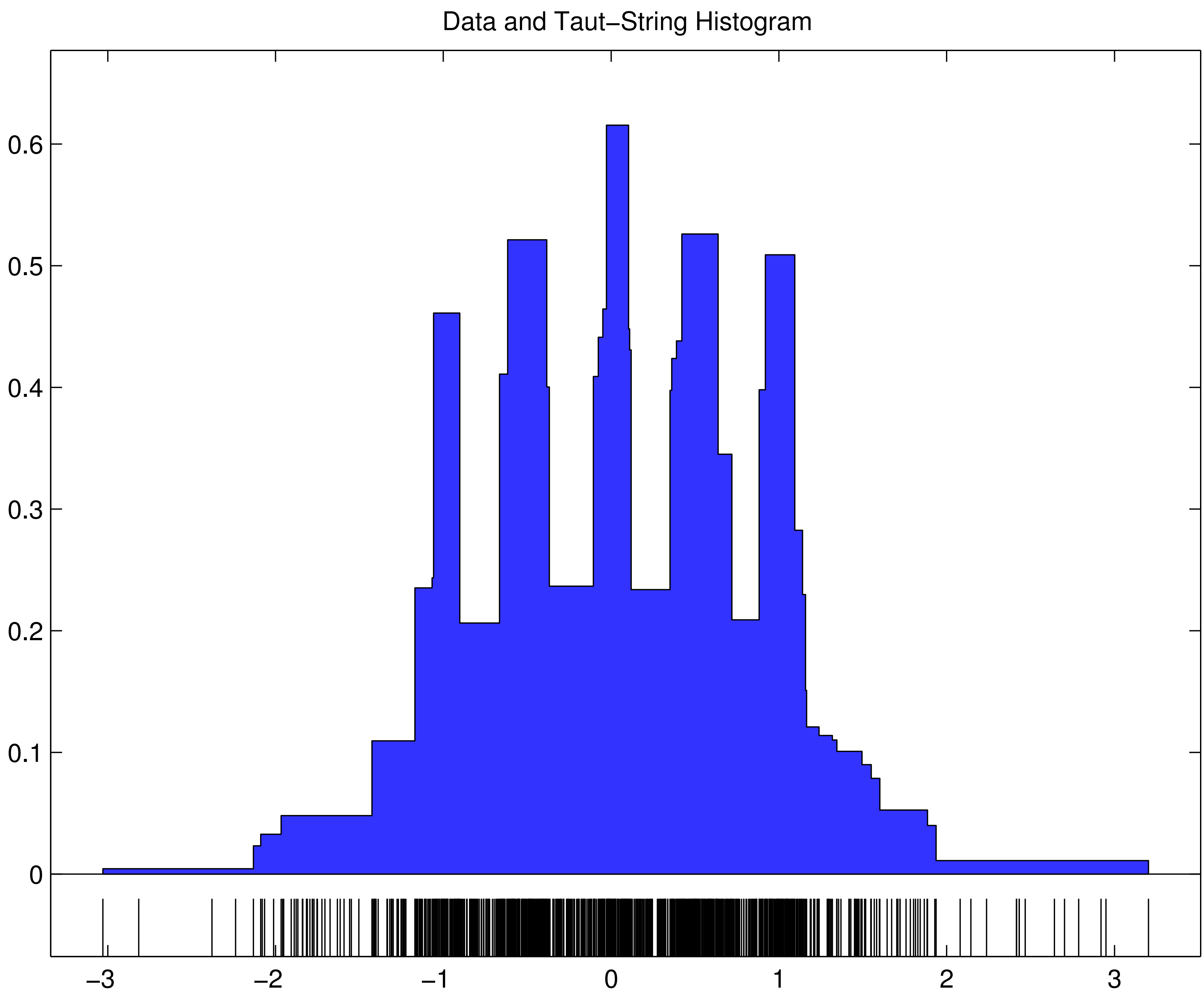
Numerical example

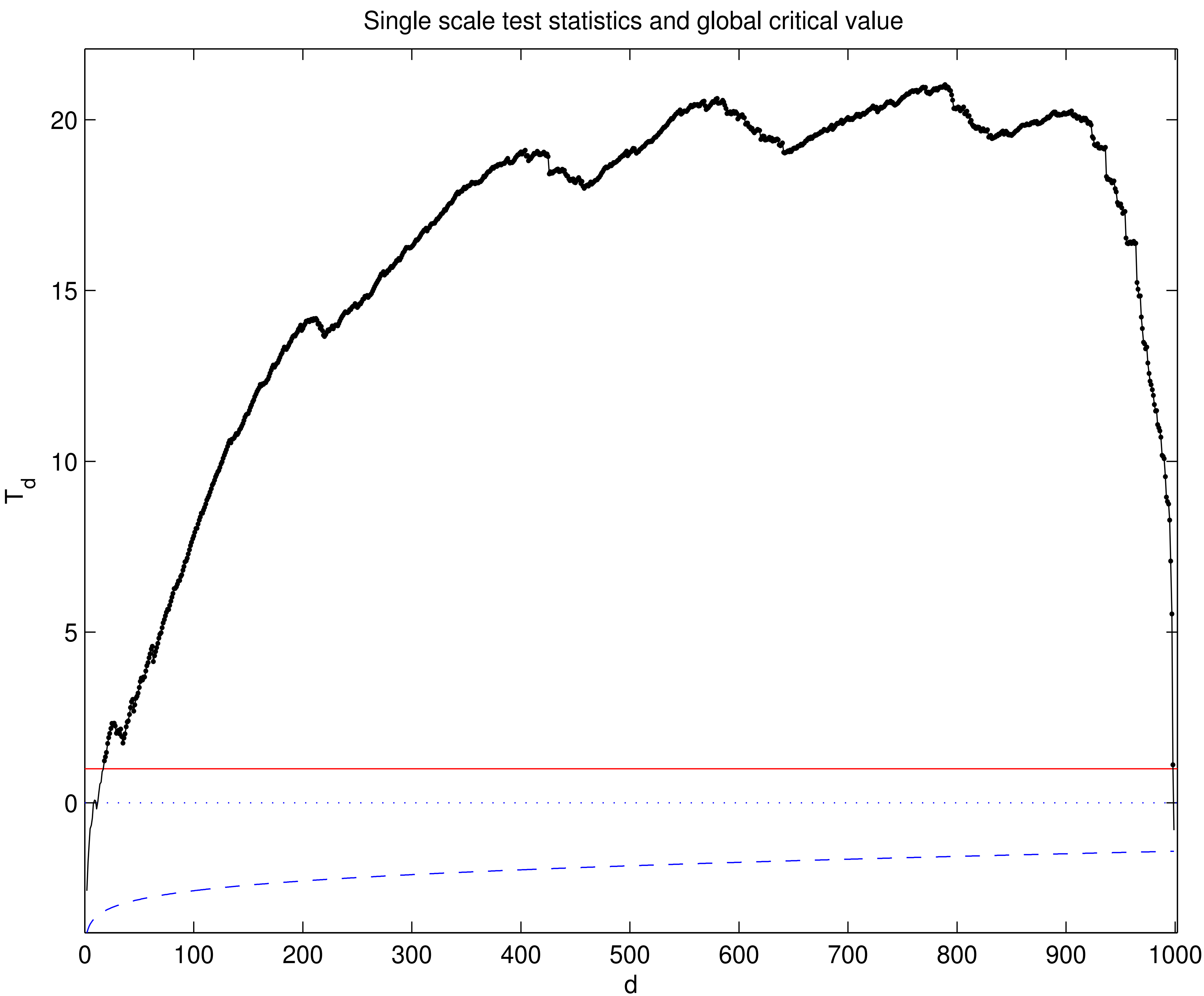
For a simulated sample of size $n = 1000$ from the "claw density" show

- a histogram estimator of f (Davies and Kovac)
- the function

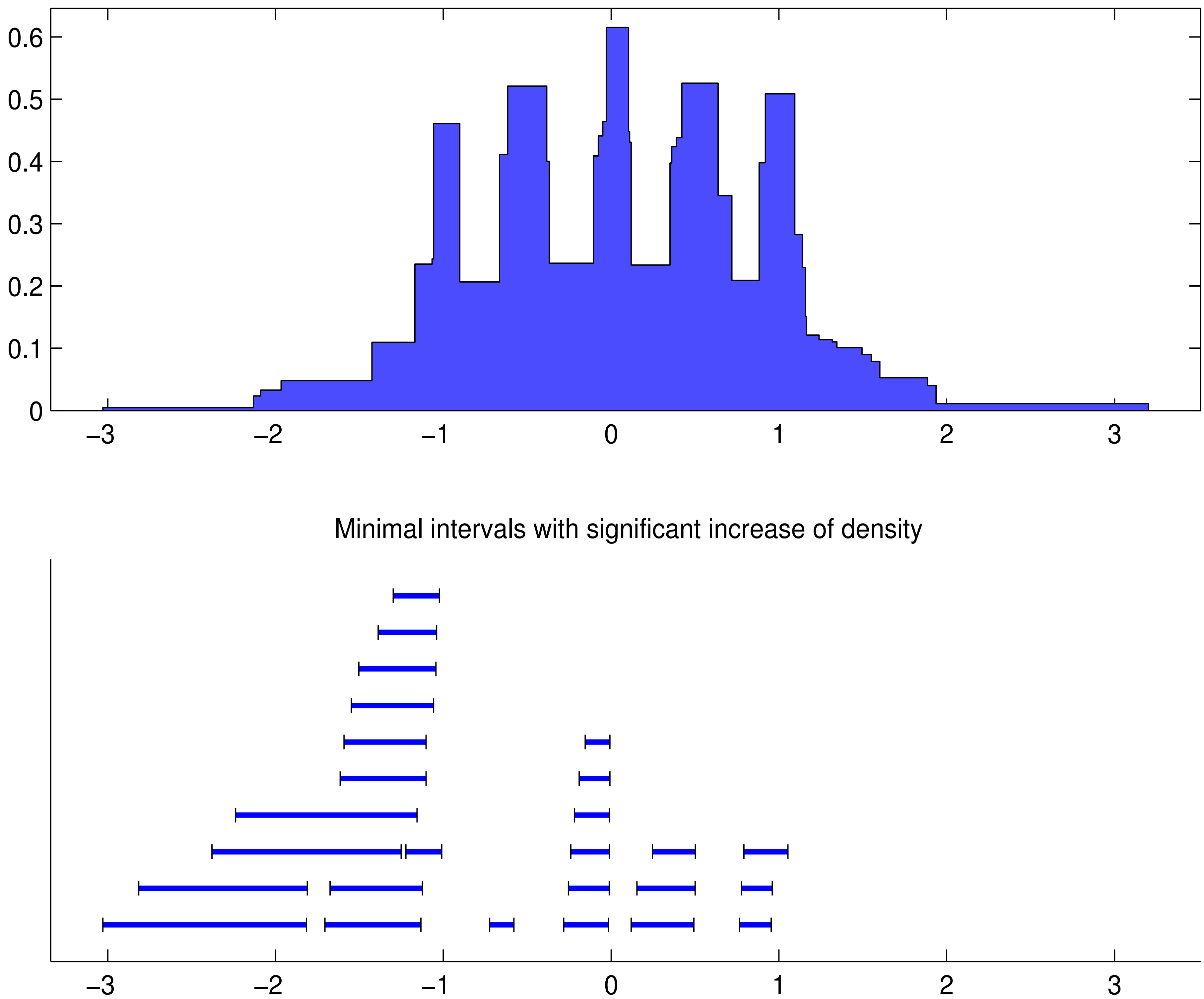
$$d \;\longmapsto\; \max_{k-j=d} \left(\left| T_{jk}(X) \right| - G\left(\frac{d}{n+1}\right) \right)$$

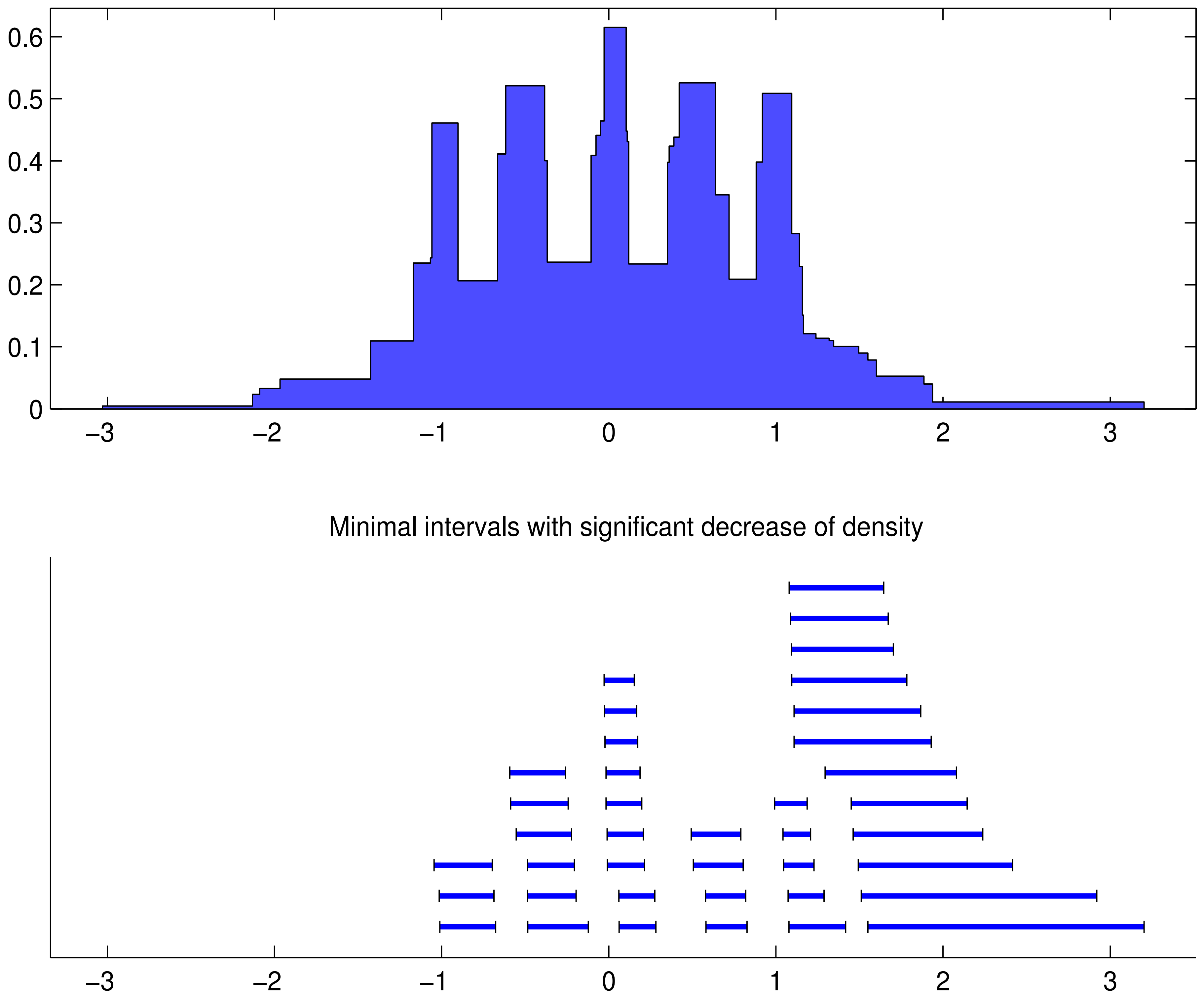
- minimal intervals with essential increases/decreases or local extrema

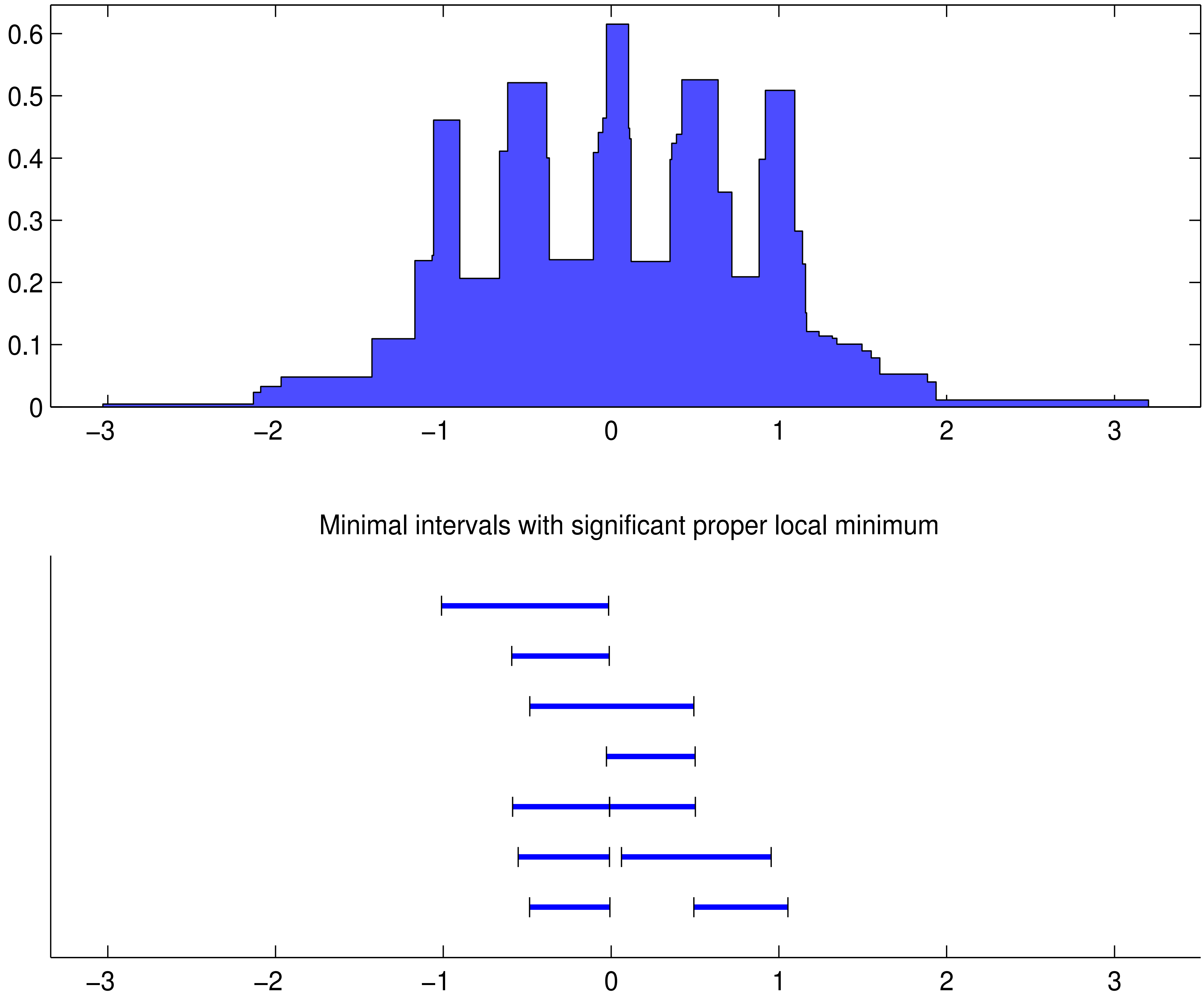

Data and Taut−String Histogram
0.6
0.5
0.4
0.3
0.2
0.1
0
−3
−2
−1
0
1
2
3



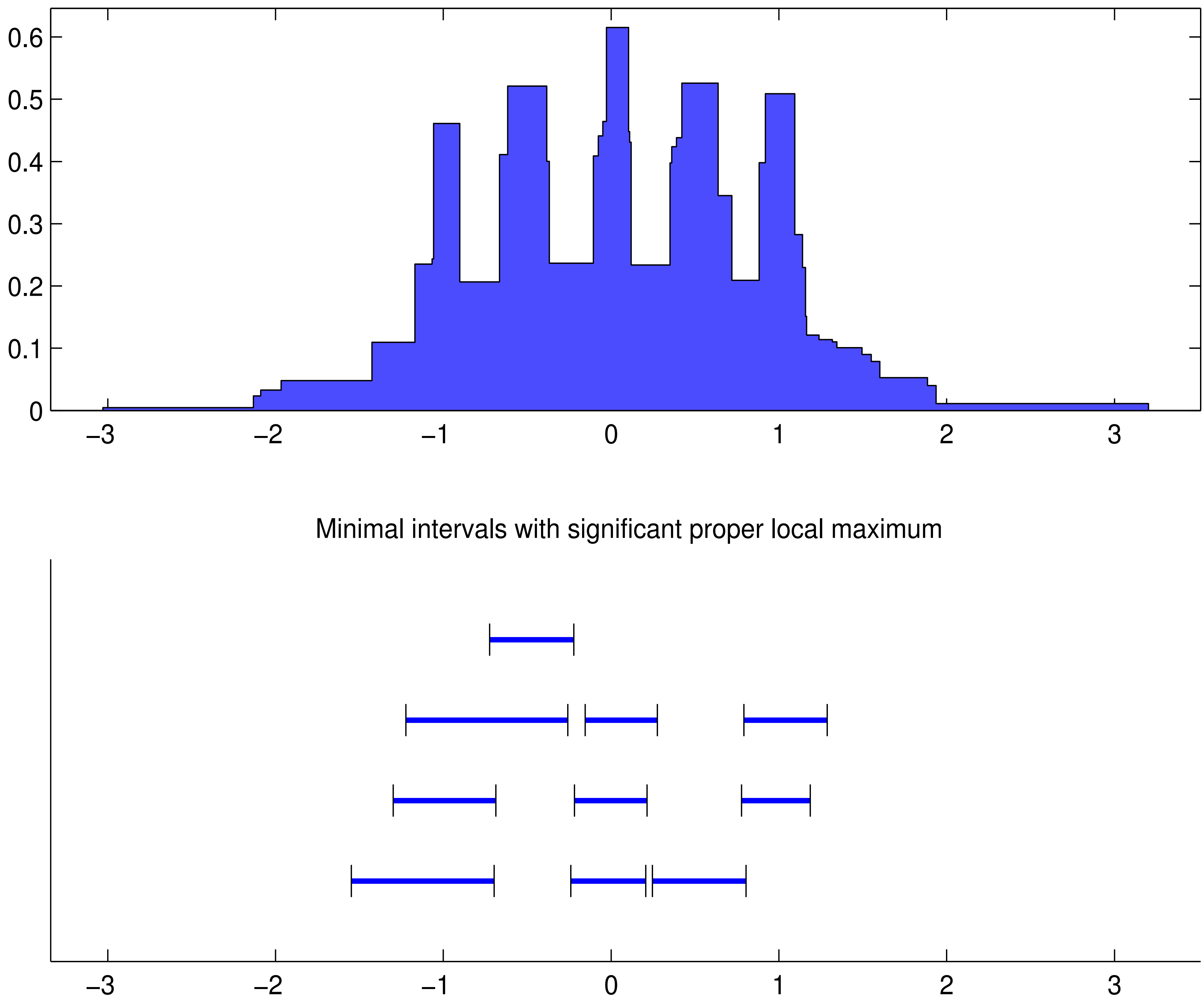
Single scale test statistics and global critical value
T_d
d
0
100
200
300
400
500
600
700
800
900
1000
0
5
10
15
20


0.6
0.5
0.4
0.3
0.2
0.1
0
–3
–2
–1
0
1
2
3
Minimal intervals with significant increase of density
–3
–2
–1
0
1
2
3


0.6
0.5
0.4
0.3
0.2
0.1
0
–3
–2
–1
0
1
2
3
Minimal intervals with significant decrease of density
–3
–2
–1
0
1
2
3



0.6
0.5
0.4
0.3
0.2
0.1
0
−3
−2
−1
0
1
2
3
Minimal intervals with significant proper local minimum
−3
−2
−1
0
1
2
3


0.6
0.5
0.4
0.3
0.2
0.1
0
−3
−2
−1
0
1
2
3
Minimal intervals with significant proper local maximum
−3
−2
−1
0
1
2
3

II.3 Auxiliary results

Theorem 1 follows from general results about stochastic processes.

Here a simple version:

$$Z_n \ : \ \text{stochastic process on finite set } \Pi_n \subset \Pi$$

$$\Pi \ := \ \{(s,t) : 0 \le s < t \le 1\}$$

Define

$$T_n \ := \ \max_{(s,t)\in\Pi_n} \left(\frac{|Z_n(s,t)|}{\sqrt{t-s}} - G(t-s) \right)$$

Theorem A

Assumption 1 (subgaussian variables):

$$\mathbb{P}\left\{\frac{|Z_n(s,t)|}{\sqrt{t-s}} \geq \eta\right\} \;\leq\; 2\exp\left(-\frac{\eta^2}{2}\right)$$

Assumption 2 (subexponential increments):

$$\mathbb{P}\left\{\frac{|Z_n(s,t) - Z_n(u,v)|}{\sqrt{|s-u| + |t-v|}} \geq \eta\right\} \;\leq\; K\exp\left(-\frac{\eta}{K}\right)$$

Then

$$\sup_n \; \mathbb{P}\left\{T_n \geq \eta\right\} \;\longrightarrow\; 0 \quad \text{as } \eta \to \infty\,.$$

Theorem B

Suppose that the assumptions of Theorem A hold. In addition suppose that

- "$\Pi_n \to \Pi$" .

- the finite-dimensional distributions of Z_n (suitably extended) converge to those of a centered gaussian process Z_∞ on Π.

- $\mathrm{Cov}(Z_\infty(s,t), Z_\infty(u,v)) = 0$ whenever $t \le u$,
  $\mathrm{Var}(Z_\infty(s,t)) = t - s$.

Then

$$\boldsymbol{T_n} \to_{\mathcal{L}} \boldsymbol{T_\infty} \quad \text{and} \quad \mathbf{0} \le \boldsymbol{T_\infty} < \infty .$$

Theorems A–B apply to

$$\Pi_n := \left\{ \left(\frac{j}{n+1}, \frac{k}{n+1} \right) : 0 \le j < k \le n+1 \right\}$$

$$Z_n \left(\frac{j}{n+1}, \frac{k}{n+1} \right) := \sqrt{\frac{k-j-1}{n+1}}\, T_{jk}(\boldsymbol{U})$$

II.4 Detecting convexity/concavity of $\log f$

Local parametric models

Local coordinates:

$$\langle x \rangle_{jk} \; := \; \beta \left(\frac{x - X_j}{X_k - X_j} \right) , \quad [X_j, X_k] \to [-1, 1] \, .$$

Parametric models:

$$g_{\theta}(y) \; := \; 1\{-1 < y < 1\} \, G_{\theta} \; \exp(\theta y) \, ,$$

$$g_{\theta,\eta}(y) \; := \; 1\{-1 < y < 1\} \, G_{jk,\theta,\eta} \exp(\theta y + \eta y^2) \, .$$

Local score test statistics

Norming constants $A_\theta, B_\theta, C_\theta$ s.t.

$$\begin{aligned}\int \Big(A_\theta y^2 - B_\theta y - C_\theta\Big) y^j g_\theta(y)\,dy &= 0 \quad \text{for } j = 0, 1\,,\\ \int \Big(A_\theta y^2 - B_\theta y - C_\theta\Big)^2 g_\theta(y)\,dy &= 1\,.\end{aligned}$$

For $k - j > 2$:

$$\boldsymbol{T_{jk} := \sqrt{\frac{1}{k-j-1}} \sum_{i=j+1}^{k-1} \Big(A_{\widehat{\theta}} \langle X_i \rangle_{jk}^2 - B_{\widehat{\theta}} \langle X_i \rangle_{jk} - C_{\widehat{\theta}}\Big)}$$

with

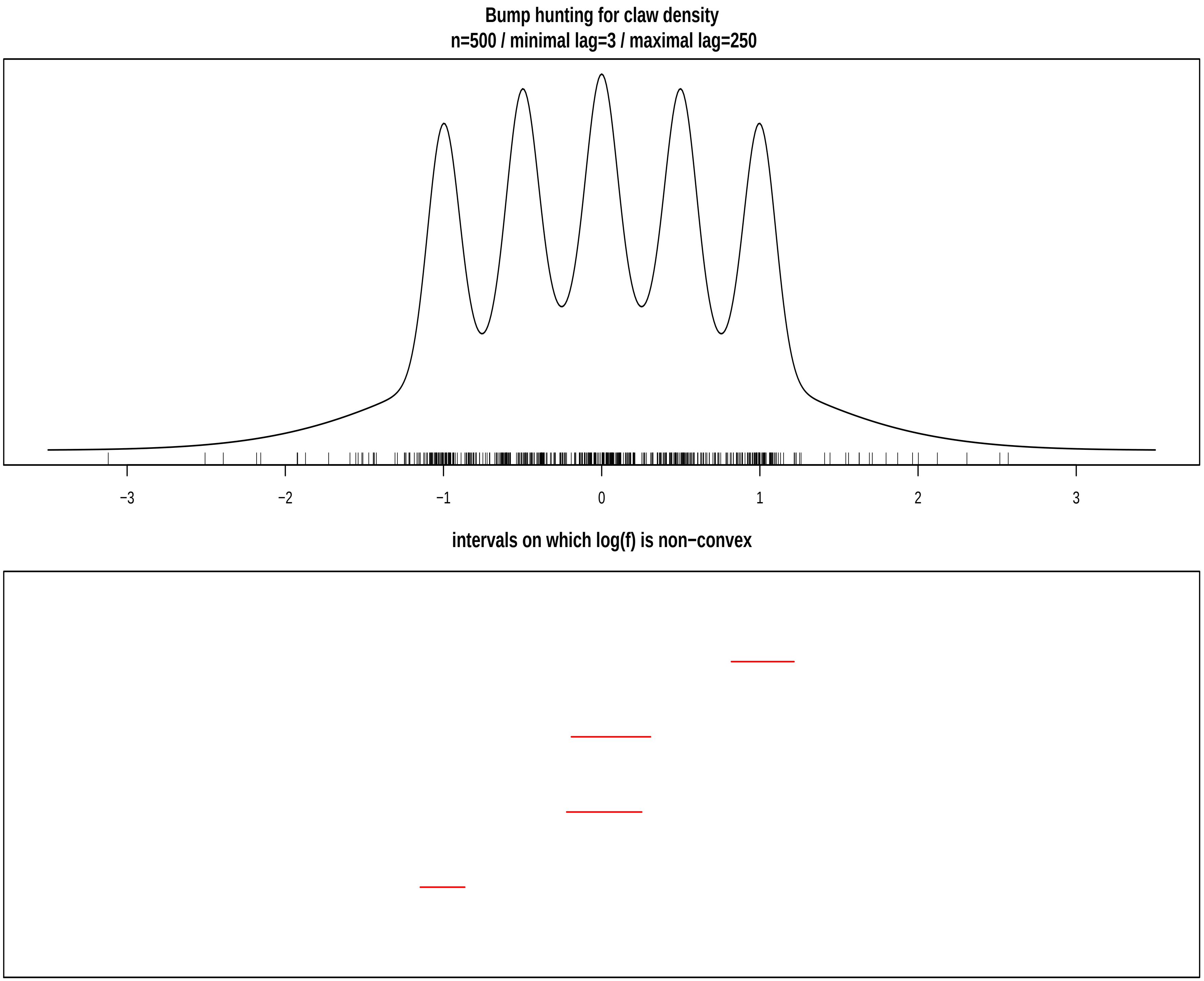
$$\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{jk} := \text{ local MLE } \ldots\,.$$

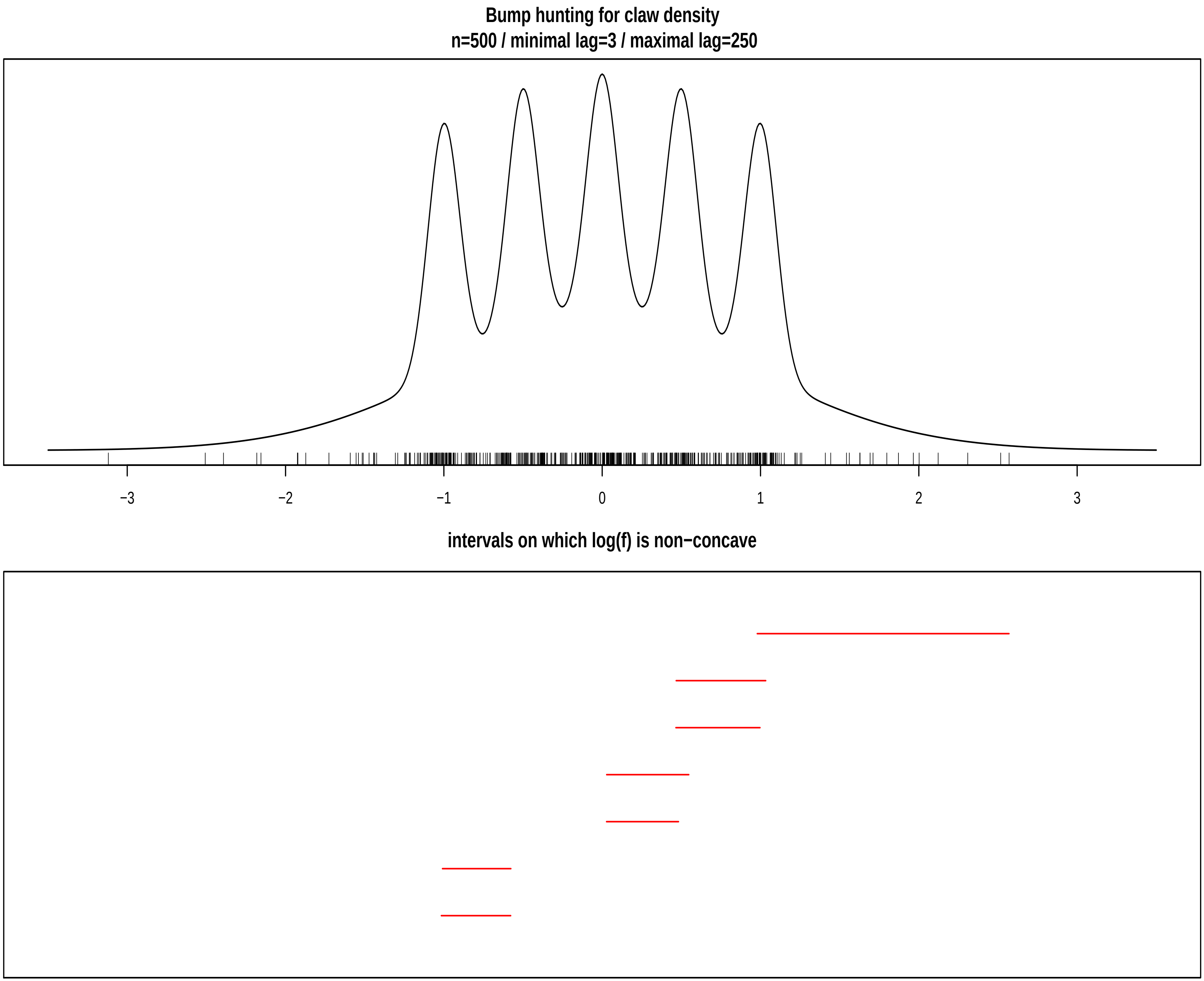
So far some theory about the local power of the single tests ...

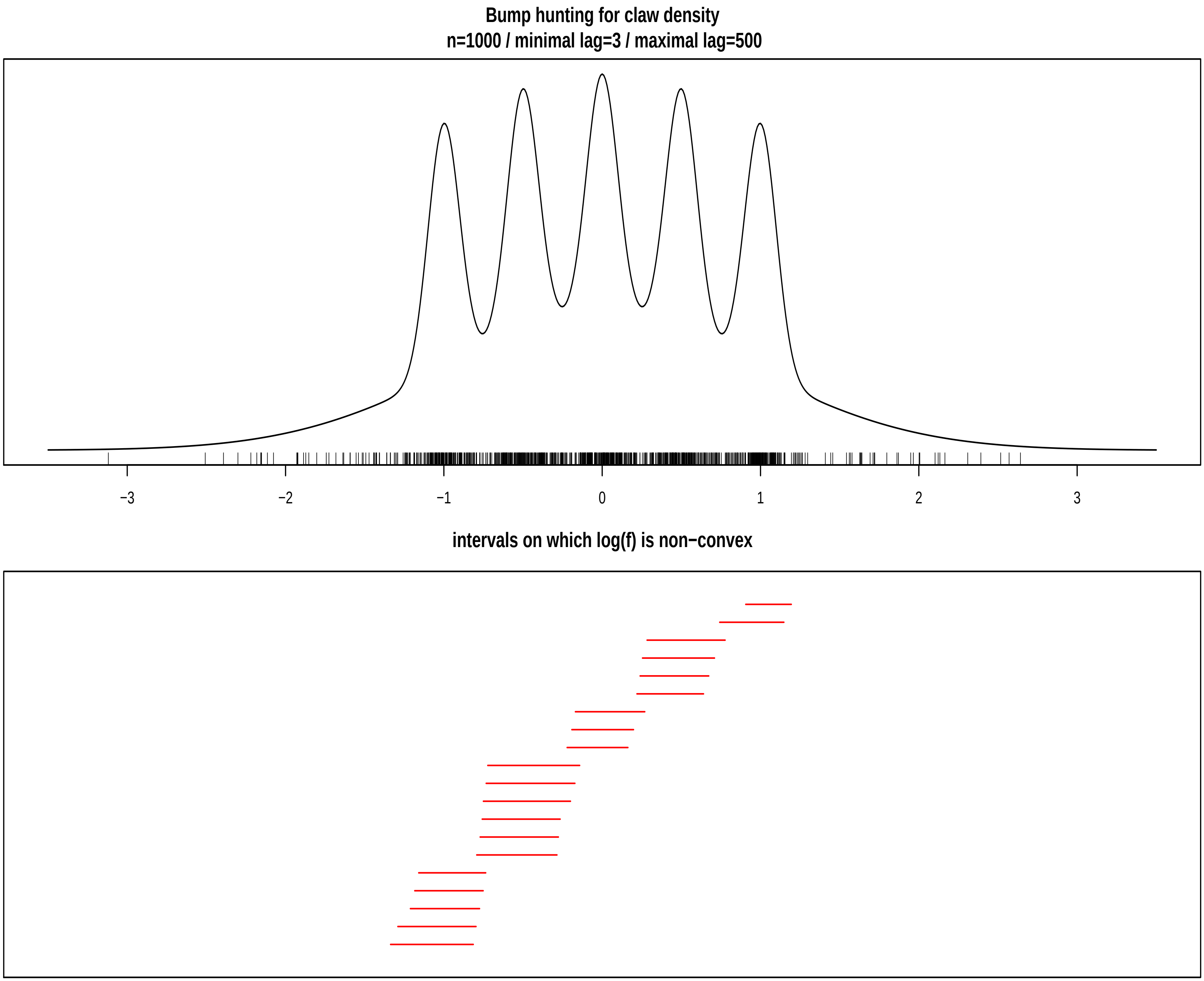
Conjecture that simulating critical values from uniform distribution yields asymptotically valid multiple test ...

Simulations and numerical examples indicate that the method works ...

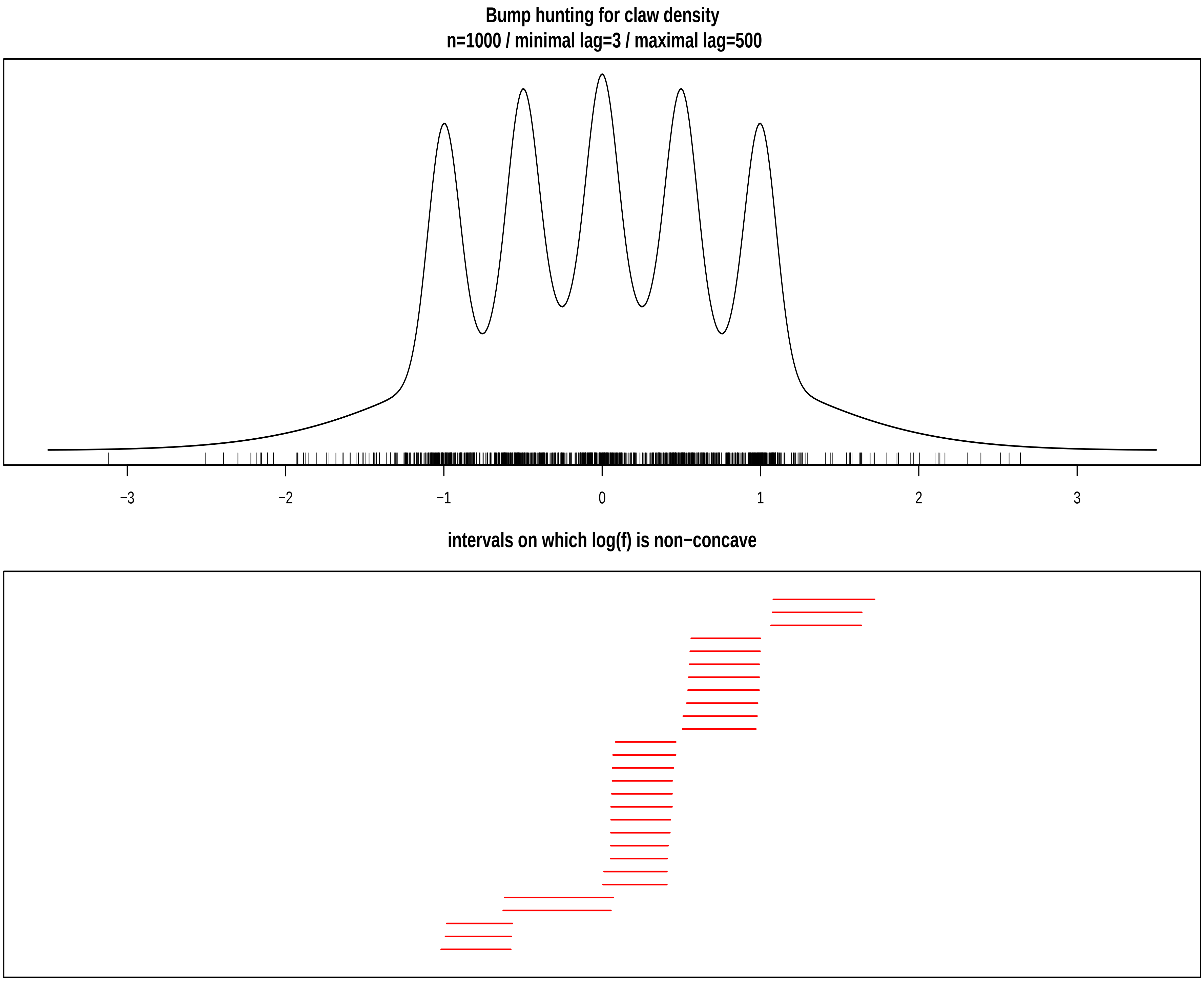
Main difficulty: Apparently no useful transformation group connected to the parametric model $(g_\theta)_{\theta \in \mathbb{R}}$ ...


Bump hunting for claw density
n=500 / minimal lag=3 / maximal lag=250
−3
−2
−1
0
1
2
3
intervals on which log(f) is non−convex


Bump hunting for claw density
n=500 / minimal lag=3 / maximal lag=250
−3
−2
−1
0
1
2
3
intervals on which log(f) is non−concave



Bump hunting for claw density
n=1000 / minimal lag=3 / maximal lag=500
−3
−2
−1
0
1
2
3
intervals on which log(f) is non−convex


Bump hunting for claw density
n=1000 / minimal lag=3 / maximal lag=500
−3
−2
−1
0
1
2
3
intervals on which log(f) is non−concave