Some statistical challenges of topological inference in the 1D case

Max Wardetzky[1] Ulrich Bauer[2]
Axel Munk[1] Hannes Sieling[1]

[1]Georg-August-Universität Göttingen

[2]TU München

Setup

Consider one-dimensional signal $f : [0, 1] \to \mathbb{R}$ with k modes.
Suppose f is observed by a finite number of measurements:

$$Y_i = f(t_i) + \epsilon_i, \quad 0 = t_0 < t_1 < \cdots < t_n = 1 \ .$$

Setup

Consider one-dimensional signal $f : [0, 1] \to \mathbb{R}$ with k modes. Suppose f is observed by a finite number of measurements:

$$Y_i = f(t_i) + \epsilon_i, \quad 0 = t_0 < t_1 < \cdots < t_n = 1 \, .$$

Question: With what probability can we infer the number of modes of f from the observations (Y_i)?

Setup

Consider one-dimensional signal $f : [0,1] \to \mathbb{R}$ with k modes.
Suppose f is observed by a finite number of measurements:

$$Y_i = f(t_i) + \epsilon_i, \quad 0 = t_0 < t_1 < \cdots < t_n = 1 \, .$$

Question: With what probability can we infer the number of modes of f from the observations (Y_i)?

Noise (ϵ_i) independently distributed with mean zero s.t. for some $\kappa > 0$, $v > 0$ and all $m \geq 2$:

$$\mathbb{E}\,|\epsilon_i|^m \leq v m! \kappa^{m-2}/2 \ \text{ for all } i = 1, \ldots, n.$$

Setup (cont'd)

- Not on our agenda: *First regularize (filter) data, then perform topological inference* (Bubenik, Carlsson, Chazal, Cohen-Steiner, Guibas, Kim, Mémoli, Mérigot, Oudot, Sheehy, ...).

Setup (cont'd)

- Not on our agenda: *First regularize (filter) data, then perform topological inference* (Bubenik, Carlsson, Chazal, Cohen-Steiner, Guibas, Kim, Mémoli, Mérigot, Oudot, Sheehy, ...).
- Hard to analyze effect of filtering from statistical perspective *without* a priori assumptions on data or oracles.

Setup (cont'd)

- Not on our agenda: *First regularize (filter) data, then perform topological inference* (Bubenik, Carlsson, Chazal, Cohen-Steiner, Guibas, Kim, Mémoli, Mérigot, Oudot, Sheehy, ...).
- Hard to analyze effect of filtering from statistical perspective *without* a priori assumptions on data or oracles.
- Goal: Statistical bounds on number of modes of f inferred from data Y only, *without* reconstructing f along the way.

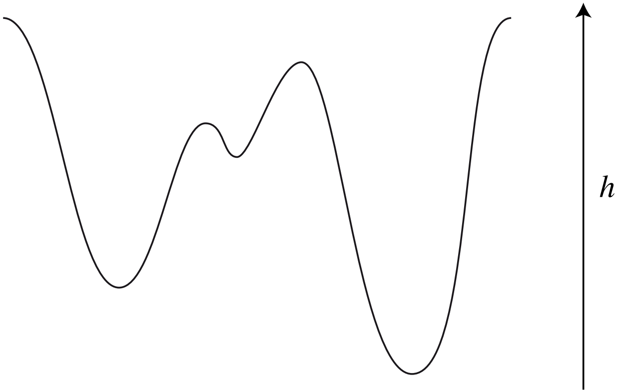
Persistent homology [Edelsbrunner et al., 2002]

Investigate change of homology for sublevel sets

Persistent homology [Edelsbrunner et al., 2002]

Investigate change of homology for sublevel sets

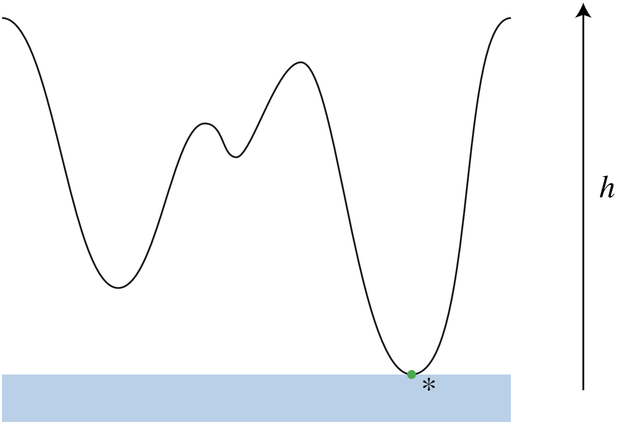
Example: connected components in 1D



Persistent homology [Edelsbrunner et al., 2002]

Investigate change of homology for sublevel sets

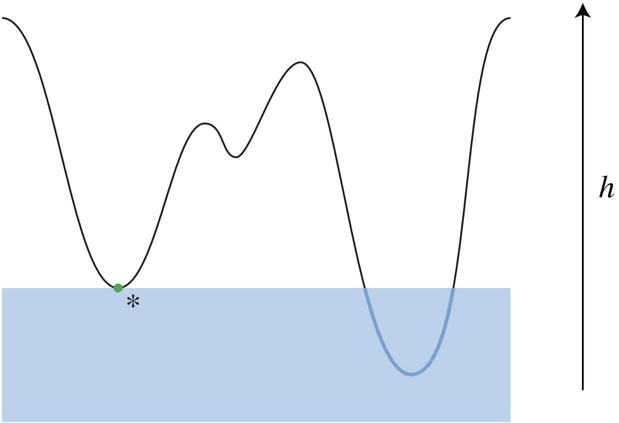
Example: connected components in 1D



Persistent homology [Edelsbrunner et al., 2002]

Investigate change of homology for sublevel sets

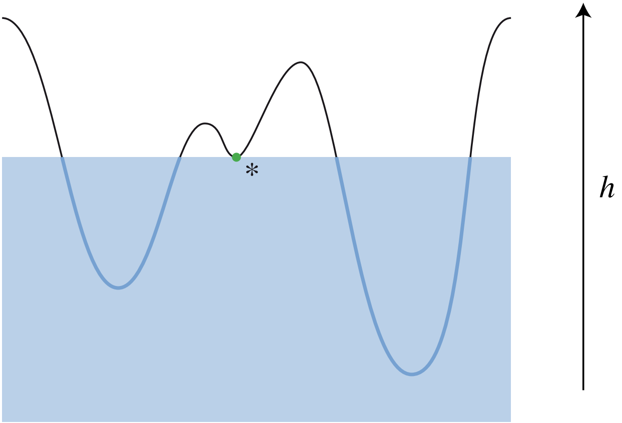
Example: connected components in 1D



Persistent homology [Edelsbrunner et al., 2002]

Investigate change of homology for sublevel sets

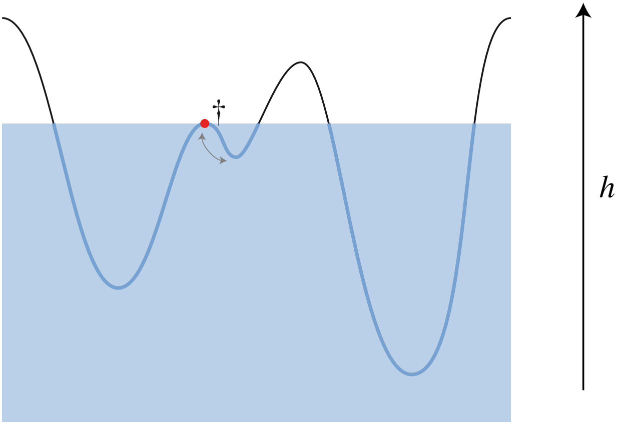
Example: connected components in 1D



Persistent homology [Edelsbrunner et al., 2002]

Investigate change of homology for sublevel sets

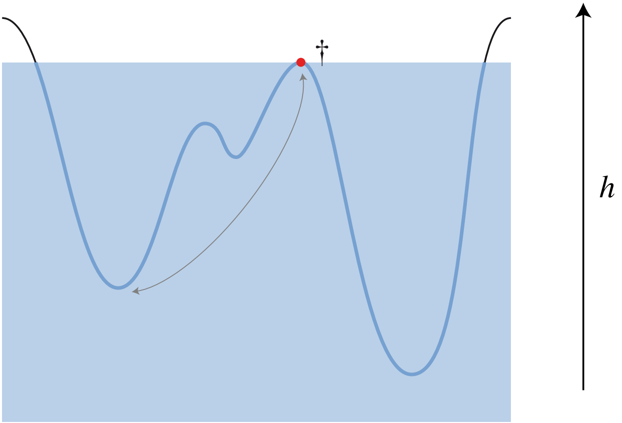
Example: connected components in 1D



Persistent homology [Edelsbrunner et al., 2002]

Investigate change of homology for sublevel sets

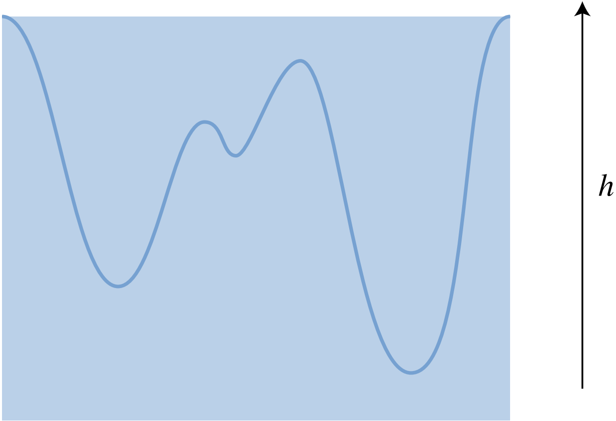
Example: connected components in 1D



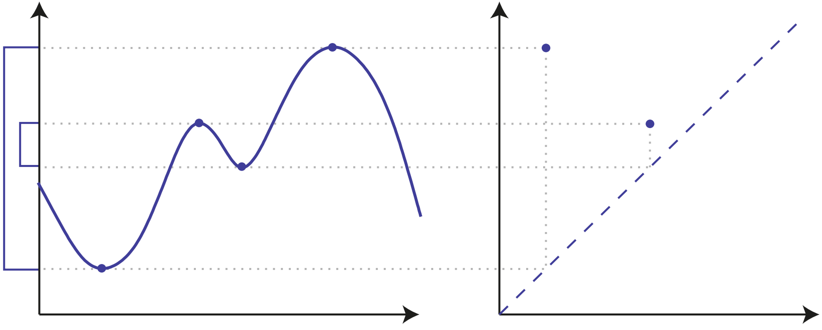
Persistent homology [Edelsbrunner et al., 2002]

Investigate change of homology for sublevel sets

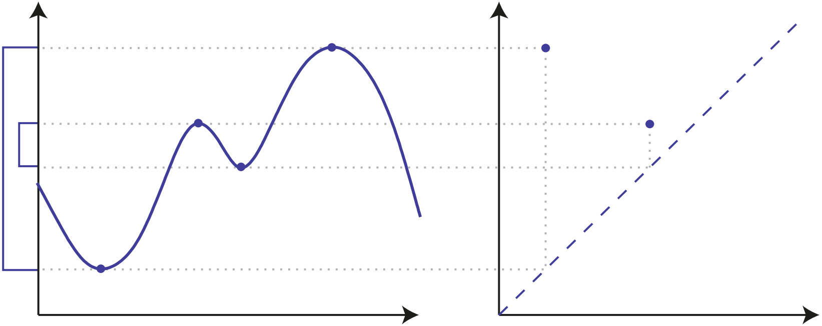
Example: connected components in 1D



Persistence diagrams [Cohen-Steiner et al., 2005]

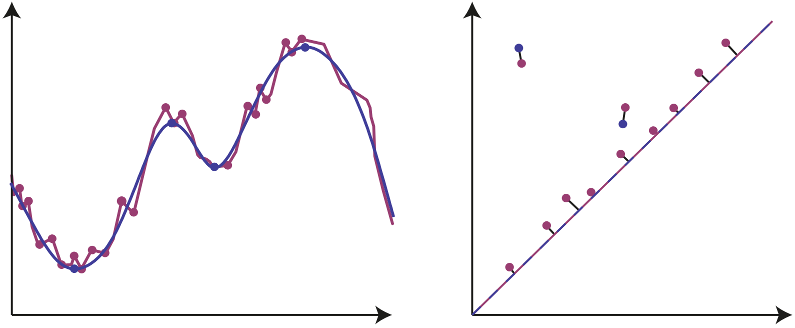


Persistence diagrams [Cohen-Steiner et al., 2005]



persistence of pair of critical points = death - birth

Persistence diagrams [Cohen-Steiner et al., 2005]



persistence of pair of critical points = death - birth

Persistence signatures

Order persistence values from highest to lowest:

$$s_{0,\infty}(f) \geq s_{1,\infty}(f) \geq s_{2,\infty}(f) \geq \cdots \geq 0 \geq 0 \ldots$$

Persistence signatures

Order persistence values from highest to lowest:

$$s_{0,\infty}(f) \geq s_{1,\infty}(f) \geq s_{2,\infty}(f) \geq \cdots \geq 0 \geq 0 \ldots$$

Lemma (Bauer)

Let X denote the space of piecewise constant functions on (some) equipartition of $[0,1]$. Let $X_k \subset X$ denote the set of functions with at most k inner maxima (= modes). Then

$$s_{k,\infty}(f) = 2 \cdot \operatorname{dist}_\infty(f, X_k) \ .$$

Persistence signatures

Order persistence values from highest to lowest:

$$s_{0,\infty}(f) \geq s_{1,\infty}(f) \geq s_{2,\infty}(f) \geq \cdots \geq 0 \geq 0 \ldots$$

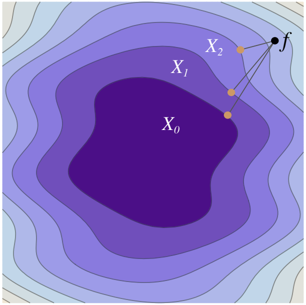
Lemma (Bauer)

Let X denote the space of piecewise constant functions on (some) equipartition of $[0, 1]$*. Let* $X_k \subset X$ *denote the set of functions with at most k inner maxima (= modes). Then*

$$s_{k,\infty}(f) = 2 \cdot \mathrm{dist}_\infty(f, X_k) \; .$$

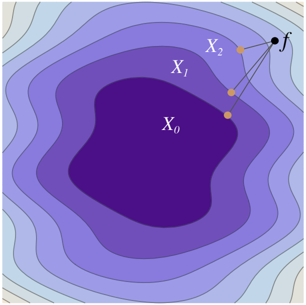
Note: Stability implies that $|s_{k,\infty}(f) - s_{k,\infty}(g)| \leq 2\|f - g\|_\infty \; \forall k$.
This gives *upper* bound $s_{k,\infty}(f) \leq 2 \cdot \mathrm{dist}_\infty(f, X_k)$.

Persistence signatures



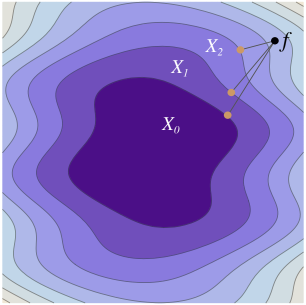
Interpret persistence signatures as distance to set of functions with at most k modes.

Different metrics – different signatures



Define *metric signature* $s_k(f) := \text{dist}(f, X_k)$ with respect to *some* metric d on X.

Different metrics – different signatures



Define *metric signature* $s_k(f) := \mathrm{dist}(f, X_k)$ with respect to *some* metric d on X.

Call d *descriptive* if for every f with at least $k + 1$ modes $s_k > 0$.

Different metrics – different signatures

Lemma (Stability of metric signatures)

For all metrics and all k: $|s_k(f) - s_k(g)| \leq d(f, g)$.

Different metrics – different signatures

Lemma (Stability of metric signatures)

For all metrics and all k: $|s_k(f) - s_k(g)| \leq d(f, g)$.

Proof.

Distance to sets is 1-Lipschitz. □

Different metrics – different signatures

Lemma (Stability of metric signatures)

For all metrics and all k: $|s_k(f) - s_k(g)| \leq d(f, g)$.

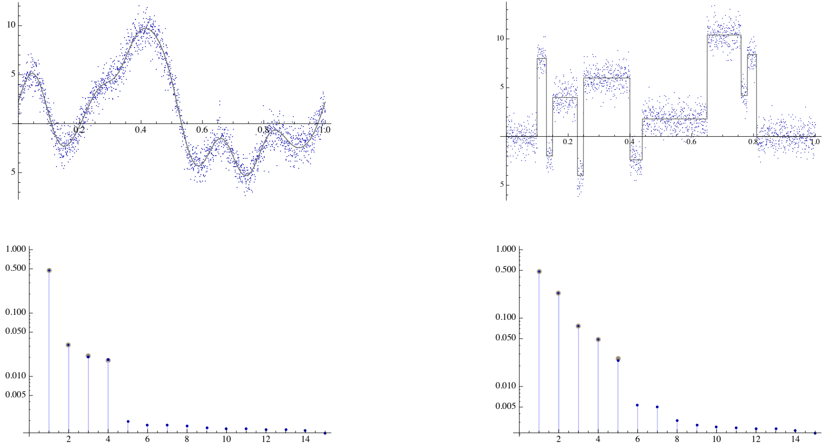
Proof.

Distance to sets is 1-Lipschitz. □

Consider the following descriptive metrics:

- Persistence signatures: $d_\infty(f, g) = sup_x|f(x) - g(x)|$.
- Kolmogorov signatures: $d_K(f, g) = sup_x|F(x) - G(x)|$ for antiderivatives F and G (with $F(0) = G(0) = 0$).

Thresholding metric signatures



Thresholding metric signatures

Empirical signatures: $s_0(Y) \geq s_1(Y) \geq s_2(Y) \geq \cdots$ Define:

$$k_q(Y) := \max \left\{ j : s_{j-1}(Y) \geq q \right\}$$

Thresholding metric signatures

Empirical signatures: $s_0(Y) \geq s_1(Y) \geq s_2(Y) \geq \cdots$ Define:

$$k_q(Y) := \max\left\{j : s_{j-1}(Y) \geq q\right\}$$

Example 1:

$$f(x) = \begin{cases} 1 \text{ if } x \in [1/3, 2/3)\,, \\ 0 \text{ else}\,. \end{cases}$$

Thresholding metric signatures

Empirical signatures: $s_0(Y) \geq s_1(Y) \geq s_2(Y) \geq \cdots$ Define:

$$k_q(Y) := \max\left\{j : s_{j-1}(Y) \geq q\right\}$$

Example 1:

$$f(x) = \begin{cases} 1 \text{ if } x \in [1/3, 2/3)\ , \\ 0 \text{ else}\ . \end{cases}$$

Extreme value theory for i.i.d. normal noise: Largest value on $[1/3, 2/3)$ approaches $1 + \sqrt{2\log(n)}$, lowest value on complement approaches $-\sqrt{2\log(n)}$ with $\mathbb{P} \to 1$.

Thresholding metric signatures

Empirical signatures: $s_0(Y) \geq s_1(Y) \geq s_2(Y) \geq \cdots$ Define:

$$k_q(Y) := \max \left\{ j : s_{j-1}(Y) \geq q \right\}$$

Example 1:

$$f(x) = \begin{cases} 1 \text{ if } x \in [1/3, 2/3)\,, \\ 0 \text{ else}\,. \end{cases}$$

Observation: Thresholding persistence signatures at $q(n) = \sqrt{2\log(n)}$ and thresholding Kolmnogorov signatures at $q = 1/2$ detects single mode of f with $\mathbb{P} \to 1$.

Thresholding metric signatures

Example 2: (decreasing signal to noise ratio)

$$f_n(x) = \begin{cases} \delta_n \text{ if } x \in [1/3, 2/3)\ , \\ 0 \text{ else}\ . \end{cases}$$

Thresholding metric signatures

Example 2: (decreasing signal to noise ratio)

$$f_n(x) = \begin{cases} \delta_n \text{ if } x \in [1/3, 2/3)\ , \\ 0 \text{ else}\ . \end{cases}$$

Theorem (Bauer, Munk, Sieling, W.)

Let $\delta_n \sqrt{n} \to \infty$ and $\delta_n \sqrt{\log(n)} \to 0$. Then there exists no successful thresholding strategy for persistence signatures:

$$\limsup_{n\to\infty} \mathbb{P}\left(k^{\infty}_{q_n}(Y) = 1\right) < 1$$

for every possible thresholding sequence (q_n).

Thresholding metric signatures

Example 2: (decreasing signal to noise ratio)

$$f_n(x) = \begin{cases} \delta_n \text{ if } x \in [1/3, 2/3) , \\ 0 \text{ else} . \end{cases}$$

Theorem (Bauer, Munk, Sieling, W.)

Let $\delta_n \sqrt{n} \to \infty$ and $\delta_n \to 0$. Then thresholding Kolmogorov signatures at $q(n) = \delta_n/2$ detects single mode of f with $\mathbb{P} \to 1$.

Thresholding metric signatures

Example 3: (needle in heystack)

$$f_n(x) = \begin{cases} (1+\varepsilon)\sqrt{2\log(n)} \text{ if } x \in [j/n, (j+1)/n)\,, \\ 0 \text{ else}\,. \end{cases}$$

for $\varepsilon > 0$ and some j that is *not known a priori*.

Thresholding metric signatures

Example 3: (needle in heystack)

$$f_n(x) = \begin{cases} (1+\varepsilon)\sqrt{2\log(n)} \text{ if } x \in [j/n, (j+1)/n)\,, \\ 0 \text{ else}\,. \end{cases}$$

for $\varepsilon > 0$ and some j that is *not known a priori*.

Observations:

- Sup-norm thresholding $(Y_1, \ldots, Y_n)$ minimax efficient for detecting single mode of f [Donoho/Jin, Ingster/Suslina].
- No thresholding known for persistence or Kolmogorov.

Empirical Kolmogorov signatures

Theorem (Bauer, Munk, Sieling, W.)

Let $\delta > 0$*. Then*

$$\mathbb{P}\left(\max_{k\in\mathbb{N}_0} |s_k(Y) - s_k(f)| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2 n}{2\nu + 2\kappa\delta}\right) .$$

Moreover, for given probability $\alpha \in (0, 1)$*, one can construct non-asymptotic confidence bands:*

$$\mathbb{P}\left(s_k(f) \in \left[(s_k(Y) - \tau_n(\alpha))_+ \,, s_k(Y) + \tau_n(\alpha)\right] \text{ for all } k \in \mathbb{N}_0\right) \geq 1 - \alpha \,,$$

where $(x)_+ = \max(0, x)$ *and* $\tau_n(\alpha)$ *can be explicitly computed. Asymptotically:* $\tau_n(\alpha) \approx 1/\sqrt{n}$*.*

Empirical Kolmogorov signatures

Remarks:

- These are "honest" (non-asymptotic) confidence bands.
- No a priori assumption on f required.

Thresholding K-signatures – overest. modes

Theorem (Bauer, Munk, Sieling, W.)

Let f have at most k modes, and let $\alpha \in (0, 1)$. Then

$$\mathbb{P}\left(k_{\tau_n(\alpha)}(Y) > k\right) \leq \alpha \ ,$$

i.e., $\tau_n(\alpha)$ controls the probability of overestimating the number of modes of f.

Thresholding K-signatures – overest. modes

Theorem (Bauer, Munk, Sieling, W.)

Let f have at most k modes, and let $\alpha \in (0,1)$. Then

$$\mathbb{P}\left(k_{\tau_n(\alpha)}(Y) > k\right) \leq \alpha \ ,$$

i.e., $\tau_n(\alpha)$ controls the probability of overestimating the number of modes of f.

Fact: $\tau_n(\alpha)$ is independent of the number and magnitude of the modes of f. In this sense the result is universal.

Thresholding K-signatures – underest. modes

Remarks:

- Obtaining a universal result in opposite direction, i.e., controlling the probability of *underestimating* the number of modes, is more delicate.

Thresholding K-signatures – underest. modes

Remarks:

- Obtaining a universal result in opposite direction, i.e., controlling the probability of *underestimating* the number of modes, is more delicate.
- Without a priori information on the "smallest scales" of f, no method can provide a control for their underestimation [Donoho].

Thresholding K-signatures – underest. modes

Remarks:

- Obtaining a universal result in opposite direction, i.e., controlling the probability of *underestimating* the number of modes, is more delicate.
- Without a priori information on the "smallest scales" of f, no method can provide a control for their underestimation [Donoho].
- *Only* possible to provide a bound for underestimating those signatures of f that are larger than a certain threshold.

Thresholding K-signatures – underest. modes

Theorem (Bauer, Munk, Sieling, W.)

Let $\alpha \in (0,1)$. *Then*

$$\mathbb{P}\left(k_{\tau_n(\alpha)}(Y) < k_{2\tau_n(\alpha)}(f)\right) \leq \alpha \ .$$

Let f *have at most* k *modes. Then one has two-sided bound:*

$$\mathbb{P}\left(k_{2\tau_n(\alpha)}(f) \leq k_{\tau_n(\alpha)}(Y) \leq k\right) \geq 1 - \alpha \ .$$

Thresholding K-signatures – underest. modes

Theorem (Bauer, Munk, Sieling, W.)

Let $\alpha \in (0, 1)$. *Then*

$$\mathbb{P}\left(k_{\tau_n(\alpha)}(Y) < k_{2\tau_n(\alpha)}(f)\right) \leq \alpha \ .$$

Let f *have at most* k *modes. Then one has two-sided bound:*

$$\mathbb{P}\left(k_{2\tau_n(\alpha)}(f) \leq k_{\tau_n(\alpha)}(Y) \leq k\right) \geq 1 - \alpha \ .$$

Fixing α, one has $\tau_n(\alpha) \approx 1/\sqrt{n} \Rightarrow \exists C$ such that asymptotically by thresholding at $C/\sqrt{n}$, it can be guaranteed that all signatures of f above a certain threshold get detected with $\mathbb{P} \geq 1 - \alpha$.

Thresholding K-signatures – estimating modes

So far: No a priori information about f. But now:

Thresholding K-signatures – estimating modes

So far: No a priori information about f. But now:

Theorem (Bauer, Munk, Sieling, W.)

Let f have at most k modes and assume $s_{k-1}(f) \geq \epsilon$. Then

$$\mathbb{P}\left(k_{\epsilon/2}(Y) = k\right) \geq 1 - 2\exp\left(-\frac{\epsilon^2 n}{8v + 4\kappa\epsilon}\right).$$

Thresholding K-signatures – estimating modes

So far: No a priori information about f. But now:

Theorem (Bauer, Munk, Sieling, W.)

Let f have at most k modes and assume $s_{k-1}(f) \geq \epsilon$. Then

$$\mathbb{P}\left(k_{\epsilon/2}(Y) = k\right) \geq 1 - 2\exp\left(-\frac{\epsilon^2 n}{8v + 4\kappa\epsilon}\right).$$

Number of modes of f can be estimated correctly from empirical signatures with $\mathbb{P} \to 1$ *under the assumption* of a lower bound on magnitude (in the Kolmogorov norm) of the smallest mode of f. This is independent of the number of modes of f.

Computing Kolmogorov signatures – Taut strings

Definition (Taut strings)

Let $f \in L^\infty[a, b]$ with antiderivative F. The taut string U_α is the minimizer of

$$\int_a^b \sqrt{1 + U'_\alpha(t)^2}\, \mathrm{d}t$$

subject to $U_\alpha(a) = F(a), \;\; U_\alpha(b) = F(b), \;\; \|U - F\|_\infty \leq \alpha.$

Computing Kolmogorov signatures – Taut strings

Definition (Taut strings)

Let $f \in L^\infty[a, b]$ with antiderivative F. The taut string U_α is the minimizer of

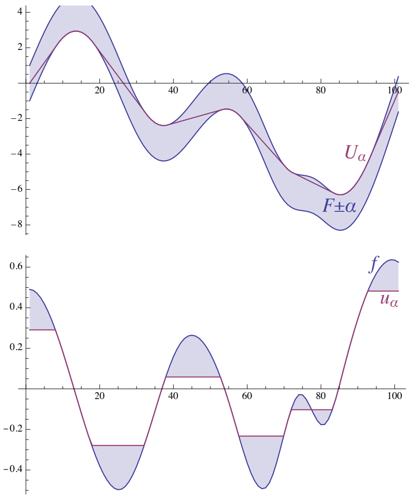
$$\int_a^b \sqrt{1 + U_\alpha'(t)^2}\, \mathrm{d}t$$

subject to $U_\alpha(a) = F(a), \;\; U_\alpha(b) = F(b), \;\; \|U - F\|_\infty \leq \alpha.$

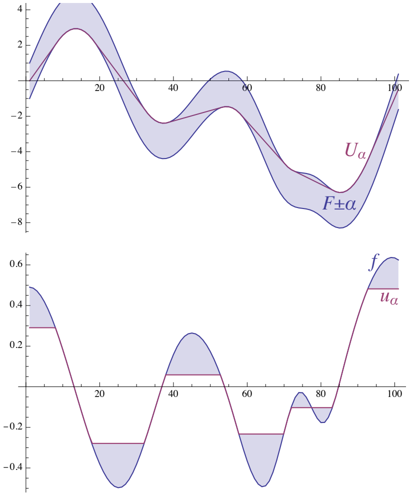
Theorem (Bauer, Munk, Sieling, W.)

The function $u_\alpha = U_\alpha'$ minimizes the number of modes among all functions u with $d_{\mathrm{Kol}}(f, u) \leq \alpha$.

Taut strings continued

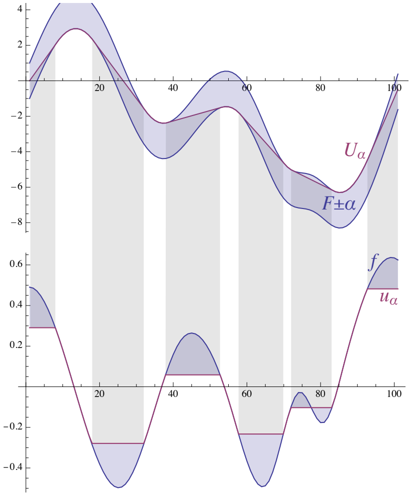


Taut strings continued



Observations:

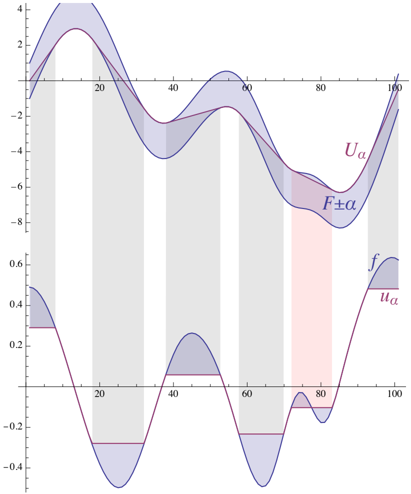
Taut strings continued



Observations:

- u_α coincides with f apart from some intervals, on which it is constant.

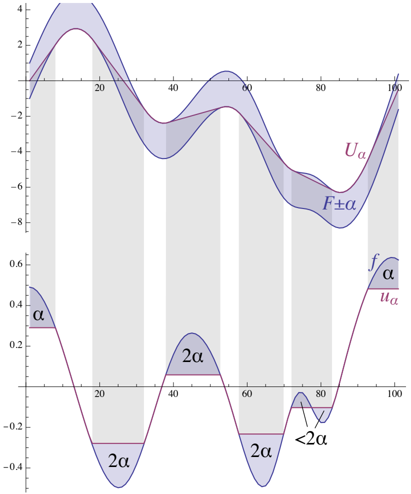
Taut strings continued



Observations:

- u_α coincides with f apart from some intervals, on which it is constant.
- New cancelation of critical points occurs for $\alpha_k = s_k^{\mathrm{Kol}}$.

Taut strings continued



Observations:

- u_α coincides with f apart from some intervals, on which it is constant.
- New cancelation of critical points occurs for $\alpha_k = s_k^{\text{Kol}}$.
- Kolmogorov signatures can be computed in $O(n \log(n))$.

Summary

- Honest confidence bands for Kolmogorov signatures.

Summary

- Honest confidence bands for Kolmogorov signatures.
- Universal bound on over-estimating modes.

Summary

- Honest confidence bands for Kolmogorov signatures.
- Universal bound on over-estimating modes.
- Exact estimation possible exponentially fast given smallest Kolmogorov signature of f.

Summary

- Honest confidence bands for Kolmogorov signatures.
- Universal bound on over-estimating modes.
- Exact estimation possible exponentially fast given smallest Kolmogorov signature of f.
- Distance-based analogy for persistence breaks for dim>2.

Summary

- Honest confidence bands for Kolmogorov signatures.
- Universal bound on over-estimating modes.
- Exact estimation possible exponentially fast given smallest Kolmogorov signature of f.
- Distance-based analogy for persistence breaks for dim>2.
- Kolmogorov approach breaks for dim>1.

Summary

- Honest confidence bands for Kolmogorov signatures.
- Universal bound on over-estimating modes.
- Exact estimation possible exponentially fast given smallest Kolmogorov signature of f.
- Distance-based analogy for persistence breaks for dim>2.
- Kolmogorov approach breaks for dim>1.
- But: Statistical questions persist for higher dim.

Summary

- Honest confidence bands for Kolmogorov signatures.
- Universal bound on over-estimating modes.
- Exact estimation possible exponentially fast given smallest Kolmogorov signature of f.
- Distance-based analogy for persistence breaks for dim>2.
- Kolmogorov approach breaks for dim>1.
- But: Statistical questions persist for higher dim.
- Reference: Bauer, Munk, Sieling, W.: *Persistence Barcodes versus Kolmogorov Signatures: Detecting Modes of One-Dimensional Signals*. Found Comput Math, 2017.

Thank you for your attention!