Scalar field analysis with aberrant noise

Mickaël Buchet

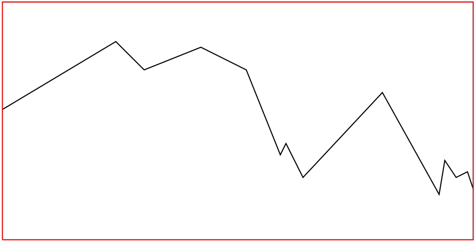
joint work with F. Chazal, T. Dey, F. Fan,
S. Oudot and Y. Wang

How many peaks do you see?

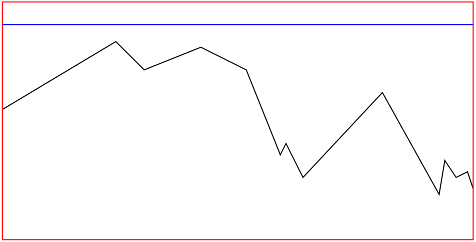
How many peaks do you see?

How many peaks do you see?

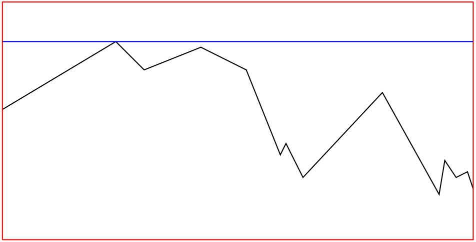
Persistence diagram



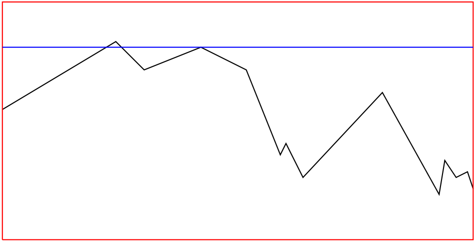
Persistence diagram



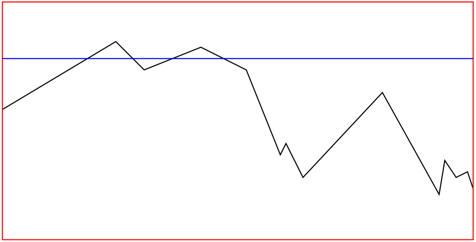
Persistence diagram



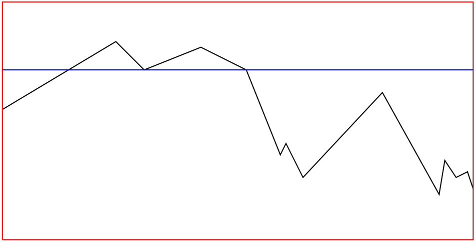
Persistence diagram



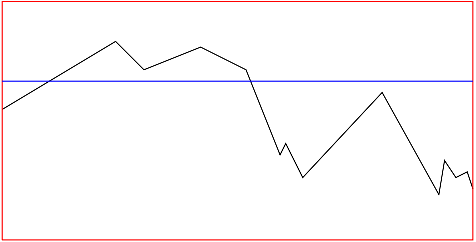
Persistence diagram



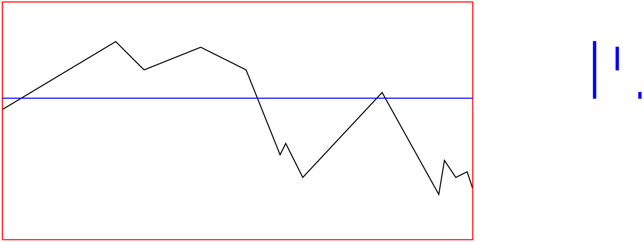
Persistence diagram



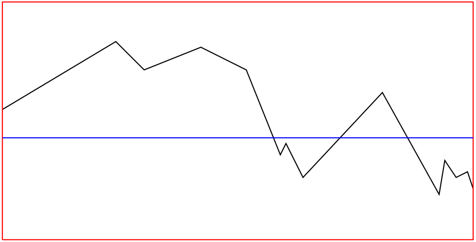
Persistence diagram

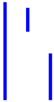


Persistence diagram

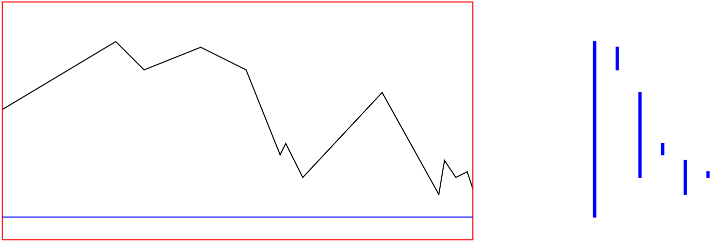


Persistence diagram

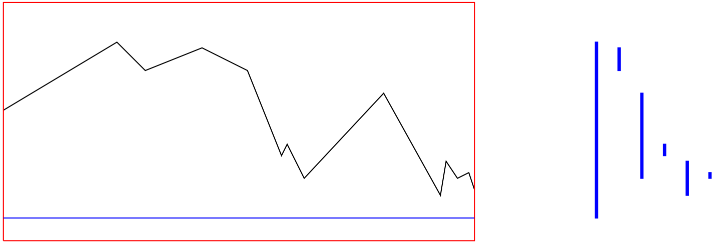


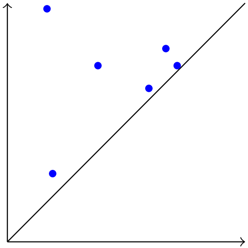


Persistence diagram

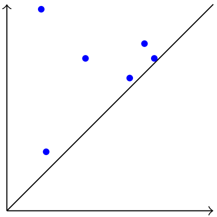


Persistence diagram

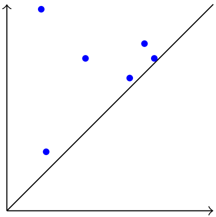


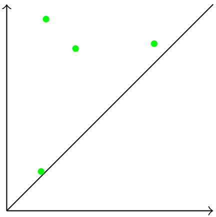


Comparison between persistence diagrams

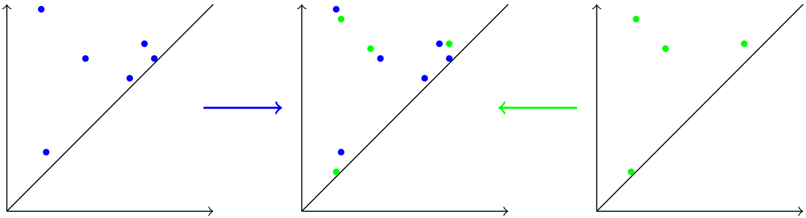


Comparison between persistence diagrams

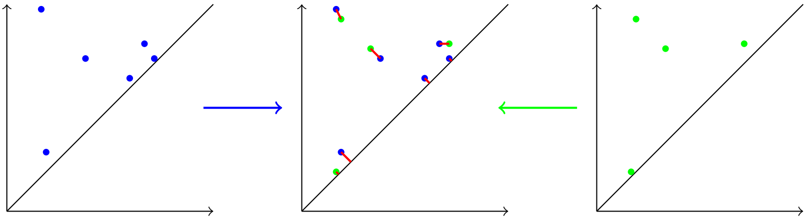




Comparison between persistence diagrams



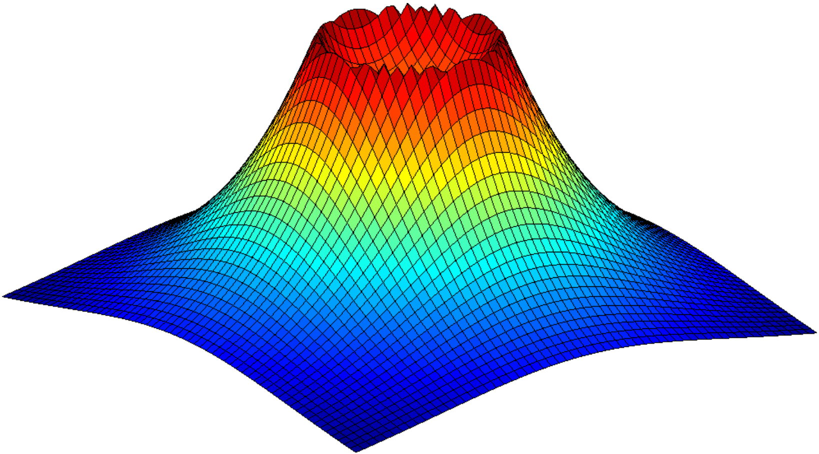
Comparison between persistence diagrams



Bottleneck distance:

$$d_B(D, E) = \inf_{b \in \mathcal{B}} \max_{x \in D} ||x - b(x)||_\infty$$

Higher dimensions



Applicative setting

The *ground truth*:

- a sub-manifold M of $\mathbb{R}^d$
- a c-Lipschitz function $f : \mathrm{M} \mapsto \mathbb{R}$

Applicative setting

The *ground truth*:

- a sub-manifold M of $\mathbb{R}^d$
- a c-Lipschitz function $f : \mathrm{M} \mapsto \mathbb{R}$

What we really know:

- a set of points $P \in \mathbb{R}^d$
- a function $\tilde{f} : P \mapsto \mathbb{R}$

Applicative setting

The *ground truth*:

- a sub-manifold M of $\mathbb{R}^d$
- a c-Lipschitz function $f : \mathrm{M} \mapsto \mathbb{R}$

What we really know:

- a set of points $P \in \mathbb{R}^d$
- a function $\tilde{f} : P \mapsto \mathbb{R}$

Approximate the persistence diagram D of f by a diagram $\hat{D}$

Previous work

From Chazal, Guibas, Oudot and Skraba (DCG'11)
Under some conditions on ϵ and the geometry of M,

Theorem

If P is an ϵ Riemannian sample of M *and* $||f|_P - \tilde{f}||_\infty \leq \xi$, *then:*

Previous work

From Chazal, Guibas, Oudot and Skraba (DCG'11)
Under some conditions on ϵ and the geometry of M,

Theorem

If P is an ϵ Riemannian sample of M *and* $||f|_P - \tilde{f}||_\infty \leq \xi$, *then:*

$$d_B(D, \hat{D}) \leq 4c\epsilon + \xi$$

Previous work

From Chazal, Guibas, Oudot and Skraba (DCG'11)
Under some conditions on ϵ and the geometry of M,

Theorem

If P is an ϵ Riemannian sample of M and $||f|_P - \tilde{f}||_\infty \leq \xi$, then:

$$d_B(D, \hat{D}) \leq 4c\epsilon + \xi$$

Theorem

If P is an ϵ Riemannian sample of M and the pairwise distances between points of P are known with precision ν, then:

Previous work

From Chazal, Guibas, Oudot and Skraba (DCG'11)
Under some conditions on ϵ and the geometry of M,

Theorem

If P is an ϵ Riemannian sample of M and $||f|_P - \tilde{f}||_\infty \leq \xi$, then:

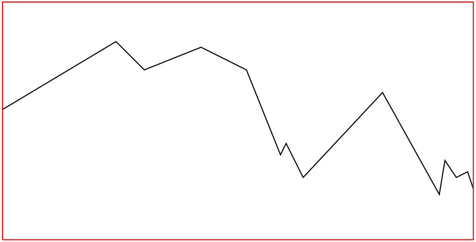
$$d_B(D, \hat{D}) \leq 4c\epsilon + \xi$$

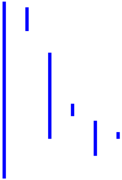
Theorem

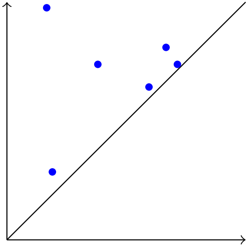
If P is an ϵ Riemannian sample of M and the pairwise distances between points of P are known with precision ν, then:

$$d_B(D, \hat{D}) \leq (4\epsilon + 2\nu)c$$

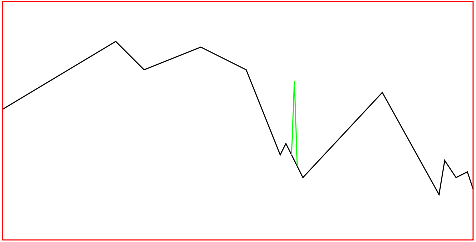
A bad example

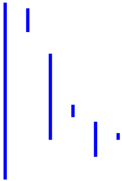


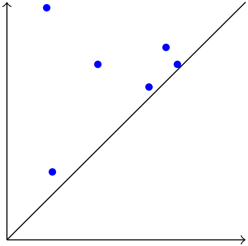




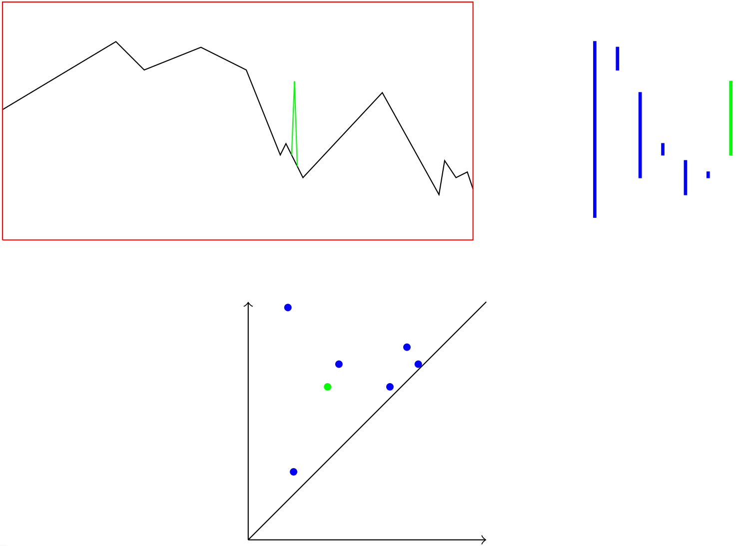
A bad example







A bad example



Filtered simplicial complex

Classical algorithms to compute a persistence diagram work on a filtered simplicial complex.

Filtered simplicial complex

Classical algorithms to compute a persistence diagram work on a filtered simplicial complex.

A simplicial complex is a set X of simplices such that for any simplex $\sigma \in X$, all facets of σ are also in X.

Filtered simplicial complex

Classical algorithms to compute a persistence diagram work on a filtered simplicial complex.

A simplicial complex is a set X of simplices such that for any simplex $\sigma \in X$, all facets of σ are also in X.

We say that X is filtered when there exists a family of simplicial complexes $\{X_\alpha\}_{\alpha\in\mathbb{R}}$ such that for all $\alpha < \beta$, $X_\alpha \subset X_\beta \subset X$.

The Rips complex

Topologies of $\{f^{-1}(]-\infty,\alpha)\}$ and $\{\tilde{f}^{-1}(]-\infty,\alpha])\}$ are completely different:

The Rips complex

Topologies of $\{f^{-1}(]-\infty,\alpha)\}$ and $\{\tilde{f}^{-1}(]-\infty,\alpha])\}$ are completely different:

The Rips complex

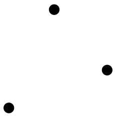
Topologies of $\{f^{-1}(]-\infty,\alpha)\}$ and $\{\tilde{f}^{-1}(]-\infty,\alpha])\}$ are completely different:

Giving thickness by building a Rips complex:

The Rips complex

Topologies of $\{f^{-1}(]-\infty,\alpha)\}$ and $\{\tilde{f}^{-1}(]-\infty,\alpha])\}$ are completely different:

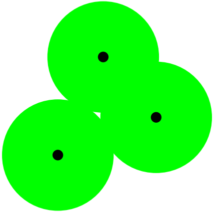
Giving thickness by building a Rips complex:



The Rips complex

Topologies of $\{f^{-1}(]-\infty, \alpha)\}$ and $\{\tilde{f}^{-1}(]-\infty, \alpha])\}$ are completely different:

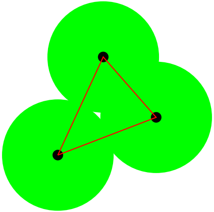
Giving thickness by building a Rips complex:



The Rips complex

Topologies of $\{f^{-1}(]-\infty,\alpha)\}$ and $\{\tilde{f}^{-1}(]-\infty,\alpha])\}$ are completely different:

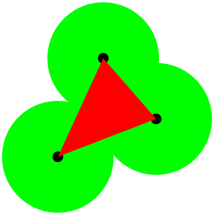
Giving thickness by building a Rips complex:



The Rips complex

Topologies of $\{f^{-1}(]-\infty,\alpha)\}$ and $\{\tilde{f}^{-1}(]-\infty,\alpha])\}$ are completely different:

Giving thickness by building a Rips complex:



Nested Rips complexes

Sometimes, there is no parameter such that the Rips complex capture the topology of M.

Nested Rips complexes

Sometimes, there is no parameter such that the Rips complex capture the topology of M.

Solution: take two different parameters $\delta < \delta'$ and look at the image of $H_*(R_\delta)$ by the morphism induced by the inclusion $R_\delta \hookrightarrow R_{\delta'}$.

Nested Rips complexes

Sometimes, there is no parameter such that the Rips complex capture the topology of M.

Solution: take two different parameters $\delta < \delta'$ and look at the image of $H_*(R_\delta)$ by the morphism induced by the inclusion $R_\delta \hookrightarrow R_{\delta'}$.

Effect of "cleaning" the persistence diagram.

Filtration by functional values

We study the scalar field f and not the manifold M.

Filtration by functional values

We study the scalar field f and not the manifold M.

We work for fixed parameters δ and δ' and we use the values of $\tilde{f}$ to build the filtration:

$$P_\alpha = \tilde{f}^{-1}(]-\infty, \alpha])$$

Filtration by functional values

We study the scalar field f and not the manifold M.

We work for fixed parameters δ and δ' and we use the values of $\tilde{f}$ to build the filtration:

$$P_\alpha = \tilde{f}^{-1}(]-\infty, \alpha])$$

$$\{R_\delta(P_\alpha) \hookrightarrow R_{\delta'}(P_\alpha)\}_{\alpha \in \mathbb{R}}$$

Theoretical guarantees

From Chazal, Guibas, Oudot and Skraba (DCG'11)
Let $\varrho(\mathrm{M})$ be the strong convexity radius of M.

Theoretical guarantees

From Chazal, Guibas, Oudot and Skraba (DCG'11)
Let $\varrho(\mathrm{M})$ be the strong convexity radius of M.

Theorem

If P is an ϵ Riemannian sample of M, $||f|_P - \tilde{f}||_\infty \leq \xi$ *and* $\epsilon < \frac{1}{4}\varrho(\mathrm{M})$*:*

$$\forall \delta \in [2\epsilon, \frac{1}{2}\varrho(\mathrm{M})[, d_B(\mathrm{Dgm}(f), \mathrm{Dgm}(R_\delta(P_\alpha) \hookrightarrow R_{2\delta}(P_\alpha))) \leq 2c\delta + \xi$$

Theoretical guarantees

From Chazal, Guibas, Oudot and Skraba (DCG'11)
Let $\varrho(\mathrm{M})$ be the strong convexity radius of M.

Theorem

If P is an ϵ Riemannian sample of M, $||f|_P - \tilde{f}||_\infty \leq \xi$ *and* $\epsilon < \frac{1}{4}\varrho(\mathrm{M})$*:*

$$\forall \delta \in [2\epsilon, \frac{1}{2}\varrho(\mathrm{M})[, d_B(\mathrm{Dgm}(f), \mathrm{Dgm}(R_\delta(P_\alpha) \hookrightarrow R_{2\delta}(P_\alpha))) \leq 2c\delta + \xi$$

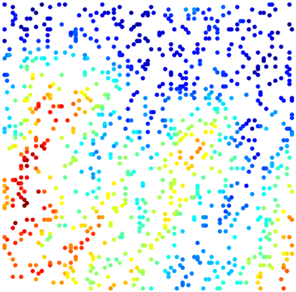
Theorem

If P is an ϵ Riemannian sample of M *and the distance between points of P are given by a function $\tilde{d}$ such that* $\frac{d_\mathrm{M}(x,y)}{\lambda} \leq \tilde{d}(x,y) \leq \nu + \mu\frac{d_\mathrm{M}(x,y)}{\lambda}$*, then:*

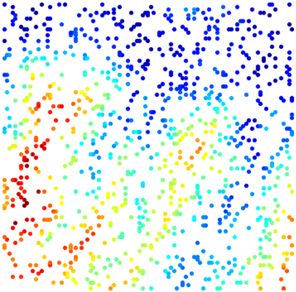
$$\forall \delta \geq \nu + 2\mu\frac{\epsilon}{\lambda},\ \delta' \in [\nu + 2\mu\delta, \frac{1}{\lambda}\varrho(\mathrm{M})[,$$

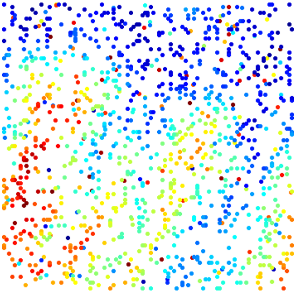
$$d_B(\mathrm{Dgm}(f), \mathrm{Dgm}(R_\delta(P_\alpha) \hookrightarrow R_{\delta'}(P_\alpha)) \leq c\lambda\delta'$$

Sources of noise

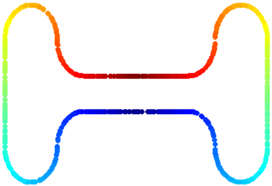


Sources of noise

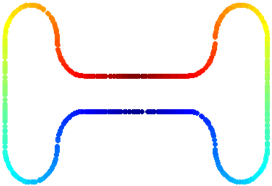


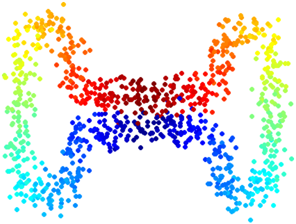


Sources of noise

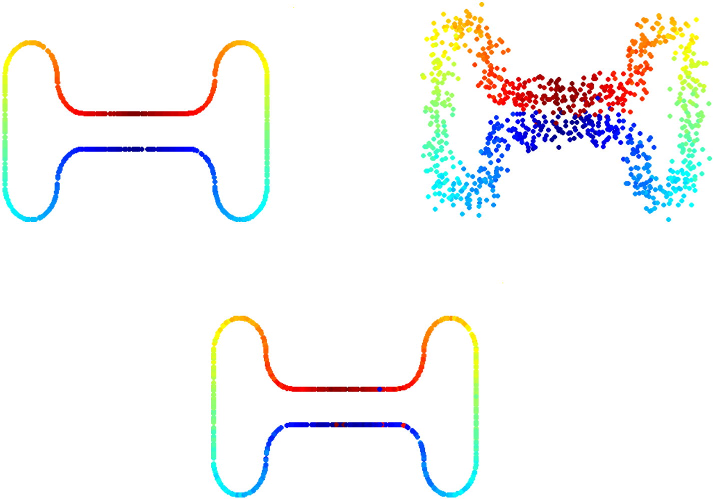


Sources of noise





Sources of noise



Discrepancy

We compute new functionnal values for points :

Discrepancy

We compute new functionnal values for points :

For every point p in P:

Discrepancy

We compute new functionnal values for points :

For every point p in P:

1. Build the set $NN_k(p)$ of its k nearest neighbors.

Discrepancy

We compute new functionnal values for points :

For every point p in P:

1. Build the set $NN_k(p)$ of its k nearest neighbors.
2. Find the subset Y of k' values in $\tilde{f}(NN_k(p))$ with the smallest variance.

Discrepancy

We compute new functionnal values for points :

For every point p in P:

1. Build the set $NN_k(p)$ of its k nearest neighbors.
2. Find the subset Y of k' values in $\tilde{f}(NN_k(p))$ with the smallest variance.
3. Fix the new function value $\hat{f}(p)$ as the barycenter of Y.

A variant using the median

We compute new functionnal values for points :

For every point p in P:

1. Build the set $NN_k(p)$ of its k nearest neighbors.
2. Fix the new function value $\hat{f}(p)$ as the median of $\tilde{f}(NN_k(p))$.

Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

Asymptotic behaviour

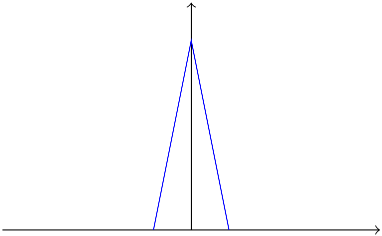
When $k \to \infty$ and $\frac{k}{n} \to 0$.

Probability distribution around the correct value:

Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

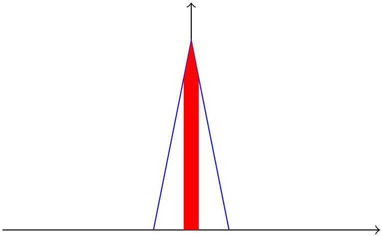
Probability distribution around the correct value:



Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

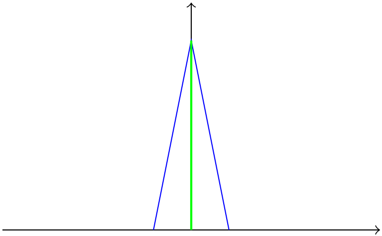
Probability distribution around the correct value:



Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

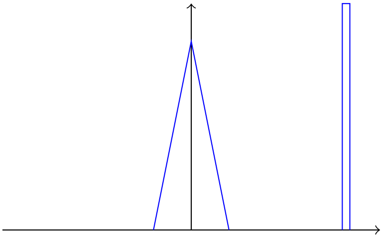
Probability distribution around the correct value:



Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

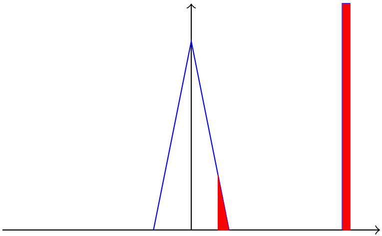
Probability distribution around the correct value:



Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

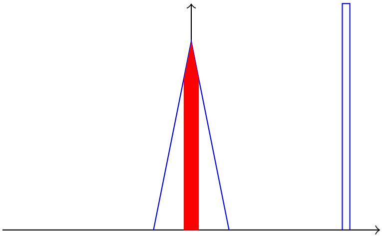
Probability distribution around the correct value:



Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

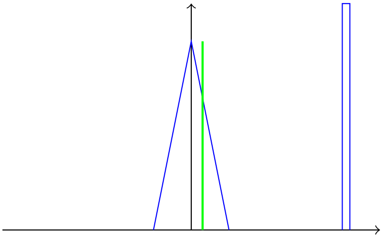
Probability distribution around the correct value:



Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

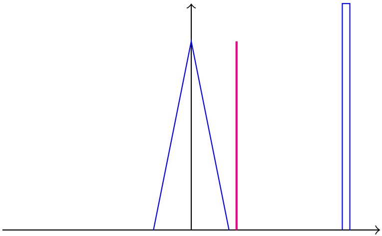
Probability distribution around the correct value:



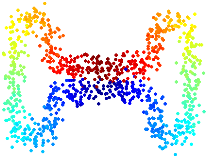
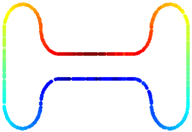
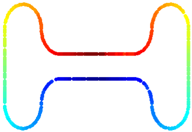
Asymptotic behaviour

When $k \to \infty$ and $\frac{k}{n} \to 0$.

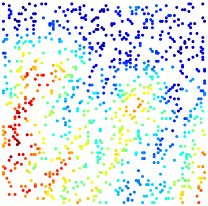
Probability distribution around the correct value:



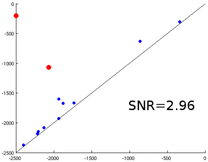
Bone

	Noisy input	k-NN regression	Discrepancy
			
Max	16.23	3.18	0.37
Mean	0.349	.204	.097

Persistence of topographic map

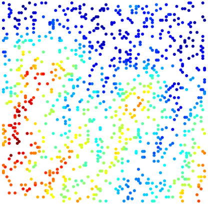


Topographic map

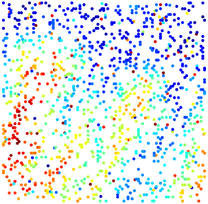


Original persistence diagram

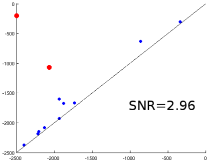
Persistence of topographic map



Topographic map

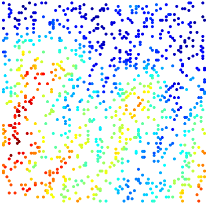


Noisy topographic map

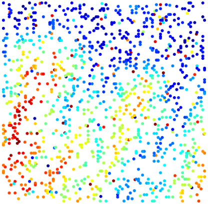


Original persistence diagram

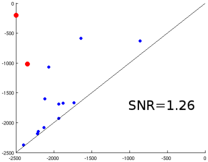
Persistence of topographic map



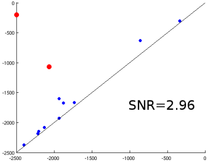
Topographic map



Noisy topographic map

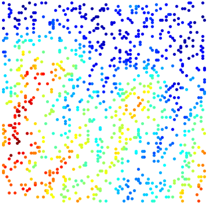


Noisy persistence diagram

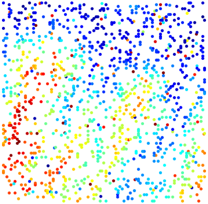


Original persistence diagram

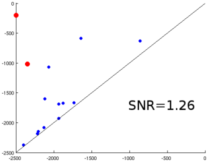
Persistence of topographic map



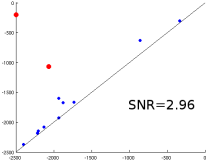
Topographic map



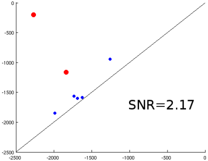
Noisy topographic map



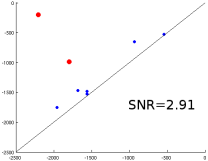
Noisy persistence diagram



Original persistence diagram



k-NN persistence diagram



Discrepancy persistence diagram

Images

No noise

Images

No noise

40% outliers

Images

No noise

40% outliers

kNN

Images

No noise

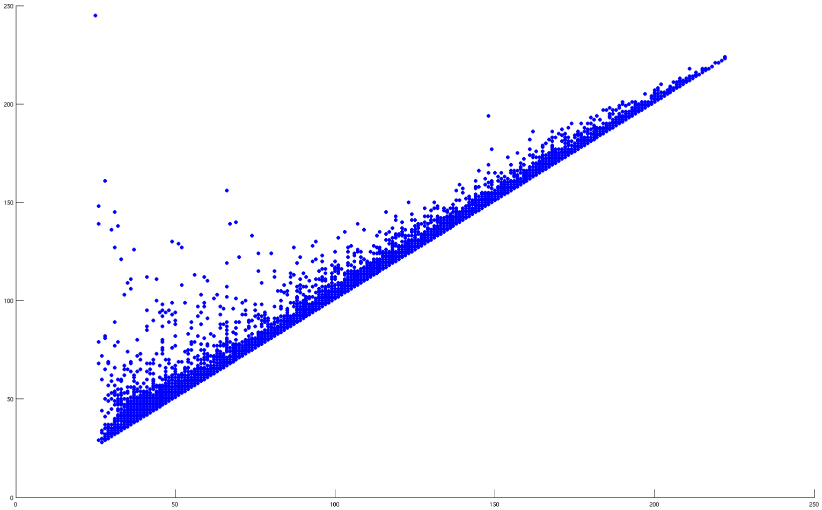
40% outliers

kNN

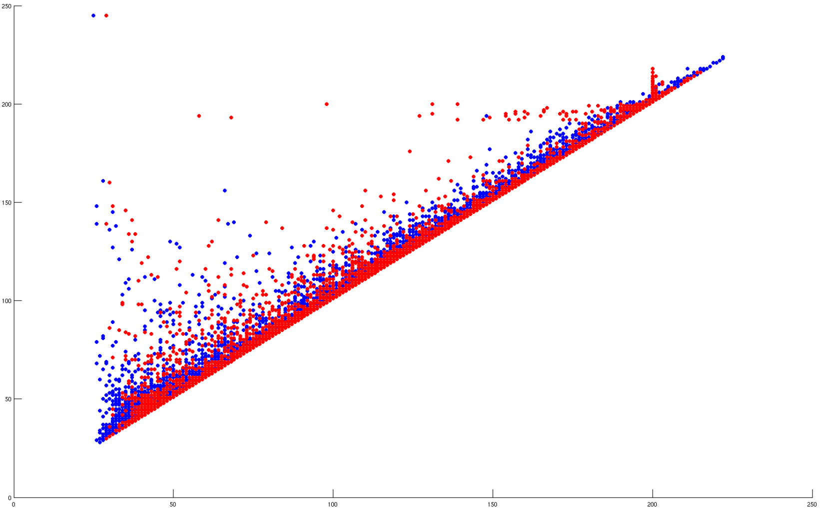
Discrepancy

Median

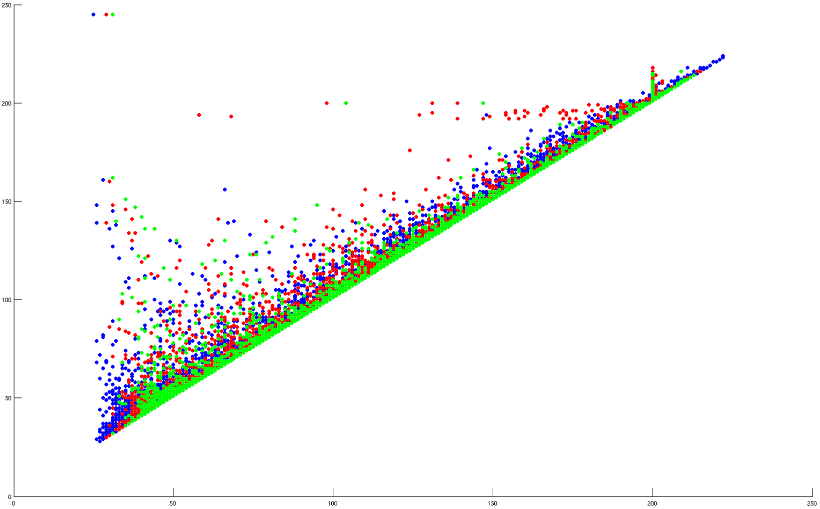
Lena's diagrams



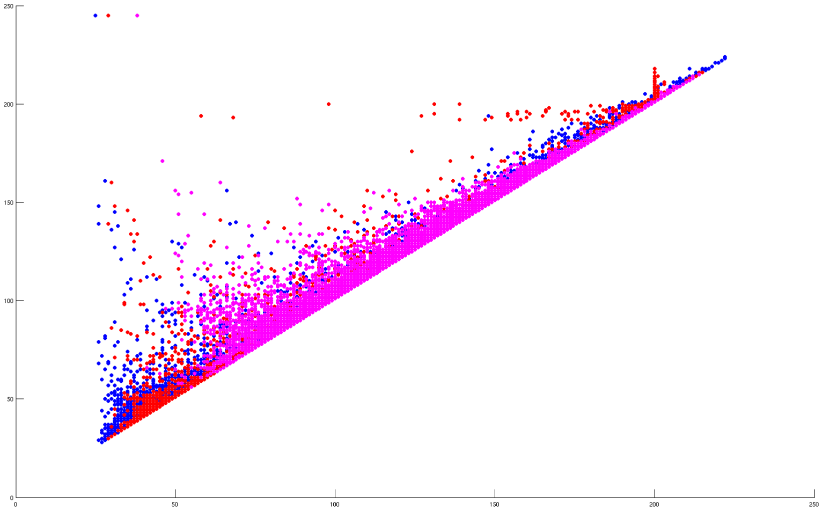
Lena's diagrams



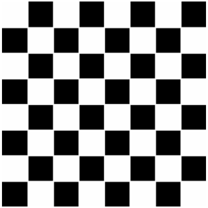
Lena's diagrams



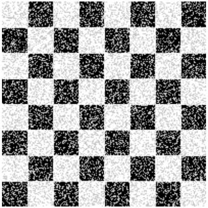
Lena's diagrams



Chessboard

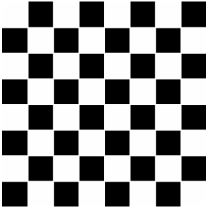


No noise

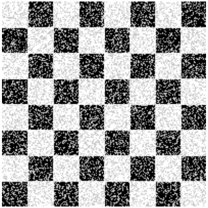


30% outliers

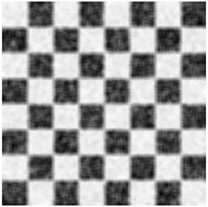
Chessboard



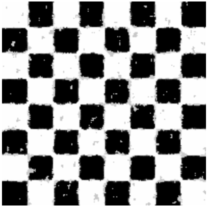
No noise



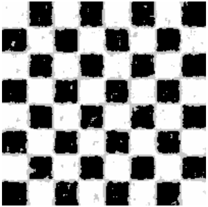
30% outliers



kNN



Discrepancy



Median

Building a complex with good properties

We need a complex that has the correct geometric structure to analyze the scalar field.

Building a complex with good properties

We need a complex that has the correct geometric structure to analyze the scalar field.

- Use of the super-level sets of a density estimator
- Use of the sub-level sets of a distance-like function.

Distance to measure in a nutshell

The distance to a measure is a distance-like function designed to cope with outliers.

$$d_{\mu,m}(p) = \sqrt{\frac{1}{k} \sum_{x \in NN_k(p)} ||p - x||^2}$$

Distance to measure in a nutshell

The distance to a measure is a distance-like function designed to cope with outliers.

$$d_{\mu,m}(p) = \sqrt{\frac{1}{k} \sum_{x \in NN_k(p)} ||p - x||^2}$$

- Easy to compute pointwise.

Distance to measure in a nutshell

The distance to a measure is a distance-like function designed to cope with outliers.

$$d_{\mu,m}(p) = \sqrt{\frac{1}{k} \sum_{x \in NN_k(p)} ||p - x||^2}$$

- Easy to compute pointwise.
- Guarantees of geometric inference [Chazal, Cohen-Steiner, Mérigot, 2011]

Distance to measure in a nutshell

The distance to a measure is a distance-like function designed to cope with outliers.

$$d_{\mu,m}(p) = \sqrt{\frac{1}{k} \sum_{x \in NN_k(p)} ||p - x||^2}$$

- Easy to compute pointwise.
- Guarantees of geometric inference
  [Chazal, Cohen-Steiner, Mérigot, 2011]
- Properties of a density estimator
  [Biau, Chazal, Cohen-Steiner, Devroye, Rodrigues, 2011]

A complete noise model

Three conditions:

A complete noise model

Three conditions:

1. Dense sampling:

$$\forall x \in \mathrm{M},\ d_{\mu,m}(x) \leq \epsilon$$

A complete noise model

Three conditions:

1. Dense sampling:

$$\forall x \in \mathrm{M},\ d_{\mu,m}(x) \leq \epsilon$$

2. No cluster of noise:

$$r = \sup\{l \in \mathbb{R} | \forall x,\ d_{\mu,m}(x) < l \implies d(x, \mathrm{M}) \leq d_{\mu,m}(x) + \epsilon\}$$

A complete noise model

Three conditions:

1. Dense sampling:

$$\forall x \in \mathrm{M},\ d_{\mu,m}(x) \leq \epsilon$$

2. No cluster of noise:

$$r = \sup\{l \in \mathbb{R} | \forall x,\ d_{\mu,m}(x) < l \implies d(x, \mathrm{M}) \leq d_{\mu,m}(x) + \epsilon\}$$

3. For any point close to M, most of the neighboring values are good:

$$\forall p \in d_{\mu,m}^{-1}(]-\infty, \eta]),\ |\{q \in NN_k(p) |\, |\tilde{f}(q) - f(\pi(p))| \leq s\} \geq k'$$

Theoretical results

Theorem

If P is a set verifying the previous conditions anf f is a c-Lipschitz fucntion then:

$$\forall \delta \in [2\eta + 6\epsilon, \frac{\varrho(\mathrm{M})}{2}],\ \delta' \in [2\eta + 2\epsilon + \frac{2R_M}{R_M - (\eta + \epsilon)}\delta, \frac{R_M - (\eta + \epsilon)}{R_M}\varrho(M)],$$

$$d_B(\mathrm{Dgm}(f), \hat{D}) \leq \left(\frac{cR_M\delta'}{R_M - (\eta + \epsilon)} + \xi s \right)$$

with $\xi = 1$ for the median and $\xi = 1 + 2\sqrt{\frac{k-k'}{2k'-k}}$ for the discrepancy.

Take home

Take home

- A versatile and model free algorithm for functional denoising

Take home

- A versatile and model free algorithm for functional denoising
- Scalar field analysis with noise in both the geometry and the functional values

Take home

- A versatile and model free algorithm for functional denoising
- Scalar field analysis with noise in both the geometry and the functional values

But...

Take home

- A versatile and model free algorithm for functional denoising
- Scalar field analysis with noise in both the geometry and the functional values

But...

- The algorithm needs some parameters.

Take home

- A versatile and model free algorithm for functional denoising
- Scalar field analysis with noise in both the geometry and the functional values

But...

- The algorithm needs some parameters.
- Heuristics exist but there is no general method to choose their value.