

Multivariate Topological Data Analysis

Peter Bubenik

Cleveland State University

November 20, 2008, Duke University
joint work with Gunnar Carlsson (Stanford), Peter Kim and
Zhiming Luo (Guelph), and Moo Chung (Wisconsin–Madison)

Ideal Truth
(Parameter)

Observed Reality
(Statistic)

Function on a compact
manifold

measurement

Sampled Data

Filter using level sets

Topological Description
via Persistent Homology

Ideal Truth
(Parameter)

Observed Reality
(Statistic)

Function on a compact
manifold

measurement

Sampled Data

Filter using level sets

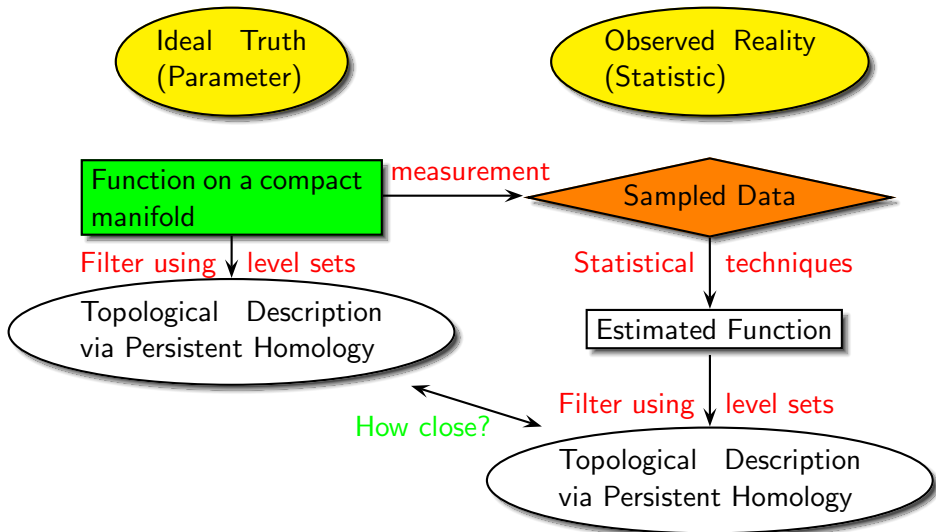
Topological Description
via Persistent Homology

Filtered Simplicial Complex

How close?

Topological Description
via Persistent Homology

Framework



Persistent homology describes the homological features which persist as a single parameter changes.

Here, we take this parameter to be a **threshold on the function** on the space from which we are sampling.

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ and assume that if $f'(x) = 0$, then $f''(x) \neq 0$, ie, f has non-degenerate critical points.

This means each critical point is a local minimum or a local maximum.

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ and assume that if $f'(x) = 0$, then $f''(x) \neq 0$, ie, f has non-degenerate critical points.

This means each critical point is a local minimum or a local maximum.

Define the sublevel sets $\mathbb{R}_{f \leq t} = f^{-1}(-\infty, t]$, $t \in \mathbb{R}$.

Remark

As t increases, the topology of $\mathbb{R}_{f \leq t}$ does not change as long as we do not pass a critical value.

At the critical points we have the following effects:

- At a local minimum, a new component is added.
- At a local maximum, two components are merged.

At the critical points we have the following effects:

- At a local minimum, a new component is added.
- At a local maximum, two components are merged.

Pairing critical points: Pair a local maximum, with the higher (younger) of the two local minimum associated with the two components which it joins.

At the critical points we have the following effects:

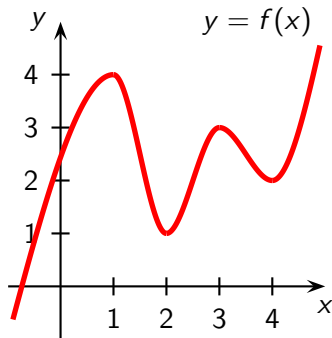
- At a local minimum, a new component is added.
- At a local maximum, two components are merged.

Pairing critical points: Pair a local maximum, with the higher (younger) of the two local minimum associated with the two components which it joins.

If we graph all pairs of critical points we obtain the (Reduced) Persistence Diagram.

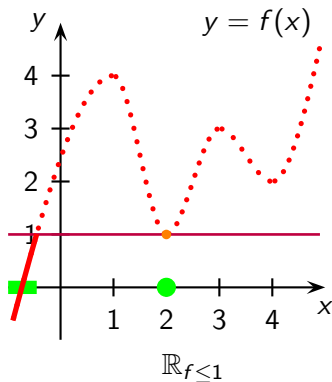
Persistence Diagram for Functions on \mathbb{R}

For example:



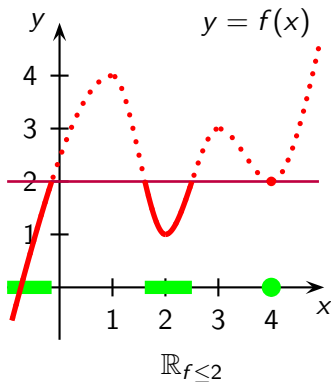
Persistence Diagram for Functions on \mathbb{R}

For example:



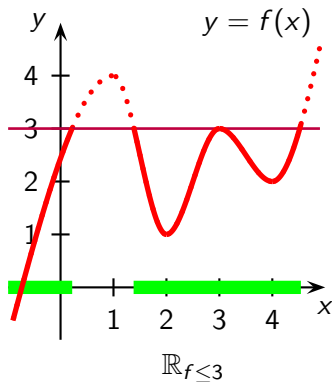
Persistence Diagram for Functions on \mathbb{R}

For example:



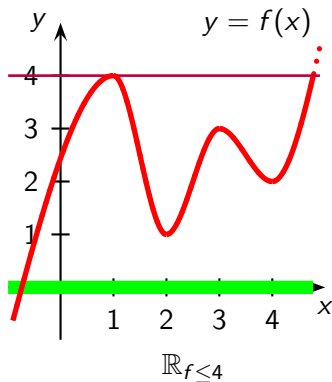
Persistence Diagram for Functions on \mathbb{R}

For example:



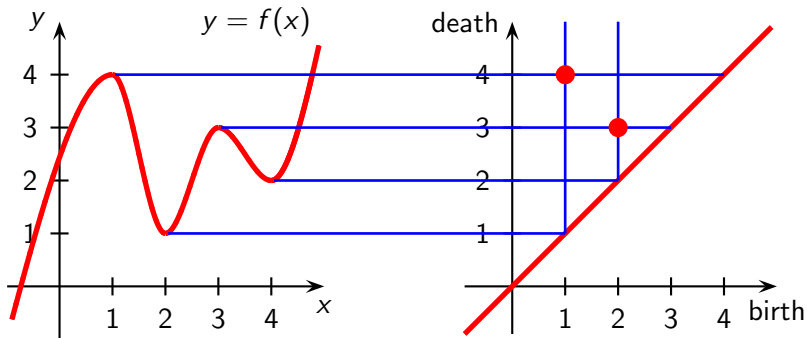
Persistence Diagram for Functions on \mathbb{R}

For example:



Persistence Diagram for Functions on \mathbb{R}

For example:



Persistent homology for densities on manifolds

Let (\mathbb{M}, g) be a compact, connected Riemannian manifold, with a dominating measure ν (the invariant measure which in local coordinates is $\sqrt{|g|}dx_1 \dots dx_d$).

Let $f : \mathbb{M} \rightarrow [0, \infty]$ such that $\int_{\mathbb{M}} f d\nu = 1$.

This probability density gives an increasing filtration of \mathbb{M} by sublevel sets

$$\mathbb{M}_{f \leq r} = \{x \in \mathbb{M} \mid f(x) \leq r\}.$$

Persistent homology for densities on manifolds

Let (\mathbb{M}, g) be a compact, connected Riemannian manifold, with a dominating measure ν (the invariant measure which in local coordinates is $\sqrt{|g|}dx_1 \dots dx_d$).

Let $f : \mathbb{M} \rightarrow [0, \infty]$ such that $\int_{\mathbb{M}} f d\nu = 1$.

This probability density gives an increasing filtration of \mathbb{M} by sublevel sets

$$\mathbb{M}_{f \leq r} = \{x \in \mathbb{M} \mid f(x) \leq r\}.$$

This induces an increasing filtration on $C_*(\mathbb{M})$,

the **Morse filtration**: $\mathcal{F}_r(C_*(\mathbb{M})) = C_*(\mathbb{M}_{f \leq r})$,

from which we can calculate the **persistent homology** of f .

The statistical viewpoint

We assume that $f = f_\theta$ belongs to a family of densities

$$\{f_\theta \mid \theta \in \Theta\}$$

where θ is the **parameter** and Θ is the **parameter space** which can be finite dimensional (parametric), or, infinite dimensional (nonparametric).

Goal

Find an estimate $\hat{\theta}$ of θ so that the persistent homology of $f_{\hat{\theta}}$ is close to the persistent homology of f_θ .

Functions on \mathbb{M} – Review of Morse Theory

Let $f : \mathbb{M} \rightarrow \mathbb{R}$ be a Morse function with distinct critical values $t_0 < t_1 < \cdots < t_k$. Recall that $M_{f \leq t} = f^{-1}(\infty, t]$.

Functions on \mathbb{M} – Review of Morse Theory

Let $f : \mathbb{M} \rightarrow \mathbb{R}$ be a Morse function with distinct critical values $t_0 < t_1 < \cdots < t_k$. Recall that $M_{f \leq t} = f^{-1}(\infty, t]$.

The index p of f at the critical point associated to the critical value t_j is the number of negative eigenvalues of the Hessian.

Functions on \mathbb{M} – Review of Morse Theory

Let $f : \mathbb{M} \rightarrow \mathbb{R}$ be a Morse function with distinct critical values $t_0 < t_1 < \dots < t_k$. Recall that $M_{f \leq t} = f^{-1}(\infty, t]$.

The index p of f at the critical point associated to the critical value t_j is the number of negative eigenvalues of the Hessian.

Morse theory says that $\mathbb{M}_{f \leq t_j}$ is homotopy equivalent to $\partial \mathbb{M}_{f \leq t_{j-1}}$ together with an attached p -dimensional cell.

Consequently, β_p increases by one (positive) or β_{p-1} decreases by one (negative).

Functions on \mathbb{M} – Review of Morse Theory

Let $f : \mathbb{M} \rightarrow \mathbb{R}$ be a Morse function with distinct critical values $t_0 < t_1 < \dots < t_k$. Recall that $M_{f \leq t} = f^{-1}(\infty, t]$.

The index p of f at the critical point associated to the critical value t_j is the number of negative eigenvalues of the Hessian.

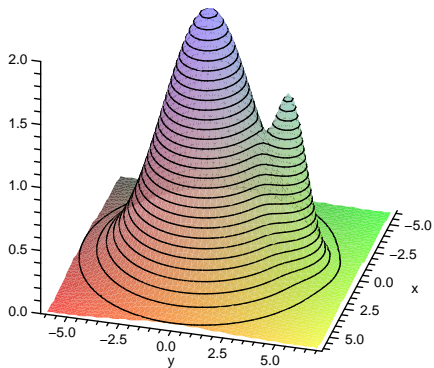
Morse theory says that $\mathbb{M}_{f \leq t_j}$ is homotopy equivalent to $\partial \mathbb{M}_{f \leq t_{j-1}}$ together with an attached p -dimensional cell.

Consequently, β_p increases by one (positive) or β_{p-1} decreases by one (negative).

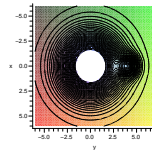
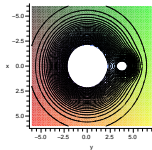
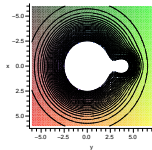
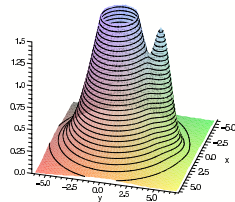
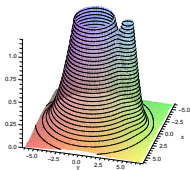
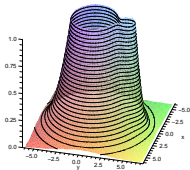
Pairing of the positive critical points of index p and the negative critical points of index $p + 1$ produces the degree- p Persistence Diagram, denoted by $D_p(f)$.

A Function

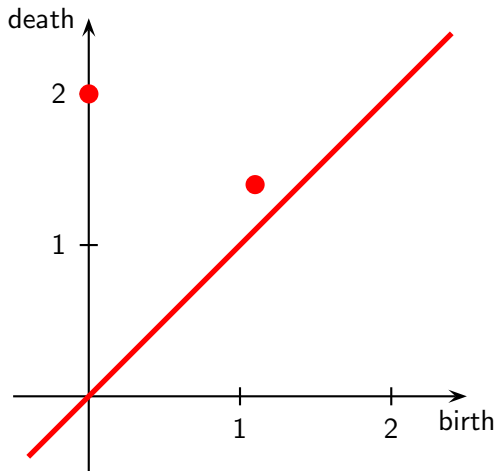
For example



Sublevel sets and H_1



Persistence Diagram of the function



We now want to define a metric on the space of Persistence Diagrams.

We now want to define a metric on the space of Persistence Diagrams.

Let $f, g : \mathbb{M} \rightarrow \mathbb{R}$ be two Morse functions, with associated Persistence Diagrams $D_p(f)$ and $D_p(g)$.

Definition

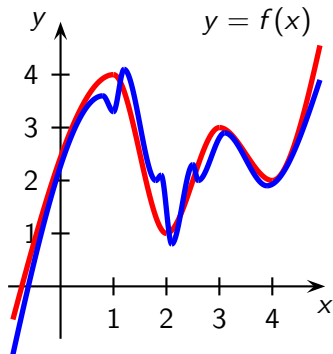
The Bottleneck distance is given by

$$d_B(D_p(f), D_p(g)) = \inf_{\eta} \sup_x \|x - \eta(x)\|_{\infty},$$

where the infimum is taken over all bijections $\eta : D_p(f) \rightarrow D_p(g)$ and the supremum is taken over all points $x \in D_p(f)$.

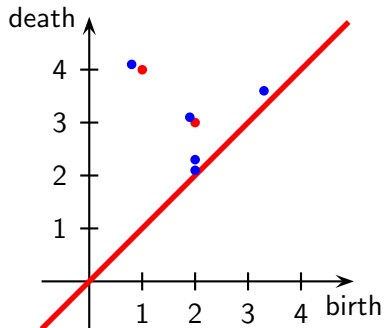
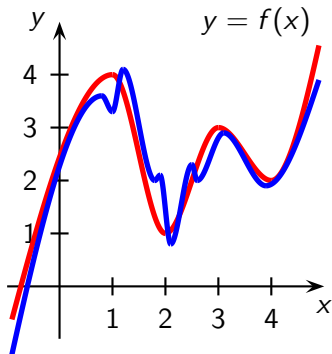
Bottleneck Distance

For example:



Bottleneck Distance

For example:



The following fundamental result bounds the bottleneck distance for persistence diagrams with the better-known supremum norm.

Theorem (Cohen-Steiner, Edelsbrunner, Harer)

$$d_B(D_p(f), D_p(g)) \leq \|f - g\|_\infty$$

The following fundamental result bounds the bottleneck distance for persistence diagrams with the better-known supremum norm.

Theorem (Cohen-Steiner, Edelsbrunner, Harer)

$$d_B(D_p(f), D_p(g)) \leq \|f - g\|_\infty$$

For us, this result is crucial as it allows us to connect topology to statistics.

Take f to be an unknown function and \hat{f} to its statistical estimator. Then,

$$d_B(D_p(f), D_p(\hat{f})) \leq \|f - \hat{f}\|_\infty$$

Nonparametric regression

Let \mathbb{M} be a manifold. We would like to be able to solve the following **Nonparametric Regression Problem**.

Problem

We assume that there exists a function $f : \mathbb{M} \rightarrow \mathbb{R}$ such that

$$y = f(x) + \epsilon, \quad x \in \mathbb{M}$$

where ϵ is a normal random variable with mean zero and variance $\sigma^2 > 0$. Given a sample $(y_1, x_1), \dots, (y_n, x_n)$, find an estimator \hat{f} of f .

We want to find an estimator \hat{f} that minimizes $\|f - \hat{f}\|_\infty$ and calculate the asymptotics as $n \rightarrow \infty$.

The precise setup

Let \mathbb{M} be a compact, connected m -dimensional Riemannian manifold with Riemannian metric $\rho(\cdot, \cdot)$ (given by geodesic distance) and volume element $d\omega$.

Let \mathbb{M} be a compact, connected m -dimensional Riemannian manifold with Riemannian metric $\rho(\cdot, \cdot)$ (given by geodesic distance) and volume element $d\omega$.

As the parameter space we take the Hölder class of functions on \mathbb{M}

$$\Lambda(\beta, L) = \left\{ f : \mathbb{M} \rightarrow \mathbb{R} \mid |f(x) - f(y)| \leq L\rho(x, y)^\beta, x, y \in \mathbb{M} \right\},$$

where $0 < \beta \leq 1$.

Definition

The expected loss (risk) of an estimator \tilde{f}_ε is given by

$$\mathbb{E}_f \|\tilde{f}_\varepsilon - f\|_\infty.$$

Theorem (B-C-C-K-L)

For the regression model,

$$\sup_{f \in \Lambda(\beta, L)} \mathbb{E}_f \|\tilde{f} - f\|_\infty \geq C \psi_n$$

as $n \rightarrow \infty$, where $\psi_n = \left(\frac{\log(n)}{n} \right)^{\beta/(2\beta+d)}$ and

$$C = L^{d/(2\beta+d)} \left(\frac{\sigma^2 \text{vol}(\mathbb{M})(\beta+d)d^2}{\text{vol}(S^{d-1})\beta^2} \right)^{\frac{\beta}{(2\beta+d)}}$$

In fact, this lower bound can be achieved.

Partition \mathbb{M} by $A_i \subset M$, $i = 1, \dots, N$ and let

$$\hat{f}_\varepsilon(x) = \sum_{i=1}^N \hat{a}_i I_{A_i}(x).$$

In fact, this lower bound can be achieved.

Partition \mathbb{M} by $A_i \subset M$, $i = 1, \dots, N$ and let

$$\hat{f}_\varepsilon(x) = \sum_{i=1}^N \hat{a}_i I_{A_i}(x).$$

By a suitable choice of A_i, \hat{a}_i , $i = 1, \dots, N$ (constructive) one has

Theorem (B-C-C-K-L)

Using this estimator,

$$\sup_{f \in \Lambda(\beta, L)} \mathbb{E} \|\hat{f}_\varepsilon - f\|_\infty \leq C\psi_\varepsilon$$

as $\varepsilon \rightarrow 0$.

Corollary

In the white noise model

$$\mathbb{E}d_B(D_p(\hat{f}_\varepsilon), D_p(f)) \leq C\psi_\varepsilon$$

as $\varepsilon \rightarrow 0$.

- Topological features of functions can be described using topological persistence.
- For nonparametric regression problems, statistical estimators can be used to recover the topology of the function.