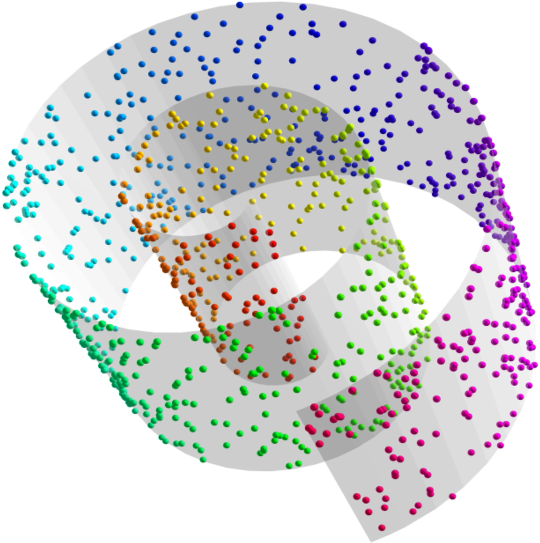
Estimating the Reach of a Manifold

Jisu Kim (Carnegie Mellon University)
Joint work with Eddie Aamari, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, Larry Wasserman

2018-03-10

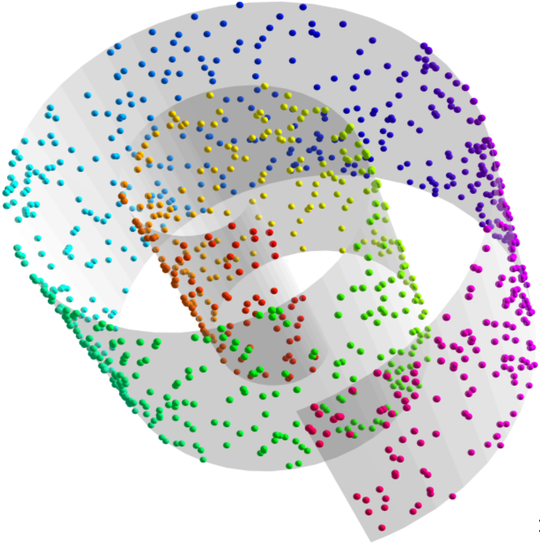
Manifold learning finds an underlying manifold to reduce dimension.



[1]

[1] http://www.skybluetrades.net/blog/posts/2011/10/30/machine-learning/

Positive reach is a common regularity assumption for manifold learning.



2

[2] http://www.skybluetrades.net/blog/posts/2011/10/30/machine-learning/

Positive reach is the minimal regularity assumption in geometric inference.

- The reach is a key parameter in:
    - Manifold learning
    - Homology inference
    - Volume estimation
    - Manifold clustering
    - Dimension estimation and reduction

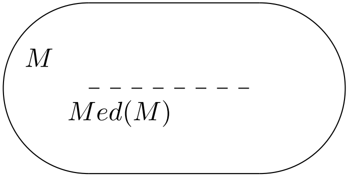
Estimating the reach of a manifold is studied.

E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman.
Estimating the Reach of a Manifold.
ArXiv e-prints, May 2017.

- The concept of reach is introduced and geometric condition for how reach is attained is studied.
- Reach estimator is presented with its statistical efficiency.
- The upper and lower bounds on the minimax rate for estimating the reach is presented.

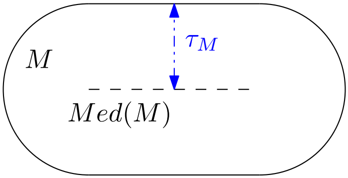
The medial axis of a set M is the set of points that have at least two nearest neighbors on the set M.

- $$Med(M) = \{z \in \mathbb{R}^D : \text{ there exists } p \neq q \in M \text{ with } \|p - z\| = \|q - z\| = d(z, M)\}.$$



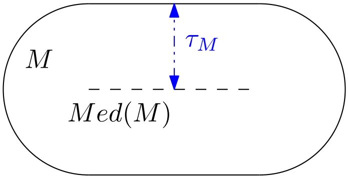
The reach of M, denoted by τ_M, is the minimum distance from $Med(M)$ to M.

- $$\tau_M = \inf_{x \in Med(M), y \in M} \|x - y\|.$$



The reach τ_M gives the maximum offset size of M on which the projection is well defined.

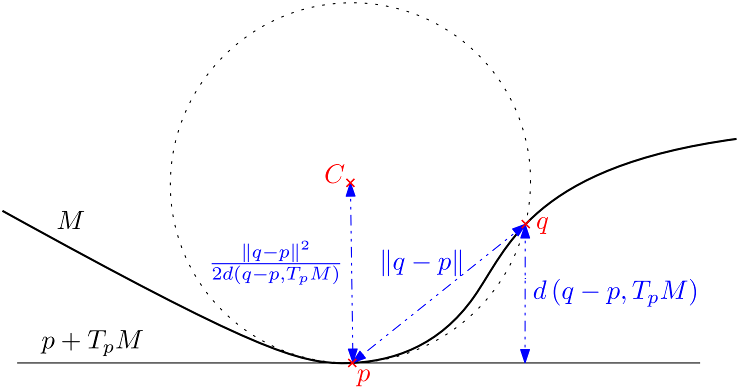
- $$\tau_M = \inf_{x \in Med(M), y \in M} \|x - y\| .$$



The reach τ_M gives the maximum radius of a ball that you can roll over M.

- When $M \subset \mathbb{R}^D$ is a manifold,

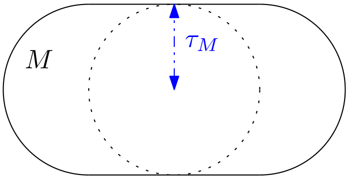
$$\tau_M = \inf_{q \neq p \in M} \frac{\|q-p\|^2}{2d(q-p, T_pM)}.$$



The reach τ_M gives the maximum radius of a ball that you can roll over M.

- When $M \subset \mathbb{R}^D$ is a manifold,

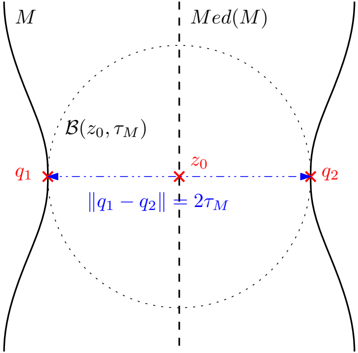
$$\tau_M = \inf_{q \neq p \in M} \frac{\|q - p\|^2}{2d(q - p, T_p M)}.$$



The bottleneck is a geometric structure where the manifold is nearly self-intersecting.

Definition

(Definition 3.1. in [1]) A pair of points (q_1, q_2) in M is said to be a bottleneck of M if there exists $z_0 \in Med(M)$ such that $q_1, q_2 \in \mathcal{B}(z_0, \tau_M)$ and $\|q_1 - q_2\| = 2\tau_M$.

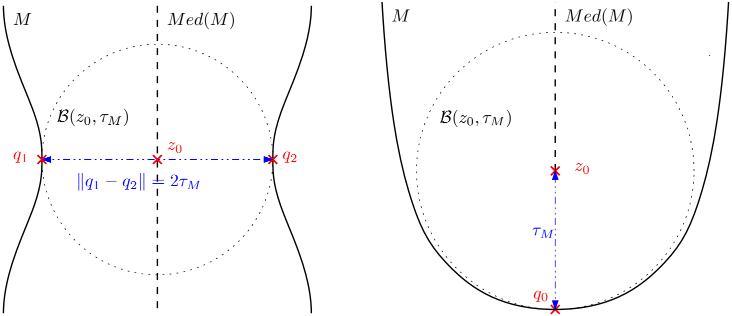


The reach is attained either from the bottleneck (global case) or the area of high curvature (local case).

Theorem

(Theorem 3.4 in [1]) At least one of the following two assertions holds:

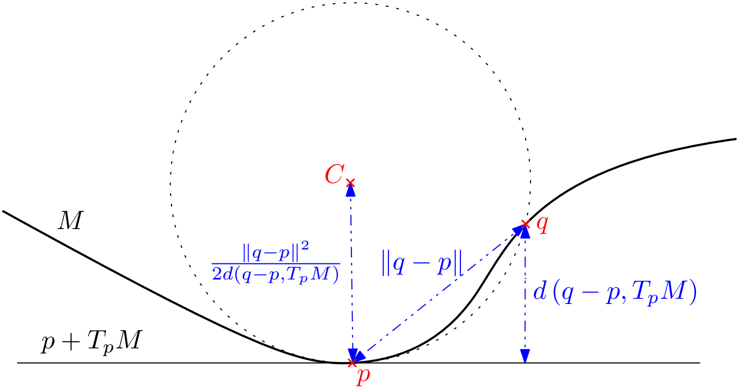
- *(Global Case)* M *has a bottleneck* $(q_1, q_2) \in M^2$.
- *(Local case) There exists* $q_0 \in M$ *and an arc-length parametrized* γ_0 *such that* $\gamma_0(0) = q_0$ *and* $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$.



The reach τ_M gives the maximum radius of a ball that you can roll over M.

- When $M \subset \mathbb{R}^D$ is a manifold,

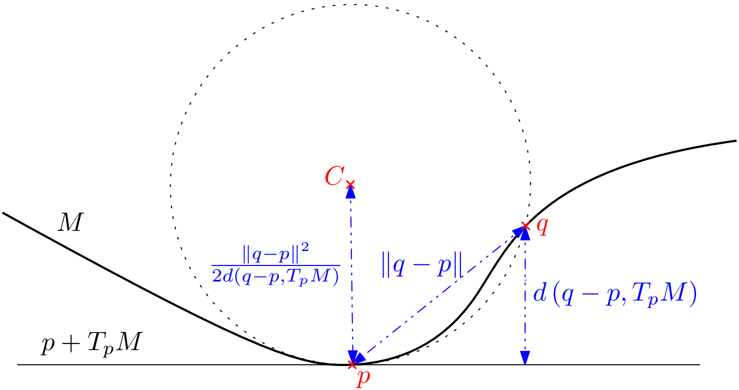
$$\tau_M = \inf_{q \neq p \in M} \frac{\|q - p\|^2}{2d(q - p, T_p M)}.$$



We define the reach estimator $\hat{\tau}$ as the maximum radius of a ball that you can roll over the point cloud.

- Let $\mathbb{X} = \{x_1, \ldots, x_n\}$ be a finite point cloud, then the reach estimator $\hat{\tau}$ is a plugin estimator as

$$\hat{\tau}(\mathbb{X}) = \inf_{x_i \neq x_j \in \mathbb{X}} \frac{\|x_j - x_i\|^2}{2d(x_j - x_i, T_{x_i}M)}.$$



The statistical efficiency of the reach estimator $\hat{\tau}$ is analyzed through its risk.

- The risk of the estimator $\hat{\tau}$ is the expected loss the estimator.

$$\mathbb{E}_{P^{(n)}}\left[\ell\left(\hat{\tau}(\mathbb{X}),\ \tau_M\right)\right].$$

  - $\mathbb{X} = \{X_1, \ldots, X_n\}$ is drawn from a fixed distribution P with its support M.
  - The loss function used is $\ell(\tau, \tau') = \left|\frac{1}{\tau} - \frac{1}{\tau'}\right|^p$, $p \geq 1$.

The risk of the reach estimator $\hat{\tau}$ is analyzed.

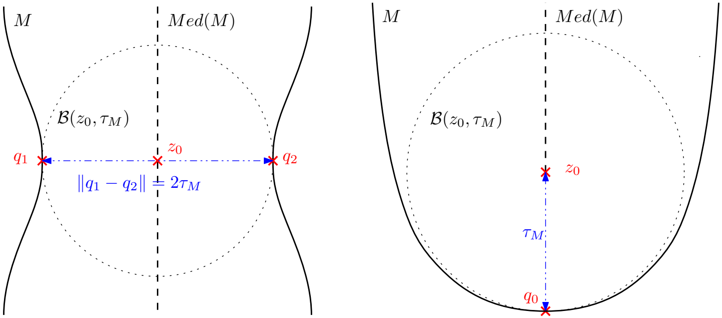
- The risk of the estimator $\hat{\tau}$ is the expected loss the estimator

$$\mathbb{E}_{P^{(n)}}\left[\left|\frac{1}{\tau_M}-\frac{1}{\hat{\tau}(\mathbb{X})}\right|^p\right].$$

  - $\mathbb{X}=\{X_1,\ldots,X_n\}$ is drawn from a fixed distribution P with its support M.
  - The loss function used is $\ell(\tau,\tau')=\left|\frac{1}{\tau}-\frac{1}{\tau'}\right|^p$, $p\geq 1$.

The reach estimator has the risk of $O\left(n^{-\frac{2p}{3d-1}}\right)$.

- The reach estimator has the risk of $O\left(n^{-\frac{p}{d}}\right)$ for the global case.
- The reach estimator has the risk of $O\left(n^{-\frac{2p}{3d-1}}\right)$ for the local case.

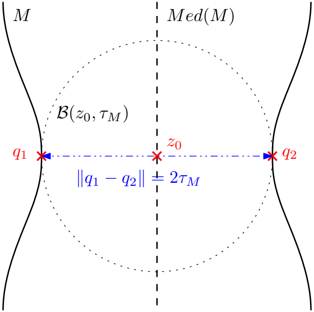


The reach estimator has the maximum risk of $O\left(n^{-\frac{p}{d}}\right)$ for the global case.

Proposition

(Proposition 4.3 in [1]) Assume that the support M has a bottleneck. Then,

$$\mathbb{E}_{P^n}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathbb{X})}\right|^p\right] \lesssim n^{-\frac{p}{d}}.$$

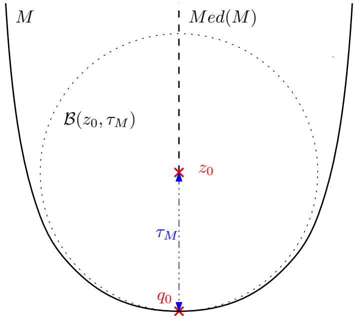


The reach estimator has the maximum risk of $O\left(n^{-\frac{2p}{3d-1}}\right)$ for the local case.

Proposition

(Proposition 4.7 in [1]) Suppose there exists $q_0 \in M$ and a geodesic γ_0 with $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$. Then,

$$\mathbb{E}_{P^n}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathbb{X})}\right|^p\right] \lesssim n^{-\frac{2p}{3d-1}}.$$



The statistical difficulty of the reach estimation problem is analyzed by the minimax rate.

- Minimax rate is the risk of an estimator that performs best in the worst case, as a function of sample size.
-
$$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\ell \left(\hat{\tau}_n(\mathbb{X}),\ \tau_M \right) \right].$$
  - $\mathbb{X} = \{X_1, \ldots, X_n\}$ is drawn from a fixed distribution P with its support M, where P is contained in set of distributions $\mathcal{P}$.
  - An estimator $\hat{\tau}_n$ is any function of data $\mathbb{X}$.
  - The loss function used is $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^p$, $p \geq 1$.

The statistical difficulty of the reach estimation problem is analyzed by the minimax rate.

- Minimax rate is the risk of an estimator that performs best in the worst case, as a function of sample size.
- $$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n(\mathbb{X})} \right|^p \right].$$
  - $\mathbb{X} = \{X_1, \ldots, X_n\}$ is drawn from a fixed distribution P with its support M, where P is contained in set of distributions $\mathcal{P}$.
  - An estimator $\hat{\tau}_n$ is any function of data $\mathbb{X}$.
  - The loss function used is $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^p$, $p \geq 1$.

The maximum risk of our estimator provides an upper bound on the minimax rate.

$$
\begin{aligned}
R_n &= \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n(\mathbb{X})} \right|^p \right] \\
&\leq \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathbb{X})} \right|^p \right]}_{\text{the maximum risk of our estimator}}
\end{aligned}
$$

Minimax rate is upper bounded by $O\left(n^{-\frac{2p}{3d-1}}\right)$.

Theorem

(Theorem 5.1 in [1])

$$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n}\right|^p\right] \lesssim n^{-\frac{2p}{3d-1}}.$$

Le Cam's lemma provides a lower bound based on the reach difference and the statistical difference of two distributions.

- Total variance distance between two distributions is defined as

$$TV(P, P') = \sup_{A \in \mathcal{B}(\mathbb{R}^D)} \left| P(A) - P'(A) \right| .$$

Lemma

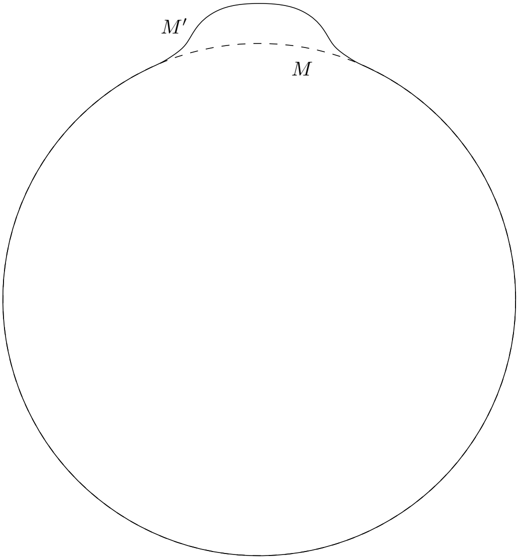
(Lemma 5.2 in [1]) Let $P, P' \in \mathcal{P}$ with respective supports M and M'. Then

$$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n} \right|^p \right] \gtrsim \left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^p \left(1 - TV(P, P')\right)^{2n} .$$

Two distributions P, P' are found so that their reaches differ but they are statistically difficult to distinguish.

- $$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n} \right|^p \right] \gtrsim \left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^p \left(1 - TV(P, P')\right)^{2n}.$$
- The lower bound measures how hard it is to tell whether the data is from distributions with different reaches.
- P and P' are found so that $\left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^p$ is large while $(1 - TV(P, P'))^{2n}$ is small.

P is a distribution supported on a sphere while P' is a distribution supported on a bumped sphere.



Mimimax rate is lower bounded by $\Omega\left(n^{-\frac{p}{d}}\right)$.

Proposition

(Proposition 5.6 in [1])

$$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n}\right|^p\right] \gtrsim n^{-\frac{p}{d}}.$$

Thank you!

Geometric assumptions are imposed to avoid an arbitrary complicated manifold.

- We let $\mathcal{M}^{d,D}_{\tau_{\min},L}$ denote the set of compact d-dimensional submanifolds $M \subset \mathbb{R}^D$ without boundary such that
  - the reach τ_M of M is lower bounded by $\tau_{\min}$, i.e. $\tau_M \geq \tau_{\min}$.
  - every arc-length parametrized geodesic γ on M has 3^{rd} derivative bounded by L, i.e. $\|\gamma'''(0)\| \leq L$.

Statistics assumptions are imposed to avoid an arbitrary complicated distribution.

- We let $\mathcal{Q}^{d,D}_{\tau_{\min},L,f_{\min}}$ denote the set of distributions Q having support $M \in \mathcal{M}^{d,D}_{\tau_{\min},L}$ and with a density $f = \frac{dQ}{dvol_M}$ satisfying $f \geq f_{\min} > 0$ on M.

We assume tangent spaces are observed with the data.

- Data takes the form $(X_1, T_{X_1}M), \ldots, (X_n, T_{X_n}M)$, where $X_1, \ldots, X_n$ are i.i.d. from a distribution $Q \in \mathcal{Q}^{d,D}_{\tau_{\min},L,f_{\min}}$.
- We let the corresponding distribution of $(X, T_X M)$ be P, and let $\mathcal{P}^{d,D}_{\tau_{\min},L,f_{\min}}$ be the set of distributions P.