Intrinsic Dimensionality Estimation for Data Sets

Yoon-Mo Jung, Jason Lee, Anna V. Little, Mauro Maggioni
Department of Mathematics, Duke University

Lorenzo Rosasco
Center for Biological and Computational Learning, MIT

September 1, 2009

Problem: We consider a novel approach for estimating the intrinsic dimensionality of high-dimensional point clouds. Assuming that the points are sampled from a k-dimensional data set corrupted by D-dimensional noise, with $k << D$, we estimate dimensionality via a new multiscale algorithm that generalizes PCA. The algorithm exploits the low-dimensional structure of the data, so that its power depends on k rather than D.

Dimensionality estimation is important in many applications in machine learning, including:

1. signal processing
2. discovering number of variables in linear models
3. molecular dynamics
4. genetics
5. financial data

PCA Approach

Counting number of "significant" singular values is classical technique in dimensionality estimation. When data is linear and noiseless, this method cannot fail.

Idea:

- Consider data points $x^1, x^2 \ldots x^n$ in $\mathbb{R}^D$.
- Form normalized data matrix:

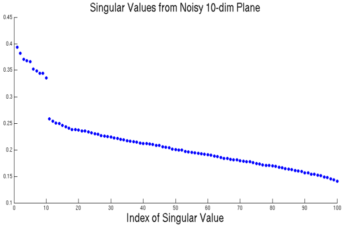
$$X = \frac{1}{\sqrt{n}} \begin{bmatrix} -x^1- \\ -x^2- \\ \ldots\ldots \\ -x^n- \end{bmatrix}$$

- Let $C = X^T X$ (the covariance matrix).
- Compute singular values of X ($\sigma_i(X) = \sqrt{\lambda_i(C)}, i = 1 \ldots D$).

Issues with PCA Approach

- Finite sample case is not completely understood; how many data points do we need for accurate results?
- Noise confuses the dimensionality.

Example:
Sample 1000 points from 10-dim plane in $\mathbb{R}^{100}$; corrupt with Gaussian noise of level $\sigma = .2$ ($.2\ N(0, I_{100})$ added to each point)



- Non-linear data results in overestimation of the dimensionality.

Model: Manifold plus Noise

1. Let $\mathcal{M}$ be manifold of dimension k embedded in $\mathbb{R}^D$ (bounded curvature).
2. Let $x^1, x^2, ..., x^n$ be n samples.
3. Suppose data is corrupted by D-dimensional noise:
   $\tilde{x}^n = x^n + \sigma\eta^n$ (e.g. $\eta \sim N(0, I_D)$)
4. Let:
   $$\tilde{X}_n = \begin{bmatrix} -\tilde{x}^1- \\ -\tilde{x}^2- \\ \cdots\cdots \\ -\tilde{x}^n- \end{bmatrix}$$
   be the corresponding noisy data matrix.
5. Goal: Estimate the dimensionality k w.h.p. from $\tilde{X}_n$.

Multiscale Algorithm to Estimate Pointwise Dimensionaliy

Fix z. Specify scale:

- Let $X(r) = \mathcal{M} \bigcap \mathcal{B}_z(r)$
- Let $X_n(r) = X_n \bigcap \mathcal{B}_z(r)$
- Let $\tilde{X}_n(r) = \tilde{X}_n \bigcap \mathcal{B}_z(r)$

Multiscale Algorithm to Estimate Pointwise Dimensionaliy

Fix z. Specify scale:

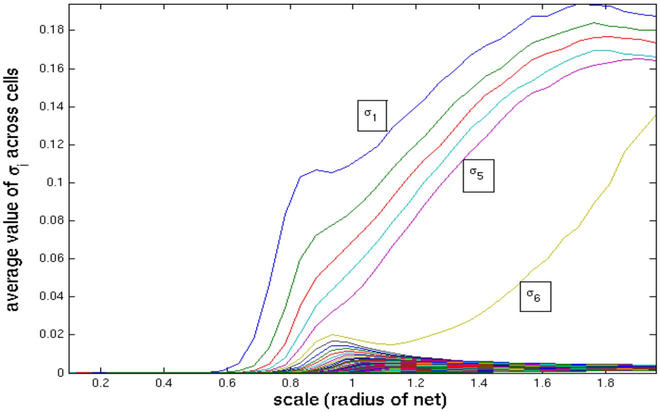
- Let $X(r) = \mathcal{M} \bigcap \mathcal{B}_z(r)$
- Let $X_n(r) = X_n \bigcap \mathcal{B}_z(r)$
- Let $\tilde{X}_n(r) = \tilde{X}_n \bigcap \mathcal{B}_z(r)$

Algorithm:

1. Let $\{\sigma_i^r\}_{i=1}^D$ be the singular values of $\tilde{X}_n(r)$.
2. Classify the σ_i as follows:
   - linear growth in r: tangent plane singular value
   - quadratic growth in r: curvature singular value
   - no growth in r: noise singular value
3. Dimensionality at z = number of tangent plane σ_i's

Example: Growth of Singular Values

- Consider $\mathbb{S}^5$ embedded in $\mathbb{R}^{100}$
- Take 1000 noisy samples ($\sigma = .05$)



Outline of Analysis, I

1. Approximate the data set by a linear manifold $X^{||}(r)$ and a normal correction $X^{\perp}(r)$. It turns out that $\mathrm{cov}(X(r)) = \mathrm{cov}(X^{||}(r)) + O(\kappa^2 r^4)$, with $||\mathrm{cov}(X(r))|| \sim O(r^2)$.
   $\longrightarrow$ *upper bound on r* to avoid distortion due to curvature

Outline of Analysis, I

1. Approximate the data set by a linear manifold $X^{||}(r)$ and a normal correction $X^{\perp}(r)$. It turns out that $\mathrm{cov}(X(r)) = \mathrm{cov}(X^{||}(r)) + O(\kappa^2 r^4)$, with $||\mathrm{cov}(X(r))|| \sim O(r^2)$.
   $\longrightarrow$ *upper bound on r* to avoid distortion due to curvature
2. Apply sampling theorems for covariance matrices to bound distance between $\mathrm{cov}(X_n^{||}(r))$ and $\mathrm{cov}(X^{||}(r))$
   $\longrightarrow$ need $O(k \log k)$ points
   $\longrightarrow$ *lower bound on r* so that $X_n^{||}(r)$ contains enough points, i.e. $O(k \log k)$ w.h.p.

Outline of Analysis, I

1. Approximate the data set by a linear manifold $X^{||}(r)$ and a normal correction $X^{\perp}(r)$. It turns out that $\mathrm{cov}(X(r)) = \mathrm{cov}(X^{||}(r)) + O(\kappa^2 r^4)$, with $||\mathrm{cov}(X(r))|| \sim O(r^2)$.
   $\longrightarrow$ *upper bound on r* to avoid distortion due to curvature
2. Apply sampling theorems for covariance matrices to bound distance between $\mathrm{cov}(X_n^{||}(r))$ and $\mathrm{cov}(X^{||}(r))$
   $\longrightarrow$ need $O(k \log k)$ points
   $\longrightarrow$ *lower bound on r* so that $X_n^{||}(r)$ contains enough points, i.e. $O(k \log k)$ w.h.p.
3. Add ambient noise and bound w.h.p. its effect on the spectrum of $X_n^{||}(r)$, using results from random matrix theory and matrix perturbation.
   $\longrightarrow$ *lower bound on r* so that the tangent plane structure is distinguishable from the noise.

Outline of Analysis, II

1. Natural normalization: $\mathbb{E}[||\eta||^2_{\mathbb{R}^D}] = O(1)$ (e.g. $\sigma = \sigma_0 D^{-\frac{1}{2}}$). Under the niceness assumptions $\kappa = O(1)$ and $\sigma_0 = O(1)$, the algorithm succeeds w.h.p. with *only $O(k \log k)$ samples, independently of D*.
2. If $\mathbb{E}[||\eta||^2_{\mathbb{R}^D}]$ grows with D (e.g. linearly as when $\eta \sim \mathcal{N}(0, I_D)$), then for D large enough the algorithm fails w.h.p.
3. Consistency ($n \to +\infty$) of the algorithm follows trivially from our analysis with niceness assumptions on the noise and curvature.
4. The random matrix scaling limit ($n \to +\infty$, $D \to +\infty$, $\frac{n}{D} \to \gamma$) is a particular case of our analysis.

Comparison with other algorithms

Our algorithm:

- Requires $O(k \log k)$ points (under niceness assumptions on noise and curvature)
- Finite sample guarantees
- Only input: $\tilde{X}_n$
- Discovers correct scale using multiscale approach

Comparison with other algorithms

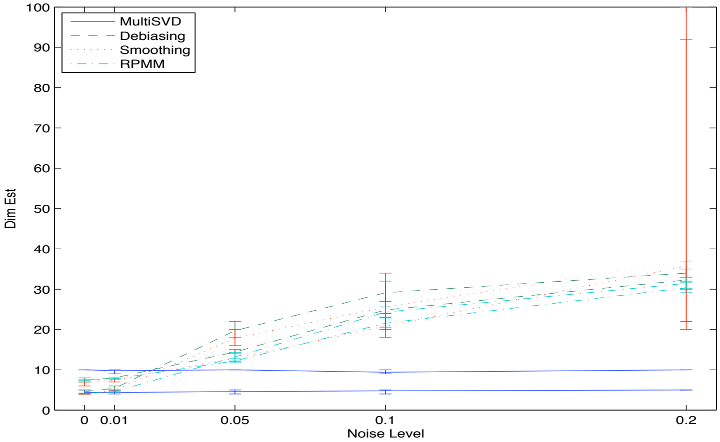
Our algorithm:

- Requires $O(k \log k)$ points (under niceness assumptions on noise and curvature)
- Finite sample guarantees
- Only input: $\tilde{X}_n$
- Discovers correct scale using multiscale approach

Other algorithms:

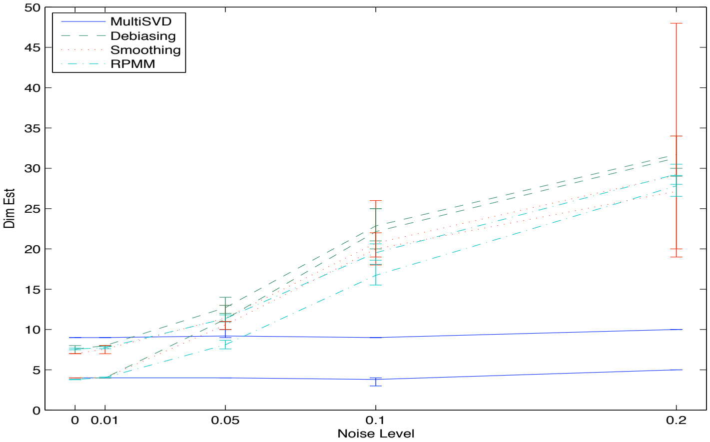
- Volume based (they require $O(2^k)$ points)
- Typically, no finite sample guarantees (at most consistent)
- Sensitive to noise
- Some involve many parameters
- Require user to specify correct scale (such as number of nearest neighors to consider)

$\mathbb{Q}^5(D = 100, n = 500)$ and $\mathbb{Q}^{10}(D = 100, n = 500)$



De-biasing algorithm of Carter, Hero, and Raich; Smoothing algorithm of Carter and Hero; Regularized Poisson Mixture Model Algorithm of Haro, Randall, and Sapiro

$\mathbb{S}^4(D = 100, n = 500)$ and $\mathbb{S}^9(D = 100, n = 500)$



De-biasing algorithm of Carter, Hero, and Raich; Smoothing algorithm of Carter and Hero; Regularized Poisson Mixture Model Algorithm of Haro, Randall, and Sapiro

Future Research

Short-term:

- Tuning algorithm
- Extending results to manifolds of different dimensionalities
- Kernelization

Long-term (employing techniques in various applications):

- Molecular Dynamics
- Genetics
- Financial data