Estimating the Intrinsic Dimension of High-Dimensional Data Sets

Anna V. Little
Department of Mathematics, Jacksonville University

Collaborators: M. Maggioni (advisor, Duke University), L. Rosasco, Y. Jung, J. Lee, G. Chen

April 17, 2015
1st Annual Workshop on Data Sciences
Tennessee State University, Nashville
**Problem**: Given a high-dimensional point cloud consisting of samples from a \( k \)-dimensional data set corrupted by \( D \)-dimensional noise, with \( k \ll D \), we estimate the intrinsic dimension via a new multiscale algorithm that generalizes PCA.

**Notation**: 
- \( n \rightarrow \) sample size
- \( D \rightarrow \) ambient dimension
- \( k \rightarrow \) intrinsic dimension

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
Example: Database of Hand Images

PCA: Classic Technique for Dimension Estimation

*When data is linear and noiseless, this method cannot fail.*

Given: \( n \) mean-zero samples \( \{x_1, \ldots, x_n\} \) in \( \mathbb{R}^D \).

Define a (centered) data matrix and empirical covariance matrix:

\[
X_n = \frac{1}{\sqrt{n}} \begin{bmatrix} -x_1 \ldots -x_n \end{bmatrix} \quad \rightarrow \quad C_n := X_n^T X_n
\]

Computes the eigenvalues of \( C_n \): \( \sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_D^2 \).

Intrinsic dimension = number of “large” eigenvalues.
Advantages:
1. Simple
2. Low sample-size requirements

Disadvantages:
1. Finite sample case is not completely understood; how many data points do we need for accurate results?
2. Noise confuses the dimensionality.
3. Fails on nonlinear data.

Example: $S^5$, $\sigma_i^2(\text{cov}(S^5)) = \frac{1}{6}$ for $1 \leq i \leq 6$
**Solution: Multiscale PCA**

Many of these issues can be addressed by computing the singular values *locally*:

(Local PCA first developed by Fukunaga and Olsen, 1971)

- Cover data set with a net of cells.
- Compute the singular values in each local cell.
- Repeat procedure with larger and larger nets.
Example:

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

1. Let $x_1, x_2, ..., x_n$ be $n$ samples from a $k$-dimensional set $\mathcal{M}$ embedded in $\mathbb{R}^D$.

2. Suppose data is corrupted by $D$-dimensional noise:

$$\tilde{x}_i = x_i + \sigma \eta_i$$

(e.g. $\eta \sim N(0, I_D)$)

$$\tilde{X}_n = \begin{bmatrix} -\tilde{x}_1 \\ -\tilde{x}_2 \\ \vdots \\ -\tilde{x}_n \end{bmatrix}$$

3. Goal: Estimate the dimensionality $k$ w.h.p. from $\tilde{X}_n$.

*Multiscale Notation:*

Fix center $z$  \[\begin{cases} X(r) = \mathcal{M} \cap B_z(r) \\ \tilde{X}_n(r) = \tilde{X}_n \cap B_{\tilde{z}}(r) \end{cases}\]
Algorithm to Estimate Dimensionality

Fix $z$. Let $\{\sigma_i^2(r)\}_{i=1}^{D}$ be the squared singular values of $\tilde{X}_n(r)$.

1. Estimate noise level; discard small scales where noise dominates.

2. Classify the $\sigma_i^2$ as follows:
   - linear growth in $r$: tangent plane squared singular value
   - quadratic growth in $r$: curvature squared singular value
   - no growth in $r$: noise squared singular value

3. Dimensionality at $z =$ number of tangent plane $\sigma_i^2$'s
Recall sphere example:

- $S^5$ embedded in $\mathbb{R}^{100}$
- 1000 noisy samples ($\sigma = .05$)
Constraints to Good Range of Scale

- **Curvature** If $r$ is chosen too large, the data will no longer appear linear, and PCA will overestimate the dimension.  
  $\rightarrow$ *upper bound on $r$*

- **Sample size** If $r$ is chosen too small, one could fail to have $O(k \log k)$ samples in each local cell, and PCA will underestimate the dimension due to lack of samples.  
  $\rightarrow$ *lower bound on $r$*

- **Noise** If $r$ is chosen too small relative to the size of the noise, the noise dominates and the $k$-dimensional structure is not detectable.  
  $\rightarrow$ *lower bound on $r$*
Main Idea:

For $D$ large, if:

\[
\sigma \sqrt{D} \quad \lesssim \quad r \quad \lesssim \quad \frac{1}{\kappa} \quad \text{and} \quad n \quad \gtrsim \quad \frac{\text{vol}(\mathcal{M}) \cdot k \log k}{\text{vol}(X(r_-))} \quad \text{sampling}
\]

then $\Delta_k = \sigma_k^2(r) - \sigma_{k+1}^2(r)$ is the largest gap w.h.p.

Note:

1. One needs $\mathbb{E}[||\eta||^2_{\mathcal{R}D}] = O(1)$ (e.g. $\sigma = \sigma_0 D^{-\frac{1}{2}}$) for the algorithm to succeed w.h.p.

2. Consistency ($n \to +\infty$) follows trivially from our analysis with niceness assumptions on the noise and curvature.

3. The random matrix scaling limit ($n \to +\infty$, $D \to +\infty$, $\frac{n}{D} \to \gamma$) is a particular case of our analysis.
**Idea of Proof:**

1. Approximate the data set by a linear manifold $X^{\parallel}(r)$ and a normal correction $X^\perp(r)$.

   $\rightarrow ||\text{cov}(X^{\parallel}(r))|| \sim O\left(\frac{1}{k}r^2\right)$

   $\rightarrow ||\text{cov}(X^\perp(r))|| \sim O\left(\frac{\kappa^2}{k}r^4\right)$

2. Bound covariance matrix perturbations due to curvature, sampling, and noise.

   $\rightarrow$ *Sampling Theorems for Covariance Matrices*

   $\rightarrow$ *Random Matrix Theory*

   $\rightarrow$ *Concentration of Measure in High Dimensions*

3. Conclude that $\max_i \Delta_i = \Delta_k$ w.h.p.
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 neighbors to consider)
$S^5(n = 250, D = 100, \sigma)$
$S(n = 250, D = 100, \sigma)$
Future Research & Extensions

- Extending results to collections of manifolds of different dimensionalities
- Proving why competing algorithms perform poorly with noise
- Use results to improve dimensionality reduction algorithms
- Employing techniques in various applications
  - Molecular Dynamics
  - Genetics
  - Financial data
- Developing a similar multiscale spectral approach for estimating the number of clusters in a data set.
Thank you!
Questions?

Email: alittle2@ju.edu