Estimating Intrinsic Dimension from Nearest Neighbor Distances

Liza Levina
Department of Statistics
University of Michigan

Joint work with Peter Bickel (UC Berkeley)
Why estimate dimension?

- Many types of modern data are extremely high-dimensional (gene expression, imaging, finance, etc)
- A lot of inference / intuition breaks down; but many methods still work
- In most cases, the data are
  - embedded in a very high-dimensional space
  - can be efficiently summarized in a space of a much lower dimension

Dimensionality reduction

- Traditional methods: PCA, multidimensional scaling
- Recent development: nonlinear manifolds (LLE, Isomap, others)
- Not clear how to determine the reduced dimension.
Manifold Projection Methods

**Black box:** $n$ points in $\mathbb{R}^p$ in $\Rightarrow$ $n$ points in $\mathbb{R}^m$ out, with $m < p$.

**Intuition:** *Locally*

- everything is *linear*
- geodesic distances between *neighbors* are preserved

**Major Algorithms**

- Locally Linear Embedding (Roweis & Saul 2000)
- Isomap (Tenenbaum et al. 2000)
- Laplacian Eigenmaps (Belkin & Niyogi 2002)
- Hessian Eigenmaps (Donoho & Grimes 2003)
Dimension Estimation Methods

- **Eigenvalue methods** (local or global PCA, dimension = the number of eigenvalues greater than a given threshold).
  - Global PCA cannot handle nonlinear manifolds
  - Local PCA is unstable

- **Nearest neighbor (NN) methods**

  If $X_1, \ldots, X_n$ are an i.i.d. sample from a density $f(x)$ in $\mathbb{R}^m$, then

  $$\frac{k}{n} \approx f(x)V(m)T_k(x)^m$$

  - $V(m)$ is the volume of the unit sphere in $\mathbb{R}^m$,
  - $T_k(x)$ is the Euclidean distance from $x$ to its $k$-th nearest neighbor.

  Estimate dimension by regressing $\log \bar{T}_k$ on $\log k$ (Pettis et al. 1979)

  - ignores dependence in $T_k$
• Fractal methods
  – Correlation dimension estimated by regressing log of
    \[
    C'_n(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbf{1}\{\|X_i - X_j\| < r\}.
    \]
    on log \( r \) over the linear part (Grassberger & Procaccia 1983)
  – Capacity dimension and packing numbers (Kégl 2002)

_Unresolved issues_

? Behavior as a function of sample size \( n \) and dimension \( m \)

? Bias and variance

? Comparisons between methods
A Maximum Likelihood Estimator of Intrinsic Dimension

**Idea:** fix a point $x$, assume $f(x) \approx \text{const}$ in a small sphere, and treat the observations as a homogeneous Poisson process.

- $X_i = g(Y_i) \in \mathbb{R}^p$; $Y_i$ are sampled from an unknown density $f$ on $\mathbb{R}^m$, with unknown $m \leq p$; $g$ is a smooth manifold mapping.

- At $x$, approximate the binomial process $\{N(t), 0 \leq t \leq R\}$

$$N(t) = \sum_{i=1}^{n} 1\{\|X_i - x\| \leq t\}$$

by a Poisson process with rate $\lambda(t) = f(x) V(m) m t^{m-1}$ and log-likelihood (letting $f(x) = e^{\theta}$)

$$L(m, \theta) = \int_0^R \log \lambda(t) \, dN(t) - \int_0^R \lambda(t) \, dt$$
• Nice exponential family; MLE for $m$

$$\hat{m}_R(x) = \left[ \frac{1}{N(R,x)} \sum_{j=1}^{N(R,x)} \log \frac{R}{T_j(x)} \right]^{-1}$$

• More convenient in practice: fix $k$ NN

$$\hat{m}_k(x) = \left[ \frac{1}{k - 1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right]^{-1}$$

• Unless local or cluster estimates are desired, average over points

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_k(X_i), \quad \hat{m} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{m}_k$$

• The MLE of $\theta = \log f(x)$ can be used to estimate entropy.
Asymptotic Bias and Variance

$m$ fixed, $n \rightarrow \infty$, $k \rightarrow \infty$, and $k/n \rightarrow 0$.

- To a first order approximation

  $$E(\hat{m}_k(x)) = m$$

  $$\text{Var}(\hat{m}_k(x)) = \frac{m^2}{k-2}$$

- When averaging over observations,

  $$E\hat{m} = E\hat{m}_k = m, \quad \text{Var}(\hat{m}_k) = O(1/n)$$

(The argument for variance is heuristic).
MLE estimator as a function of $k$

(a) 5-d normal for several $n$.  (b) Several $m$-d normals with $n = 1000$.

- Same pattern for points in a cube, on a sphere, on the "Swiss roll", etc
- **Bias** for large $k$ decreases with sample size, increases with dimension
- From now on, use the range $k_1 = 10$ to $k_2 = 20$ for averaging
Data near a manifold

• **Dimension vs. scale:** is a line plus noise 1-d or 2-d?

• Study by simulating 5-d Gaussian with mean 0, and covariance

\[ \sigma_{ij} = \begin{cases} 
1, & i = j \\
\rho, & i \neq j 
\end{cases} \]

• \( \rho = 0 \ldots 1 \Rightarrow m = 5 \ldots 1 \)

Findings

• **Only \( \rho \)** very close to 1 affects dimension

• The answer **depends on the sample size**
Data near a manifold

MLE of dimension

$n=2000$
$n=1000$
$n=500$
$n=100$
Comparing Methods

Methods:

1. The MLE
2. The regression estimator \((\log \bar{T}_k \text{ on } \log k)\)
3. The correlation dimension

- Shown on \(m\)-spheres with \(n = 1000\) uniform points
- Same pattern holds on other sets
- Ranges of parameters are fixed throughout
- MLE has smallest variance; best balance of bias and variance
Comparing methods

Mean ± 2 SD

- MLE
- Regression
- Corr.dim.
## Image Data Examples

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sample size ( n )</th>
<th>MLE</th>
<th>Regression</th>
<th>Corr. dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hands</td>
<td>481</td>
<td>3.1</td>
<td>2.5</td>
<td>3.9 / 19.7</td>
</tr>
<tr>
<td>Faces</td>
<td>698</td>
<td>4.3</td>
<td>4.0</td>
<td>3.5</td>
</tr>
</tbody>
</table>

- **Hands**: video of rotation (front, back, side views)
- **Faces**: illumination, vertical + horizontal orientation
Why is dimension estimation hard?

- Bias is quite inferior to what the asymptotics suggest for larger $m$.
- To assess sample size, do asymptotics as both $m$ and $n$ get large.
- Our result depends on the number of observations with $k$ neighbors at a distance $\leq R$ tending to $\infty$ as $R \to 0$ at some rate.
- It can be shown that we need
  \[ nV(m) R^m \to \infty \]
- Since $V(m) = \pi^{m/2} [\Gamma(m/2 + 1)]^{-1} \to 0$ as $m^{-m/2}$, enormous samples are needed for larger $m$.

**But:** for many real datasets the dimension is relatively small, and then the estimator is very reliable.