Maximum Likelihood Estimation of Intrinsic Dimension

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

Picking the right dimension is important:

- $m$ too small $\Rightarrow$ important data features are “collapsed”
- $m$ too large $\Rightarrow$ the projections become noisy and/or unstable

Major Algorithms

- **Locally Linear Embedding** (Roweis & Saul 2000), **Laplacian Eigenmaps** (Belkin & Niyogi 2002), **Hessian Eigenmaps** (Donoho & Grimes 2003): dimension is provided by the user
- **Isomap** (Tenenbaum et al. 2000): MDS error curves can be “eyeballed” to estimate dimension
- **Charting** (Brand 2002): heuristic estimate equivalent to the “regression” estimator below.
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.

  Regression estimator: estimate $m$ 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

  \[
  C_n(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1\{\|X_i - X_j\| < r\}.
  \]

  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 $\lambda(t) = f(x)V(m)mt^{m-1}$ 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

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

• For an asymptotically unbiased estimator, replace $k - 1$ with $k - 2$.

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

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_k(X_i)$$
Estimate of the entropy

- Have the MLE of $\theta(x) = \log f(x)$ (2nd parameter):

  $\exp(\hat{\theta}_R(x)) = N(R, x) [V(\hat{m}_R(x))]^{-1} R^{-\hat{m}_R(x)}$
  $\exp(\hat{\theta}_k(x)) = (k - 1) [V(\hat{m}_k(x))]^{-1} T_k(x)^{-\hat{m}_k(x)}$

- $\hat{\theta}$ can be used to estimate entropy of $f$:

  $J(f) = \int f(x) \log f(x) \, dx$

  $\hat{J}(f) = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}(X_i)$

- Entropy can potentially be used for computing mutual information (hence classification)

**Computational cost:** finding $k$ NN for every point.
Asymptotic Bias and Variance

\( m \) fixed, \( n \to \infty, k \to \infty, \) and \( k/n \to 0. \)

Asymptotically unbiased estimator:

\[
\hat{m}_k(x) = \left[ \frac{1}{k-2} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right]^{-1} = (k - 2)mY^{-1}
\]

where \( Y = m \sum_{j=1}^{k-1} \log(T_k/T_j). \)

Condition on \( T_k \) and assume the Poisson approximation is exact:

- \( (T_j/T_k)^m \) are distributed as \( k - 1 \) order statistics of Uniform(0,1)
- \( m \log(T_k/T_j) \) are distributed as \( k - 1 \) order statistics of Exponential(1)
- \( Y \) is Gamma\( (k - 1, 1) \), and \( EY^{-1} = 1/(k - 2). \)
Hence to a first order approximation

\[ E(\hat{m}_k(x)) = m \]
\[ \text{Var}(\hat{m}_k(x)) = \frac{m^2}{k - 3} \]

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).

May choose not to correct the bias (for larger \( m \) we almost never have a large enough sample size)
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
Choosing the neighborhood size

- Can average over a “reasonable” range $k_1 \ldots k_2$

$$\hat{m} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{m}_k$$

- Choose automatically: for the “right” $k$ estimates agree over different sample sizes
  1. Divide points into 4 random sets (size $n/4$ each)
  2. Compute estimates $\hat{m}_i(k)$, $i = 1 \ldots 4$ on $n/4$, $n/2$, $3n/4$, $n$ points
  3. Compute the standard deviation of the 4 estimates $\hat{m}_i(k)$ and let

$$\hat{k} = \arg \min_k \text{SD}(\hat{m}_i(k))$$
Error function for choosing $k$

Sphere in $\mathbb{R}^5$, $n = 1000$. 
Neighborhood size as a function of dimension and sample size

Spheres in $\mathbb{R}^d (m = d - 1)$

- $k$ increases with $n$, decreases with $m$; SD $\approx 5 \ldots 10$
Fixed vs. adaptive neighborhood size

Mean(SD) of estimated dimension for spheres in $\mathbb{R}^d$

1st line: average over $k = 10 \ldots 20$; 2nd line: $k$ chosen adaptively

<table>
<thead>
<tr>
<th>d</th>
<th>Sample size $n$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>1.00(0.015)</td>
</tr>
<tr>
<td></td>
<td>1.00(0.018)</td>
</tr>
<tr>
<td>5</td>
<td>3.83(0.08)</td>
</tr>
<tr>
<td></td>
<td>3.82(0.10)</td>
</tr>
<tr>
<td>10</td>
<td>7.51(0.18)</td>
</tr>
<tr>
<td></td>
<td>7.58(0.28)</td>
</tr>
<tr>
<td>20</td>
<td>12.96(0.30)</td>
</tr>
<tr>
<td></td>
<td>13.25(0.64)</td>
</tr>
</tbody>
</table>

- Adaptive $k$ has slightly less bias, slightly more variance for higher dimensions
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
- If \( \rho \approx 1 \), smaller sample sizes lead to lower dimension estimates
Data near a manifold

MLE of dimension

\(1 - \rho\) (log scale)
Comparing Methods

- Methods:
  1. The MLE
  2. The regression estimator \((\log T_k, \log k)\)
  3. The correlation dimension

- Shown on \(m\)-spheres with \(n = 1000\) uniform points

- Similar pattern 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.
## Popular Dataset Examples

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Data dim.</th>
<th>Sample size</th>
<th>MLE</th>
<th>Regression</th>
<th>Corr. dim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss roll</td>
<td>3</td>
<td>1000</td>
<td>2.0 (0.03)</td>
<td>1.8 (0.03)</td>
<td>2.0 (0.24)</td>
</tr>
<tr>
<td>Faces</td>
<td>$64 \times 64$</td>
<td>698</td>
<td>4.8</td>
<td>4.0</td>
<td>3.5</td>
</tr>
<tr>
<td>Hands</td>
<td>$480 \times 512$</td>
<td>481</td>
<td>2.9</td>
<td>2.5</td>
<td>3.9 / 19.7</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. Consider

\[ M(n, k, R) = \sum_{i=1}^{n} 1 \{ \text{there exist } j_l \neq i, \ 1 \leq l \leq k, \text{ such that } |X_{j_l} - X_i| \leq R \} . \]

- It can be shown that we need, for all $k, n, R > 0$

\[ \frac{1}{n} EM(n, k, R) \geq \delta(R) > 0 \]
With some approximations, this is essentially equivalent to

\[ nV(m)R^m \rightarrow \infty \]

Since

\[ V(m) = \pi^{m/2}[\Gamma(m/2 + 1)]^{-1} \asymp m^{-m/2}, \]

enormous samples are needed for larger \( m \).

This problem is shared by all estimators

Calibration on known datasets has been proposed

For many real datasets the dimension is relatively small, and then the estimator is very reliable.
Some unanswered questions

● What is the distribution of the estimator?

● What is the effect of noise on dimension estimation and manifold projections in general?

● What if there are multiple dimensions/manifolds within one dataset?

● How much can all of this help in classification?