Dimensionality Reduction of Data Sets

Anthony Grebe
under the direction of
Professor Victor Wickerhauser

October 9, 2014
Main Idea

- Finite data set in high-dimensional space ($\mathbb{R}^d$)
- Data lie on $p$-dimensional manifold ($p \ll d$)
Main Idea

- Finite data set in high-dimensional space ($\mathbb{R}^d$)
- Data lie on $p$-dimensional manifold ($p \ll d$)

Diffusion Mapping

- Used to map data from $\mathbb{R}^d$ to $\mathbb{R}^p$, $p \ll d$
Diffusion Mapping

- Used to map data from $\mathbb{R}^d$ to $\mathbb{R}^p$, $p \ll d$

Image source: “Diffusion Maps” by Coifman and Lafon
Consider a random walk on the data set between points \((x_i)\)
Algorithm

- Consider a random walk on the data set between points \((x_i)\)

\[
P_{ij} = \exp \left( -\frac{||x_i - x_j||^2}{\epsilon} \right) \theta(d - ||x_i - x_j||)
\]

- Transition probability decreases with distance
Consider a random walk on the data set between points \((x_i)\)

\[
P_{ij} = \exp \left( -\frac{||x_i - x_j||^2}{\epsilon} \right) \theta(d - ||x_i - x_j||)
\]

- Transition probability decreases with distance
- \(P = \text{row-stochastic transition matrix}\)
Algorithm

- Euclidean distance in $\mathbb{R}^d$ (original set)
**Algorithm**

- Euclidean distance in $\mathbb{R}^d$ (original set)
- Transition probability
Algorithm

- Euclidean distance in $\mathbb{R}^d$ (original set)
- Transition probability
- Diffusion distance

$$D(x_i, x_j)^2 = \sum_{k=1}^{n} \frac{(P_{ik} - P_{jk})^2}{\pi(x_k)}$$
Algorithm

- Euclidean distance in $\mathbb{R}^d$ (original set)
- Transition probability
- Diffusion distance

$$D(x_i, x_j)^2 = \sum_{k=1}^{n} \frac{(P_{ik} - P_{jk})^2}{\pi(x_k)}$$

- Diffusion space
Algorithm

- Euclidean distance in $\mathbb{R}^d$ (original set)
- Transition probability
- Diffusion distance

$$D(x_i, x_j)^2 = \sum_{k=1}^{n} \frac{(P_{ik} - P_{jk})^2}{\pi(x_k)}$$

- Diffusion space
  - Found by eigendecomposition of $P$
**Algorithm**

- Euclidean distance in $\mathbb{R}^d$ (original set)
- Transition probability
- Diffusion distance

\[ D(x_i, x_j)^2 = \sum_{k=1}^{n} \frac{(P_{ik} - P_{jk})^2}{\pi(x_k)} \]

- Diffusion space
  - Found by eigendecomposition of $P$
  - Truncation after $p$ eigenvalues
Examples
Dimensionality of Manifold

\[ p_{\text{known in sample data sets, not real life}} \]

Suffices to estimate \( p \) at a point \( x \).

Nearest neighbors of \( x \) are \( x_1, x_2, \ldots, x_q \).

\( q \) is an upper bound for \( p \).

\[ M = \left( \begin{array}{cccc}
(x_1 - x) & (x_2 - x) & \cdots & (x_q - x)
\end{array} \right)^T \in \mathbb{R}^{q \times d} \]
Dimensionality of Manifold

- $p$ known in sample data sets, not real life
Dimensionality of Manifold

- $p$ known in sample data sets, not real life
- Suffices to estimate $p$ at a point $x$
Dimensionality of Manifold

- $p$ known in sample data sets, not real life
- Suffices to estimate $p$ at a point $x$
- Nearest neighbors of $x$ are $x_1, x_2, \ldots, x_q$
  - $q = \text{upper bound for } p$
Dimensionality of Manifold

- $p$ known in sample data sets, not real life
- Suffices to estimate $p$ at a point $x$
- Nearest neighbors of $x$ are $x_1, x_2, \ldots, x_q$
  - $q = \text{upper bound for } p$

$$M = \begin{bmatrix} (x_1 - x)^T \\ (x_2 - x)^T \\ \vdots \\ (x_q - x)^T \end{bmatrix} \in \mathbb{R}^{q \times d}$$
Dimensionality of Manifold

- Singular value decomposition (SVD) of $M$
Dimensionality of Manifold

- Singular value decomposition (SVD) of $M$

  $$M = U\Sigma V^{-1}$$

  - $U, V$ are orthogonal
  - $\Sigma$ is diagonal
Triangulations

In the plane, the division of the convex hull of a set of points into triangles simplifies data set (triangles are well understood).
In the plane, the division of the convex hull of a set of points into triangles
Triangulations

- In the plane, the division of the convex hull of a set of points into triangles
- Simplifies data set (triangles are well understood)
Triangulations
Triangulations
Triangulations

BAD

GOOD
Delaunay Triangulations

- No point is inside the circumcircle of any triangle.
- Consequence: Maximizes the minimum angle.
- Exists and is unique for points in "general position".

Delaunay Triangulations

- No point is inside the circumcircle of any triangle

Delaunay Triangulations

- No point is inside the circumcircle of any triangle
- Consequence: Maximizes the minimum angle

Delaunay Triangulations

- No point is inside the circumcircle of any triangle
- Consequence: Maximizes the minimum angle
- Exists and is unique for points in “general position”

Higher Dimensions
Higher Dimensions

- Relatively easy to generalize Delaunay triangulations to $\mathbb{R}^n$
Higher Dimensions

- Relatively easy to generalize Delaunay triangulations to $\mathbb{R}^n$
- More difficult to generalize to curved manifolds
  - Need points to be sufficiently dense