Nonlinear Dimensionality Reduction: Conclusions

Emil Eirola

Time Series Prediction and Chemoinformatics Group
Adaptive Informatics Research Centre
Helsinki University of Technology

November 20, 2007

Outline

1. Summary
   - Intrinsic dimensionality estimation
   - Dimensionality reduction
   - Latent variable separation
2. Taxonomy of methods
   - Distance preserving methods
   - Topology preserving methods
   - By kernel and algorithm
3. Methodology
   - Data flow
   - Other considerations

Summary

The types of problems discussed on this course are

- Intrinsic dimensionality estimation
- Dimensionality reduction
- Latent variable separation

for high-dimensional data sets.

Summary

The types of problems discussed on this course are

- Intrinsic dimensionality estimation
- Dimensionality reduction
- Latent variable separation

for high-dimensional data sets.

Principal component analysis (PCA) **can** be used for all three, but we can do better.

Intrinsic dimensionality estimation

How many parameters are needed to approximate the manifold?

Instead of PCA we can use fractal dimensionality measures

- Capacity dimension ("box-counting")
- Correlation dimension

Also iterative testing—"trial-and-error"—methods can be effective.

Dimensionality reduction

How to find a representative projection to a lower-dimensional space?

PCA was improved by

- Better cost function/optimisation $\Rightarrow$ Sammon's NLM (1969)

Dimensionality reduction

How to find a representative projection to a lower-dimensional space?

PCA was improved by

- Better cost function/optimisation $\Rightarrow$ Sammon's NLM (1969)
- Stochastic techniques $\Rightarrow$ CCA (Demartines & Hérault, 1995)

Dimensionality reduction

How to find a representative projection to a lower-dimensional space?

PCA was improved by

- Better cost function/optimisation $\Rightarrow$ Sammon's NLM (1969)
- Stochastic techniques $\Rightarrow$ CCA (Demartines & Hérault, 1995)
- Geodesic distances and graph distances $\Rightarrow$ Isomap (Tenenbaum, 1998)

Dimensionality reduction

How to find a representative projection to a lower-dimensional space?

PCA was improved by

- Better cost function/optimisation ⇒ Sammon's NLM (1969)
- Stochastic techniques ⇒ CCA (Demartines & Hérault, 1995)
- Geodesic distances and graph distances ⇒ Isomap (Tenenbaum, 1998)
- Topology considerations ⇒ SOM (Kohonen, 1982)

Dimensionality reduction

How to find a representative projection to a lower-dimensional space?

PCA was improved by

- Better cost function/optimisation ⇒ Sammon's NLM (1969)
- Stochastic techniques ⇒ CCA (Demartines & Hérault, 1995)
- Geodesic distances and graph distances ⇒ Isomap (Tenenbaum, 1998)
- Topology considerations ⇒ SOM (Kohonen, 1982)
- Data-driven methods ⇒ Isotop (Lee, 2002)

Latent variable separation

What are the paremeters which describe the manifold?

PCA has been improved into

- Projection pursuit (PP, Kruskal, 1972)
- Blind source separation and independent component analysis (BSS/ICA, Jutten & Hérault, 1980s)

Of the methods on this course only generative topographic mapping (GTM) is effective for variable separation.

Taxonomy: Distance preserving methods

- Euclidan
  - Multidimensional scaling (MDS), equivalent to PCA
  - Sammon's nonlinear mapping (NLM)
  - Curvilinear component analysis (CCA)
- Geodesic
  - Isomap
  - Geodesic NLM (GNLM)
  - Curvilinear distance analysis (CDA)
- Other
  - Kernel PCA (KPCA)
  - Semidefinite embedding (SDE)

Taxonomy: Topology preserving methods

- Predefined lattice
  - Self-organising maps (SOM)
  - Generative topographic mapping (GTM)
- Data-driven lattice
  - Locally linear embedding (LLE)
  - Laplacian eigenmaps (LE)
  - Isotop

Taxonomy: By kernel and algorithm

Distance pres.	MDS algorithm	NLM alg.	CCA alg.
Euclidean	metric MDS	NLM	CCA
Geodesic	Isomap	GNLM	CDA
Commute time	LE		
Fixed kernel	KPCA		
Optimised kernel	SDE		

Topology pres.	ANN-like	MLE by EM	Spectral
Predefined lattice	SOM	GTM	
Data-driven lattice	Isotop		LLE

Artificial **N**eural **N**etwork, **M**aximum **L**ikelihood **E**stimation, **E**xpectation **M**aximation

Data flow

Five steps to success:

1. Data selection

Data flow

Five steps to success:

1. Data selection
2. Calibration/normalization

Data flow

Five steps to success:

1. Data selection
2. Calibration/normalization
3. Linear dimensionality reduction by PCA

Data flow

Five steps to success:

1. Data selection
2. Calibration/normalization
3. Linear dimensionality reduction by PCA
4. Nonlinear dimensionality reduction

   or

   latent variable separation

Data flow

Five steps to success:

1. Data selection
2. Calibration/normalization
3. Linear dimensionality reduction by PCA
4. Nonlinear dimensionality reduction
   or
   latent variable separation
5. Visualization/modeling/classification/prediction/etc.

Methodology considerations based on data set size N

- Large data set, $N > 2000$
  - Probably too computationally heavy for most methods
  - Consider reducing the size by sampling or vector quantization
- Medium-sized set, $200 < N \leq 2000$
  - The NLDR methods will generally work ok
- Small data set, $N \leq 200$
  - Problems are likely
  - PCA can still be used safely

Methodology considerations based on dimensionality D

- Very high dimensionality, $D > 50$
  - Some methods can "get confused"
  - Use PCA first for reducing dimensionality and denoising without significantly losing information
- High dimensionality, $5 < D \leq 50$
  - Probably ok, but proceed with caution
- Low dimensionality, $D \leq 5$
  - The NLDR methods can be used safely

Methodology considerations based on intrinsic dimension

Target dimension d *vs.* intrinsic dimension iD

- if $d \gg iD$
  - Anything will work, so use PCA
- if $d \approx iD$
  - Use some NLDR method
  - If the manifold is highly curved it would be good to have $d = iD + 1$ or $iD + 2$.
- if $d < iD$
  - E.g., for visualisation
  - Risky business
  - Methods based on eigenvectors are relatively safer, since they converge better and you can choose which eigenvectors to use.
  - SOM or NeRV