Characteristics of an Analysis Method

Elia Liitiäinen (elia.liitiainen@hut.fi)

T-61.6050
Helsinki University of Technology, Finland

September 23, 2007

Introduction

- Basic concepts are presented.
- The PCA method is analyzed.
- Motivation for more sophisticated methods is given.

Outline

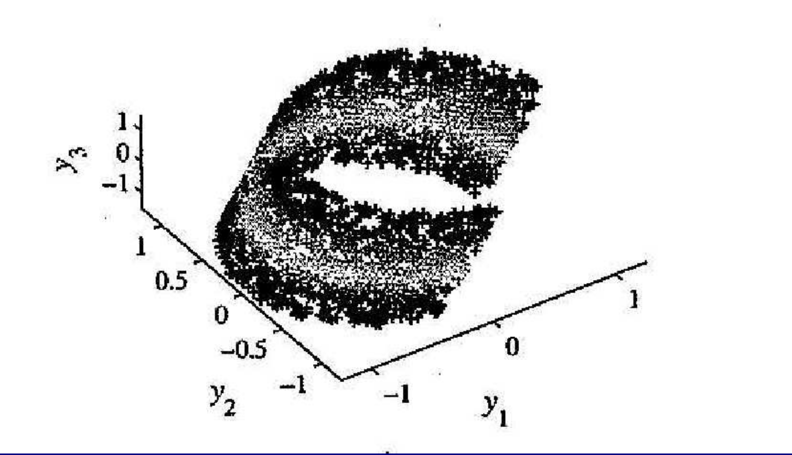
Basic Requirements

- Estimation of the embedding dimension.
- Dimensionality reduction
- Separation of latent variables.

Instrinsic Dimensionality

- Let us assume that the data is in $\Re^D$.
- The basic assumption: the data can be embedded into $\Re^P$ with $P < D$.

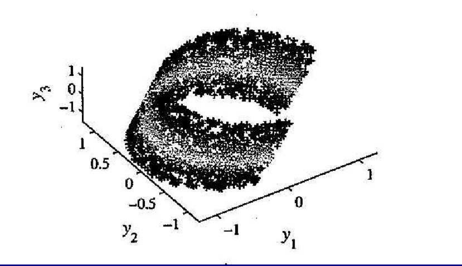
Example: A Low Dimensional Manifold

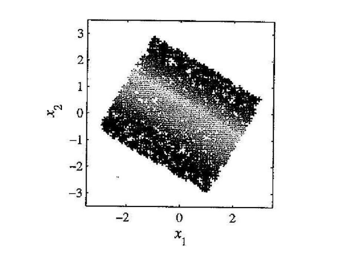


Latent Variables vs. Dimensionality Reduction

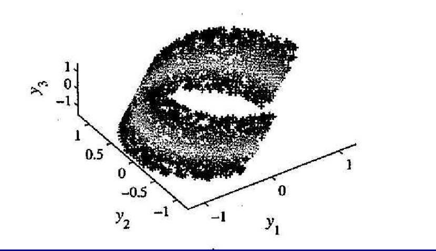
- When extracting latent variables, a model generating the data is assumed (eg. ICA).
- The embedding into a lower dimensional space is done under this constraint.
- Dimensionality reduction is easier: any low dimensional representation is a solution.
- DR is less interpretable?

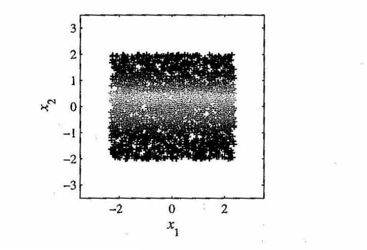
Example: Dimensionality Reduction





Example: Recovery of Latent Variables





- Dimensionality reduction under an independence constraint.

Fundamental Issues

- Many dimensionality reduction algorithms assume that the data is generated by a model.
- For example, in PCA it is assumed that a number of latent variables explain the data in a linear way.
- For the same model, different algorithms are possible.

The Criterion

- The dimensionality reduction is often done using a criterion that is optimized.
- One possible idea is to measure distance preservation.
- One may either try to preserve the distances between points or alternatively the topology.

Projection as a criterion

- Let $\mathcal{P} : \Re^D \rightarrow \Re^P$ be a projection.
- $\mathcal{P}^{-1}$ denotes the reconstruction $\Re^P \rightarrow \Re^D$.
- A common criterion in dimensionality reduction is

$$E[\|y - \mathcal{P}^{-1}\mathcal{P}(x)\|^2].$$

Derivation of PCA (1)

- The basic model behind PCA is

$$y = Wx$$

  with y a random variable in $\Re^D$ and W a $D \times P$ matrix.
- The sample $(X_i, Y_i)_{i=1}^N$ is available; mean centering is assumed.
- Normalization/scaling is done according to prior knowledge.

Derivation of PCA (2)

- Assume that W has orthonormal columns.
- The projection criterion leads to

$$\min_W E[\|y - WW^T y\|^2].$$

- This corresponds to finding the subspace which allows best possible reconstruction.

Derivation of PCA (3)

- The optimization problem can be written equivalently as

$$\max_W E[y^T W W^T y].$$

- Let Y be the matrix of samples as column vectors.
- Empirical approximation leads to

$$\max_W \mathrm{tr}[Y^T W W^T Y].$$

Derivation of PCA (4)

- Singular value decomposition $Y = V\Sigma U^T$ yields

$$\max_W \operatorname{tr}[U\Sigma V^T W W^T V \Sigma U^T].$$

- The solution is taking the columns of V corresponding to the largest singular values, which can be written as $W = VI_{D\times P}$.
- The reconstruction error depends on the singular values $\sigma_{P+1}, \ldots, \sigma_D$.

Relation to the Covariance Matrix

- Let C_y be the covariance matrix of the observations.
- Finding the projection V is equivalent to finding the eigenvectors $V_1, \dots, V_P$ corresponding to the biggest eigenvalues.
- The eigenvectors are the directions of maximal variance.

Choosing the embedding dimensionality

- A simple method is to plot sorted eigenvalues.
- After some point, the decrease is neglible.
- This often fails; other choices include Akaike's information criterion and other complexity penalization methods.
- It is also possible to put a threshold: for example, require that 95% of the variance is preserved.

Example: Determination of Instrinsic Dimensionality

Choose $X \sim N(0, I) \in \Re^2$,

$$A = [0.1\ 0.2; 0.4\ 0.2; 0.3\ 0.3; 0.5\ 0.1; 0.1\ 0.4]$$

and

$$Y = AX + \epsilon$$

with $\epsilon \sim N(0, 0.1I)$.

Example: Determination of Instrinsic Dimensionality (2)

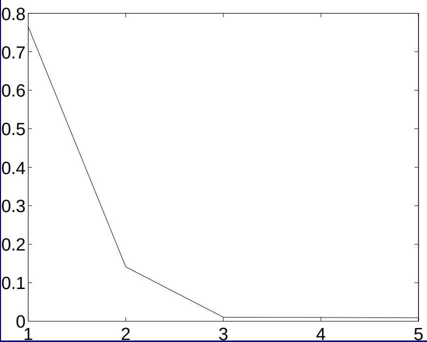
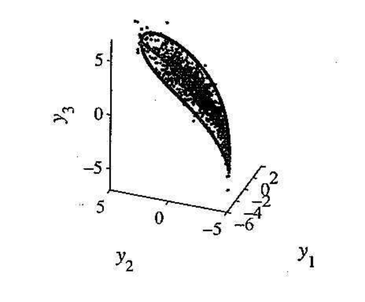


Figure: Eigenvalues of the covariance matrix.

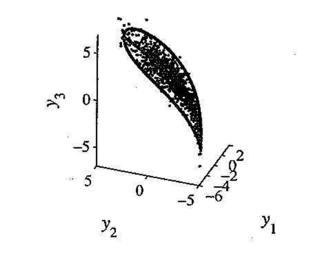
- The first two contain most of the variance.

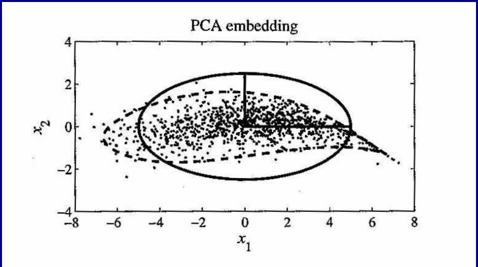
PCA for nonlinear data (1)

- The model: $y = \begin{bmatrix} 4\cos(\frac{1}{4}x_1) \\ 4\sin(\frac{1}{4}x_1) \\ x_1 + x_2 \end{bmatrix}$.



PCA for nonlinear data (2)





- The reconstructed surface would be a plane.

DR vs. generative (latent variable) models

- It is possible to model the data using latent variables and estimate the parameters.
- In practice, it is simpler to directly learn a projection.

Local Dimensionality Reduction

- A nonlinear manifold is locally approximately linear.
- It is possible to derive a local PCA as a generalization to the nonlinear case.

Other Issues

- Batch vs. online algorithm
- Local maximas $<->$ global optimization (PCA)

Conclusion

- Many dimensionality reduction methods are based on the assumption that the data is approximately on a manifold.
- PCA solves the linear case, but fails in nonlinear problems.
- Thank you for your attention.