Introduction to dimensionality reduction

Gianluca Campanella

Contents

Dimensionality reduction

Dimensionality reduction

Idea

- Identify correlated columns
- Replace them with a new column that ‘encapsulates’ the others

Example

- { car, cat, truck, van }

$\rightarrow$ { cat, vehicle }

Dimensionality reduction

Why?

- 'True' dimensionality is lower
- Too many correlated variables $\rightarrow$ collinearity
- Difficult to visualise

How?

- Project onto a lower-dimensional space...
- ...while retaining (most of) some property

Manifold learning

Multidimensional scaling (MDS)

Aim

- Project onto a lower-dimensional space...
- ...while retaining most of the distance structure

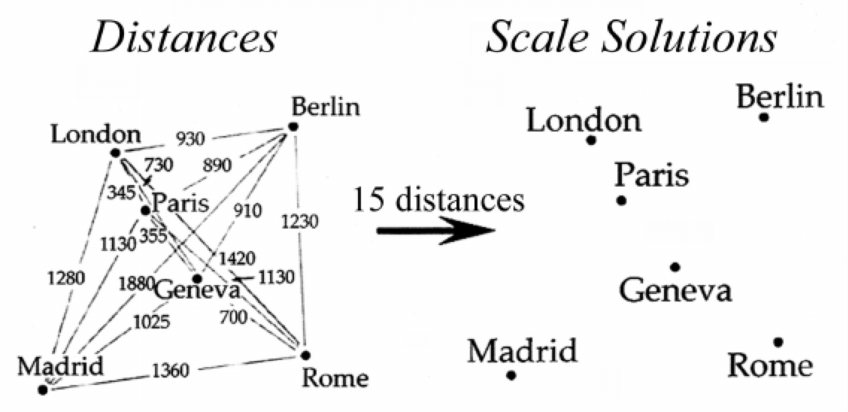
Method

- Input: dissimilarity matrix (not necessarily a metric)
- Find a 'close' representation (squared loss)

Limitations

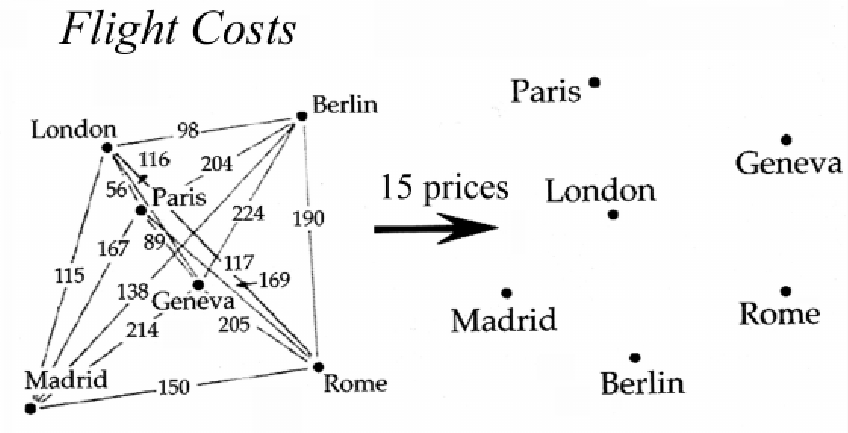
- Somewhat slow (numerical optimisation)
- Embeddings are not necessarily unique or 'optimal'

Multidimensional scaling (MDS)



From Cutting et al. (2013)

Multidimensional scaling (MDS)



From Cutting et al. (2013)

PCA and PLS

Principal component analysis (PCA)

Aim

- Project onto a lower-dimensional space…
- …while retaining most of the correlation structure

Method

- Eigendecomposition of covariance/correlation matrix
- Typically using singular value decomposition (SVD)

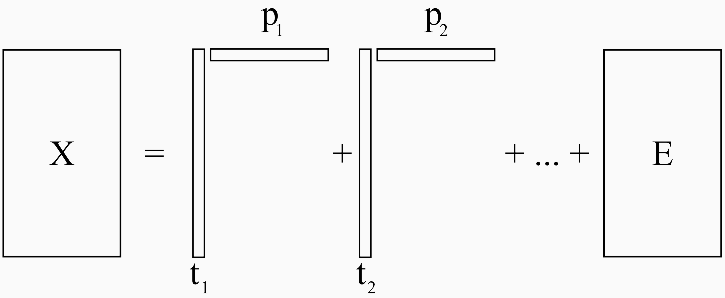
Limitations

- Unsupervised method $\rightarrow$ outcome is disregarded
- PCs may not be explanatory of $\mathbf{Y}$ (noise-driven)

Principal component analysis (PCA)

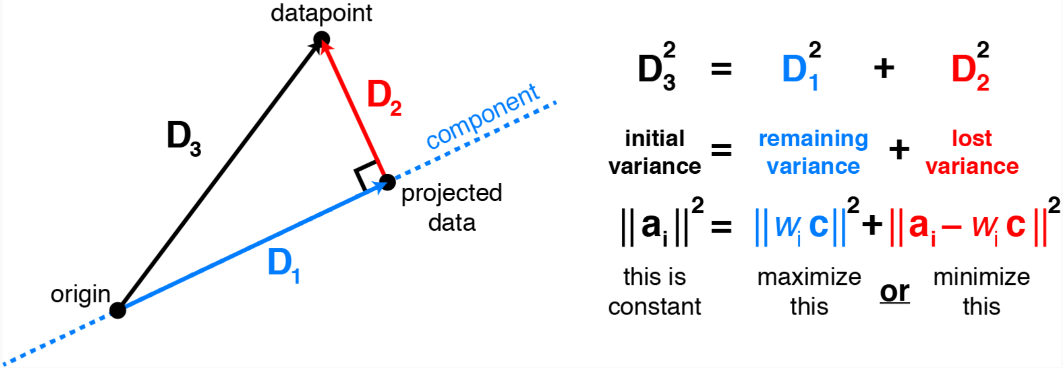
Model

- Defined by the 'direction' vectors p_i (loadings)
- Loadings are oriented in such a way that the project data t_i (scores) have maximum variance



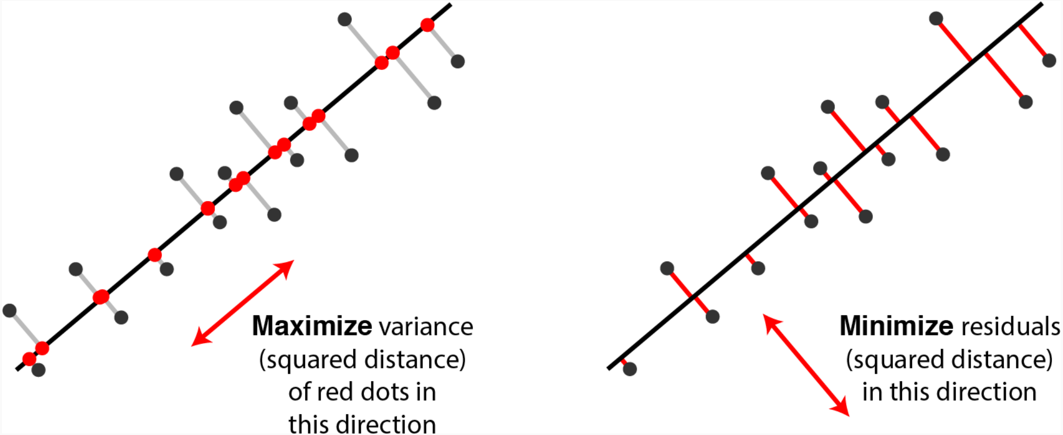
From *Process Improvement Using Data*

Principal component analysis (PCA)



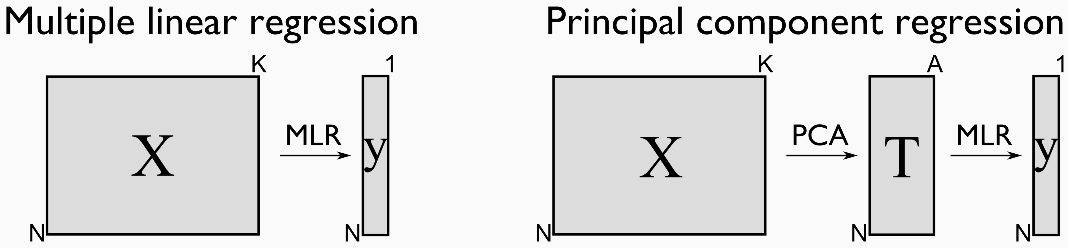
From Alex Williams' blog

Principal component analysis (PCA)



From Alex Williams' blog

Partial least squares (PLS) regression



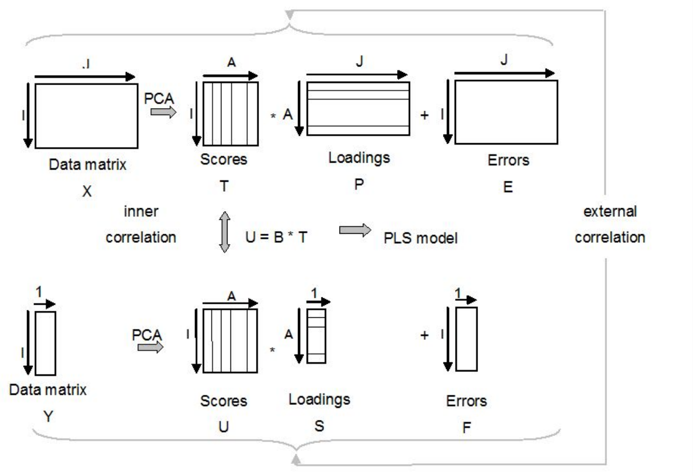
From *Process Improvement Using Data*

Advantages

- Single-step model
- Components capture variability in **X** and **Y**

→ Fewer components, more compact model

Partial least squares (PLS) regression



From Böhm et al. (2013)