Principal Component Analysis

Giorgos Korfiatis

Alfa-Informatica
University of Groningen

Seminar in Statistics and Methodology, 2007

What Is PCA?

- Dimensionality reduction technique
- Aim: Extract relevant info from confusing data sets
- Similar to Factor Analysis, SVD
- Used in various domains (neuroscience, comp graphics, sociolinguistics, dialectology, ...)
- Employs matrix algebra concepts

Dim Reduction

- When numerous variables involved
- Question whether they have something in common
- Are they independent?
- Or do they measure the same ‘underlying’ variable?
- To what extent a variable contributes to the underlying one?
- Aim: Reduce number of variables in a meaningful way

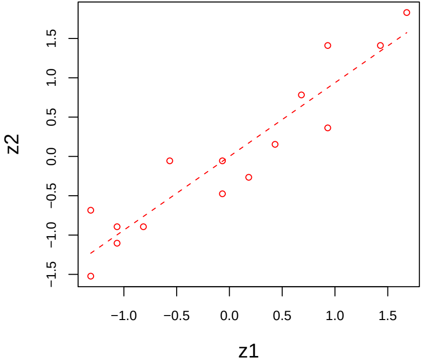
A Toy Example

X_1	X_2	z_1	z_2
7.0	10.0	−0.81497	−0.89393
10.0	12.0	−0.06653	−0.4749
12.0	15.0	0.43243	0.15364
13.0	18.0	0.68191	0.78219
16.0	21.0	1.43036	1.41073
14.0	16.0	0.93139	0.36316
6.0	10.0	−1.06445	−0.89393
11.0	13.0	0.18295	−0.26539
6.0	9.0	−1.06445	−1.10344
14.0	21.0	0.93139	1.41073
5.0	7.0	−1.31393	−1.52247
10.0	14.0	−0.06653	−0.05587
17.0	23.0	1.67984	1.82976
5.0	11.0	−1.31393	−0.68441
8.0	14.0	−0.56549	−0.05587

Table: Measurements (X) and standardized scores (z)

- 15 subjects measured on 2 variables (X_1 and X_2)
- z facilitate computations
- $z = \left(X - \bar{X}\right) / s$
- Values seem to correlate...

A Toy Example

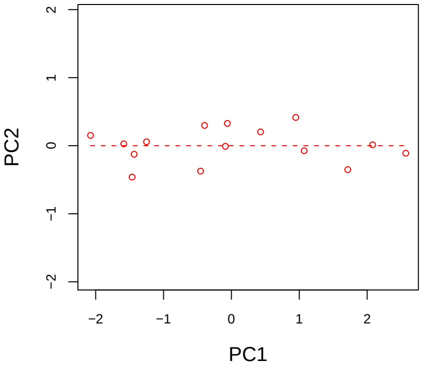


- Correlated, $r = 0.937$
- Perhaps one variable is enough
- But which one?
- Better to combine both somehow

Employing PCA

- Attempts to uncover the underlying variable(s)
- New variables called principal components
- Principal components are sorted
    - First: max part of variance
    - Second: max part of the remaining variance
    - ...
- Scores on PCs should not correlate
- PCs are orthogonal

Employing PCA



- Like rotating data points to fit the X axis
- Actually a matrix transformation
- We may ignore PC2

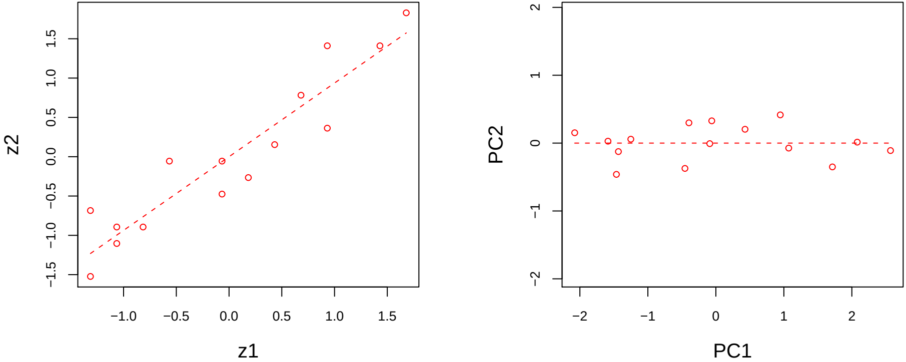
Some Matrix Algebra...

- We have the correlation matrix $R = \begin{bmatrix} 1.000 & 0.937 \\ 0.937 & 1.000 \end{bmatrix}$
- We can compute the eigenvalues of the matrix

$$\lambda_1 = 1.937 \quad \lambda_2 = 0.063$$

- Notice that sum of λ equals sum of variance (the diagonal)
- Represent 'contribution' of the dimensions
- E.g. if $\lambda_1 = 2,\ \lambda_2 = 0$, variables would be dependent
- Eigenvalues correspond to eigenvectors, used to transform the data

Data Transformation



- Initial data matrix multiplied by eigenvector matrix
- PC values are in different space than initial variables!

A Bigger Example

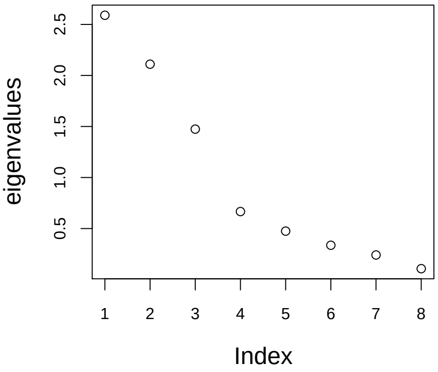
Grades of students on school courses

		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Spanish	X_1	1.00							
German	X_2	**0.65**	1.00						
Maths	X_3	0.01	0.04	1.00					
Physics	X_4	−0.07	0.13	**0.65**	1.00				
History	X_5	0.14	0.22	−0.03	−0.34	1.00			
English	X_6	**0.78**	**0.59**	0.04	0.21	−0.04	1.00		
Chemistry	X_7	0.14	0.14	**0.66**	**0.50**	0.03	0.11	1.00	
Geography	X_8	0.12	0.12	0.32	0.08	**0.38**	−0.05	0.12	1.00

Table: Correlation matrix

Three groups: $\{X_1, X_2, X_6\}$, $\{X_3, X_4, X_7\}$, $\{X_5, X_8\}$

Picking Components



- Calculate the eigenvalues
- Eigenvalue $\lambda \leftrightarrow$ PC
- $\lambda \sim$ variance explained by PC
- Keep those larger than 1 *or*
- Keep those before the ‘elbow’ *or*
- Keep those for 70% to 80% of variance (sum of λs)

Contribution of Initial Variables

Variable		PC_1	PC_2	PC_3
Spanish	X_1	-0.439	-0.407	0.057
German	X_2	-0.438	-0.324	-0.002
Maths	X_3	-0.353	0.485	-0.139
Physics	X_4	-0.334	0.452	0.228
History	X_5	-0.060	-0.221	-0.669
English	X_6	-0.449	-0.313	0.287
Chemistry	X_7	-0.375	0.371	-0.070
Geography	X_8	-0.183	0.072	-0.625
Var explained		32.4%	26.4%	18.4%

Table: Correlations of variables and PCs (loadings)

- Columns are the eigenvectors actually
- 3 groups expected: $\{X_1, X_2, X_6\}$, $\{X_3, X_4, X_7\}$, $\{X_5, X_8\}$
- But this is not very clear...

Cleaning the Picture: Rotation

Variable	PC_1	PC_2	PC_3
X_1	-0.597	~ 0.0	~ 0.0
X_2	-0.533	~ 0.0	-0.105
X_3	~ 0.0	0.601	-0.124
X_4	~ 0.0	0.559	0.235
X_5	~ 0.0	-0.127	-0.693
X_6	-0.591	~ 0.0	0.178
X_7	~ 0.0	0.525	~ 0.0
X_8	~ 0.0	0.177	-0.630

Table: Correlations after VARIMAX

- VARIMAX rotation: Maximizes the variance of loadings per factor
- Orthogonal rotation of loadings
- Amount of variance explained not affected

Assumptions – Limitations

- Linearity – change of basis
- Mean and variance are sufficient (variables normally distributed)
- Principal components are orthogonal
- Non-parametric method (there is a kernel PCA extension)
- Does not distinguish variance due to error (unlike Factor analysis)

Application in Dialectology

- Geographic patterns of surnames (Manni *et al.*, 2006)
- List of Dutch surnames (excluding very common and rare)
- Distance matrix of locations with respect to surname differentiation (Nei measure):

$$d_{i,j} = \sum_s n_{si} n_{sj} \Big/ \left(\sum_s n_{si}^2 \sum_s n_{sj}^2 \right)^{1/2}$$

n_{si}: frequency of surname s in location i

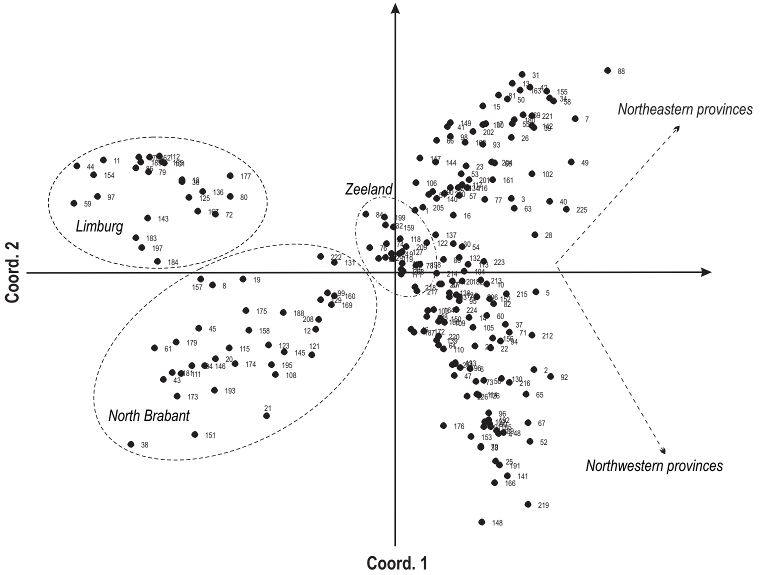
Initial Data

Loc	ℓ_1	ℓ_2	$\cdots$	ℓ_{226}
ℓ_1	0	$d_{1,2}$	$\cdots$	$d_{1,226}$
ℓ_2	$d_{2,1}$	0	$\cdots$	$d_{2,226}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
ℓ_{226}	$d_{226,1}$	$d_{226,2}$	$\cdots$	0

Table: Distance matrix

- Based on 19,910 surnames
- 226 Dutch locations
- Symmetric matrix
- Variables: distance from locations
- PCA conducted on this matrix

Plot of First Two PCs



Remarks

- Dialect distinction
  - Limburg and North Brabant clusters clear
  - North/south distinction
  - No overlap between NE and NW samples in the swarm
- 2 PCs account only for 30% of variance
- Following PCs clarify more

Conclusions

- Non-parametric method for Dim reduction
- Reduces the variable space
- Often meaningful clusters possible
- Easy to apply
- Be careful with the assumptions

References

T. Rietveld and R. van Hout
Statistical Techniques for the Study of Language and Language Behaviour.
Mouton de Gruyter, 1993.

J. Shlens
A Tutorial on Principal Component Analysis.

F. Manni, W. Heeringa and J. Nerbonne
To What Extent are Surnames Words? Comparing Geographic Patterns of Surname and Dialect Variation in the Netherlands.
Literary and Linguistic Computing, Vol. 21, No. 4, 2006.