Principal Components Analysis (PCA)

Janette Walde

janette.walde@uibk.ac.at

Department of Statistics
University of Innsbruck

Outline I

Outline II
Literature

Idea of PCA I

- Suppose that we have a matrix of data X with dimension $n \times p$, where p is large. A central problem in multivariate data analysis is **dimension reduction**: Is it possible to describe, with accuracy, the values of p variables with a smaller number $r < p$ of new variables (variable reduction)?

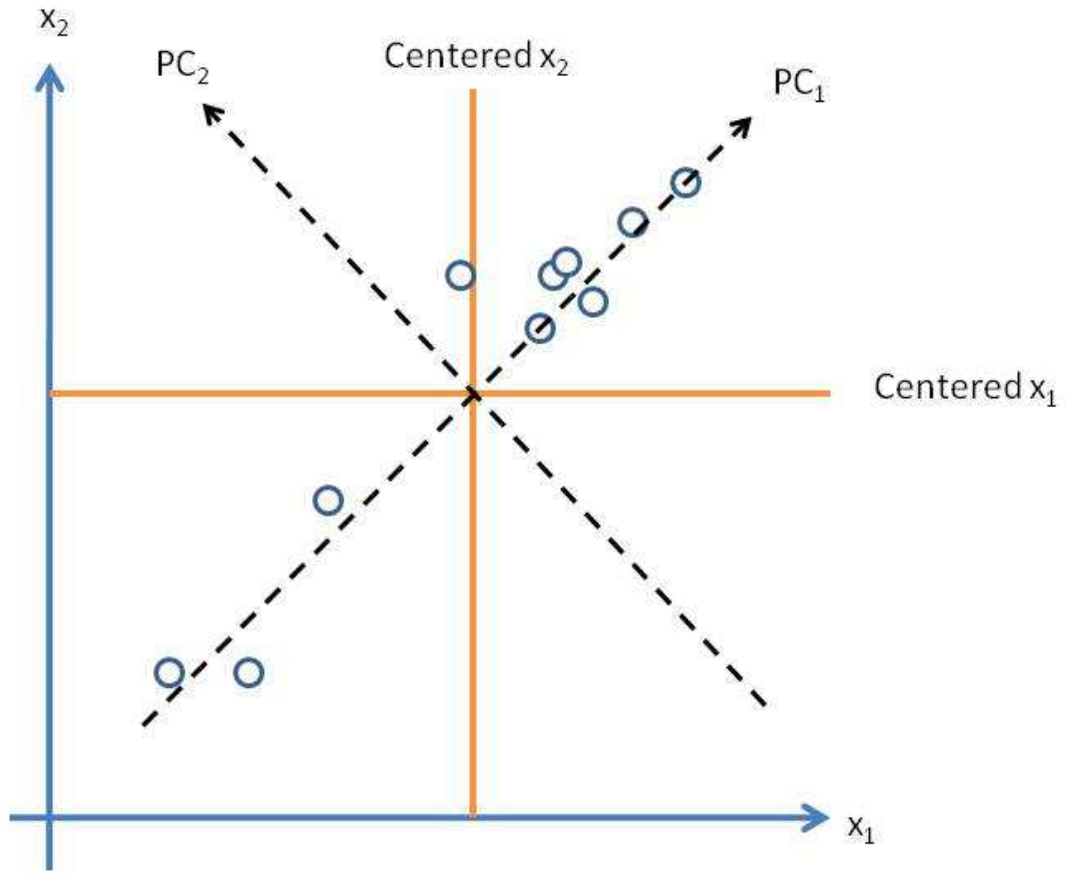
- The **principal component analysis approach** consists on providing an adequate representation of the information with a smaller number of variables constructed as linear combinations of the originals (centered).

Idea of PCA II

- We begin by identifying a group of variables whose variance we believe can be represented more parsimoniously by a smaller set of components, or factors. The end result of the principal components analysis will tell us which variables can be represented by which components, and which variables should be retained as individual variables because the factor solution does not adequately represent their information.

- Relate observed variables to latent variables.

PCA Principle - Axis Rotation



Notation: Linear combinations of variables

For $i = 1, ..., n$ objects (observation units) PCA transforms $j = 1$ to p variables $(X_1, X_2, ..., X_p)$ into $k = 1, ..., p$ new uncorrelated variables $(Z_1, Z_2, ..., Z_p)$ called principal components or factors:

$$z_{ik} = c_{1k}x_{i1} + c_{2k}x_{i2} + ... + c_{pk}x_{ip}$$

z_{ik}...value or score for component k for object i
x_{ij}... values of the original variables for object i
c_{jk}... weights (coefficients) that indicate how much each original variable (j) contributes to the linear combination forming this component (k)

Linear combinations of variables II

- Depending on the analysis, these new variables are termed variously, discriminant functions, canonical functions or variates, principal components or factors.
- The derived variables are extracted so the first explains most of the variance in the original variables, the second explains most of the remaining variance, ...
- The new derived variables are independent of each other.
- Although the number of components that can be derived is equal to the number of original variables, p, we hope that the first few components summarize most of the variation of the original variables.

Eigenvalues and Eigenvectors I

When there are more than two variables the components are extracted in practice by a spectral decomposition of a covariance or correlation matrix. The matrix approach to deriving components produces two important pieces of information: Eigenvalues and Eigenvectors of the association matrix.

- Eigenvalue, also called characteristic root ($\lambda_1, ..., \lambda_p$).
  - Estimates of the eigenvalues provides measures of the amount of the original total variance explained by each of the new derived variables.
  - The sum of all the eigenvalues equals the sum of the variances of the original variables. PCA rearranges the variance in the original variables so it is concentrated in the first few new components.

Eigenvalues and Eigenvectors II

- Eigenvectors (characteristic vectors).
  - Eigenvectors are lists of coefficients or weights (c_j) showing how much each original variable contributes to each new derived variable.
  - The eigenvectors are usually scaled so that the sum of squared coefficients for each eigenvector equals one.
  - Eigenvectors are orthogonal.

Which association matrix to use?

The choice is the covariance matrix of the correlation matrix:

- The **covariance matrix** is based on mean-centered variables and is appropriate when the variables are measured in comparable units and differences in variance between variables make an important contribution to the interpretation.
- The **correlation** is based on z-standardized variables and is necessary when the variables are measured in very different units and we wish to ignore differences between variances.

The choice depends on how much we want different variances among variables to influence our results.

Requirements of PCA I

- The variables included must be metric level or dichotomous (dummy-coded) nominal level.
- The sample size must be greater than 50 (preferably 100).
- The ratio of cases to variables must be 5 to 1 or larger.
- Significance of the elements of the correlation matrix. The correlation matrix for the variables must contain 2 or more correlations of 0.30 or greater.
- Inverse of correlation matrix. Off diagonal elements of the inverse matrix of the correlation matrix should be near zero.
- Bartlett-Test (test of sphericity). The Bartlett test of sphericity is statistically significant.
  H_0: Variables are uncorrelated.
  H_1: Variables are correlated.

Requirements of PCA II

- Anti Image Matrix. The anti-image correlation matrix contains partial correlation coefficients, and the anti-image covariance matrix contains partial covariances. Most of the off-diagonal elements should be small in a good factor model. Rule of thumb: The correlation matrix is not suitable for factor analysis if the proportion of off-diagonal elements of the anti image covariance matrix being unequal to zero (> 0.09) is more than 25%.
- Kaiser-Meyer-Olkin-Criterium. The *measure of sampling adequacy* (MSA) shows to what extend the original variables belong together. If the MSA is less than 0.5 the correlation matrix is not applicable.

Interpreting the components

- The eigenvectors provide the coefficients (c_js) for each variable in the linear combination for each component.
- The further each coefficient is from zero, the greater the contribution that variable makes to that component.
- **Component loadings** are simple correlations (using Pearson's r) between the components and the original variables.
- Ideally we would like a situation where each variable loads strongly on only one component and the loadings are close to plus/minus one or zero.
- For interpretation we look at loadings in absolute value greater than 0.5.

Rotation of components I

The common situation where numerous variables load moderately on each component can sometimes be alleviated by a second rotation of the components after the initial PCA. The aim of this additional rotation is to obtain simple structure, where the coefficients within a component are as close to one or zero as possible.
There are two types of factor rotation methods:

- **Orthogonal rotations**. For example the Varimax procedure rotates the axis such that the two vertices remain 90 degrees (perpendicular) to each other. Assumes uncorrelated factors.

Rotation of components II

- **Oblique rotation** (Direct Oblimin) rotates the axis such that the vertices can have any angle (e.g., other than 90 degrees). Allows factors to be correlated. One can specify the parameter Delta to control the extent to which factors can be correlated among themselves. Delta should be 0 or negative, with 0 yielding the most highly correlated factors and large negative numbers yielding nearly orthogonal solutions (Rule of thumb: -5 is almost orthogonal).

How many components to retain?

- Interpretability. It is important to examine the interpretability of the components and make sure that those providing a biologically interpretable result are retained.
- Eigenvalues greater than one rule. If the PCA is based on a correlation matrix keep any component that has an eigenvalue greater than one.
- Scree diagram. Plot of the eigenvalues for each component and looking for an obvious break (or elbow).
- Test of eigenvalue equality. Bartlett's test when using a covariance matrix.
- Analysis of residuals.

Analysis of residuals - Q-values

- When we retain fewer than all p components, we can only estimate the original data and there will be some of the information in the original data not explained by the components - this is the residual.
- Alternatively, we can measure the difference between the observed correlations or covariances and the predicted correlations or covariances based on the less than p components - this is termed the residual correlation or covariance matrix.
- We have a residual term for each variable for each object and the sum (across variables) of squares of the residuals, often termed Q, can be derived for each object.
- Unusually large values for Q for any observation are an indication that the less than p components we have retained do not adequately represent the original data set for that object.

Application

Lovett et al. (2000) studied the chemistry of forested watersheds in the Catskill Mountains in New York State. They close 29 sites (observations) on first and second order streams and measured the concentrations of ten chemical variables (NO_3^-, total organic N, total N, NH_4^-, dissolved organic C, SO_4^{2-}, Cl^-, Ca^{2+}, Mg^{2+}, H^+), averaged over three years, and four watershed variables (maximum elevation, sample elevation, length of stream, watershed area).

Only use the chemical variables for the PCA. Preliminary checks of the data showed that one stream, Winnisook Brook, was severely acidified with a concentration of H far in excess of the other streams so this site was omitted from further analysis. Additionally, three variables (dissolved organic C, Cl and H) were very strongly skewed and were transformed to $\log_{10}$.

Descriptive Statistics

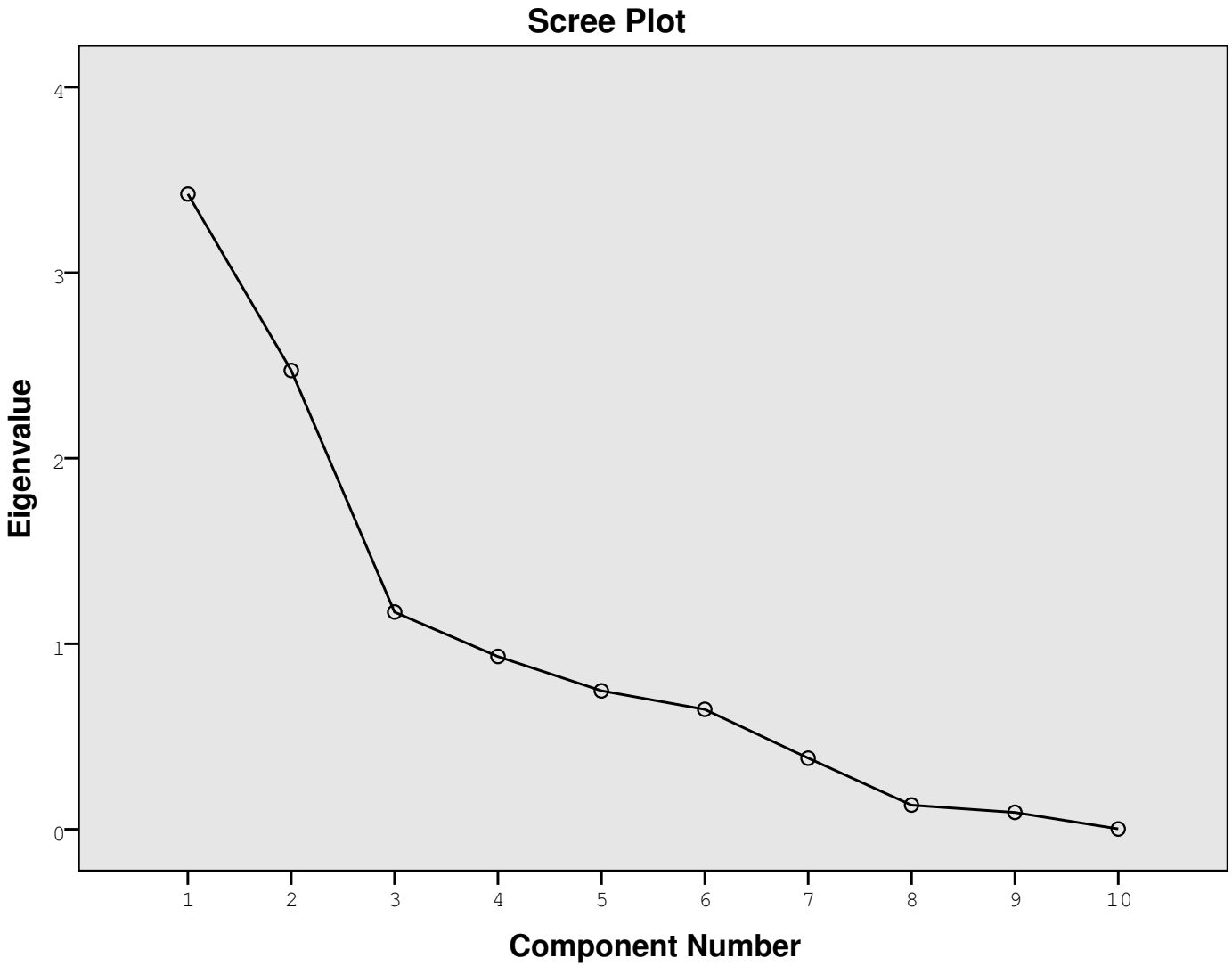
	Mean	Std. Deviation	Analysis N
NO3	22.853	8.6133	38
TON	4.971	1.2793	38
TN	27.892	8.0996	38
NH4	1.647	.7322	38
SO4	62.079	5.2195	38
CA	65.134	13.9550	38
MG	22.858	5.1234	38
LDOC	1.832	.1464	38
LCL	1.328	.1551	38
LH	-.669	.2927	38

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative
1	3.424	34.239	34.239
2	2.473	24.729	58.968
3	1.171	11.711	70.679
4	.932	9.315	79.995
5	.746	7.462	87.456
6	.646	6.464	93.920
7	.384	3.835	97.755
8	.131	1.308	99.063
9	.091	.914	99.977
10	.002	.023	100.000

Extraction Method: Principal Component Analysis.

Scree Plot


Scree Plot
Eigenvalue
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10
Component Number

Component Matrix [a]

	Component		
	1	2	3
NO3	-.483	.816	.053
TON	.272	-.471	.557
TN	-.423	.802	.166
NH4	.422	-.118	-.527
SO4	.682	.354	.262
CA	.520	.701	.087
MG	.873	.024	.326
LDOC	-.533	-.231	.608
LCL	.662	-.248	-.019
LH	-.735	-.443	.006

Extraction Method: Principal Component Analysis.

[a] 3 components extracted.

Rotated Component Matrix [a]

	Component		
	1	2	3
NO3	.046	.943	.104
TON	.175	-.578	.491
TN	.126	.893	.192
NH4	.090	-.284	-.617
SO4	.808	-.064	-.028
CA	.794	.327	-.182
MG	.817	-.448	.011
LDOC	-.324	.038	.775
LCL	.393	-.551	-.206
LH	-.801	-.002	.307

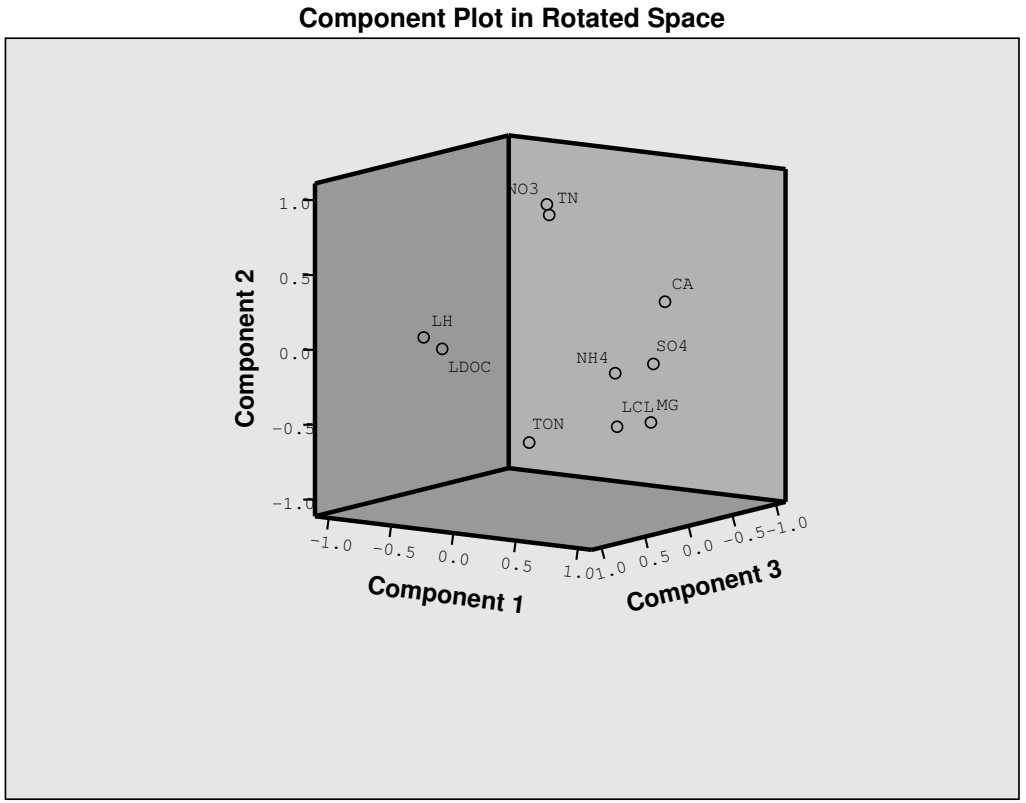
Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

[a] Rotation converged in 5 iterations.

Rotation Sums of Squared Loadings

Component	Total	% of Variance	Cumulative %
1	2.908	29.081	29.081
2	2.719	27.185	56.266
3	1.441	14.413	70.679



Communalities I

Communality is the total amount of variance an original variable shares with all other variables included in the analysis.

Two important concepts of communalities:

1. Principal components analysis assumes that the total variance of the original variables can be explained via the components and uses as starting values for the communalities 1. PCA does not explicitly estimate communalities as in the underlying theoretical model communalities are set equal to the total variance.

Communalities II

2. However, if we assume that besides a common variance each variable has a specific variance too communalities have to be less than 1. Additionally, different factor extraction methods can be applied. As starting estimates for the communalities in many cases the highest squared correlation coefficient of two variables is employed:

$$1 \geq h_j^2 \geq R_j^2 \geq max_k\{r_{jk}^2\}.$$

3. Using the theoretical concept of specific variances of each original variable and employing an appropriate factor extracting method we are conducting a Factor Analysis.

Extraction Methods in Factor Analysis I

- Principal components refers to the principal components model, in which items are assumed to be exact linear combinations of factors. The Principal components method assumes that components (factors) are uncorrelated. It also assumes that the communality of each item sums to 1 over all components (factors), implying that each item has 0 unique variance.

The remaining factor extraction methods allow the variance of each item to be composed to be a function of both item communality and nonzero unique item variance. The following are methods of Common Factor Analysis:

Extraction Methods in Factor Analysis II

- Principal axis factoring uses squared multiple correlations as initial estimates of the communalities. These communalities are entered into the diagonals of the correlation matrix, before factors are extracted from this matrix.
- ...

Factor Scores I

A factor score coefficient matrix shows the coefficients by which items are multiplied to obtain factor scores.

1. Regression Scores - Regression factor scores have a mean of 0 and variance equal to the squared multiple correlation between the estimated factor scores and the true factor values. They can be correlated even when factors are assumed to be orthogonal. The sum of squared discrepancies between true and estimated factors over individuals is minimized.
2. Bartlett Scores - Bartlett factor scores have a mean of 0. The sum of squares of the unique factors over the range of items is minimized.

Factor Scores II

3. Anderson-Rubin Scores - Anderson-Rubin factor scores are a modification of Bartlett scores to ensure orthogonality of the estimated factors. They have a mean of 0 and a standard deviation of 1.

Literature

- Lovett, G.M., Weathers, K.C., and Sobczak, W. V. (2000). Nitrogen saturation and retention in forested watersheds of the Catskill Mountains, New York. *Ecological Applications* 10, pp 73-84.