Reading

Nature Biotechnology **26**, 303 - 304 (2008)
doi:10.1038/nbt0308-303

What is principal component analysis?

Markus Ringnér[1]

Principal component analysis is often incorporated into genome-wide expression studies, but what is it and how can it be used to explore high-dimensional data?

PCA as a tool to Quantify and Visualise

1

Multivariate Analysis

Multiple Regression/two way-ANOVA/mixed models /machine learning algorithms

Ordination methods

2

What is the difference between these two pairwise correlation matrices?

	X_1	X_2	X_3	X_4	X_5
X_1	1.00	0.80	0.90	0.78	0.87
X_2	0.80	1.00	0.76	0.87	0.78
X_3	0.90	0.76	1.00	0.78	0.89
X_4	0.78	0.87	0.78	1.00	0.95
X_5	0.87	0.78	0.89	0.95	1.00

	X_1	X_2	X_3	X_4	X_5
X_1	1.00	0.87	0.96	0.04	0.05
X_2	0.87	1.00	0.95	0.03	0.07
X_3	0.96	0.95	1.00	0.04	0.05
X_4	0.04	0.03	0.04	1.00	0.84
X_5	0.05	0.07	0.05	0.84	1.00

3

What is the difference between these two pairwise correlation matrices?

	X_1	X_2	X_3	X_4	X_5
X_1	1.00	0.80	0.90	0.78	0.87
X_2	0.80	1.00	0.76	0.87	0.78
X_3	0.90	0.76	1.00	0.78	0.89
X_4	0.78	0.87	0.78	1.00	0.95
X_5	0.87	0.78	0.89	0.95	1.00

One dimension

	X_1	X_2	X_3	X_4	X_5
X_1	1.00	0.87	0.96	0.04	0.05
X_2	0.87	1.00	0.95	0.03	0.07
X_3	0.96	0.95	1.00	0.04	0.05
X_4	0.04	0.03	0.04	1.00	0.84
X_5	0.05	0.07	0.05	0.84	1.00

Two dimensions

4

Ordination analyses

- Uncover, organize and summarize the main patterns of variation in a set of variables measured over multiple observations.

- Patterns of variation are structured in a reduced space with smaller number number of dimensions.

- Reduction is possible because often variables are associated (e.g., correlated). Dimensions represent combinations (e.g., linear combinations of variables).

5

Ordination analyses

A procedure for adapting a multidimensional swarm of data points in such a way that when it is projected onto a reduced number of dimensions any intrinsic pattern will become apparent.

Adapted from Connie Clark

6

Ordination analyses – uncover and organize data; a quick example:

Species

Site	B	I	D	A	H	E	G	C
4	1	0	1	0	0	0	0	1
1	0	0	0	1	0	0	0	0
7	0	0	0	0	1	1	1	0
8	0	1	0	0	1	0	1	0
6	0	0	1	0	0	1	1	0
5	0	0	1	0	0	1	0	1
10	0	1	0	0	0	0	0	0
2	1	0	0	1	0	0	0	0
9	0	1	0	0	1	0	0	0
3	1	0	0	1	0	0	0	1

7

Ordination analyses – uncover and organize data; a quick example:

Species

Site	B	I	D	A	H	E	G	C
4	1	0	1	0	0	0	0	1
1	0	0	0	1	0	0	0	0
7	0	0	0	0	1	1	1	0
8	0	1	0	0	1	0	1	0
6	0	0	1	0	0	1	1	0
5	0	0	1	0	0	1	0	1
10	0	1	0	0	0	0	0	0
2	1	0	0	1	0	0	0	0
9	0	1	0	0	1	0	0	0
3	1	0	0	1	0	0	0	1

=

Species

Sites	A	B	C	D	E	G	H	I
1	1	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0
4	0	1	1	1	0	0	0	0
5	0	0	1	1	1	0	0	0
6	0	0	0	1	1	1	0	0
7	0	0	0	0	1	1	1	0
8	0	0	0	0	0	1	1	1
9	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	1

8

Ordination methods

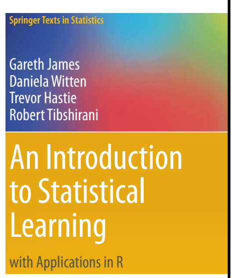
- **Principal Component Analysis (PCA)**
- Correspondence Analysis (CA)
- Principal Coordinate Analysis (PCoA)
- Discriminant Function Analysis (DFA)
- Principal Curve Analysis
- Etc, etc, etc…

Principal components analysis (PCA) is perhaps the most common technique used to summarize patterns among variables in multivariate datasets.

9

10

Some treat Principal Component Analysis (PCA) as an unsupervised learning method (an exploratory technique such as k-means)



11

Supervised *versus* unsupervised learning techniques

- Techniques for unsupervised learning are fast growing in a number of fields, particularly biology.

- A cancer researcher might assay gene expression levels in 100 patients with breast cancer. They might then look for subgroups among the breast cancer samples, or among the genes, in order to obtain a better understanding of the disease.

- A search engine might choose what search results to display to a particular individual based on the click histories of other individuals with similar search patterns. These statistical learning tasks, and many more, can be performed via unsupervised learning techniques.

Adapted from James et al. 2013

12

Supervised *versus* unsupervised learning techniques

In contrast, unsupervised learning is often much more challenging. The exercise tends to be more subjective, and there is no simple goal for the analysis, such as prediction of a response.

Unsupervised learning is often performed as part of an exploratory data analysis.

Hard to assess the results obtained given that there is no universally accepted mechanism for performing cross-validation or validating results on an independent data set; there is no way to check how the models does because we don't know the true answer—the problem is unsupervised.

Adapted from James et al. 2013

13

Examples of Principal Component Analysis

14

Principal components analysis (PCA) - example 1

A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study

Monique L Den Boer, Marjon van Slegtenhorst*, Renée X De Menezes, Meyling H Cheok, Jessica G C A M Buijs-Gladdines, Susan T C J M Peters, Laura J C M Van Zutven, H Berna Beverloo, Peter J Van der Spek, Gaby Escherich†, Martin A Horstmann†, Gritta E Janka-Schaub†, Willem A Kamps‡, William E Evans, Rob Pieters‡*

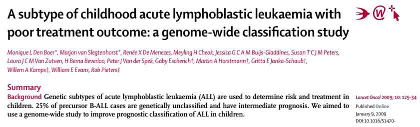
Summary
Background Genetic subtypes of acute lymphoblastic leukaemia (ALL) are used to determine risk and treatment in children. 25% of precursor B-ALL cases are genetically unclassified and have intermediate prognosis. We aimed to use a genome-wide study to improve prognostic classification of ALL in children.

Lancet Oncol 2009; 10: 125–34
Published Online
January 9, 2009
DOI:10.1016/S1470-

Quantification and Visualisation

15

Principal components analysis (PCA) - example 1


A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study

Monique L Den Boer, Marjon van Slegtenhorst*, Renée X De Menezes, Meyling H Cheok, Jessica G C A M Buijs-Gladdines, Susan T C J M Peters, Laura J C M Van Zutven, H Berna Beverloo, Peter J Van der Spek, Gaby Escherich†, Martin A Horstmann†, Gritta E Janka-Schaub†, Willem A Kamps‡, William E Evans, Rob Pieters‡*

Summary

Background Genetic subtypes of acute lymphoblastic leukaemia (ALL) are used to determine risk and treatment in children. 25% of precursor B-ALL cases are genetically unclassified and have intermediate prognosis. We aimed to use a genome-wide study to improve prognostic classification of ALL in children.

Lancet Oncol 2009; 10: 125–34
Published Online
January 9, 2009
DOI:10.1016/S1470-

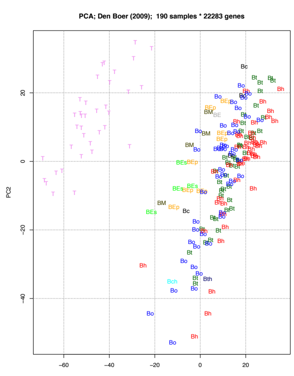
Data matrix: 190 observations by 22283 columns

Gene expression (22283 genes)

Gene expression (190 patients)

16

Principal components analysis (PCA) - example 1

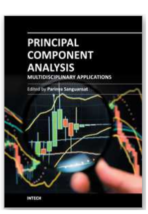


Each letter is a patient. Labels stand for different lymphoblastic leukaemia (ALL) types.

Data matrix: 190 observations by **22283** columns.

17

Principal components analysis (PCA) - example 2



PCA – A Powerful Method for Analyze Ecological Niches

Franc Janžekovič and Tone Novak
University of Maribor, Faculty of Natural Sciences and Mathematics, Department of Biology, Maribor Slovenia

18

Principal components analysis (PCA) - example 2

2.1 Environmental niche of three hymenopteran and two spider species

Between 1977 and 2004, 63 caves and artificial tunnels were ecologically investigated in Slovenia; the three most abundant Hymenoptera species found in these studies have been ecologically evaluated (details in Novak et al. 2010a). In the caves, many environmental data were collected, as follows. The following abbreviations of the environmental variables are used: Dist-E = distance from entrance; Dist-S = distance from surface; Illum = illumination; PCS = passage cross-section; Tair =air temperature; RH = relative air humidity; Tgr = ground temperature; HY = substrate moisture. The hymenopteran spatial niche breadth was originally represented by nine variables.

Data matrix: 63 observations (caves) by 9 columns

Environmental variables (9)

63 caves

PCA – A Powerful Method for Analyze Ecological Niches

Franc Janžekovič and Tone Novak
University of Maribor, Faculty of Natural Sciences and Mathematics, Department of Biology, Maribor Slovenia

19

Principal components analysis (PCA) - example 2

(pairwise correlation among environmental variables)

	1	2	3	4	5	6	7	8	9
1 Air temperature	1.00 ---								
2 *arc-sin* relative air humidity	0.15 0.133	1.00 ---							
3 Ground temperature	**0.94 <0.001**	0.18 0.079	1.00 ---						
4 *arc-sin* substrate moisture	**0.388 <0.001**	**0.59 <0.001**	**0.37 <0.001**	1.00 ---					
5 Airflow	**-0.48 <0.001**	**-0.36 <0.001**	**-0.43 <0.001**	**-0.55 <0.001**	1.00 ---				
6 Distance from entrance	**-0.34 <0.001**	0.14 0.153	**-0.41 <0.001**	0.10 0.312	0.04 0.712	1.00 ---			
7 Distance from surface	-0.02 0.837	**0.24 0.017**	-0.04 0.683	**0.46 <0.001**	-0.11 0.275	**0.67 <0.001**	1.00 ---		
8 Passage cross-section	**0.35 <0.001**	0.17 0.089	**0.23 0.025**	**0.39 <0.001**	**-0.40 <0.001**	-0.11 0.274	0.05 0.656	1.00 ---	
9 *log* illumination	**0.45 <0.001**	-0.18 0.077	**0.46 <0.001**	-0.04 0.690	-0.07 0.494	**-0.821 <0.001**	**-0.679 <0.001**	**0.37 <0.001**	1.00 ---

Table 1. Pearson correlations coefficient among nine environmental variables. Significant correlations in bold. (Upper row r, lower row p).

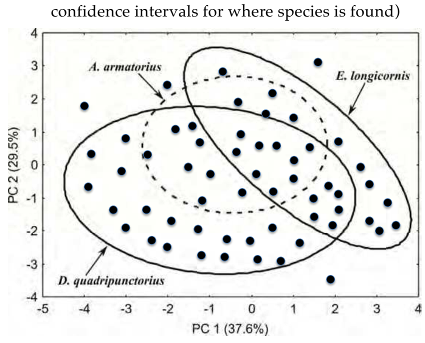
PCA – A Powerful Method for Analyze Ecological Niches

Franc Janžekovič and Tone Novak
University of Maribor, Faculty of Natural Sciences and Mathematics, Department of Biology, Maribor Slovenia

20

Principal components analysis (PCA) - example 2

(niche differences – dots represent different caves ellipsoids are confidence intervals for where species is found)



Each dot represents a cave

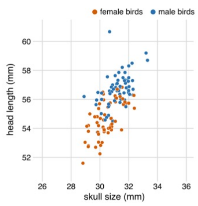
Fig. 5. Ordination of the nine environmental variables in 1st and 2nd PC axes. Ellipses (95% confidence) represent spatial niches in the three hymenopteran species.

21

Principal Component Analysis (PCA): A geometric interpretation

PCA finds the coordinate system (called principal components) that best represents the internal variability in the data, essentially re-projecting the data on these coordinate system. As such, PCA represents associations among variables (gene, environmental variables) and data points are re-projected so that the correlations among variables is maximized.

Original data



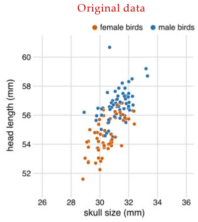
Source https://wilkelab.org/SDS375/slides/dimension-reduction-1.html#9

22

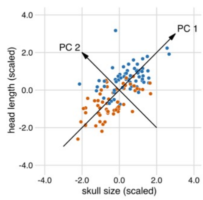
Principal Component Analysis (PCA): A geometric interpretation

PCA finds the coordinate system (called principal components) that best represents the internal variability in the data, essentially re-projecting the data on these coordinate system. As such, PCA represents associations among variables (gene, environmental variables) and data points are re-projected so that the correlations among variables is maximized.

Original data



Standardization and PCA fitting



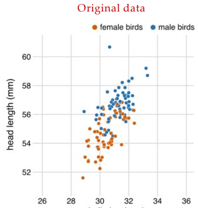
Source https://wilkelab.org/SDS375/slides/dimension-reduction-1.html#9

23

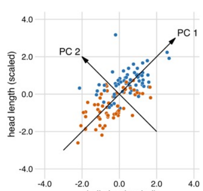
Principal Component Analysis (PCA): A geometric interpretation

PCA finds the coordinate system (called principal components) that best represents the internal variability in the data, essentially re-projecting the data on these coordinate system. As such, PCA represents associations among variables (gene, environmental variables) and data points are re-projected so that the correlations among variables is maximized.

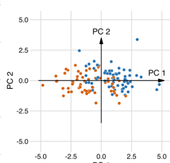
Original data



Standardization and PCA fitting



Rotation



PCA aligns their axes with directions of maximum variation in the data

Source https://wilkelab.org/SDS375/slides/dimension-reduction-1.html#9

24

Principal Component Analysis (PCA): A geometric interpretation

- PCA constructs a new coordinate system (new variables, PCs) which are linear combinations of the original data and which are defined to align the samples along their major axes of variation (assuming linearity).
- Thus, PCA determines the coordinate system that best represents the internal variability in the data, essentially re-projecting the data.

25

The association among variables need to be measured by either (in most cases):

Correlation Matrix (for variables that have different units or scales, e.g., ph, temperature).

Covariance Matrix (variables have the same units, e.g., body length & body width in cm).

Raw data when variables are in the same units (more difficult to interpret) and calculations differ (very rare to find applications in the literature); rarely used.

26

Correlation *versus* covariance

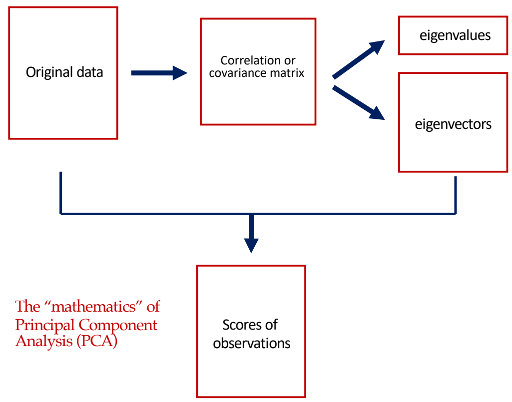
$$COV_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$X = 0 \;\&\; Y = 0 \;\therefore s_x = s_x \;\&\; s_y = s_y$$

$$COR_{xy} = \frac{COV_{xy}}{s_x s_y}$$

$$X = 0 \;\&\; Y = 0 \;\therefore s_x = 1 \;\&\; s_y = 1$$

27



28

The mathematics of Principal Component Analysis (PCA):

Eigen-analysis is a mathematical operation on a *square symmetric* matrix (e.g., pairwise correlation matrix, pairwise covariance matrix).

A *square* matrix has the same number of rows as columns.

A *symmetric* matrix is the same if you switch rows and columns.

29

square and symmetric matrix (e.g., pairwise correlation matrix)

	X_1	X_2	X_3	X_4	X_5
X_1	1.00	0.80	0.90	0.78	0.87
X_2	0.80	1.00	0.76	0.87	0.78
X_3	0.90	0.76	1.00	0.78	0.89
X_4	0.78	0.87	0.78	1.00	0.95
X_5	0.87	0.78	0.89	0.95	1.00

30

The important components of Principal Component Analysis (pun intended)

31

Principal component analysis presents three important structures:

1 – **Eigenvalues:** represent the amount of variation in the original data summarized by each principal component. The first principal component (PC-1) presents the largest amount, PC-2 presents the second largest amount, and so on.

32

Eigenvalues

	X_1	X_2	X_3	X_4	X_5
X_1	1.00	0.80	0.90	0.78	0.87
X_2	0.80	1.00	0.76	0.87	0.78
X_3	0.90	0.76	1.00	0.78	0.89
X_4	0.78	0.87	0.78	1.00	0.95
X_5	0.87	0.78	0.89	0.95	1.00

"one dimension"

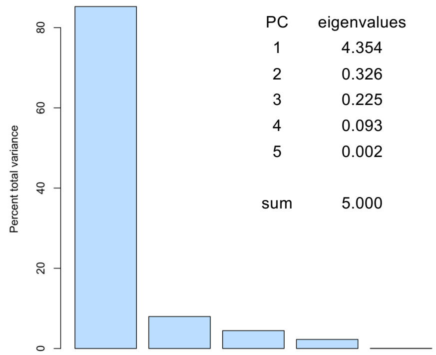
Eigenvalues:

PC	eigenvalues	%
1	4.354	0.871
2	0.326	0.065
3	0.225	0.045
4	0.093	0.019
5	0.002	0.000
sum	5.000	1.000

"Lower" dimensionality because it kept a large proportion of the variation in the data in the first PC.

33

Plot of eigenvalue contributions



PC	eigenvalues	%
1	4.354	0.871
2	0.326	0.065
3	0.225	0.045
4	0.093	0.019
5	0.002	0.000
sum	5.000	1.000

34

Eigenvalues

1.00	0.87	0.96	0.04	0.05
0.87	1.00	0.95	0.03	0.07
0.96	0.95	1.00	0.04	0.05
0.04	0.03	0.04	1.00	0.84
0.05	0.07	0.05	0.84	1.00

"two dimensions"

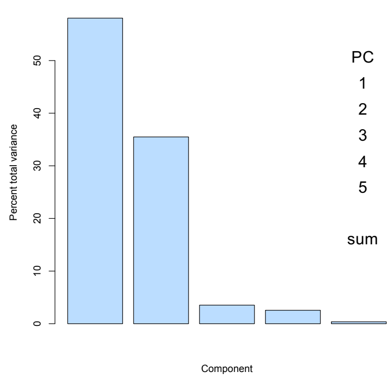
Eigenvalues:

PC	eigenvalues	%
1	2.867	0.573
2	1.827	0.365
3	0.167	0.033
4	0.124	0.025
5	0.015	0.003
sum	5.000	1.000

"Higher" dimensionality because two components are needed to summarize variation.

35

Plot of eigenvalues



PC	eigenvalues	%
1	2.867	0.573
2	1.827	0.365
3	0.167	0.033
4	0.124	0.025
5	0.015	0.003
sum	5.000	1.000

36

Principal component analysis presents three important structures:

2 - **Eigenvectors:** Each principal component is a linear function with coefficients for each variable.

- Eigenvectors contain these coefficients. High values, positive or negative, represents high association with the component.

37

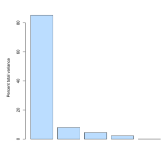
Correlation matrix

	X_1	X_2	X_3	X_4	X_5
X_1	1.00	0.80	0.90	0.78	0.87
X_2	0.80	1.00	0.76	0.87	0.78
X_3	0.90	0.76	1.00	0.78	0.89
X_4	0.78	0.87	0.78	1.00	0.95
X_5	0.87	0.78	0.89	0.95	1.00

"one dimension"

Associated eigenvectors

var	PC 1	2	3	4	5
1	0.447	-0.436	0.330	-0.687	0.170
2	0.432	0.533	0.644	0.181	-0.288
3	0.445	-0.534	0.035	0.692	0.192
4	0.450	0.489	-0.413	-0.063	0.619
5	0.462	-0.039	-0.552	-0.109	-0.684



38

Eigenvectors can be seen as regression coefficients, where the component is the dependent variable. A "one dimension" matrix has only one interpretable principal component.

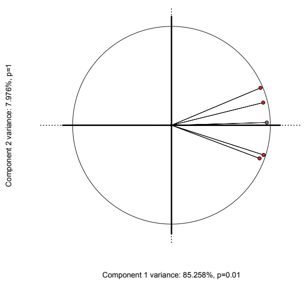
$$PC\text{-}1 = 0.447X_1 + 0.432X_2 + 0.445X_3 + 0.450X_4 + 0.462X_5$$

Unlike the numbers after =, this is not a subtraction but a hyphen stating that this is the first and second Principal Components (PC).

var	PC 1	2	3	4	5
1	0.447	0.436	0.330	-0.687	0.170
2	0.432	-0.533	0.644	0.181	-0.288
3	0.445	0.534	0.035	0.692	0.192
4	0.450	-0.489	-0.413	-0.063	0.619
5	0.462	0.039	-0.552	-0.109	-0.684

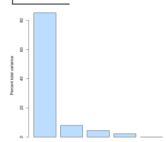
39

Eigenvectors (simulated data with 1 dimension): only first axis (PC-1) should be interpreted



PC axis

variable	1
1	0.447
2	0.432
3	0.445
4	0.450
5	0.462



40

Correlation matrix

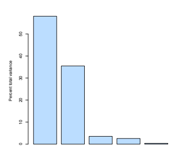
	X_1	X_2	X_3	X_4	X_5
X_1	1.00	0.87	0.96	0.04	0.05
X_2	0.87	1.00	0.95	0.03	0.07
X_3	0.96	0.95	1.00	0.04	0.05
X_4	0.04	0.03	0.04	1.00	0.84
X_5	0.05	0.07	0.05	0.84	1.00

"two dimensions"

Associated eigenvectors (only interpret the first two components (PC)

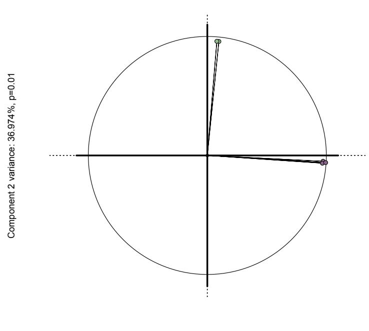
PC

var	1	2	3	4	5
1	0.569	-0.064	0.249	-0.642	0.445
2	0.567	-0.060	-0.298	0.661	0.386
3	0.585	-0.067	0.061	-0.010	-0.806
4	0.072	0.704	0.651	0.273	0.039
5	0.085	0.702	-0.650	-0.277	-0.043



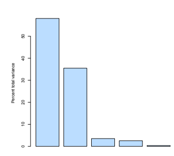
41

Eigenvectors (simulated data with 2 dimensions): only first two axis (PC-1 & PC-2) should be interpreted



PC axis

var	1	2
1	0.569	-0.064
2	0.567	-0.060
3	0.585	-0.067
4	0.072	0.704
5	0.085	0.702



42

Principal component analysis presents three important structures:

3 – **Multivariate scores:** Since each component is a linear function of the variables, when multiplying the standardized variables (in the case of correlation matrices) by the eigenvector structure, a matrix containing the position of each observation in each principal component is produced.

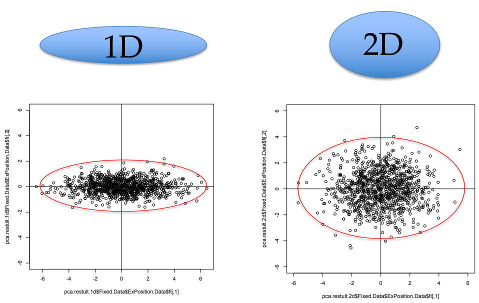
The plot of these scores in the first few dimensions, represents the main patterns of variation among the original observations (more in the empirical example).

$PC\text{-}1=0.569X_1+0.567X_2+0.585X_3+0.072X_4+0.085X_5$

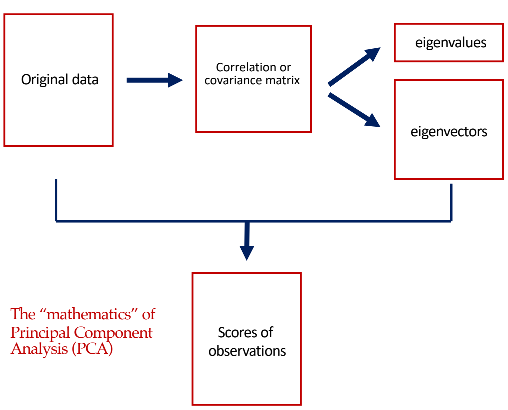
$PC\text{-}2=-0.064X_1-0.060X_2-0.067X_3+0.704X_4+0.702X_5$

43

PCA Scores: one versus two dimensions



44



The "mathematics" of Principal Component Analysis (PCA)

45

Next lecture: How many PCA dimensions? Inferential frameworks for determining number of axes to interpret and the significance of each variable on each axis (lots of work on this area).

1st) determine how many axes to interpret (i.e., how many PCs capture correlated variation in the data?).

Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Computational Statistics & Data Analysis 49 (2005) 974–997
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

How many principal components? stopping rules for determining the number of non-trivial axes revisited

Pedro R. Peres-Neto*, Donald A. Jackson, Keith M. Somers

46