Principal Component Analysis (PCA)

ADA Session 5

Dr Wenjuan Zhang

Wenjuan.zhang@wbs.ac.uk

Scenarios when PCA can be used

- A financial analyst is interested in determining the financial health of firms in a given industry. Research studies have identified **a large number** of financial ratios (about 120) that can be used for such a purpose.

- The quality control department is interested in developing **a few key composite indices from numerous pieces of information** resulting from the manufacturing process to determine if the process is or is not in control.

- The marketing manager is interested in developing a regression model to forecast sales. However, the independent variables under consideration are correlated (**multicollinearity**) among themselves.

What is Principal component analysis (PCA)

- An interdependence multivariate statistical technique

- Aims to find a way of condensing the information contained in your original variables into a smaller set of principal components without loosing much information.

- A technique for forming new variables (**principal components**) which are linear composites of the original variables.

- Principal components
  - the maximum number of principal components that can be formed is equal to the number of original variables,
  - are uncorrelated among themselves,
  - are ordered by their importance (starting with the most important),
  - We hope that first few of them contain enough information about original variables.

When we use PCA?

2 Aims of PCA

- Main aim is to represent the original data using a lower-dimension new variables, i.e. to **reduce dimension,** hence called a data-reduction technique

- Other aim is to represent the original variables via new variables that are uncorrelated, these new uncorrelated variables can be used in further analysis where multicolinearity is a problem (e.g. In regression analysis, cluster analysis...)

PCA in statistical books and software

- In some statistical software (SPSS or SAS) the Principal Component Analysis is listed under the heading Factor Analysis.
- In other software PCA has its own dedicated routine (e.g. in Splus)
- Some books talk about PCA as part of Factor Analysis (such as Hair et. al Multivariate Data Analysis, A Global Perspective, 7th edition).
- A good description of PCA ideas and math is in Bryan Manly, Multivariate Statistical Methods – A primer, 2004.
- There are similarities and differences between PCA and Factor Analysis.
- Factor Analysis will be taught in the next session, and the similarities and differences with PCA will be discussed.

How we do Principal Component Analysis: 6 stages

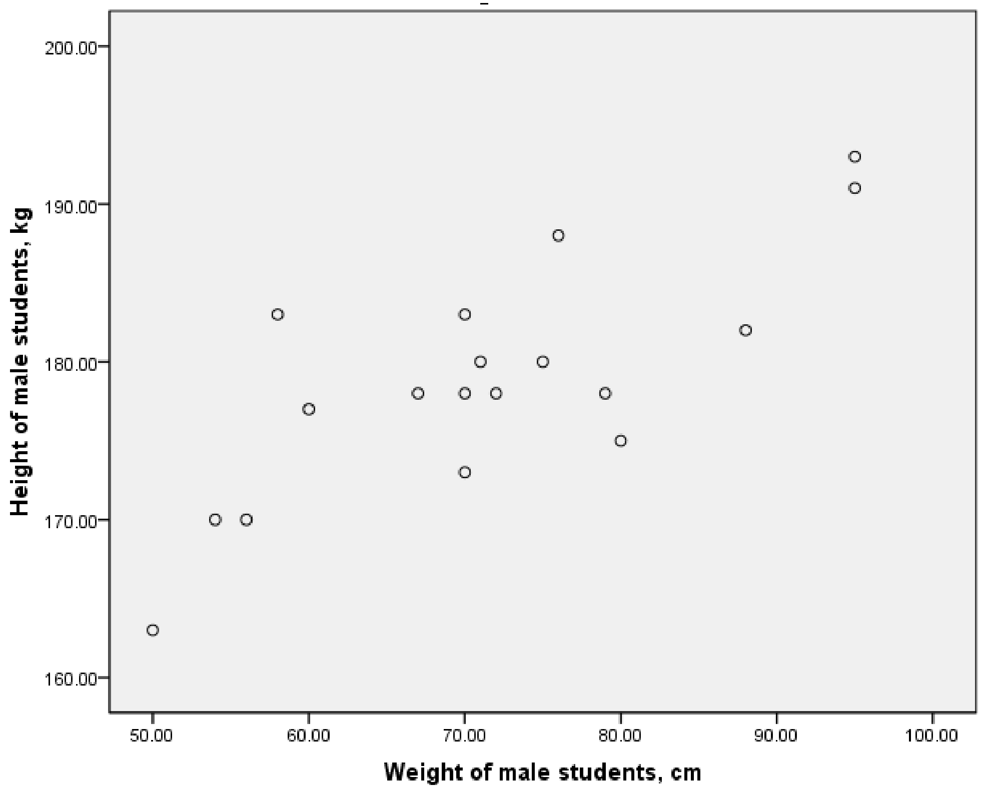
1. Objectives	Define the problem. Aim?
2. Research Design	Make pre-analysis decisions: Sample size, variables, outliers, missing values, standardization?
3. Check Assumptions	Multicollinearity?
4. Create Principal Components	Calculate the Principal Components. Number of important principal components? Can we reduce dimensionality?
5. Interpretation	Is interpretation of PCs possible?
6. Validation	Stability: outliers? Use of principal component scores for further analysis, e.g.for cluster analysis.

Example (using SPSS)

Heights & Weights of Male Students

Ht (cm)	Wt (kg)
163	50
170	54
170	56
173	70
175	80
177	60
178	67
178	70
178	72
178	79
180	71
180	75
182	88
183	58
183	70
188	76
191	95
193	95

Aim: Can we reduce the dimension, i.e. the number of the variables? Can we replace the two variables by one new variable that can be used for comparing the students?



Stage 2 Research Design

Sample size

- Should be > 50 observations, preferably > 100
- Should be > 10 times multiple of the number of variables.

Variables for PCA?

- Should be metric variables.
- Can also be ordinal variables such as responses from 7-point Likert scale used in questionnaires

Outliers?

- Need to remove outliers, because they will distort the PCA solution.
- To detect outliers we can use same measures as in the Cluster Analysis.

Stage 2 Research Design continued...

Type of data on PCA

- PCA can be either done on mean-corrected data, the calculation is based on the covariance matrix.
- If variables measured on different scales, or if different variability, then PCA should be done on **standardized** variables, i.e. PCA is done on **correlation** matrix

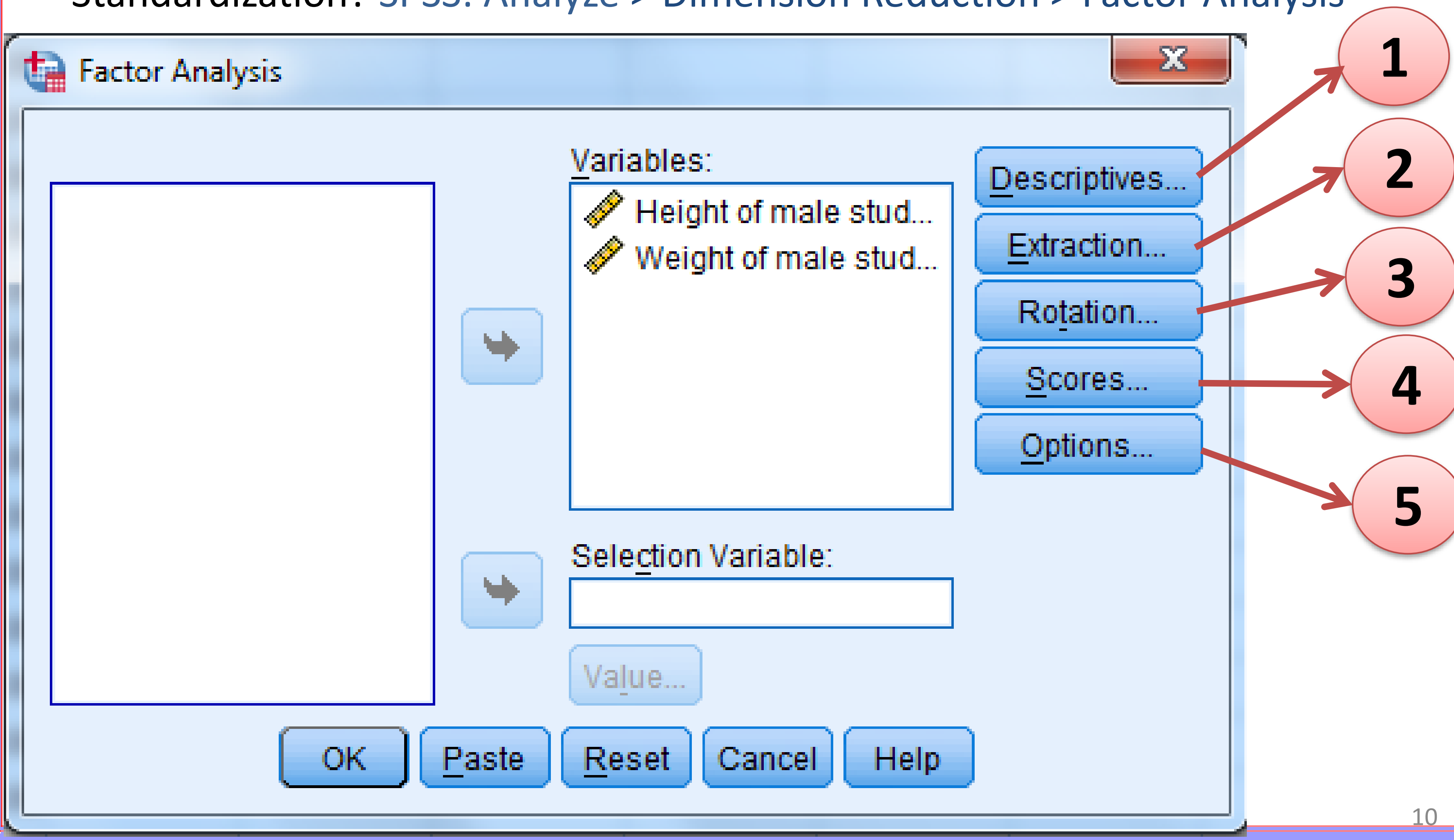
Missing values?

- If values are missing at random – not a problem.
- If values are missing NOT at random
  - E.g. If some groups of respondents tend not to answer some parts of a questionnaire
  - This is a problem, data do not have enough information about the relationship between the variables.
  - Any findings from PCA only relates to the sample at hand.

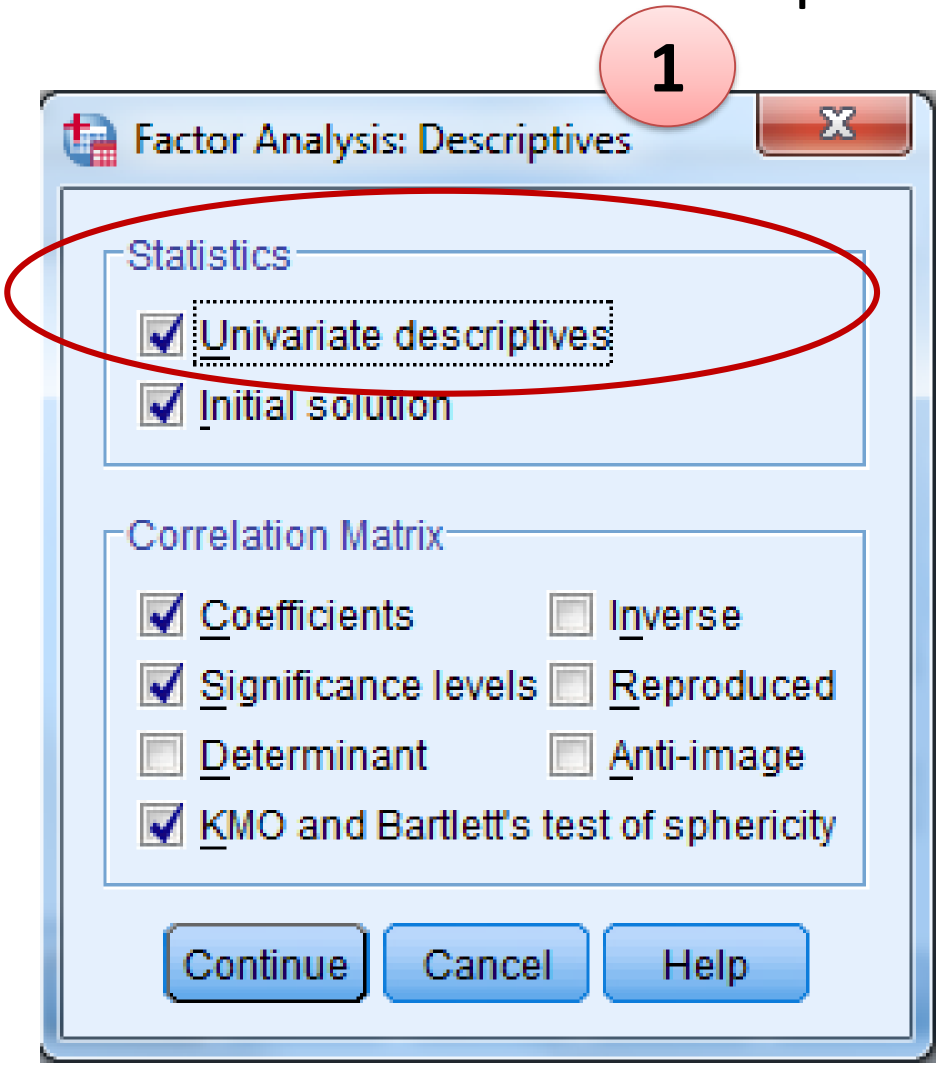
Example: Heights & Weights of Male Students

Think about research design in this example:

- Sample size? Variables? Outliers? Missing values?
- Standardization? SPSS: Analyze > Dimension Reduction > Factor Analysis



Click the 'Descriptives' button to bring up dialog box. Choose univariate descriptives.



Descriptive Statistics

	Mean	Std. Deviation	N
Height of male students, kg	178.8889	7.44303	18
Weight of male students, cm	71.4444	13.03490	18

The 2 variables are measured in different scales and there are differences in standard deviation, hence we will standardize the data i.e. we will use the correlation matrix for PCA calculations.

Stage 3 Assumptions Check

Is there substantial multicollinearity in the variables?

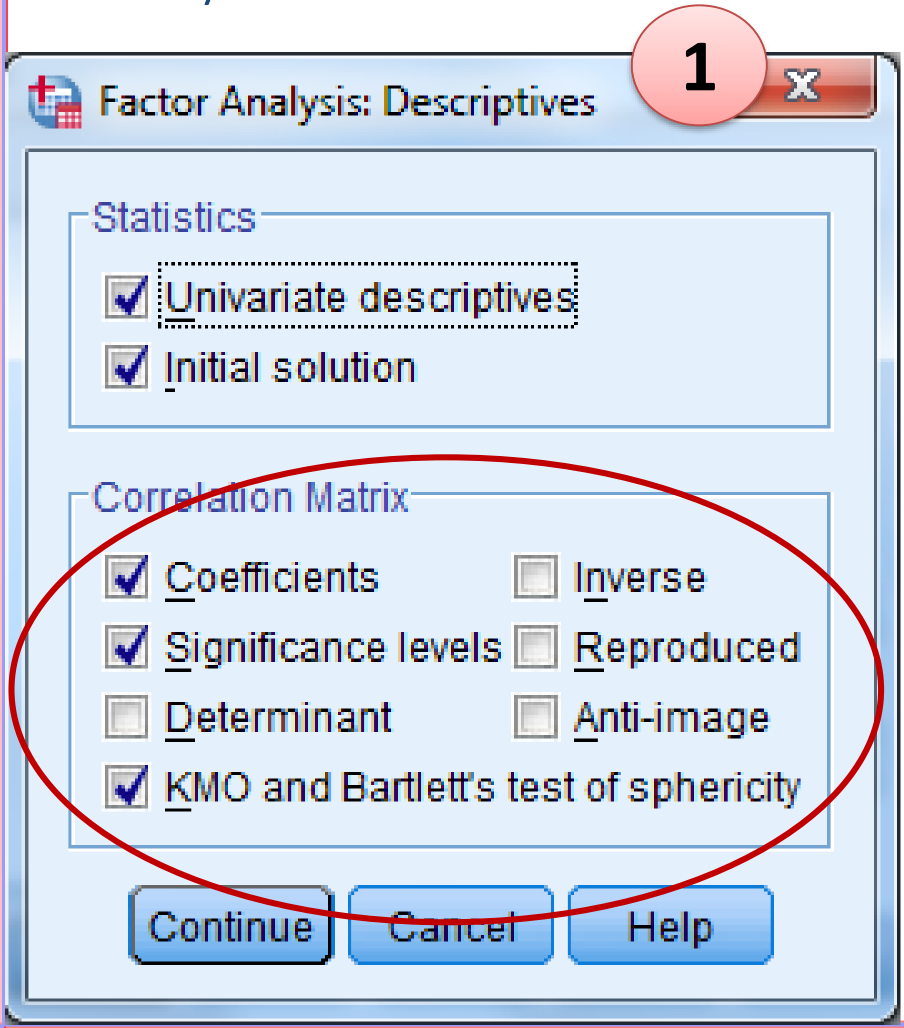
If original variables are uncorrelated then they can not be reduced to a smaller number of variables, i.e. PCA does not work. We need original variables to be highly correlated, positively or negatively.

- **Pairwise Correlations**
  - Rule of thumb: If at least one pairwise correlation > 0.8, then we conclude that these variables are highly correlated.
  - More precisely we check if correlation is significant ($p < 0.05$)
  - Or a lot of pairwise correlations that are > 0.3
- **Kaiser-Meyer-Olkin (KMO) statistic (measure)**
  - KMO is used for assessing sampling adequacy and evaluates the correlations and partial correlations to determine if the data are likely to coalesce on components (i.e. some items highly correlated, some not)
  - If KMO > 0.5 then we conclude that some variables are highly correlated
- **Bartlett's test**
  - The Bartlett's test evaluates whether or not our correlation matrix is an identity matrix (1 on the diagonal & 0 on the off-diagonal).
  - If significant (e.g. $p < .05$) then we conclude that the correlation matrix is different from diagonal.

Example: Heights & Weights of Male Students

Assumptions check: Multicollinearity?

In SPSS the Principal Component Analysis is done via Factor Analysis menu:
Analyze > Dimension Reduction > Factor Analysis > ... **Descriptives...**



Example: Heights & Weights of Male Students

Assumptions check: Multicollinearity? Does it make sense to try to use PCA to reduce dimensionality of the data?

Correlation Matrix

		Height of male students, kg	Weight of male students, cm
Correlation	Height of male students, kg	1.000	.777
	Weight of male students, cm	.777	1.000
Sig. (1-tailed)	Height of male students, kg		.000
	Weight of male students, cm	.000	

There is strong evidence that data are highly correlated because Correlation=0.777 with p-value=0.000<0.05.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.500
Bartlett's Test of Sphericity	Approx. Chi-Square	14.324
	df	1
	Sig.	.000

Bartlet test p-value=0.000<0.05 and KMO KMO=0.5. So we believe there is basis for using PCA to reduce dimensionality.

Stage 4 Calculate Principal Components

How we calculate Principal Components?

- We have *p* original variables (e.g. Height and weight.)
- PCA finds *p* new variables called *principal components* (PC) that are linear combinations of original variables.
- The first *PC1* is the most important. The importance is measured via variance. PC1 accounts for the *maximum total variance* in the data.
- The second *PC2* is the second most important. It is uncorrelated to the PC1 and it accounts for the maximum variance that is left unexplained by PC1.
- The *m*th principal component accounts for the maximum variance that has not been accounted for by the *first m-1* variables, and is uncorrelated with them. Etc, until *p* new variables (principal components) are created.
- The new variables are used to assign new values to objects, the new values are called *principal components scores*.

Stage 4 Calculate Principal Components continued...

How we calculate Principal Components? Graphical illustration of ideas

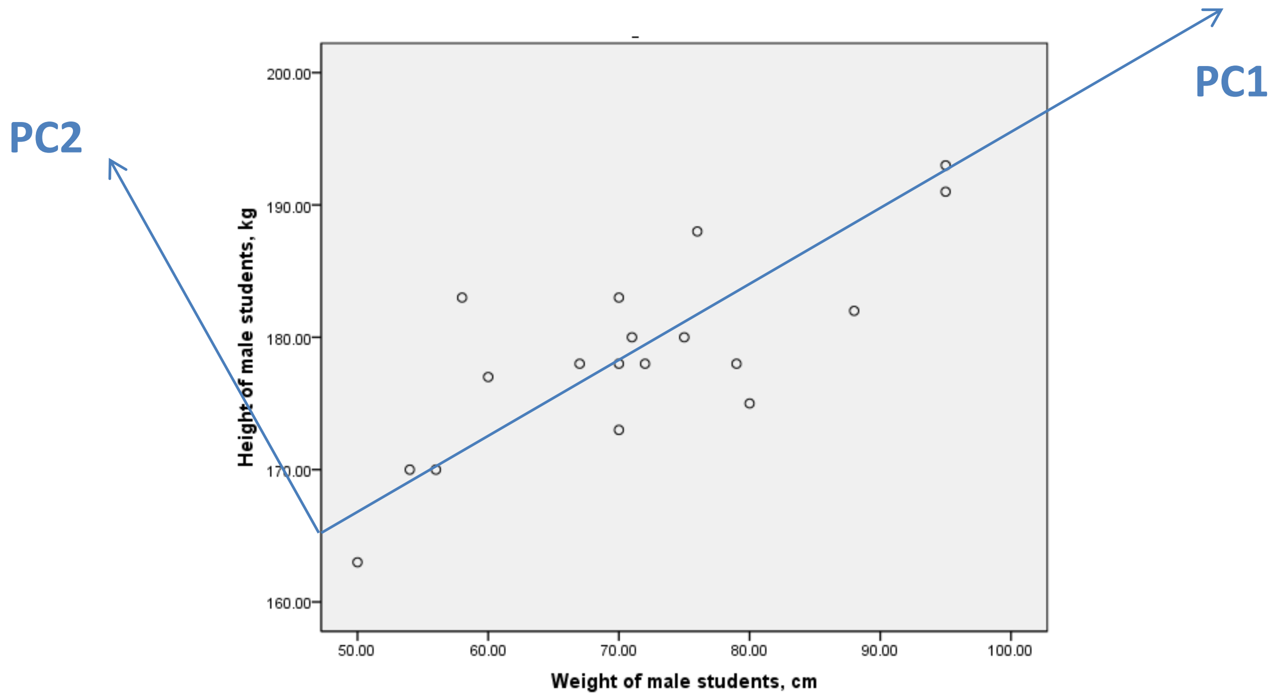
- Identify alternative axes by rotating the original axes by a certain angle
- The projection of the observations to the new axes produce new variables which are orthogonal (uncorrelated).
- The first new axis (PC1) should go in direction of the largest spread in the data.

Example (using SPSS)

Heights & Weights of Male Students

Ht (cm)	Wt (kg)
163	50
170	54
170	56
173	70
175	80
177	60
178	67
178	70
178	72
178	79
180	71
180	75
182	88
183	58
183	70
188	76
191	95
193	95

The graphical illustration of the principal components, PC1 and PC2



Stage 4 Calculate Principal Components continued...

How we calculate Principal Components? Analytical approach

Assuming that there are *p original variables*, $X_1, X_2, \ldots, X_p$, we are interested in forming the following p linear combinations:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

$$\vdots$$

$$Z_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

Where $Z_1, Z_2, \ldots, Z_p$ are the ***p principal components (PCs)***, and a_{ij} is the ***weight*** of the *j*th variable for the *i*th principal component. If original data are standardized, then the weights of PCs are calculated as eigen vectors of the correlation matrix, and then the variances of PCs are the eigen values of the correlation matrix.

Example: Heights & Weights of Male Students

Calculate Principal Components in SPSS:

-Analyze > Descriptive statistics> Descriptives...

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
Height of male students, kg	18	178.8889	7.44303	55.399
Weight of male students, cm	18	71.4444	13.03490	169.908
Valid N (listwise)	18			

Total variability=55.4+169.9=225.3

Height accounts for 24.6% of all variability (i.e. 55.4/225.3)

Weight accounts for 75.4% of total variability

Example: Heights & Weights of Male Students

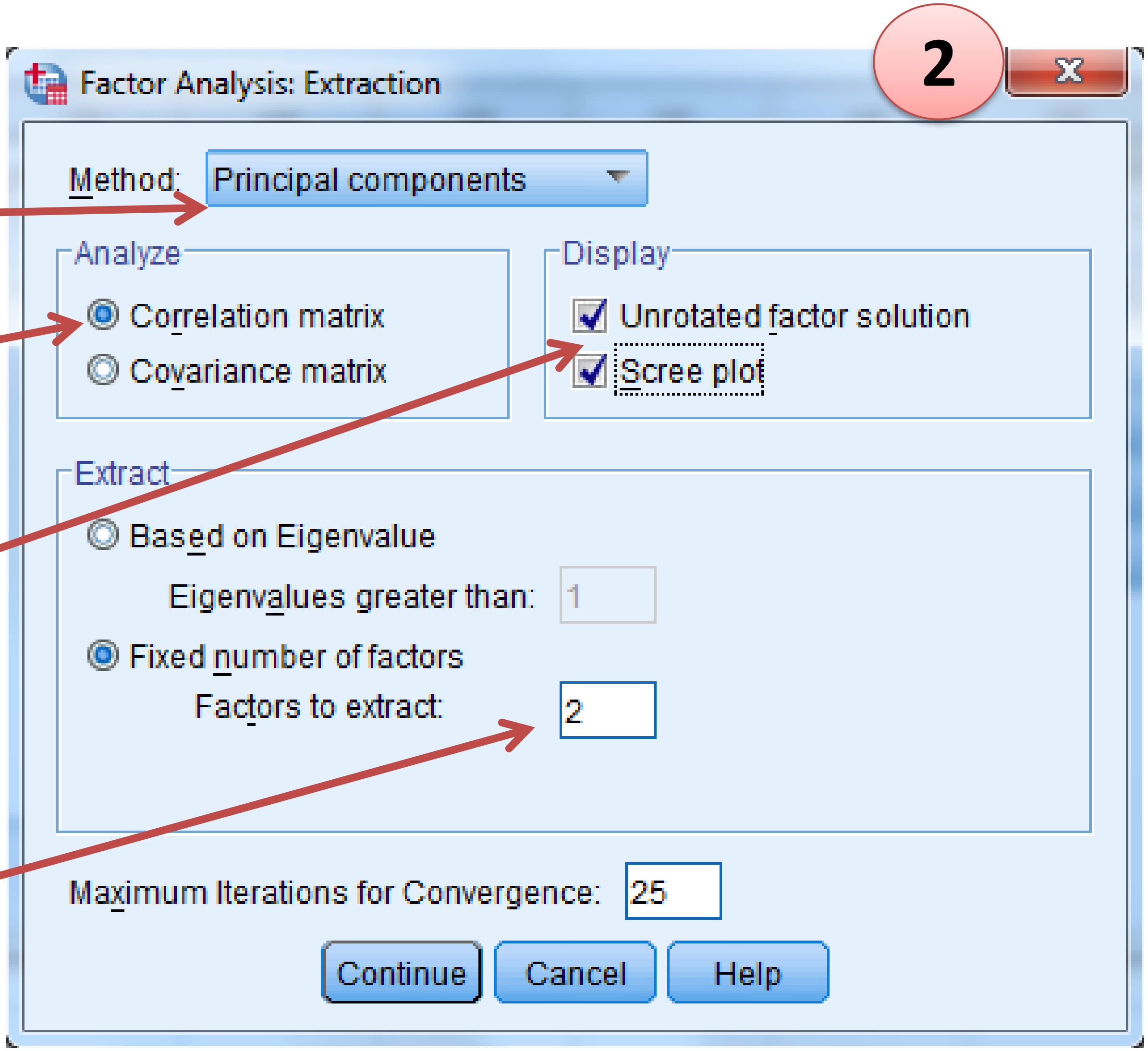
Extraction

Extraction method: Choose Principal components

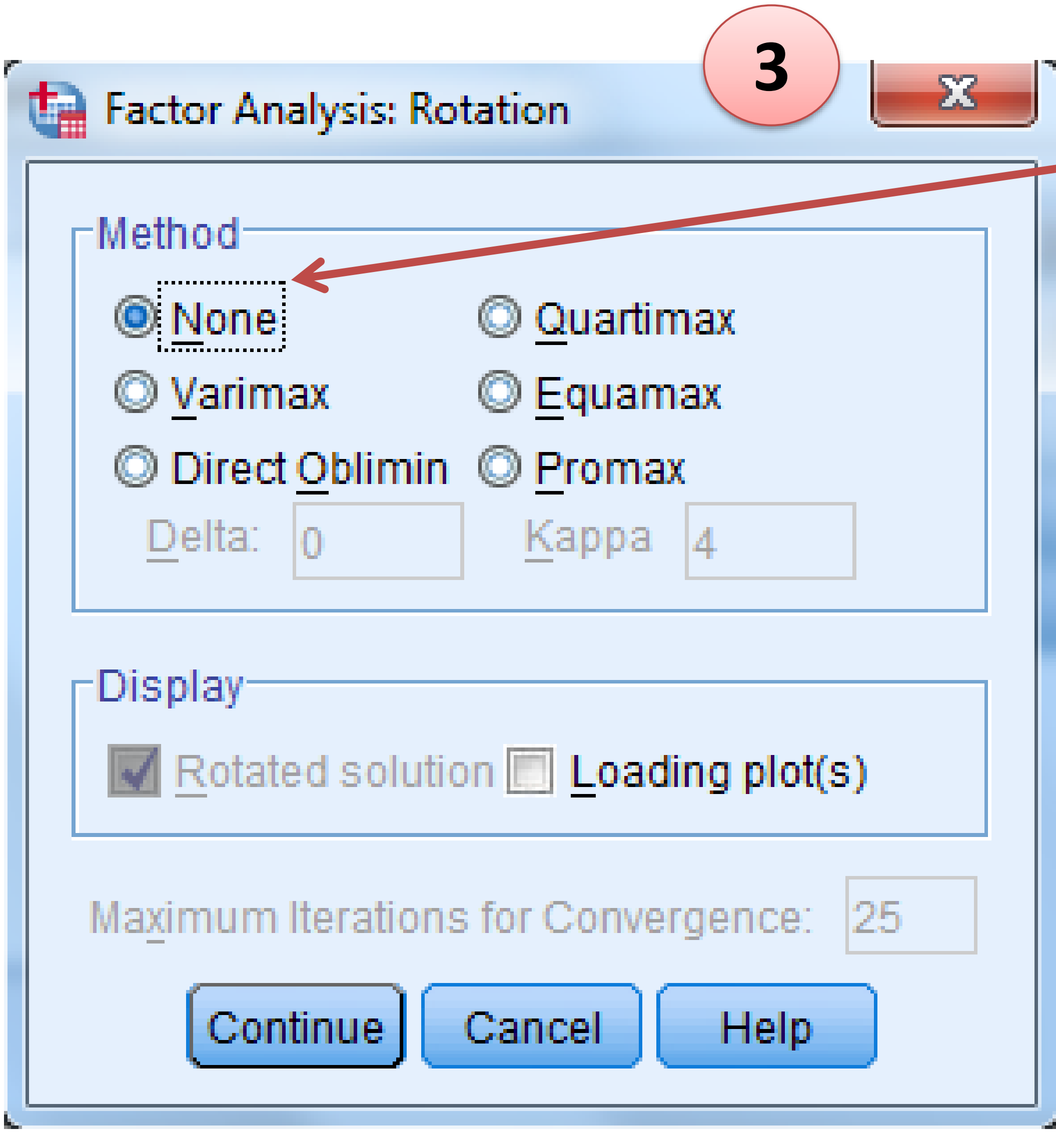
Use standardised data

Choose to display unrotated factor solution, and the scree plot

Number of factors to extract must be specified to be equal to the number of original variables.

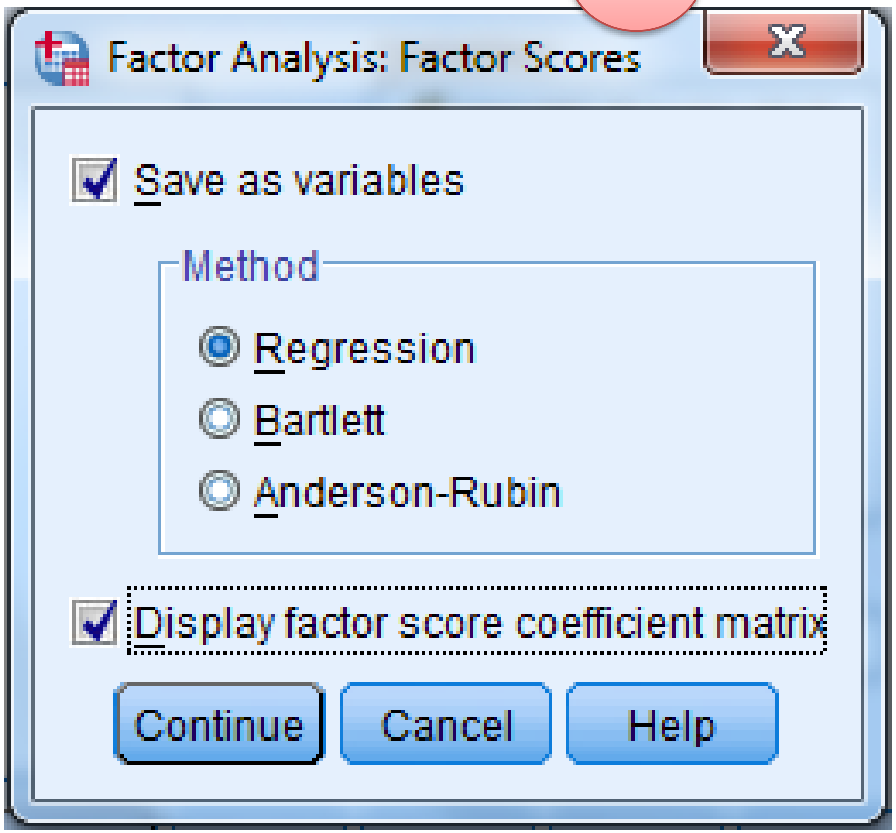


Rotation



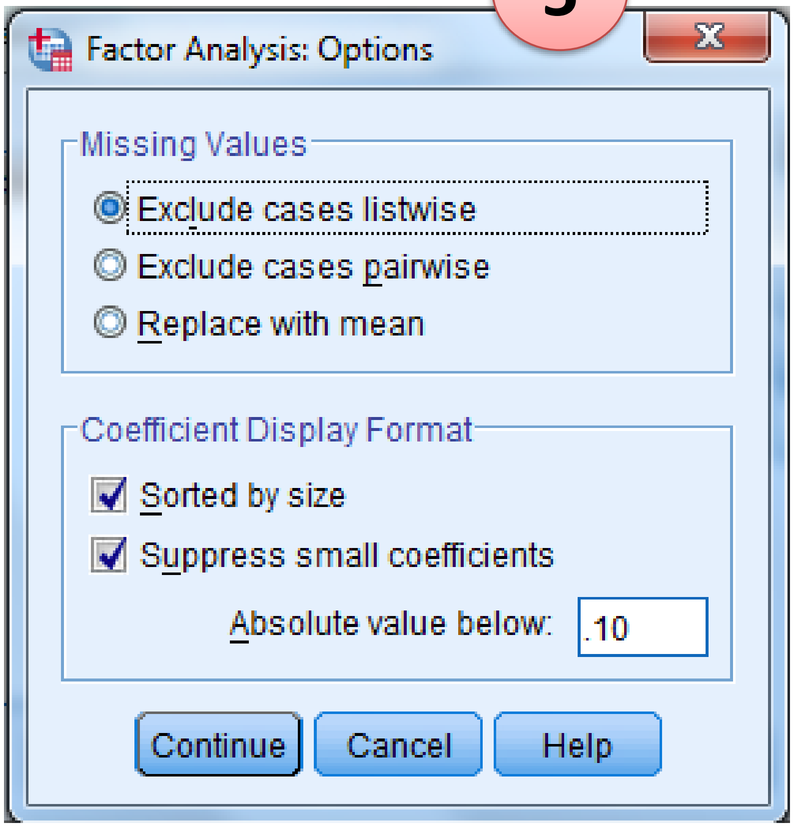
Rotation must be specified as None.

4



Factor Analysis: Factor Scores
Save as variables
Method
Regression
Bartlett
Anderson-Rubin
Display factor score coefficient matrix
Continue
Cancel
Help

5



Factor Analysis: Options
Missing Values
Exclude cases listwise
Exclude cases pairwise
Replace with mean
Coefficient Display Format
Sorted by size
Suppress small coefficients
Absolute value below:
.10
Continue
Cancel
Help

How much information/variance is explained by new variables?

Communalities

	Initial	Extraction
Height of male students, kg	1.000	1.000
Weight of male students, cm	1.000	1.000

Extraction Method: Principal Component Analysis.

100.00 % means that the total variance of the new variables is the same as the original variables.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.777	88.831	88.831	1.777	88.831	88.831
2	.223	11.169	100.000	.223	11.169	100.000

Extraction Method: Principal Component Analysis.

The first principal component accounts for 88.83%, thus if we only use PC1, we would be able to account for 88.83% of the variance of the original data.

The **eigenvalues** are the variances of the PCs, hence must be decreasing and must sum to 2. They are reported in the scree plot are the same as the variance accounted for by each new variables.

Stage 4 Calculate Principal Components continued...

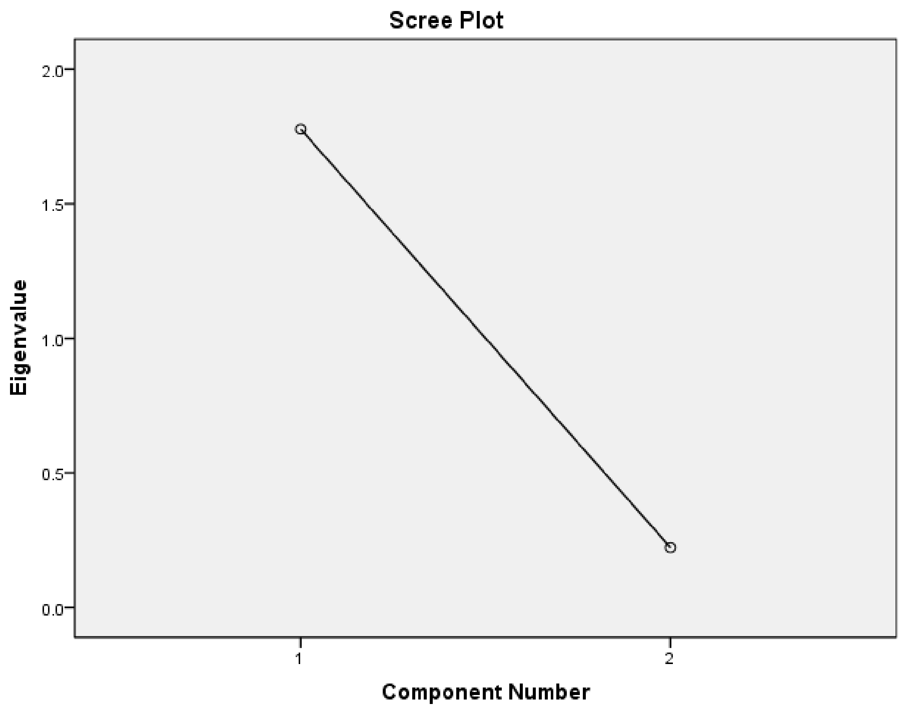
Decide on number of important principal components? Can we reduce dimensionality?

- The more PC we keep the more information about original variables we keep.
- Number of important PCs depends on our goal.
- If our goal is to reduce dimensionality,
  - Then we may want to keep first several PCs that explain certain amount of information (usually 60% or higher) contained in original variables.
  - Another rule is to keep all those PCs with eigenvalue >=1
  - Another rule is to keep those PCs at which the Scree Plot is sharply decreasing, drop those where Scree Plot flattens out.
- If our goal it to express the original variables via a new uncorrelated variables
  - Then we can keep all PCs or those who explain almost all information (Usually 90% or higher).

Example: Heights & Weights of Male Students

Number of Principal Components to retain?

Scree plot: it visualizes the variance/importance of the PCs



First eigenvalue>=1, so we can reduce dimensionality to 1 principal component (remember our goal was to reduce the dimensionality and not to express the original variables via uncorrelated variables).

Stage 5 Interpretation

Interpretation of PCs

- PC are rarely interpretable, which is a problem of PCA.
- Nevertheless we can try to see if we can do the interpretation.
- Often the PC1 can be interpreted as an average index of all original variables.

Example: Heights & Weights of Male Students

Component Matrix[a]

	Component	
	1	2
Weight of male students, cm	.943	.334
Height of male students, kg	.943	-.334

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Interpretation of PCs:

- The **Component Matrix** shows correlations between the original variables and the PCs.
- We see that both Height and Weight are highly positively correlated with PC1 suggesting that PC1 can be interpreted as an index of body size. Hence we could call PC1 as a “**body size**”.
- The variables have low correlations (< 0.4) with PC2 and hence it does not make sense to interpret the PC2. If the weights at PC2 were bigger, then PC2 could be interpreted as a measure of contrast between height and weight, or **obesity**.

Stage 6 Validation and Further Use of PCs

Validation of PCA

- How the PCA solution change if we remove outliers?

Further use of PCs in other analyses

- Use first few principal components to summarize data.
- If original variables are highly correlated then we can not use them for analyses that are sensitive to multicolinearity (such as cluster analysis or regression analysis). Then PCA can be done to express the original variables in new variables that are uncorrelated. For each object (student) the values at each principal component (i.e. the principal component scores) are calculated and these are used for the further analysis, e.g. for cluster analysis.

Example: Heights & Weights of Male Students

Further use of PCs

Component Score Coefficient Matrix

	Component	
	1	2
Height of male students, kg	.531	-1.496
Weight of male students, cm	.531	1.496

- Use the mean and standard deviation from Height and Weight i.e. Height (178.8889, 7.44303), Weight (71.4444, 13.0349).
- Hence, for the first student with height=163cm, and weight=50kg, the score at first principal component is

  0.531x(163-178.8889)/7.44303+0.531x(50-71.4444)/13.03409=-2.007

- and the score at the second principal component is

  -1.496x(163-178.8889)/7.44303+1.496x(50-71.4444)/13.03409=0.732.

Example: Heights & Weights of Male Students

Further use of PCs

Component Score Covariance Matrix

Component	1	2
1	1.000	.000
2	.000	1.000

Correlation matrix of PCs.

- Because principal components are orthogonal (independent).
- It should contain 1's on diagonal and 0'f off diagonal.
- No multicollinearity problem any more.
- Principal components can be used instead of original data in further analysis.

The principal component scores for each student.

Notice that the **first component scores** roughly increase with increasing height and weight. It gives a way of comparing students with respect to their overall body size.

Height	Weight	FAC1_19	FAC2_19
163.00	50.00	-2.00525	.73245
170.00	54.00	-1.34352	-.21548
170.00	56.00	-1.26213	.01407
173.00	70.00	-.47852	1.01791
175.00	80.00	.07102	1.76366
177.00	60.00	-.60041	-.93387
178.00	67.00	-.24424	-.33144
178.00	70.00	-.12214	.01288
178.00	72.00	-.04075	.24244
178.00	79.00	.24415	1.04586
180.00	71.00	.06111	-.27435
180.00	75.00	.22390	.18475
182.00	88.00	.89554	1.27482
183.00	58.00	-.25415	-2.36945
183.00	70.00	.23423	-.99214
188.00	76.00	.83480	-1.30851
191.00	95.00	1.82190	.26921
193.00	95.00	1.96445	-.13280

The second component scores are low and negative for slim students, and large positive for not so slim students.

Example: Heights & Weights of Male Students

Use of PC in further analysis

For the purpose of illustration: Assume the aim was not to reduce dimensionality, but rather to create uncorrelated new variables that contain same information about data, and to use them in cluster analysis.

In such case we should

- Keep both PC1 and PC2,
- Calculate principal scores for each student and for each PC1 and PC2
- Do cluster analysis on the principal scores, store the cluster memberships.
- Do the interpretation of the cluster solution via original variables.

Here, we used hierarchical clustering with Squared Euclidean distance, Ward method, Standardized variables… see next slide.

Example: Heights & Weights of Male Students

PC2 scores

PC1 scores

Clusters when original (standardized) variables were used

Clusters when PC1 and PC2 were used

	Height	Weight	FAC1_19	FAC2_19	CLU2_2	CLU2_3
1	163.00	50.00	-2.00525	.73245	1	1
2	170.00	54.00	-1.34352	-.21548	1	1
3	170.00	56.00	-1.26213	.01407	1	1
4	173.00	70.00	-.47852	1.01791	1	2
5	175.00	80.00	.07102	1.76366	1	2
6	177.00	60.00	-.60041	-.93387	1	1
7	178.00	67.00	-.24424	-.33144	1	1
8	178.00	70.00	-.12214	.01288	1	1
9	178.00	72.00	-.04075	.24244	1	1
10	178.00	79.00	.24415	1.04586	1	2
11	180.00	71.00	.06111	-.27435	1	1
12	180.00	75.00	.22390	.18475	1	1
13	182.00	88.00	.89554	1.27482	2	2
14	183.00	58.00	-.25415	-2.36945	1	1
15	183.00	70.00	.23423	-.99214	1	1
16	188.00	76.00	.83480	-1.30851	2	1
17	191.00	95.00	1.82190	.26921	2	2
18	193.00	95.00	1.96445	-.13280	2	2
19						

Beijing Olympics 2008 Decathlon Result

Name	100m	long	shot	high	400m	hurd.	disc.	pole	jave.	1500m
Clay	10.44	7.78	16.27	1.99	48.92	13.93	53.79	5.00	70.97	05:06.6
Krauchanka	10.96	7.61	14.39	2.11	47.30	14.21	44.58	5.00	60.23	04:27.5
Suarez	10.90	7.33	14.49	2.05	47.91	14.15	44.45	4.70	73.98	04:29.2
Pogorelov	11.07	7.37	16.53	2.08	50.91	14.47	50.04	5.00	64.01	05:01.6
Barras	11.26	7.08	15.42	1.96	49.51	14.21	45.17	5.00	65.40	04:29.3
Sebrle	11.21	7.68	14.78	2.11	49.54	14.71	45.50	4.80	63.93	04:49.6
Kasyanov	10.53	7.56	15.15	1.96	47.70	14.37	48.39	4.30	51.59	04:28.9
Niklaus	11.12	7.29	13.23	2.05	49.65	14.37	45.39	5.20	60.21	04:32.9
Smith	10.85	7.04	15.09	1.99	47.96	14.08	50.91	4.60	51.52	04:31.6
Schrader	10.80	7.70	13.67	1.99	48.47	14.71	40.41	4.80	60.27	04:26.8
Pahapill	11.15	7.04	14.36	2.11	50.90	14.51	49.35	4.80	67.07	04:47.0
Drozdov	11.02	7.23	16.26	2.02	51.56	15.51	47.43	5.10	62.57	04:41.3
Raja	10.89	7.29	14.79	1.96	48.98	14.06	39.83	4.80	67.16	04:49.6
Martineau	11.19	7.19	13.78	1.99	49.99	14.73	44.09	4.70	71.44	04:38.0
Garcia	10.64	7.07	15.82	1.96	49.66	13.90	36.73	4.70	65.60	05:00.5
Shubianok	11.31	6.86	14.88	1.99	50.02	14.52	45.80	4.60	62.10	04:38.1
Parkhomenka	11.29	6.99	15.49	1.93	50.71	15.06	45.27	4.70	64.60	04:45.2
Qi	11.15	7.22	13.40	1.93	49.39	14.60	46.46	4.30	63.09	04:39.3
Bertocchi	11.00	7.05	14.10	1.90	48.72	14.32	44.91	4.70	45.33	04:42.3
Addy	10.76	7.38	14.91	1.93	48.51	14.31	42.30	4.20	52.50	05:12.2
Awde	11.06	7.12	12.03	1.78	47.16	14.69	37.12	4.90	53.18	04:44.8
Sepehrzad	10.92	6.80	16.02	1.90	50.75	14.64	50.32	4.00	49.56	05:06.7
Sitar	11.21	7.25	12.41	2.05	50.10	15.03	39.25	4.00	47.23	04:37.4
Dizdarevic	11.16	7.02	13.97	1.96	52.02	15.61	39.86	4.00	43.58	04:51.4

Decathlon example

- Various time and measured events were recorded for 24 top world male athletes in Beijing Olympics 2008.
- Aim: form a measure(s) which can help us to rank the athletes. – Dimension reduction problem.

PCA on mean-corrected data

	x100m	long	shot	high	x400m	hurdles	discus	pole	javelin	x1500m
Mean	11.00	7.25	14.64	1.99	49.43	14.53	44.89	4.66	59.88	284.08
Variance	0.06	0.07	1.39	0.01	1.74	0.19	19.63	0.13	73.21	182.02
Std dev.	0.24	0.27	1.18	0.08	1.32	0.44	4.43	0.35	8.56	13.49

- Variance of the original variables vary a lot, range from 0.01 to 182

- Which results that the first principal component loads heavily on to the long distance event 1500m.

Covariance Matrix

	x100m	long	shot	high	x400m	hurdles	discus	pole	javelin	x1500m
x100m	0.0556	-0.0299	-0.0953	0.0025	0.1453	0.0544	-0.1538	0.0013	-0.0004	-0.8296
long	-0.0299	0.0705	0.0042	0.0082	-0.1511	-0.0258	0.1059	0.0290	0.6015	-0.3144
shot	-0.0953	0.0042	1.3918	0.0159	0.4597	-0.1084	2.9388	0.0734	2.9899	6.8716
high	0.0025	0.0082	0.0159	0.0060	0.0206	-0.0013	0.1039	0.0079	0.2498	-0.2140
x400m	0.1453	-0.1511	0.4597	0.0206	1.7434	0.3640	0.7502	-0.0760	0.0672	6.2172
hurdles	0.0544	-0.0258	-0.1084	-0.0013	0.3640	0.1949	-0.3387	-0.0457	-1.2729	-0.2791
discus	-0.1538	0.1059	2.9388	0.1039	0.7502	-0.3387	19.6335	0.2484	6.9532	5.5106
pole	0.0013	0.0290	0.0734	0.0079	-0.0760	-0.0457	0.2484	0.1251	1.8237	-1.0784
javelin	-0.0004	0.6015	2.9899	0.2498	0.0672	-1.2729	6.9532	1.8237	73.2083	-5.8124
x1500m	-0.8296	-0.3144	6.8716	-0.2140	6.2172	-0.2791	5.5106	-1.0784	-5.8124	182.0246

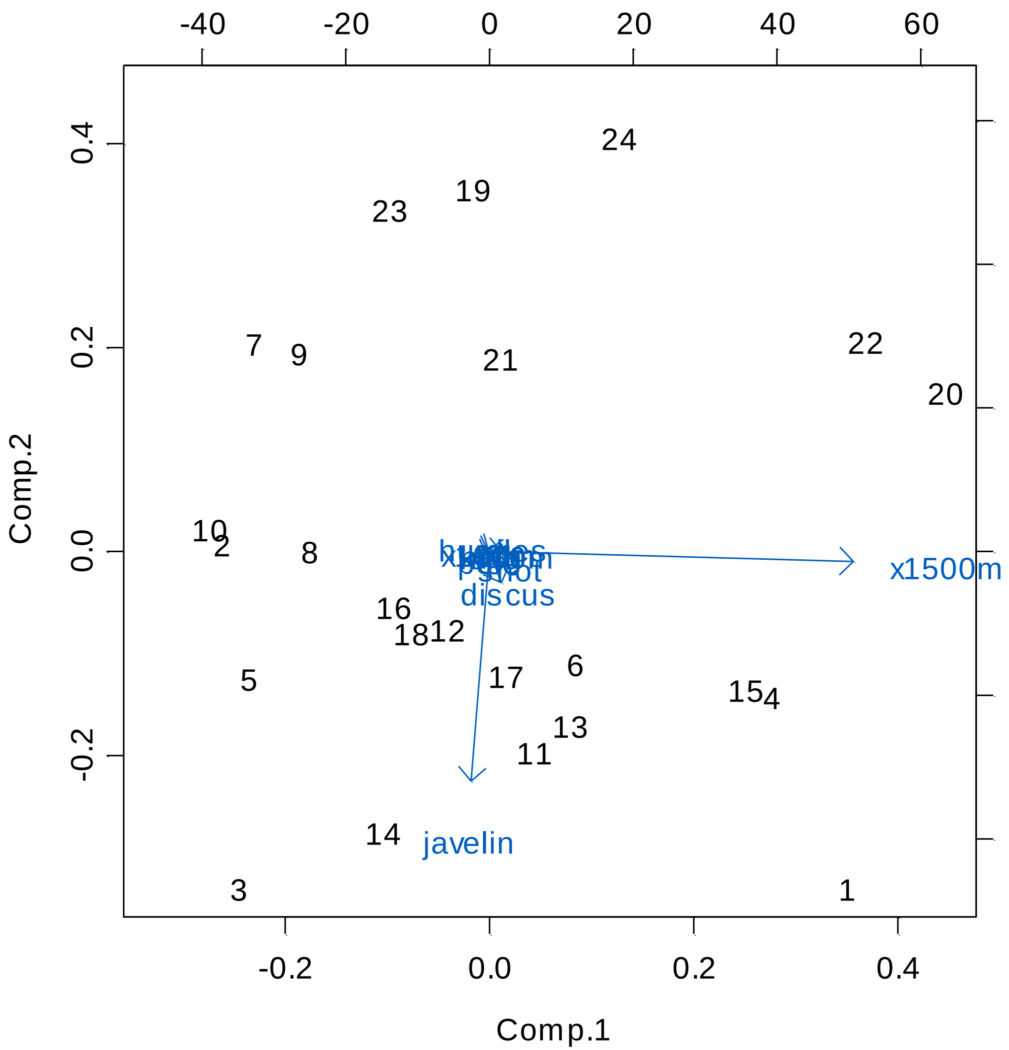
Eigenvectors

Component Score Coefficient Matrixa

	Component									
	1	2	3	4	5	6	7	8	9	10
x100m	.000	.000	.000	.020	-.031	-.097	.150	-.520	1.593	-.777
long	.000	.000	.000	-.019	.000	.568	-.132	.781	.869	-1.536
shot	.003	.007	.034	.109	1.412	.146	-.232	-.507	.733	-.174
high	.000	.000	.000	.001	.000	.012	.003	.067	.161	1.706
x400m	.003	.001	.009	.970	-.142	-.443	.438	1.701	-.360	-1.580
hurd	.000	-.001	-.001	.085	-.059	1.036	-.540	-1.214	-.594	1.233
disc	.011	.069	1.002	-.145	-.705	.009	-.050	-.167	-.241	-.114
pole	.000	.001	.000	-.008	.020	.447	1.211	-.107	-.432	.227
jave	-.031	.982	-.271	.008	-.349	-.104	-.794	-.433	-.390	.164
x1500m	.994	.068	-.138	-.369	-.521	.262	.189	-.516	.210	.765

- The loading of x1500m on Comp.1 is 0.994, which means that the first principal component is dominant by the long distance event 1500m.
- The loading of javelin on Comp.2 is 0.982, javelin loads heavily on to comp.2.

Biplot of Comp.1 vs. Comp.2
Not available in SPSS



- •Dominant by 1500m and javelin.
- •Athlete No.3 has very low scores on both comp.1 and comp.2, he is very good on javelin, and very fast at 1500m.

PCA on standardised data

Importance of components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Standard deviation	1.68	1.45	1.36	1.00	0.91	0.72	0.60	0.57	0.40	0.25
Variance	2.83	2.09	1.84	1.00	0.82	0.52	0.35	0.33	0.16	0.06
Proportion of variance	0.28	0.21	0.18	0.10	0.08	0.05	0.04	0.03	0.02	0.01
Cumulative proportion of variance	0.28	0.49	0.68	0.78	0.86	0.91	0.94	0.98	0.99	1.00

- Variances are much more homogeneous

Correlation Matrix

	x100m	long	shot	high	x400m	hurdles	discus	pole	javelin	x1500m
x100m	1.000	-0.478	-0.343	0.135	0.466	0.522	-0.147	0.015	0.000	-0.261
long	-0.478	1.000	0.013	0.399	-0.431	-0.220	0.090	0.308	0.265	-0.088
shot	-0.343	0.013	1.000	0.174	0.295	-0.208	0.562	0.176	0.296	0.432
high	0.135	0.399	0.174	1.000	0.201	-0.038	0.302	0.286	0.376	-0.204
x400m	0.466	-0.431	0.295	0.201	1.000	0.625	0.128	-0.163	0.006	0.349
hurdles	0.522	-0.220	-0.208	-0.038	0.625	1.000	-0.173	-0.293	-0.337	-0.047
discus	-0.147	0.090	0.562	0.302	0.128	-0.173	1.000	0.159	0.183	0.092
pole	0.015	0.308	0.176	0.286	-0.163	-0.293	0.159	1.000	0.603	-0.226
javelin	0.000	0.265	0.296	0.376	0.006	-0.337	0.183	0.603	1.000	-0.050
x1500m	-0.261	-0.088	0.432	-0.204	0.349	-0.047	0.092	-0.226	-0.050	1.000

Eigenvectors

Component Score Coefficient Matrix

	Component									
	1	2	3	4	5	6	7	8	9	10
x100m	-.202	.131	.353	-.213	-.098	-.195	.633	.036	1.523	-.668
long	.232	-.122	.062	.564	.390	.320	.287	.291	.692	-1.665
shot	.159	.324	-.225	-.070	-.090	.212	-.821	-.361	1.365	-.024
high	.146	.184	.289	.463	.010	-.719	-.067	-.597	-.208	1.290
x400m	-.162	.398	.044	.050	.220	-.009	-.201	-.091	-1.050	-2.341
hurd	-.253	.144	.145	.336	.197	.753	-.215	.346	.052	1.931
disc	.154	.260	-.074	.167	-.727	.248	.690	.543	-.369	.083
pole	.219	.041	.256	-.369	.143	.706	.383	-.854	-.463	.235
jave	.224	.137	.203	-.370	.334	-.218	-.294	1.158	-.157	.556
x1500m	-.011	.200	-.382	-.057	.548	-.203	.929	-.101	.139	.999

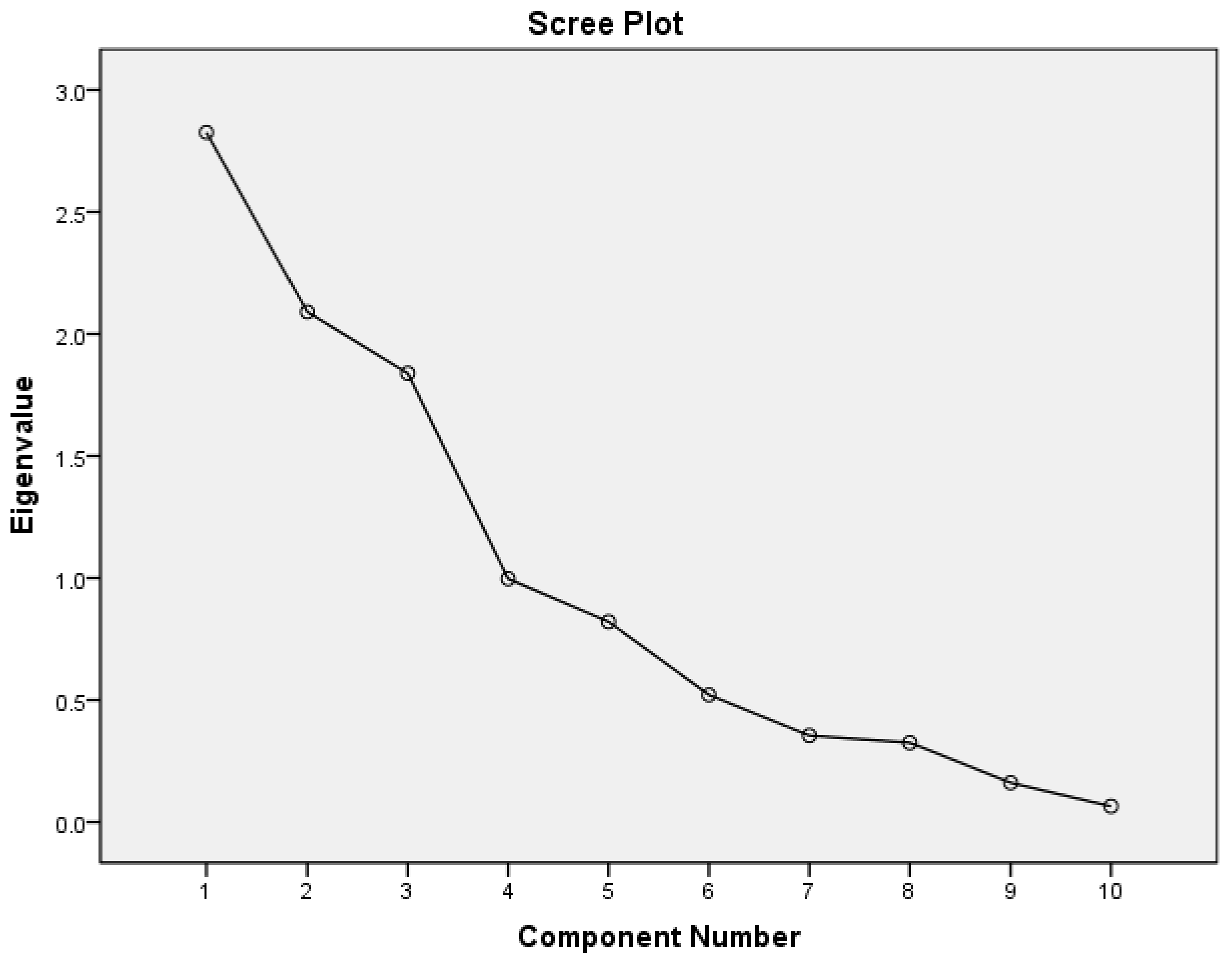
Number of Principal Components to Extract

- In the case of standardized data, retain only those components whose eigenvalues (variance) are greater than one.

- Plot the percent of variance accounted for by each principal component and look for an elbow. The plot is referred to as the scree plot. This rule can be used for both mean-corrected and standardized data.

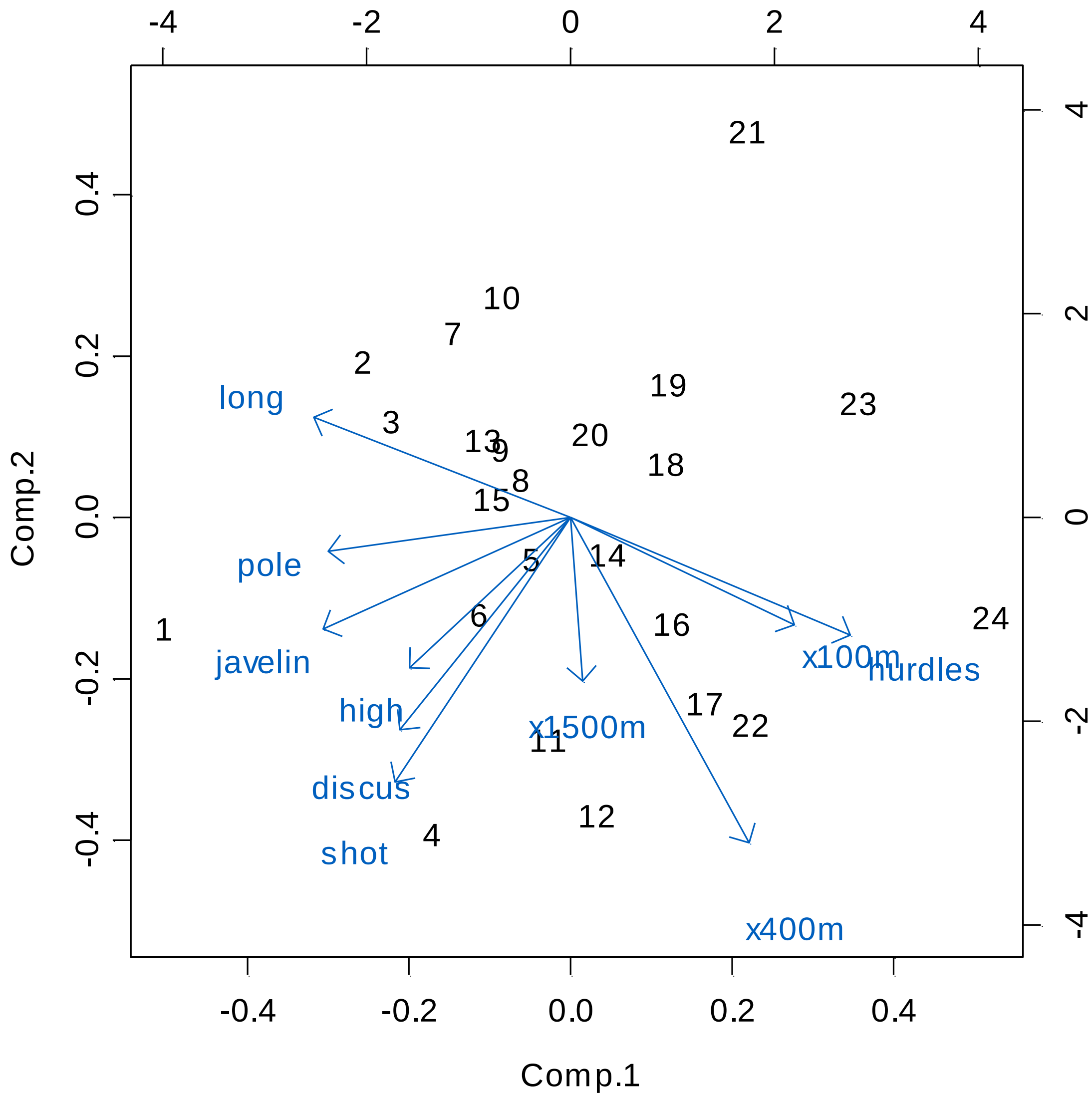
- Retain only those components that are statistically significant. (Not introduced here)

Scree plot



Scree Plot
Eigenvalue
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8
9
10
Component Number

Biplot of comp.1 vs. comp.2
Not available in SPSS



- No apparent outliers
- All events are nicely represented.

Interpret Principal Components

- Since the principal components are linear combinations of the original data, it is often necessary to interpret or provide a meaning to the linear combination.
- One can use the loadings (Eigenvectors, Component score coefficient) for interpreting the principal components. The higher the loading of a variable, the more influence it has in the formation of the principal component score and vice verse. Normally use 0.5 as the cut-off point.
- Or use component Matrix which shows correlations between the original variables and the PCs.
- In many instances the retained principal components cannot be meaningfully interpreted. In such cases researchers typically resorted to a rotation of the principal components – factor analysis.

Eigenvectors

Component Score Coefficient Matrix

	Component									
	1	2	3	4	5	6	7	8	9	10
x100m	-.202	.131	.353	-.213	-.098	-.195	.633	.036	1.523	-.668
long	.232	-.122	.062	.564	.390	.320	.287	.291	.692	-1.665
shot	.159	.324	-.225	-.070	-.090	.212	-.821	-.361	1.365	-.024
high	.146	.184	.289	.463	.010	-.719	-.067	-.597	-.208	1.290
x400m	-.162	.398	.044	.050	.220	-.009	-.201	-.091	-1.050	-2.341
hurd	-.253	.144	.145	.336	.197	.753	-.215	.346	.052	1.931
disc	.154	.260	-.074	.167	-.727	.248	.690	.543	-.369	.083
pole	.219	.041	.256	-.369	.143	.706	.383	-.854	-.463	.235
jave	.224	.137	.203	-.370	.334	-.218	-.294	1.158	-.157	.556
x1500m	-.011	.200	-.382	-.057	.548	-.203	.929	-.101	.139	.999

- PC1 is naturally interpreted as a measure of overall ‘athletic ability’. Coefficients are positive for measured events, negative for timed events.
- Remaining principal components are harder to interpret, and may just represent noise.

Is PCA the Appropriate Technique

- Depends on the objectives of the study.

- If the principal components cannot be interpreted then their subsequent use in other statistical techniques may not be very meaningful.

- If the objective is for data reduction, PCA should only be performed if the data can be represented by a fewer numbers of principal components without a substantial loss of information.

- PCA is most appropriate if the variables are interrelated, only then it is possible to reduce to a fewer new variables without much lose of information.

- Formal statistical tests are available for determining if the variables are significantly correlated among themselves, but very sensitive to sample size, not practical. (e.g. Bartlett's test for standardized data)

- In practice, researchers have used their own judgment.

Use of Principal Components Scores

- Cluster analysis
- Regression analysis
- Discriminant analysis
- Solved multicollinearity problem
- A new problem can arise due to the inability to meaningfully interpret the principal components.
- Factor analysis is sometimes preferred compare to PCA.

Computer lab session (using SPSS)

Practice example 1: Height and Weight of male students

Practice example 2: Beijing Olympcis 2008 Decathlon data

- Download “HeightAndWeight” and “Decathlon2008” data files from my.wbs ADA module page.
- These are examples used in this lecture.
- Carry out PCA follow the steps from the lecture notes.

Computer lab session (using SPSS)

Practice Example 3: Employment sectors around the globe

Download the file EmploymentSector.sav from my.wbs The data represent the employment sector profile in 15 countries around the world.

- Variables:
  - Country: Country name

  % of working population employed in each employment sector
  - AGR: Agriculture
  - MIN: Mining
  - MAN: Manufacturing
  - SPS: Social & personal services
  - TC: Transport & Communications
  - CON: Construction
  - SER: Service industries
  - FIN: Finance
  - PS: Power supplies

AIM: find out if you can reduce the dimensionality of the data.

Country	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
Belgium	3.30	0.90	27.60	0.90	8.20	19.10	6.20	26.60	7.20
Denmark	9.20	0.10	21.80	0.60	8.30	14.60	6.50	32.20	7.10
France	10.80	0.80	27.50	0.90	8.90	16.80	6.00	22.60	5.70
Ireland	23.20	1.00	20.70	1.30	7.50	16.80	2.80	20.80	6.10
Italy	15.90	0.60	27.60	0.50	10.00	18.10	1.60	20.10	5.70
Luxembourg	7.70	3.10	30.80	0.80	9.20	18.50	4.60	19.20	6.20
Netherlands	6.30	0.10	22.50	1.00	9.90	18.00	6.80	28.50	6.80
UK	2.70	1.40	30.20	1.40	6.90	16.90	5.70	28.30	6.40
Austria	12.70	1.10	30.20	1.40	9.00	16.80	4.90	16.80	7.00
Portugal	27.80	0.30	24.50	0.60	8.40	13.30	2.70	16.70	5.70
Greece	41.40	0.60	17.60	0.60	8.10	11.50	2.40	11.00	6.70
Spain	22.90	0.80	28.50	0.70	11.50	9.70	8.50	11.80	5.50
Turkey	66.80	0.70	7.90	0.10	2.80	5.20	1.10	11.90	3.20
Bulgaria	23.60	1.90	32.30	0.60	7.90	8.00	0.70	18.20	6.70
Poland	31.10	2.50	25.70	0.90	8.40	7.50	0.90	16.10	6.90

Source: Extract from Euromonitor (1979) in Manly (1994, 2nd Ed.)