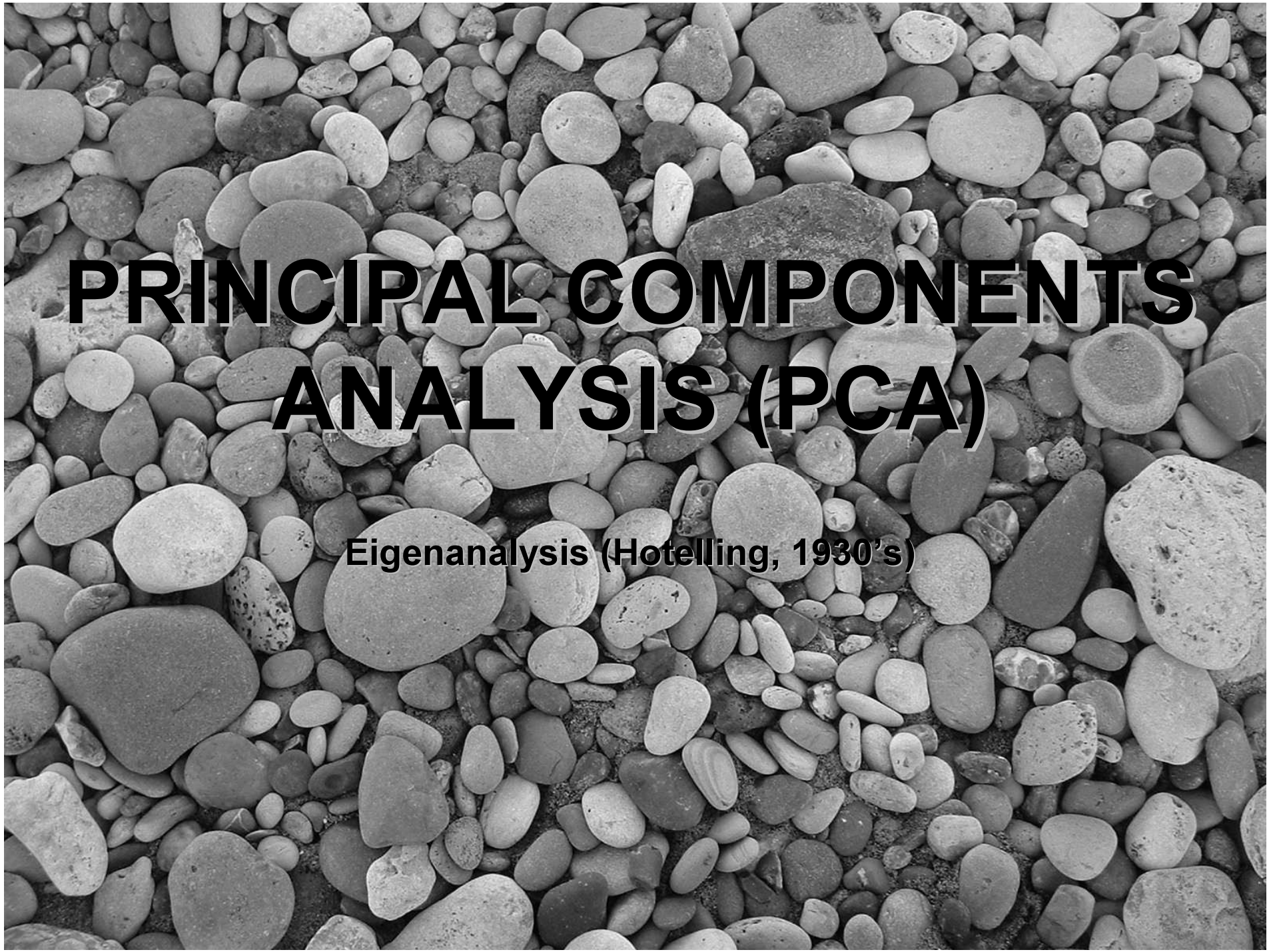

PRINCIPAL COMPONENTS ANALYSIS (PCA)
Eigenanalysis (Hotelling, 1930's)

What is it?

- PCA allows to explore the relations between multiple variables at the same time and to extract the fundamental structure of the data cloud;
- Reduces the of data set by finding new set of variables (Factor Axes = Principal Components), smaller than the original set of variables, that nonetheless retains most of the sample's information. By information I mean the variation present in the sample given by the correlations between original variables;
- These Principal Components (PCs) are uncorrelated and are ordered by the fraction of the total information each retains. The first few may represent some well separated fundamental underlying causes;
- Very useful also for screening and classification (using only main PC).


Follow the guide but… DON’T
ask too many question!!

Basic statistics

- Mean deviation about mean = mean deviation:

$$| X_i - \bar{X} |$$

- Corrected sum of squares (CSS): $\sum (X_i - \bar{X})^2$

- Sample variance (s^2): $\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ Units!!

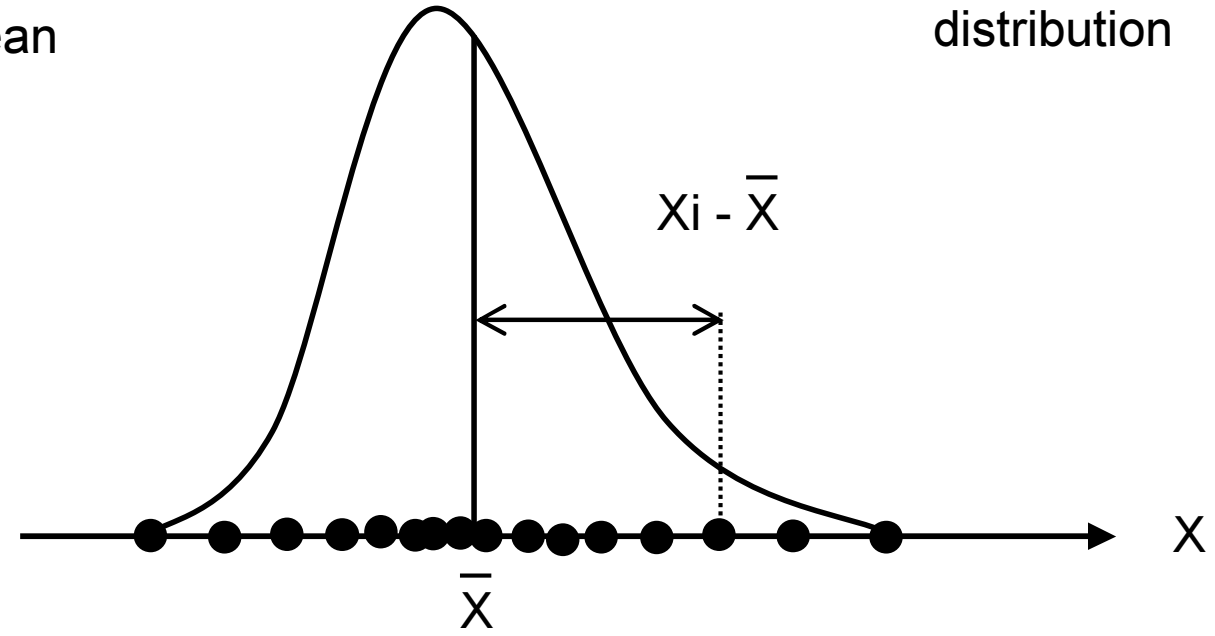
= estimate of the population variance ($\sigma 2$). n tends to underestimate $\sigma 2$ and n-1 is better. Units are square of sample units.

The square root of the variance is in the same unit as the sample and is easier to interpret =

- Standard deviation (s): $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

‘standard’ (same unit as original measurement) measure of the dispersion around the mean

Normal frequency distribution



Covariance and correlation between 2 variables

Based on the corrected sum of products (CSP) $\sum_{i=1}^{i=n} (X_i - \overline{X})\ (Y_i - \overline{Y})$

Value is fct of sample size so better to standardize:

N tends to underestimate the population value of cov because we know only sample mean. With n-1 the sample mean is closer to the population mean.

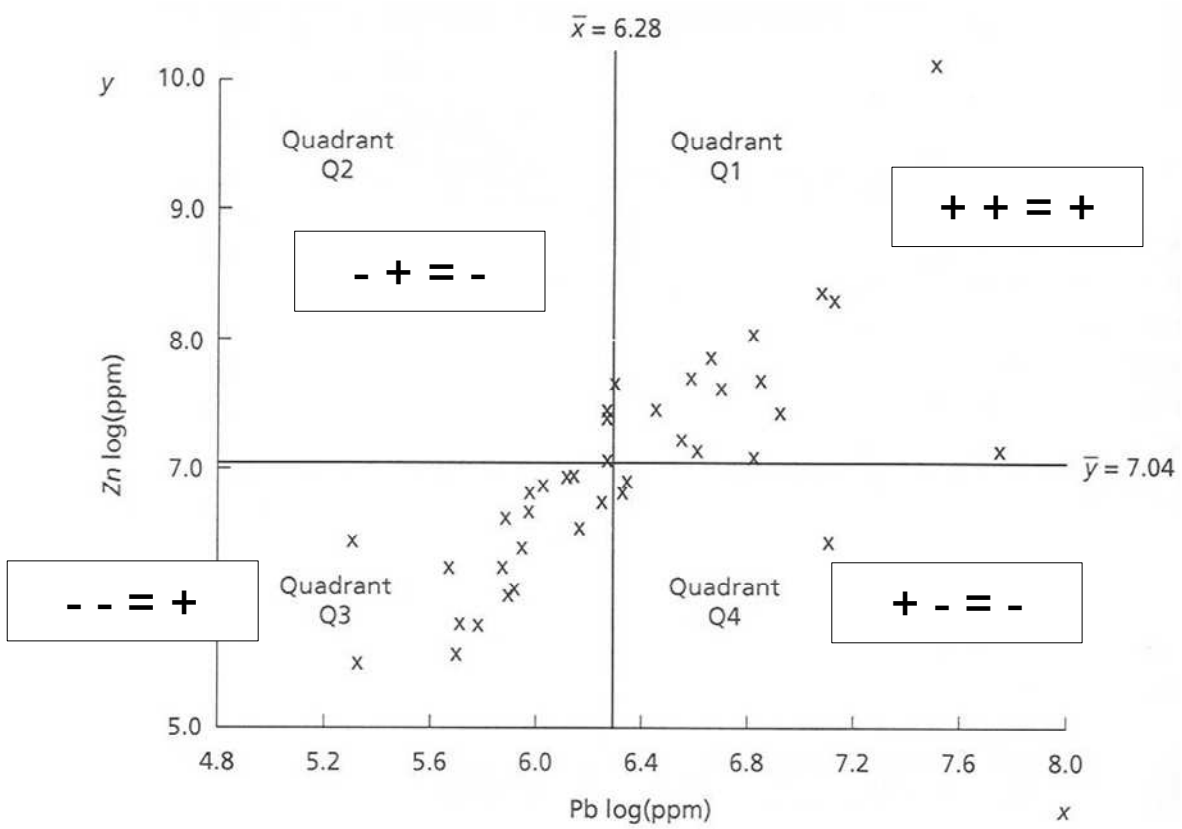
$$\text{Covariance}_{XY} = \frac{\text{CSP}}{n - 1}$$

!NB: $Cov_{xx} = var_x$

Still fct of the units of the variables: cov between L and W of a brachiopod will give a higher value if L and W are measured in mm than in cm. This effect is removed by standardization with s.

$$\text{Pearson's coefficient of correlation} = \sum_{i=1}^{i=n} \frac{(X_i - \overline{X})}{s_x} \frac{(Y_i - \overline{Y})}{s_y} \frac{1}{n - 1} = cov_{XY} / s_x s_y = r$$

Pearson's product-moment correlation coefficient, or just coefficient of correlation, is a dimensionless measure of correlation. It ranges between -1 (straight line, negative slope) to +1 (straight line, positive slope), both extremes being complete correlation. VERY sensitive to outliers! Significance is fct of sample size and can be estimated by statistical tests.



CSP has a value related to the distribution of points among quadrants around means. Here points mainly in Q1 and Q3 so CSP will be large and +. If Q2 and Q4 large -. If evenly distributed in 4 Q values will cancel out and CSP close to 0.

The magnitude of the standard deviation is, however, related to the **magnitude** of the measurements so it is difficult to assess it meaning by itself.

Regardless of the magnitude and of the shape of the distribution however:

- **75%** at least of the observations will lie within **2s** from the mean and 88.89% within 3s;
- These values rise to **95.46%** and 99.73% respectively if the observations follow a normal distribution.

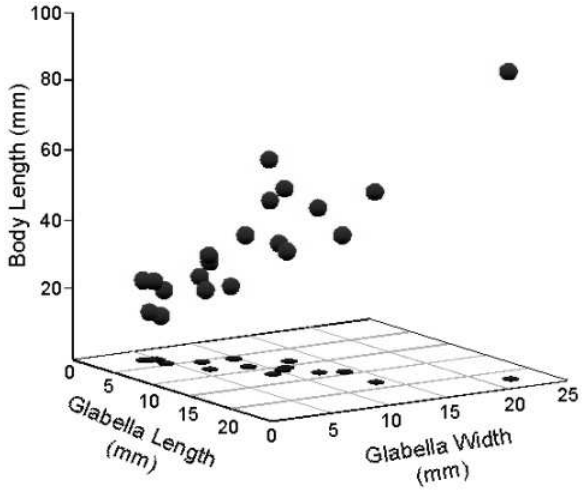
The standard deviation is used to standardize the data prior to anlysis in some cases (see later).

Concept

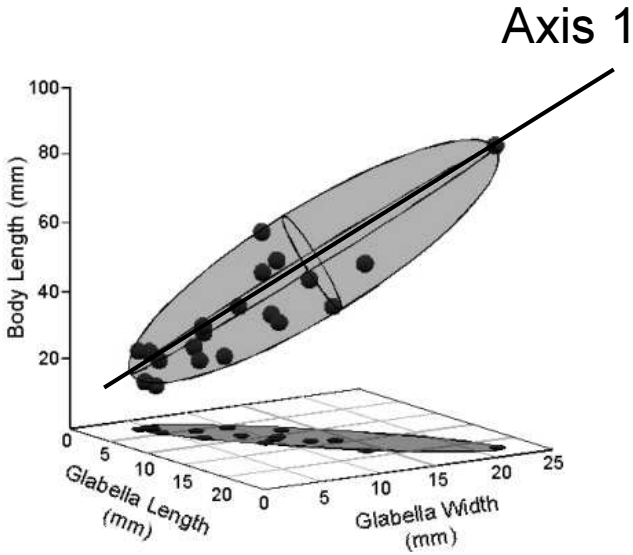
The data in A. are clearly correlated to some degree and therefore the three variables (axes) are redundant.

The reference axes can be rotated so that axis 1 now explains most of the variance. We could then use axis 1 (new variable) only to describe our data and think of an interpretation for the elongation of the data scatter along it = reduction of dimensionality.

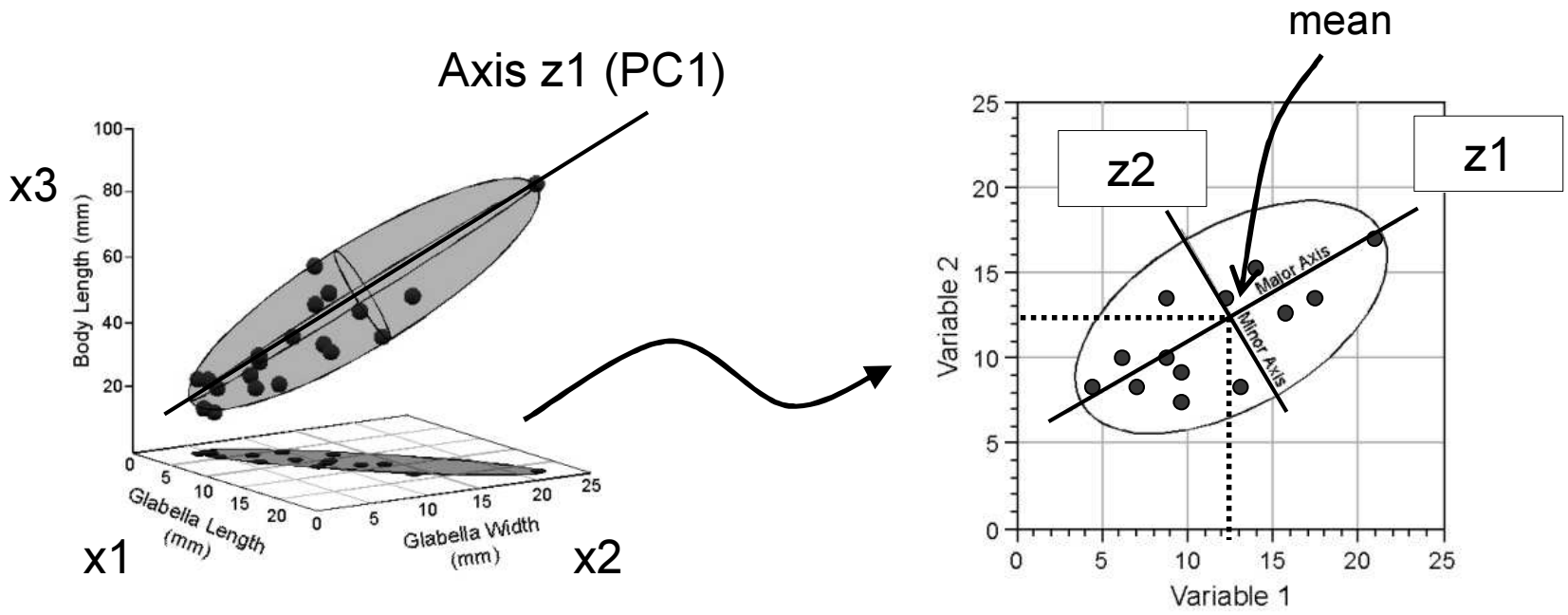
A.



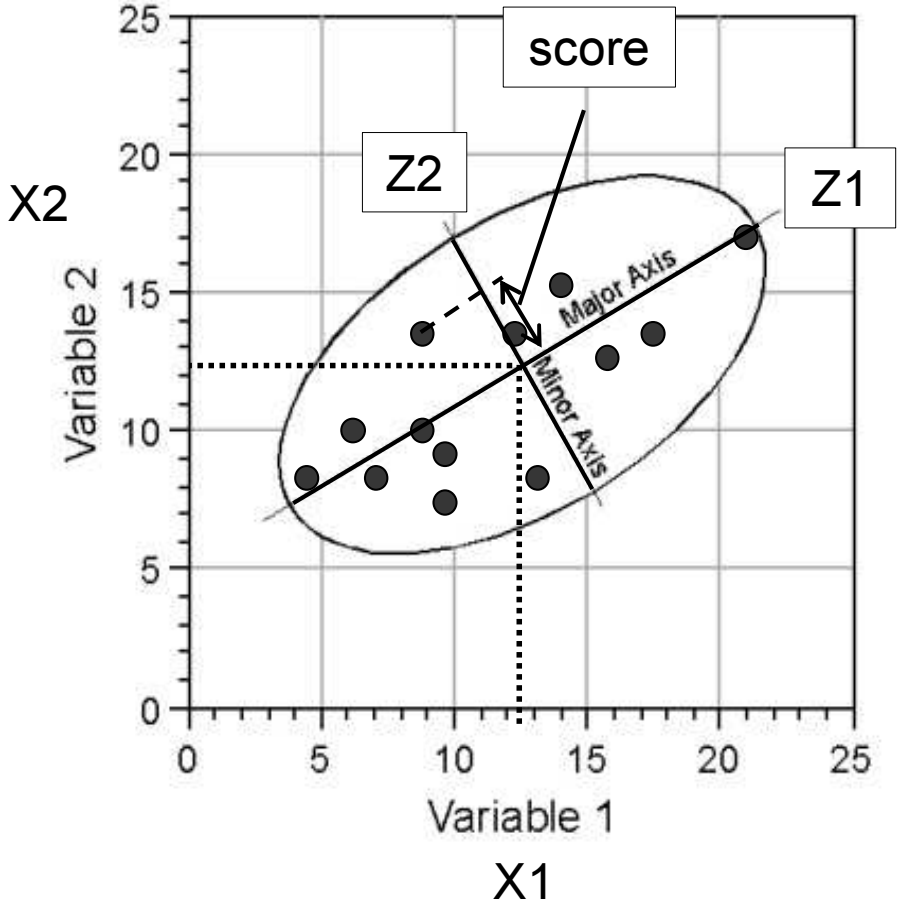
B.



Axis z1, the new variable is Principal Component 1 and is just a linear combinations of the original variables x1, x2, x3. The coordinates of our data points along the new axis can be found by projection.



The second axis explains most of the variance which is <u>not</u> accounted for by PC1 (we want that!), it is therefore perpendicular to PC1 and PC1 (Z1) and PC2 (Z2) are independent, uncorrelated.



$$Z1 \sim \mathbf{a}_1^T\mathbf{X} = a_{11}X1 + a_{21}X2$$

$$Z2 \sim \mathbf{a}_2^T\mathbf{X} = a_{12}X1 + a_{22}X2$$

= coordinates of original data points along new axes (Principal Components)

They are called PC scores and are the new data. The results of a PCA are often displayed as scatter plots of the scores along the main PC's.

$$z_1 \equiv \mathbf{a}_1^T\mathbf{x} = \sum_{i=1}^{p} a_{i1}x_i$$

T is for transposed (row vector)

In the generalized forms we have p variables X1, X2 … Xp

The vector **a**1 = (a11, a21, …, ap1)

Is chosen such that *var* (Z1) is maximum

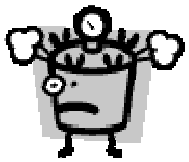
There are as many PC's as there are original variables (Xp) and their total variances are the same. Most of the variance is however hopefully concentrated in the first few PC's.

The trick is to maximize the variance of each Principal Component and to keep them uncorrelated !

Keeping the Z_k uncorrelated means that *cov* (Z_k, Z_{k+1}) = 0 k = 1 to p

And maximizing the variance of each Z could be done by increasing the values of the a_k so another condition is that $a_{1k}^2 + a_{2k}^2 + \ldots + a_{pk}^2 = 1$

Now, it can be shown (don't ask!) that…



$$\mathrm{var}[z_1] = \mathbf{a}_1^{\mathrm{T}} \mathbf{S} \mathbf{a}_1$$

Covariance matrix of the original variables X1 to Xp

In a general way then: **Z** = **AX**

A is a matrix:

$$\begin{pmatrix} a11 & \dots & a1p \\ \dots & \dots & \dots \\ ap1 & \dots & app \end{pmatrix}$$

← Vector a_1^T with weights $a_{11} \dots a_{1p}$

And because the Principal Components (Zp) are uncorrelated (by definition) the covariance matrix of the Zp =

$$\mathbf{S}_Z = \begin{pmatrix} \mathrm{var}(Z1) & 0 & \dots & 0 \\ 0 & \mathrm{var}(Z2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathrm{var}(Zp) \end{pmatrix} = \mathbf{A}\,\mathbf{S}_x\,\mathbf{A}^T$$

In linear algebra the rows of **A** (colums $\mathbf{A}^T$) are the *eigenvectors* (= sets of loadings = PC) of the matrix **S**x and the elements in the leading diagonal of **S**z are its corresponding *eigenvalues.* So the eigenvalues represent the amount of variance explained by each eigenvector (PC).

This is resolved by iterations to find the a_p (=loadings or weights) that maximize the variance of each new variable (Z_p = Principal Component). It is heavy computation as soon as p becomes greater than 3 and it requires a computer. The results of an analysis are of this type:

CASE 1: vertebrate biometry

4 variables (same units), 82 samples

Using the data on skull measurements from fossil oreodont mammals (Appendix 3.11), we obtain the following from a standard statistics package.

Eigenanalysis of the correlation matrix:

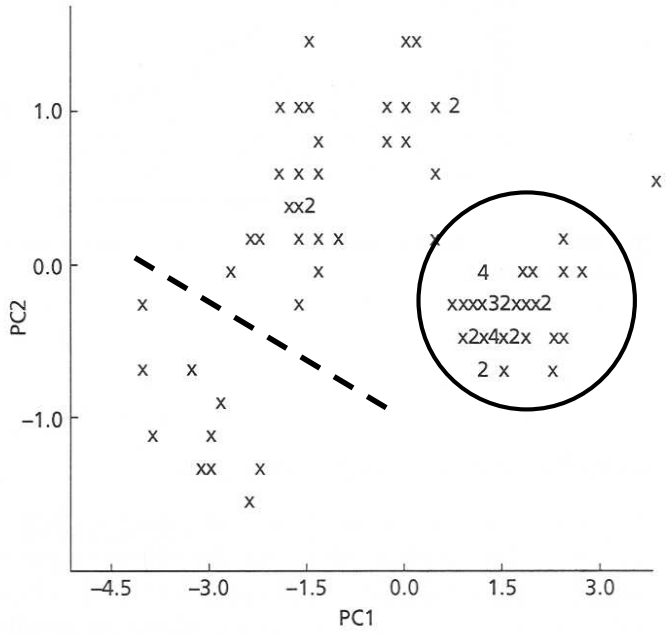
	PC1	PC2	PC3	PC4
Eigenvalue	3.4428	0.3898	0.1116	0.0557
Proportion	0.861	0.097	0.028	0.014
Cumulative	0.861	0.958	0.986	1.000

~3/4 of the remaining variance

Loadings:

Variable	PC1	PC2	PC3	PC4
Brainw.	−0.497	−0.488	0.705	−0.135
Toothl.	−0.501	−0.468	−0.598	0.414
Bullal.	−0.519	0.290	−0.307	−0.743
Bullad.	−0.482	0.677	0.226	0.508

PC1 is extremely strong (86.1% of total variance) and is decribed by loadings that are similar for each variable. It is clearly interpretable as



SIZE

CASE 2: soil geochemistry

Soils overlying dolomitised lmst and basalt with mineralized veins. Only first 6 PC's shown. PC1 = **dilution** by a major (SiO2, organic matter)? Mg neutral, both in dolom and basalts + veins? PC2 Ag and Cu + Zn, Cd and Mg?? PC3 Mg, Sr and Ca likely host carbonate signature.

9 variables, 74 samples

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	4.9035	1.4766	1.1953	0.6156	0.3732	0.1734
Proportion	0.545	0.164	0.133	0.068	0.041	0.019
Cumulative %	54.5	70.9	84.2	91.0	95.2	97.1

Loadings:

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Zn	−0.325	0.429	0.221	−0.003	0.461	0.585
Pb	−0.419	−0.151	0.067	−0.109	−0.230	0.261
Cd	−0.309	0.317	0.476	0.083	0.279	−0.619
Mg	0.064	0.650	−0.296	−0.596	−0.262	−0.006
Ca	−0.365	0.026	−0.389	0.437	−0.028	0.135
Cu	−0.360	−0.327	−0.033	−0.505	0.096	−0.255
Ag	−0.385	−0.331	−0.084	−0.316	0.110	0.105
Sr	−0.286	0.178	−0.628	0.213	0.141	−0.332
Ba	−0.357	0.145	0.280	0.187	−0.739	−0.033

CASE 3: foraminifer biometry

Measurements of the foraminifer Afrabolivina at various depths in 2 Fms (Cretaceous). Only 6 first PC's shown. PC1 = **size**, higher + scores = smaller size. PC2 shape factor? Foramen variables somewhat independent of rest.

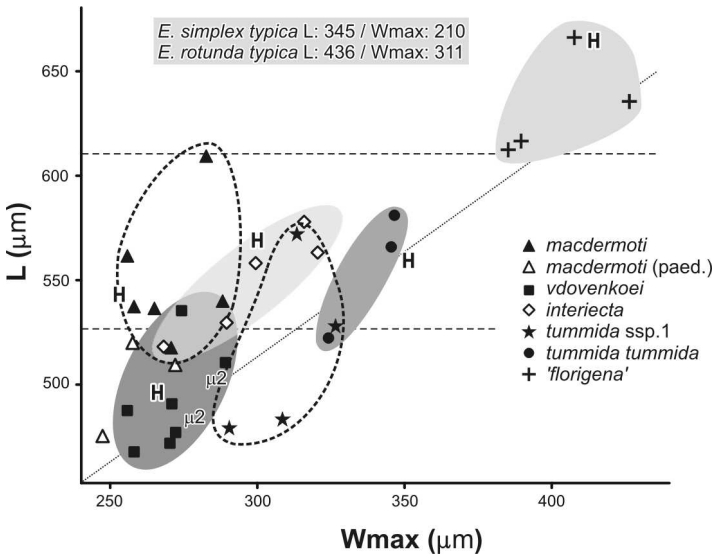
9 variables, 91 samples

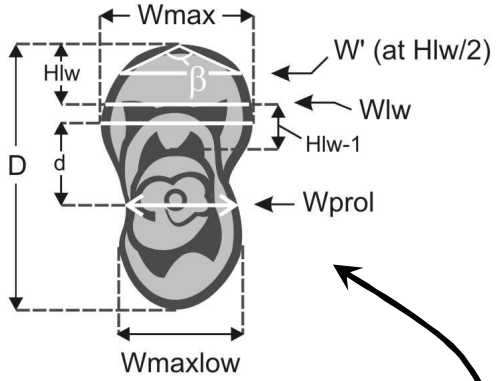
	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalue	6.1684	1.3435	0.5182	0.4109	0.2771	0.1536
Proportion	0.685	0.149	0.058	0.046	0.031	0.017
Cumulative	0.685	0.835	0.892	0.938	0.969	0.986

Loadings:

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Length	−0.327	−0.170	0.562	−0.221	−0.587	−0.360
Width	−0.384	−0.020	−0.048	−0.167	−0.101	0.634
Wdthlc	−0.394	−0.017	0.061	0.040	−0.129	0.315
Hghtlc	−0.372	0.021	0.083	0.468	0.249	−0.316
Hght2lc	−0.379	0.076	−0.012	0.402	0.203	−0.208
Diampro	−0.301	0.247	−0.729	−0.320	−0.267	−0.372
Breadth	−0.385	0.075	−0.075	0.136	0.051	0.255
Wdthfor	0.088	−0.765	−0.366	0.392	−0.341	0.039
Locfor	−0.247	−0.558	−0.001	−0.516	0.579	−0.146

CASE 4: foraminifer biometry





***E. interiecta* Vdovenko 1971**

GENSHAPE = D / (Wmax + Wmaxlow)
UMB = [(Wmax / Wprol) / d] x 100
EVOL = d / D
HSYM = (Wmax / Wmaxlow) x 100

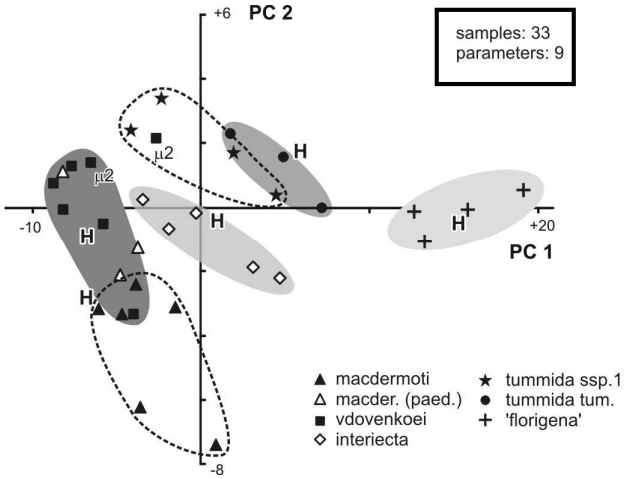
S = D / (Wmax + Wprol)
β = 2 arctg (W'/Hlw)
B = (1/β)x100 + Wmax/Wlw
EMI = S/2 + B

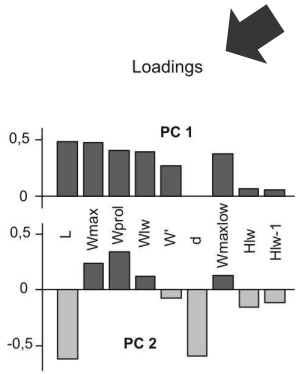
Parametres measured (9)

PC1 80.6% (eigenvalue 46.4)
PC2 and PC3 respectively
11.9% and 3.2%
(eigenvalues 6.9 and 1.8).

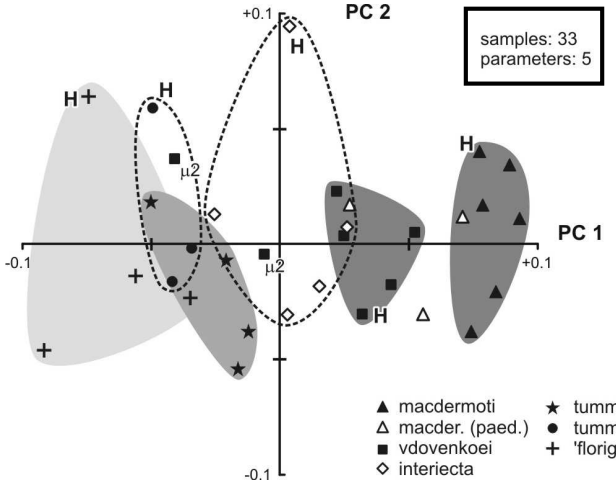
PC1 (61.8% of total variance) and
PC2 and PC3 (respectively 22.2 and
9% of variance)

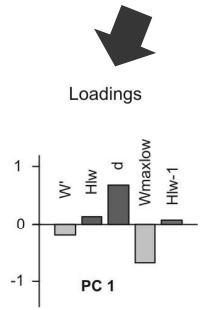
Data not standardized for size, no screening





Data all divided by D, only uncorrelated variables



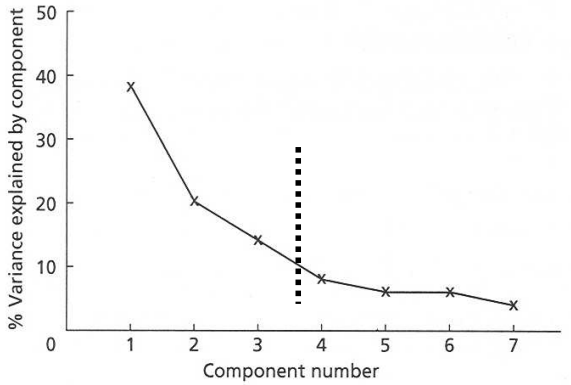


Procedure

1) Look at the correlation matrix and scatter plots of your data before considering PCA, if there is no correlation at all there is no point going further along this path;
2) Consider dropping some data if isolated outliers;
3) Standardize the data if necessary (see next slide);
4) Calculate standard statistic values and a variance-covariance or correlation matrix depending on the variables. Same units and magnitude important in their relations > **covariance** [trace variances], different units (no sense to do direct comparison between mm and ml…) or magnitude not to take into consideration for their relations > **correlation** [trace 1];
5) Perform the eigenvectors/eigenvalues iterative computation;
6) Plot scatters of the main PC's, check the loadings and scree plot; decide how many PC's are useful.

How many PC's?

1) If a new PC does not affect the the proportion of the total variance significantly > scree plot. Tendency to take too much;
2) Enough PC's should be included to explain **>= 90%** of the **total variance**. Tendency to take too few;
3) PC omitted if its variance < average of all PC's or less than 1 when the correlation matrix is used;
4) Joliffe (1986) cut-off value for eigenvalues;
5) Approximate equality of the last k eigenvalues (variances), none contains more information than any other, the important part of the data variance is therefore in the p-k PC's.

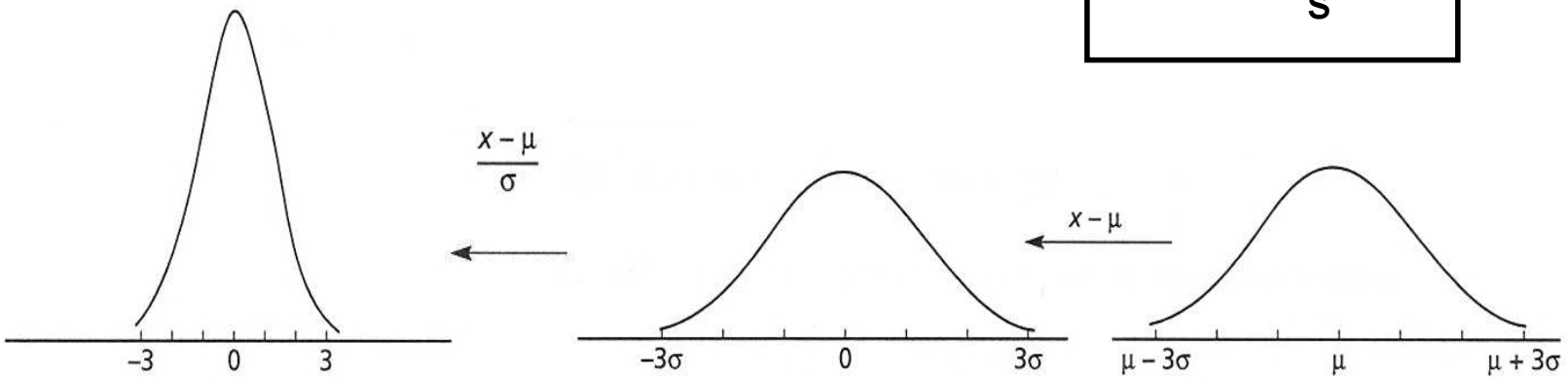


Standardization

The standard deviation can be used to scale the original measurements, expressing both distributions in terms of a common standard-deviation scale and not their original non-comparable units (mm, ml…).

Substracting the mean from every possible value and dividing by the standard deviation shifts the mean to 0 and scales the spread to give a standard deviation of 1.

$$Z = \frac{X - \overline{X}}{s}$$



Closed data

Closed data are extremely common in geology (petrology, geochemistry,microfacies, palaeoecology etc. > %, ppm). The problem is that, regardless of the underlying geological process, the value of one variable will automatically tend to affect the values of the other variable. Even worse in PCA because closure effect not linear, does not all come out in PC1…

Ex. *If the absolute amount of a major constituent (SiO2) is divided by two while keeping absolute amounts of all other constituants constant, then the % or ppm of all other constituents will inevitably increase. Any pair of minor constituents will tend to appear positively correlated.*

Solutions?

1) Any ratio between 2 variables is in principle open (x/SiO2) but not ideal because the ratio will vary with variation in the denominator (if major, see above) > chose an independent variable? Circularity or reasoning…;
2) $X' = \ln (X/Y)$ where Y is a component used as denominator for all variables produces open data (Aitchison, 1984, 1986). Changes the original data so interpretation less straightforward but the best deal!

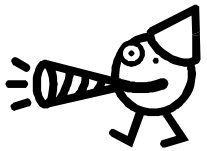
In PAST 1.33 

- The PCA routine finds the eigenvalues and eigenvectors of the variance-covariance matrix or the correlation matrix:
  - In the second option all variables are normalized by division by their standard deviations;
  - Eigenvalues are given along with percentages of variance accounted for by the corresponding PC;
  - Scatter plots of the PC's, biplots (samples, variables) and scree plots (eigenvalues) are given as well as loadings plot;
  - Another algorithm, supposedly superior to the 'classical' one is available (Singular Value Decomposition – SVD) but gives very similar results except that it centers on 0.

How good is PCA? VERY good but…

- Not adapted for non-metric data types (presence-absence, ranked data);
- Strictly speaking not a statistical technique as the results can’t be tested by objective statistical tests, like cluster analysis it is judged on the results;
- Lack of objective criteria to select the number of Principal Components to consider;
- Problem of closed data (%, ppm) and induced correlations;

References



- PAST: http://folk.uio.no/ohammer/past/

- Good websites:
  - http://www.astro.princeton.edu/~gk/A542/PCA.ppt
  - http://www.cs.rit.edu/~rsg/BIIS2005.html (lecture 7)
  - http://palaeomath.palass.org/
  - http://www.cse.csiro.au/poptools/
  - http://mathworld.wolfram.com/

Very good reference for data analysis in geology:

Swan, A.R.H. & Sandilands, M. 1995. Introduction to geological data analysis. Blackwell Science.