EXTRACTING INFORMATION FROM HIGH DIMENSIONAL DATA

Principal Component Analysis

High-dimensional descriptors

- Many different descriptors can be adopted for characterizing a set of objects under investigation:

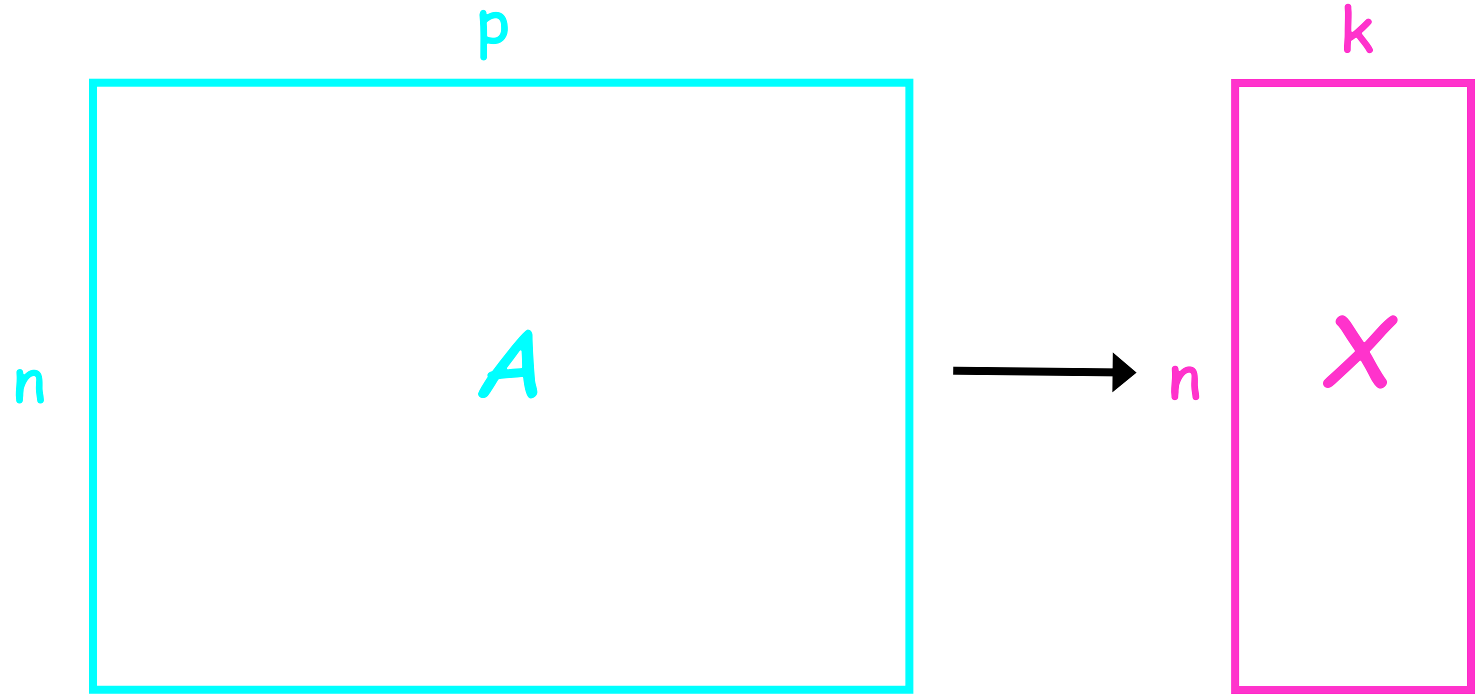
1) given a set of proteins, we can measure the residue composition (20 values), the dipeptide composition (400 values), length, average hydrophobicity..... of each sequence

2) Given a set of individuals we can measure dimensions, weight, haematic concentration of metabolites...

How can we extract a minimal set of descriptors without losing information on the variability of the set?

Data Reduction

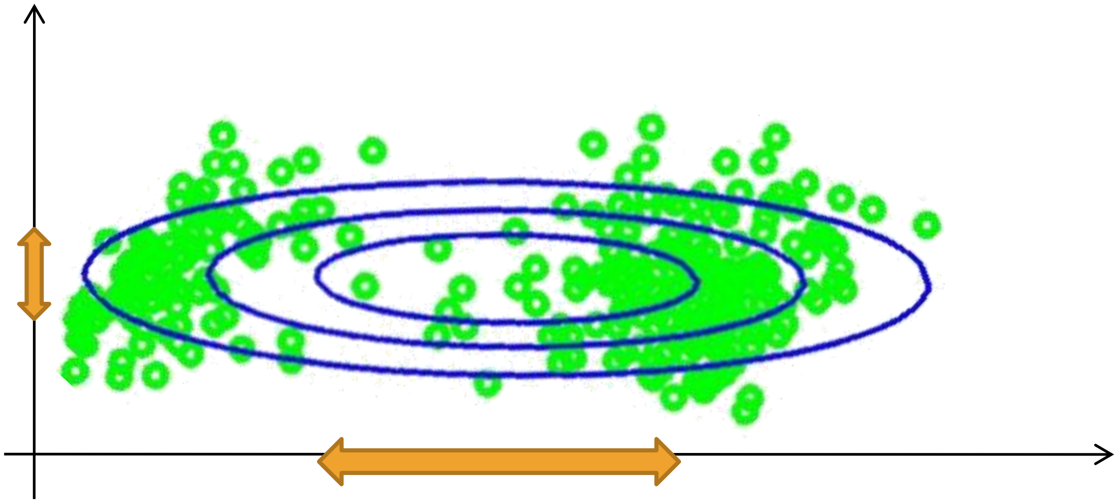
- Summarizing the **p**-dimensional description of **n** objects by a smaller set of (**k**) derived (synthetic, composite) variables.



Data Reduction

- "Residual" variation is information in *A* that is not retained in *X*
- *A* good data reduction must balance between
  - clarity of representation, ease of understanding
  - oversimplification: loss of important or relevant information.

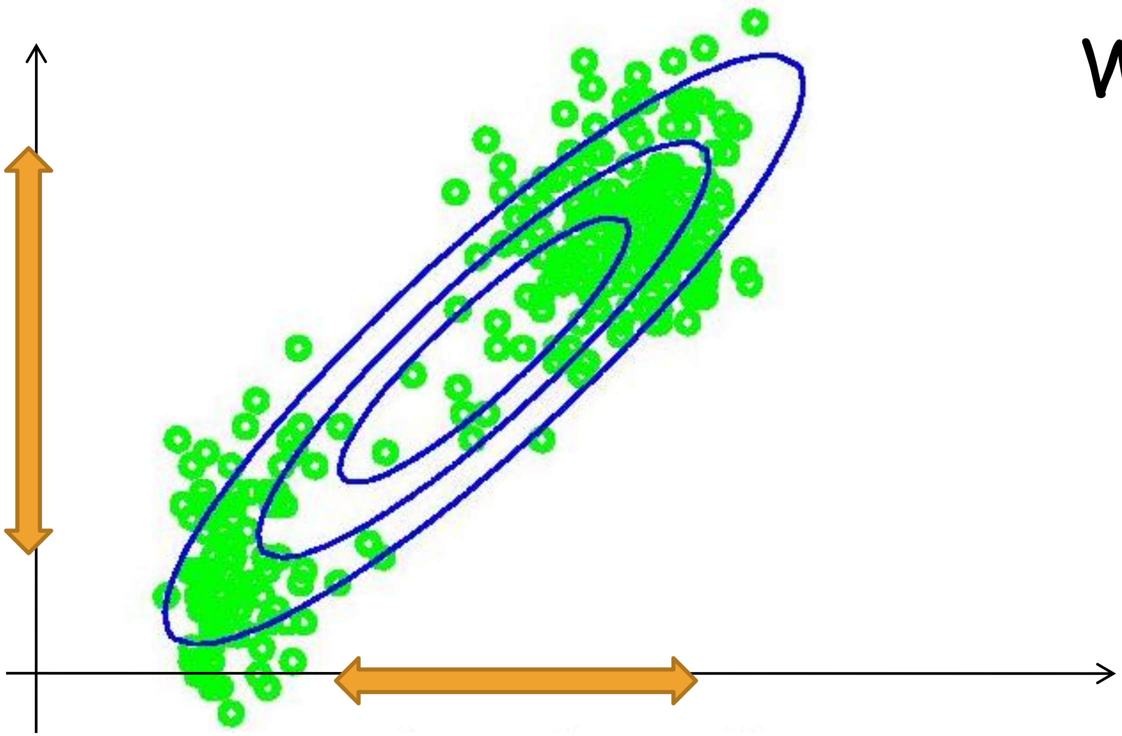
Most important descriptors



When plotting these two dimensional data it is evident that the x-direction accounts for the largest part of the variance

Directions with largest variance better describe the data set

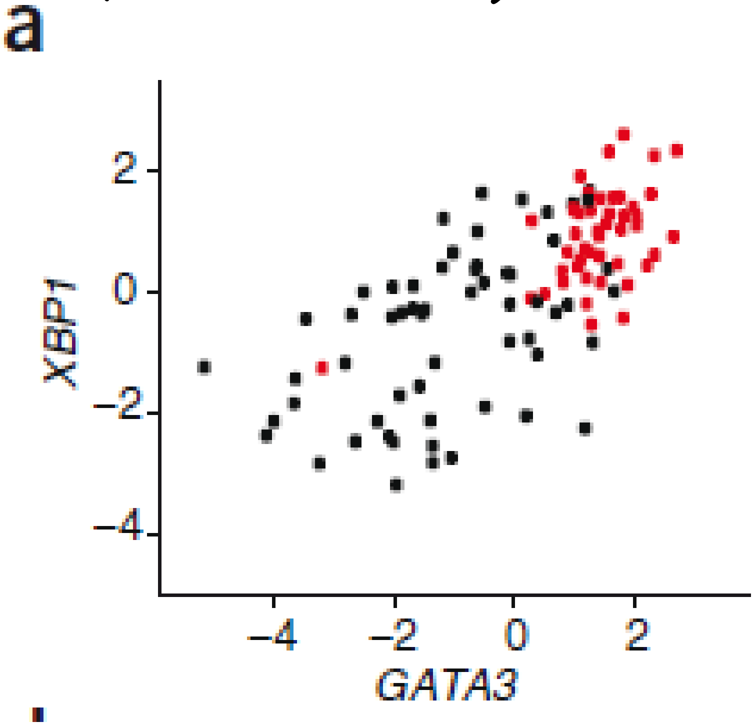
Most important descriptors



When variables are correlated the variable variance is not able to determine principal components

Example:2D

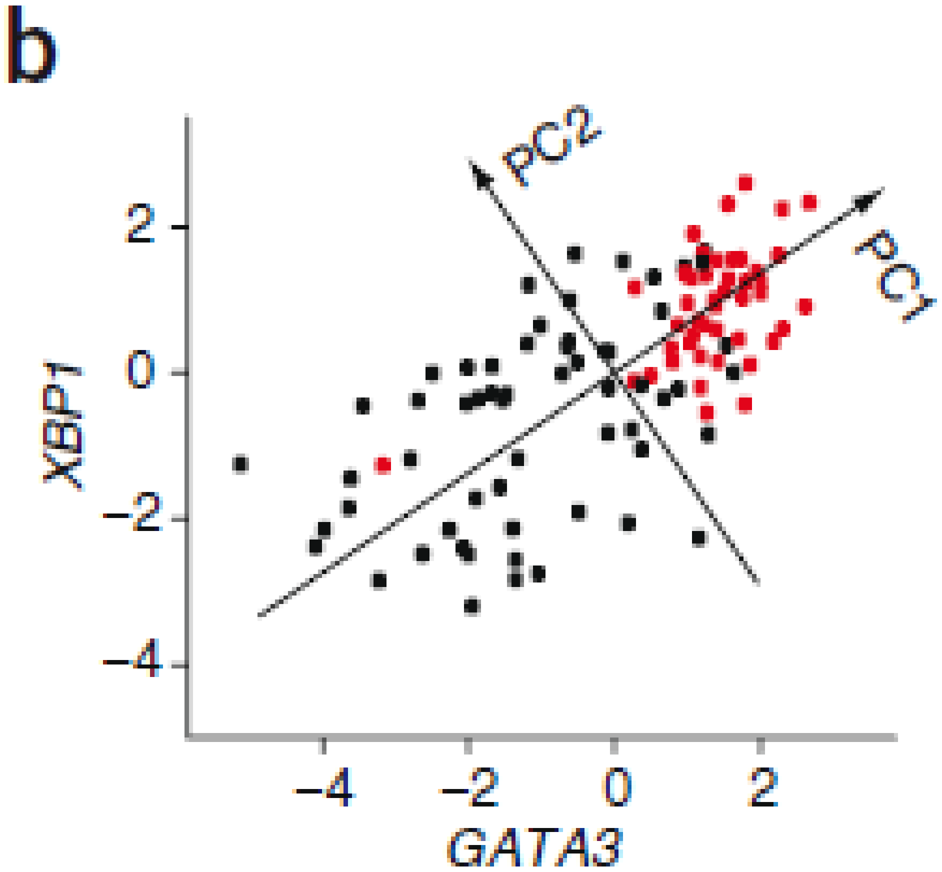
Suppose to measure the expression level of two genes (XBP1 and GATA3) in breast cancer samples expressing or not the Estrogen Receptor (ER+ in red, ER- in black)



Markus Ringnér. What is principal component analysis?
Nature Biotechnology 26, 303 - 304 (2008)

Example:2D

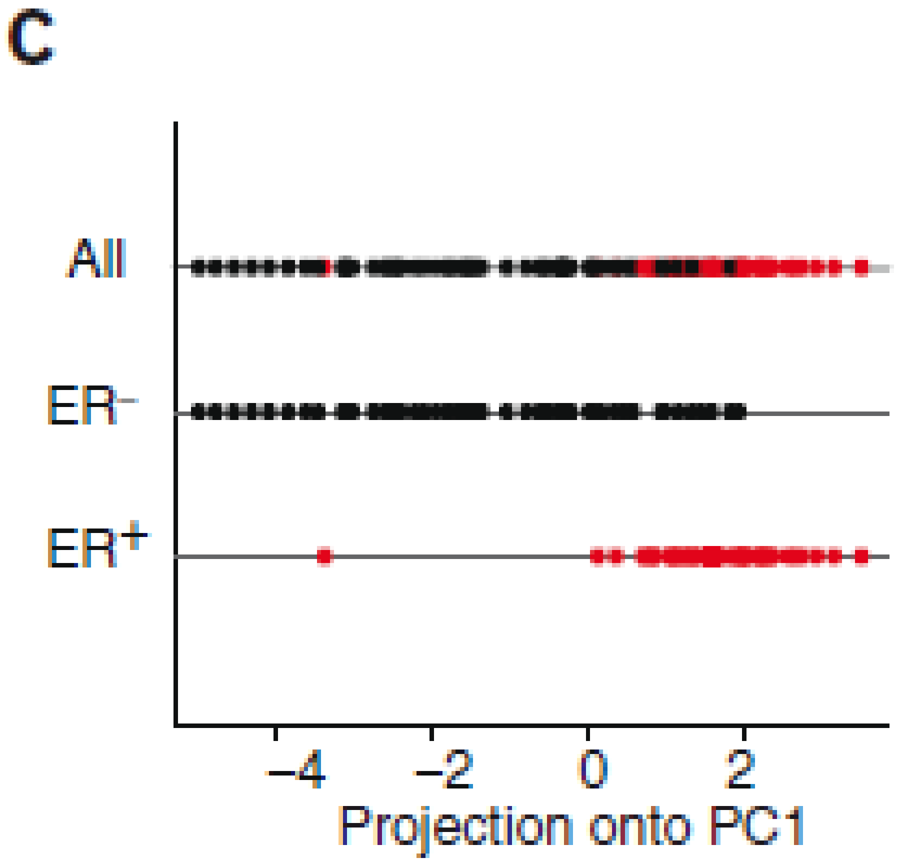
The total variance is decomposed into two orthogonal components, PCA1 and PCA2



Markus Ringnér. What is principal component analysis? Nature Biotechnology 26, 303 - 304 (2008)

Example:2D

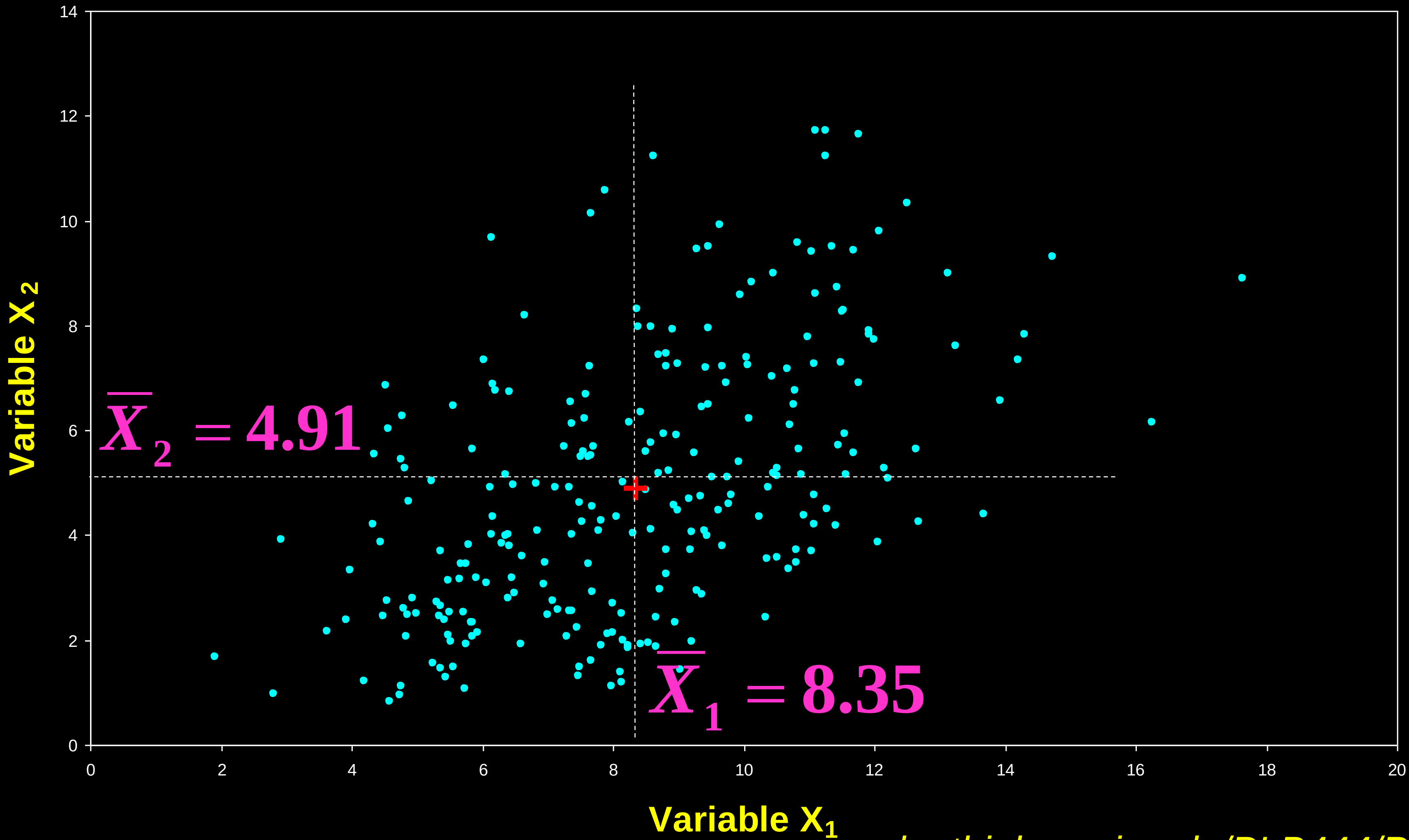
Analysis of the principal component can highlight important features in an unsupervised way



Markus Ringnér. What is principal component analysis?
Nature Biotechnology 26, 303 - 304 (2008)

2D Example of PCA

- variables X_1 and X_2 have positive covariance & each has a similar variance.

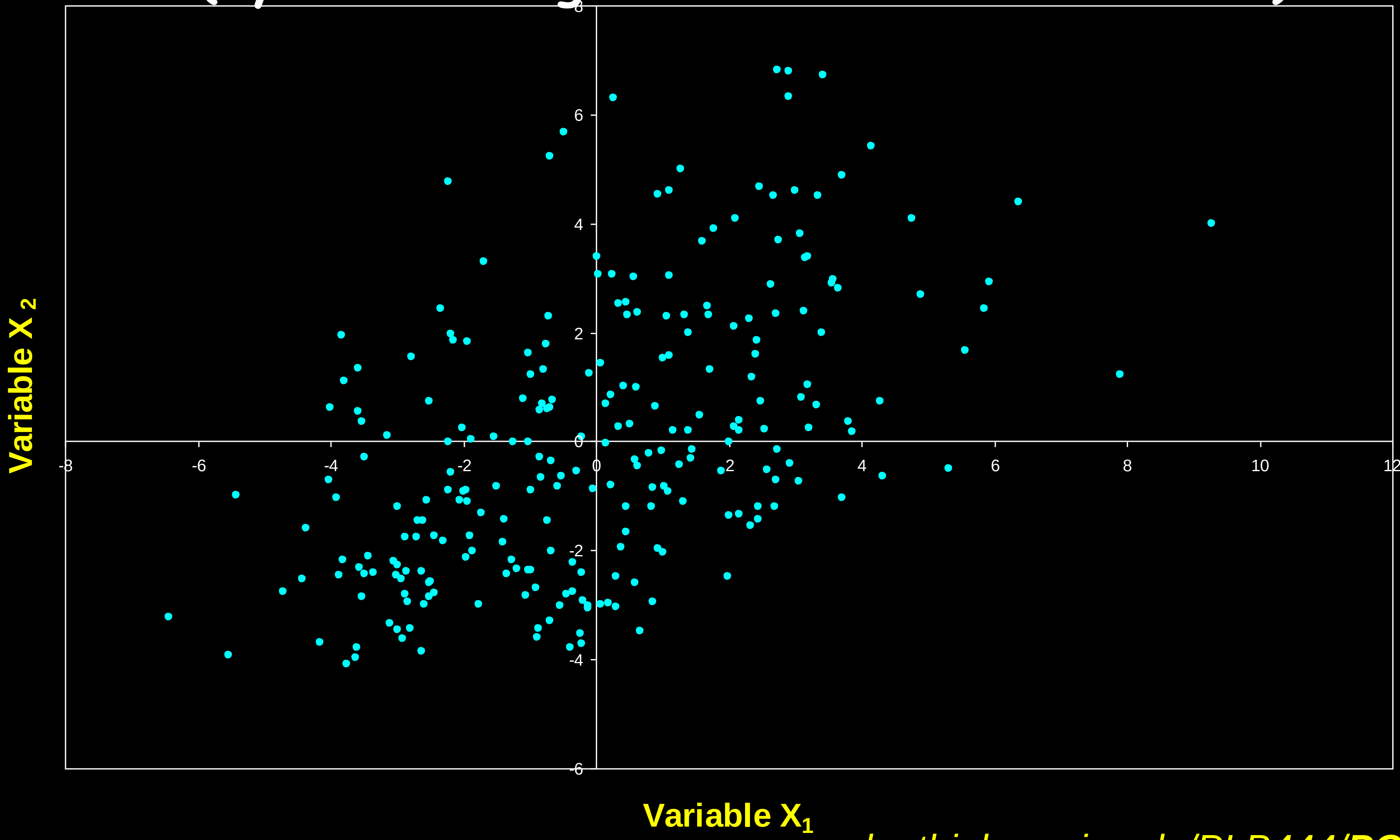


*www.plantbiology.siu.edu/PLB444/**PCA**.ppt*

$$V_1 = 6.67 \qquad V_2 = \mathbf{6.24} \qquad C_{1,2} = \mathbf{3.42}$$

Configuration is Centered

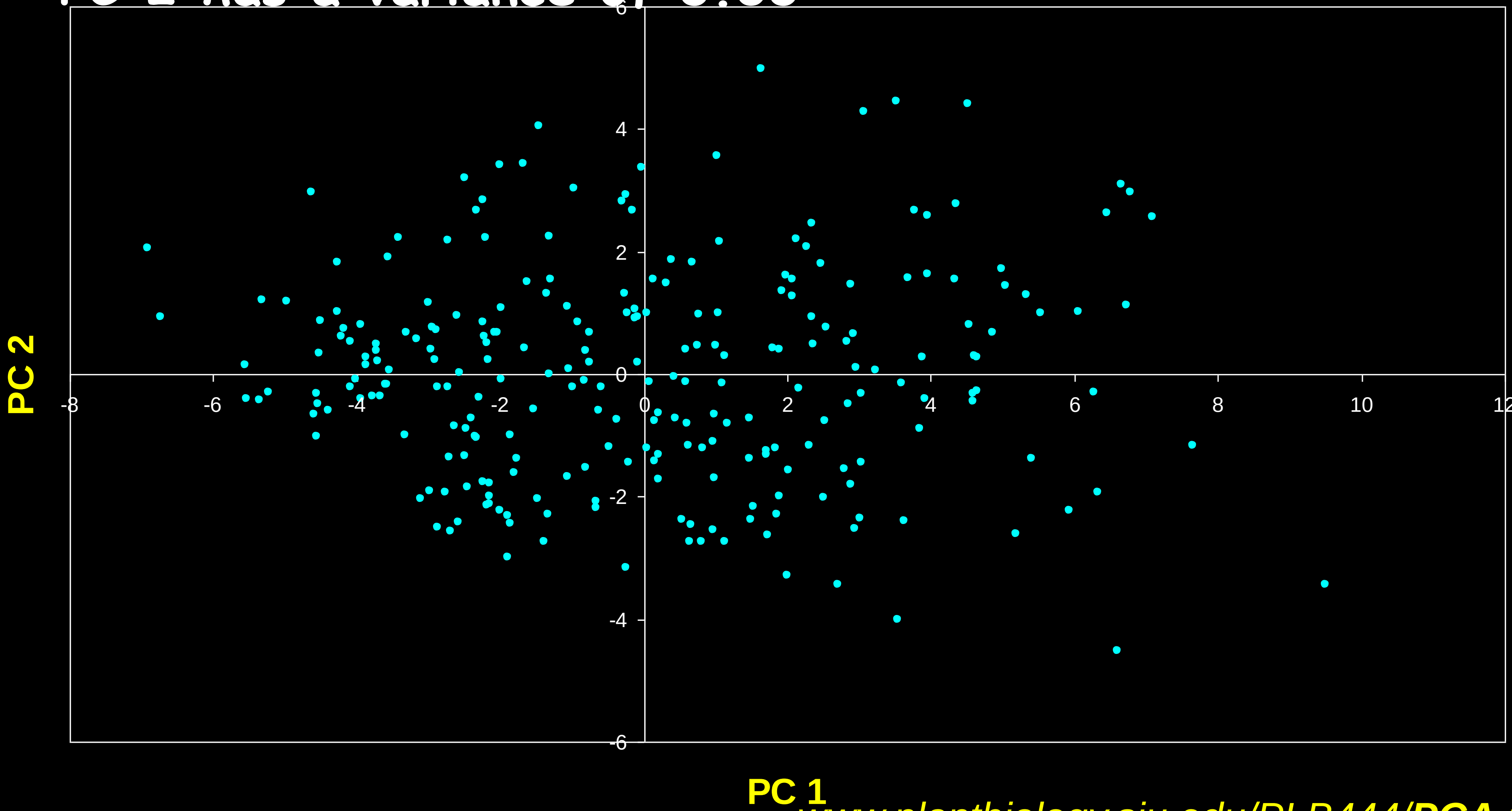
- **each variable is adjusted to a mean of zero** (by subtracting the mean from each value).



*www.plantbiology.siu.edu/PLB444/**PCA**.ppt*

Configuration is rotated→New coordinates PC1 PC2

- PC 1 and PC 2 have zero covariance.
- PC 1 has the highest possible variance (9.88)
- PC 2 has a variance of 3.03



*www.plantbiology.siu.edu/PLB444/**PCA**.ppt*

Covariance Matrix

	Variable 1	Variable 2	Variable j	Variable n
Item1	$x_1(1)$	$x_2(1)$	$x_j(1)$	$x_n(1)$
Item 2	$x_1(2)$	$x_2(2)$	$x_j(2)$	$x_n(2)$
Item i	$x_1(i)$	$x_2(i)$	$x_j(i)$	$x_n(i)$
Item m	$x_1(m)$	$x_2(m)$	$x_j(m)$	$x_n(m)$

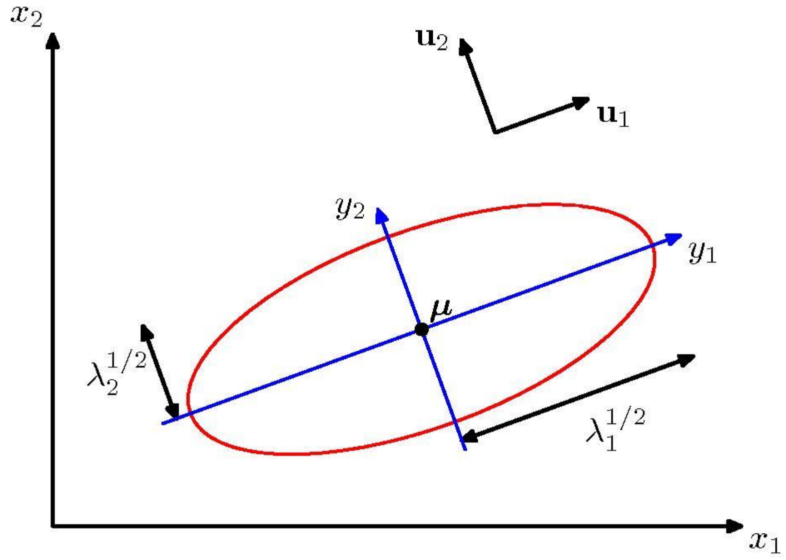
$$\mathbf{COV} = \begin{pmatrix} \sigma_1^{\;2} & \operatorname{cov}(x_1, x_2) & \operatorname{cov}(x_1, x_j) & \operatorname{cov}(x_1, x_n) \\ \operatorname{cov}(x_2, x_1) & \sigma_2^{\;2} & \operatorname{cov}(x_2, x_j) & \operatorname{cov}(x_2, x_n) \\ \operatorname{cov}(x_j, x_1) & \operatorname{cov}(x_j, x_2) & \sigma_j^{\;2} & \operatorname{cov}(x_j, x_n) \\ \operatorname{cov}(x_n, x_1) & \operatorname{cov}(x_n, x_2) & \operatorname{cov}(x_n, x_j) & \sigma_n^{\;2} \end{pmatrix}$$

Covariance and multidimensional normal distribution

$$\mathrm{N}(x\,|\,M,\Sigma)=\frac{1}{(2\pi)^{\frac{n}{2}}\left|COV\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-M)^{T}COV^{-1}(x-M)\right\}$$

M is a n-valued vector (means)

COV is a n×n symmetric matrix (covariance matrix), with determinant |COV|



Uncorrelated variables

- The set of variable is uncorrelated if all the covariances (not the variances) are null:

$$COV_{ij} = \frac{1}{m-1}\sum_{k=1}^{m}\left(x_i(k)-M_i\right)\left(x_j(k)-M_j\right)=\delta_{ij}\sigma^2{}_i$$

- that is if covariance matrix is in diagonal form.

Eigenvector equation

- *COV* is a square symmetrical matrix and can be reduced in diagonal form

$$\tilde{COV}_{ij} = \lambda_i \delta_{ij}$$

by means of the eigenvector equation

$$\mathbf{COV} \cdot \vec{u} = \lambda \cdot \vec{u}$$

$$\det\left|\mathbf{COV} - \lambda \mathbf{I}\right| = 0$$

it defines n real eigenvalues λ_i

and n real-m-valued orthonormal eigenvectors u_i

Matrix transformation

- The column normalized eigenvectors *u_i* are orthogonal and define the n x n unitary matrix U

$$U = \begin{pmatrix} U_{11} & U_{21} \\ U_{12} & U_{22} \end{pmatrix} = \begin{pmatrix} u_1(1) & u_2(1) \\ u_1(2) & u_2(2) \end{pmatrix}$$

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}$$

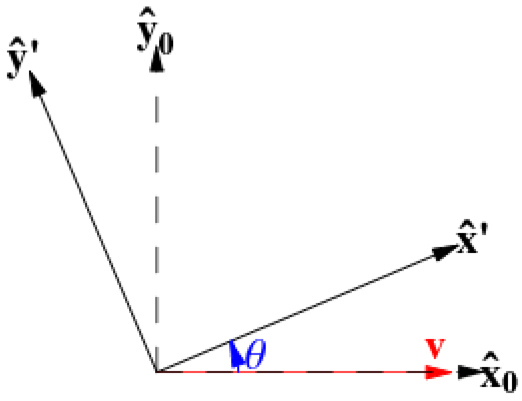
$$\mathbf{U}^{-1} = \mathbf{U}^T$$

Matrix transformation

- U defines an orthogonal rotation and/or reflection of the coordinate axes (preserving norms and angles)

$$\vec{\tilde{x}} = \mathrm{U}^{\mathrm{T}}\vec{x}$$

$$\vec{\tilde{x}}^{T}\vec{\tilde{x}} = \vec{\tilde{x}}^{T}\mathrm{U}\mathrm{U}^{\mathrm{T}}\vec{\tilde{x}} = \vec{x}^{T}\vec{x}$$



$$U = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Try for example a counterclokwise rotation of 45° and transform the point (1,1)

Matrix transformation

- Given the matrix A, the eigenvalues define the diagonal matrix Λ and the eigenvectors define the unitary matrix U so that:

$$\Lambda = U^T A U$$

Prove with matrix:

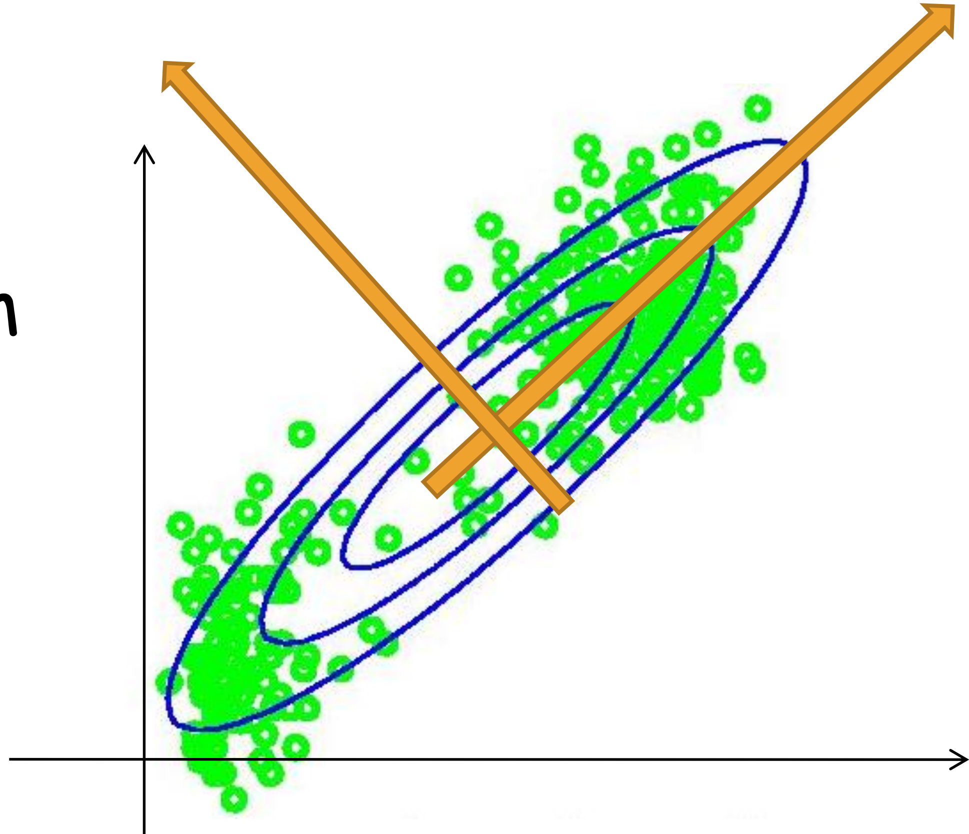
$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

Matrix transformation

- The diagonal matrix :

$$\mathbf{U}^T \cdot \mathbf{COV} \cdot \mathbf{U} = \mathbf{C\tilde{O}V}$$

- U then defines a coordinate rotation so that in the new system variables are not correlated

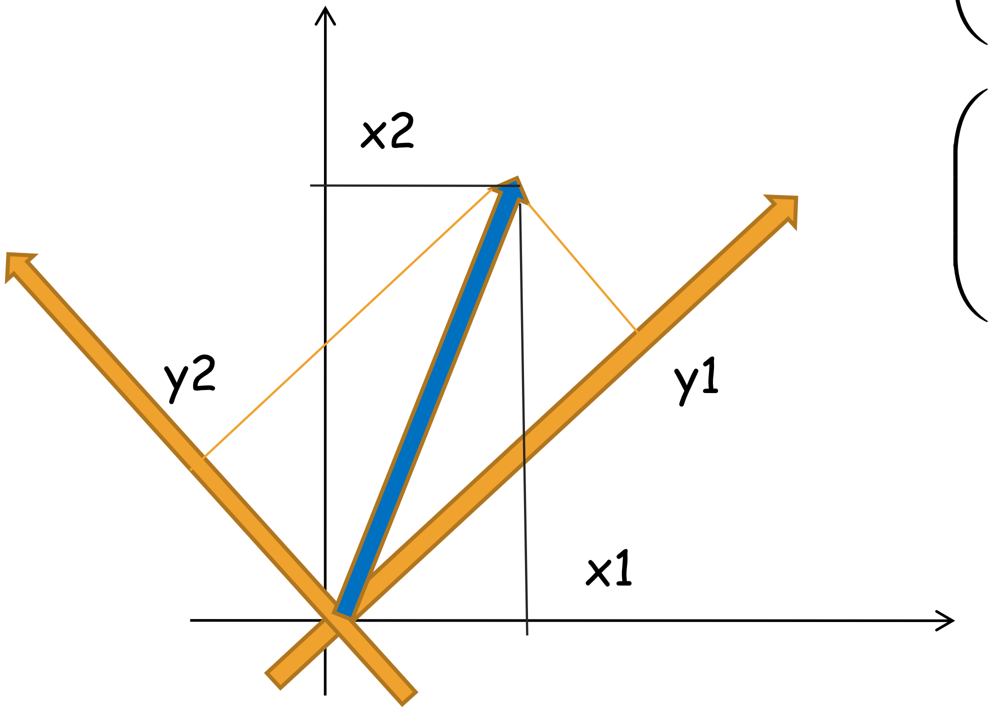


Matrix transformation

- □ Coordinate transformation

$$\begin{pmatrix} x1 \\ x2 \end{pmatrix} = \mathbf{U} \begin{pmatrix} y1 \\ y2 \end{pmatrix}$$

$$\begin{pmatrix} y1 \\ y2 \end{pmatrix} = \mathbf{U}^T \begin{pmatrix} x1 \\ x2 \end{pmatrix}$$



Matrix transformation

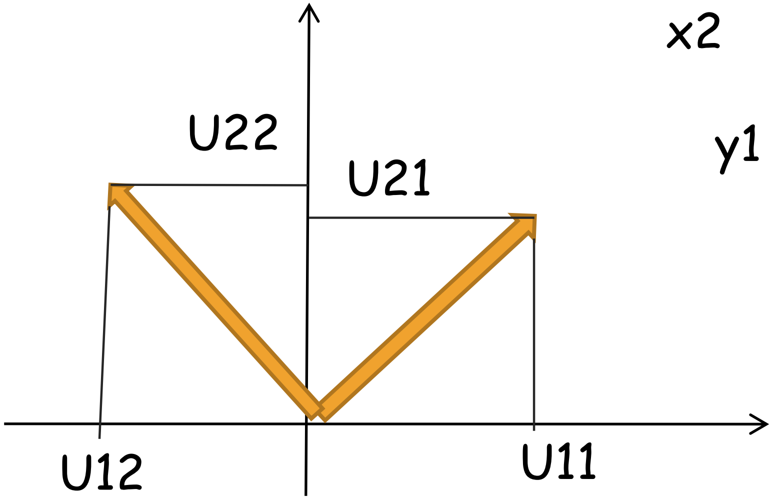
In particular the old coordinates of new axes, called LOADINGS, are the eigenvectors

$$\begin{pmatrix} x1 \\ x2 \end{pmatrix} = \mathbf{U}\begin{pmatrix} y1 \\ y2 \end{pmatrix}$$

$$\begin{pmatrix} y1 \\ y2 \end{pmatrix} = \mathbf{U}^{T}\begin{pmatrix} x1 \\ x2 \end{pmatrix}$$

$$\mathbf{U}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix}$$

$$\mathbf{U}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} U_{12} \\ U_{22} \end{pmatrix}$$



Principal Component Analysis

- Given any item x, represented by an m-valued vector $\vec{x}^T = (x_1, x_2, ..x_m)$

it can be expressed with the n ordered principal components, y, by using the basis U

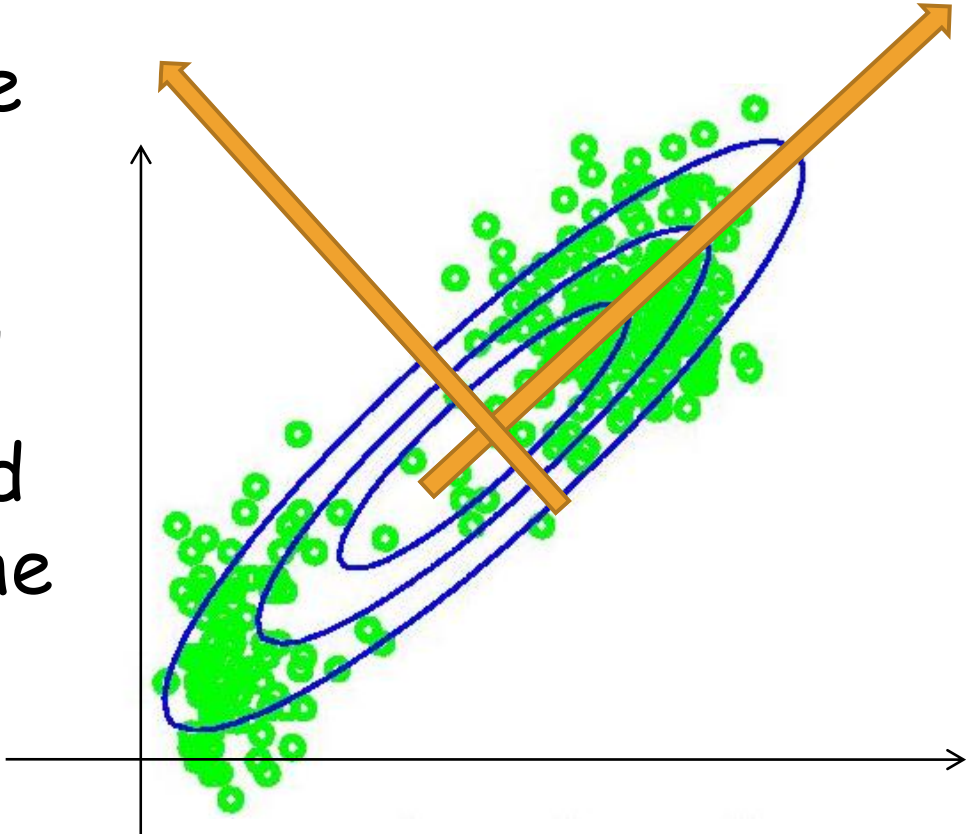
$$\vec{x} = \sum_{i=1}^{n} y_i \cdot \vec{u}_i$$

Where

$$y_i = \vec{x}^T \cdot \vec{u}_i = \vec{u}_i^T \cdot \vec{x}$$

Matrix transformation

- The eigenvalues measure the variances along the new coordinate axes.
- Usually the percentage of the total variance accounted by any coordinate is reported
- Coordinates are sorted from the highest to the lowest eigenvalue



Principal Component Analysis

Given a set of m items described with n variables:

1) Compute the mean of each variable
2) Subtract the mean to any measure
3) Compute the covariance matrix
4) Diagonalize the covariance matrix
5) Sort the eigenvalues from the highest to the lowest: the corresponding eigenvectors define the 1st,2nd....jth principal components
6) For each item i, the j-th component results from the scalar product of the i-th variable vector with the j-th eigenvector

Why “principal“?

- Given a **n**-dimensional representation of **m** observed samples and you want to represent data in a **p**-dimensional space, with p<n (e.g., for representing them in 2- or 3-dimensional space). Which is the best choice, when only linear transformations are allowed?

- **To use the first p principal components.**

Why "principal"? Minimum-error formulation

- We have to search for a complete orthonormal m-dimensional basis V

$$\vec{x}_k = \sum_{i=1}^{n} a_{ki} \cdot \vec{v}_i = \sum_{i=1}^{n} (\vec{x}_k{}^T \vec{v}_i)\vec{v}_i$$

and we have to use p-dimensions in that coordinate basis to approximate the points:

$$\tilde{\vec{x}}_k = \sum_{i=1}^{p} b_{ki} \cdot \vec{v}_i + \sum_{i=p+1}^{n} c_i \cdot \vec{v}_i$$

where c_i are independent of k

Why "principal"? Minimum-error formulation

- The goal is to minimize the error

$$E = \frac{1}{m}\sum_{k=1}^{m}\left\|\tilde{\vec{x}}_k - \vec{x}_k\right\|^2$$

For any base V, the error is minimized when:

$$b_{ki} = a_{ki} = \vec{x}_k^{\,T}\vec{v}_i$$

$$c_i = \bar{\vec{x}}_k^{\,T}\vec{v}_i$$

Then:

$$\vec{x}_k - \tilde{\vec{x}}_k = \sum_{i=p+1}^{n}\left[\left(\vec{x}_k - \bar{\vec{x}}_k\right)^T \vec{v}_i\right]\cdot\vec{v}_i$$

Why "principal"? Minimum-error formulation

$$E = \frac{1}{n}\sum_{k=1}^{m}\sum_{i=p+1}^{n}\left(\vec{x}_k{}^T\vec{v}_i - \bar{\vec{x}}_k{}^T\vec{v}_i\right)^2 = \sum_{i=p+1}^{n}\left(\vec{v}_i{}^T\mathbf{COV}\vec{v}_i\right)$$

The error is minimized, imposing the normalization of each v (via Lagrange's multipliers)

$$MINIMIZE_{\vec{v}_i}\left\{\sum_{i=p+1}^{n}\left(\vec{v}_i{}^T\mathbf{COV}\vec{v}_i\right)+\lambda_i\left(1-\vec{v}_i{}^T\vec{v}_i\right)\right\}$$

Then

$$\mathbf{COV}\vec{v}_i = \lambda_i\vec{v}_i \qquad E = \sum_{i=p+1}^{n}\lambda_i$$

choosing the lowest eigenvalues

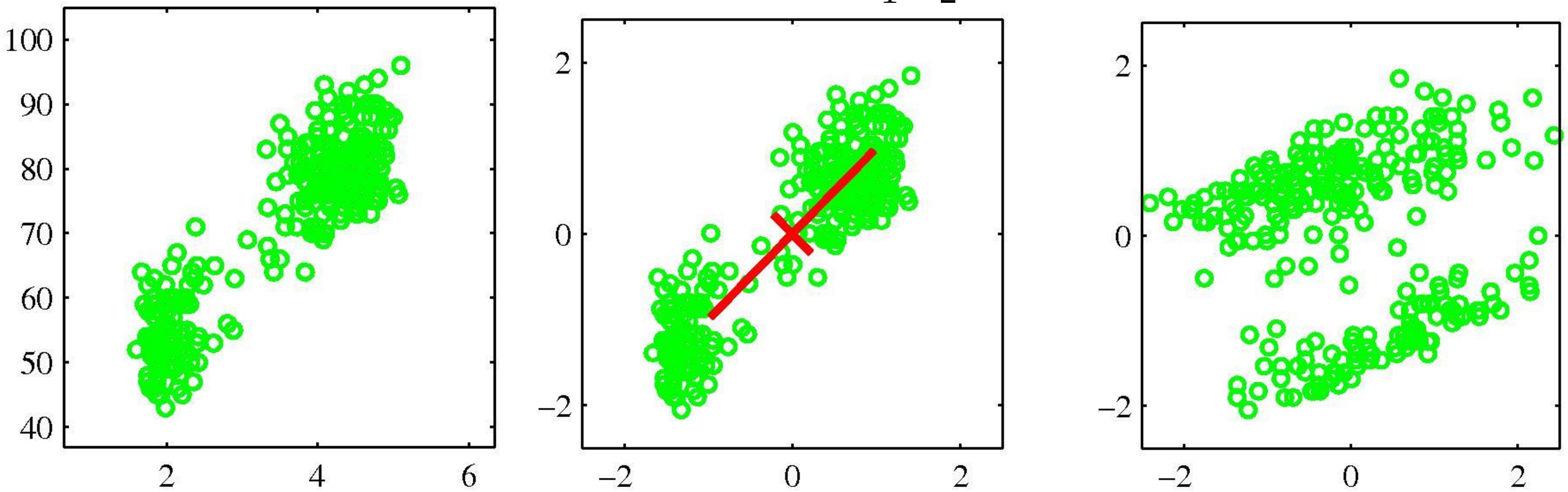
Covariance vs Correlation

- using covariances among variables only makes sense if they are measured in the same units
- even then, variables with high variances will dominate the principal components
- these problems are generally avoided by standardizing each variable to unit variance and zero mean.

Correlation PCA

- □ The correlation matrix can be chosen instead of the covariance. It gives a zero-mean, unit-covariance plot

$$corr(x_1, x_2) = \frac{1}{m-1}\sum_{i=1}^{m}\frac{(x_{1i}-M_1)(x_{2i}-M_1)}{\sigma_1\sigma_2}$$



An ecological example

- data from research on habitat definition in the endangered Baw Baw frog
- 16 environmental and structural variables measured at each of 124 sites
- correlation matrix used because variables have different units

Philoria frosti

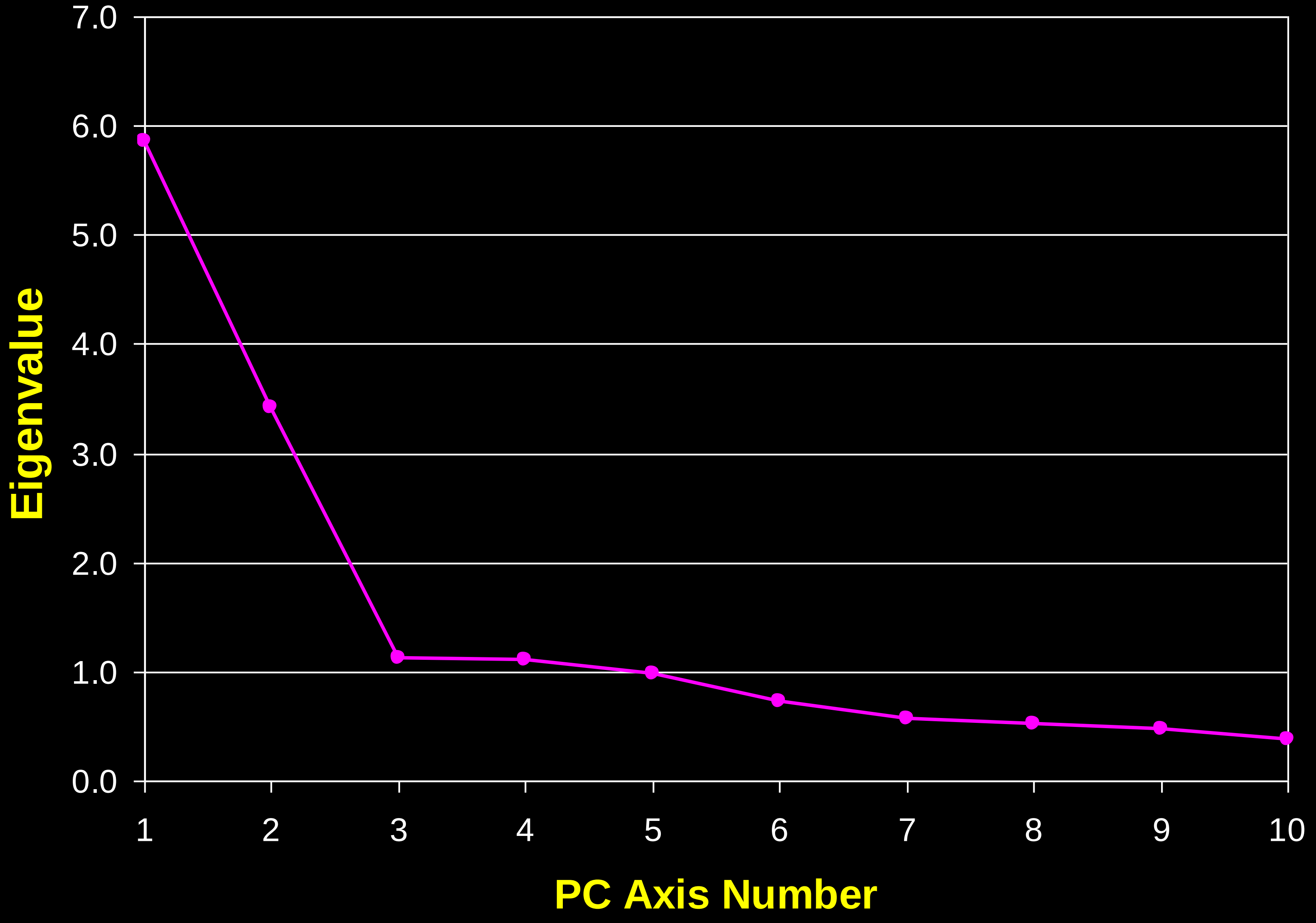
www.plantbiology.siu.edu/PLB444/**PCA**.ppt

Eigenvalues

Axis	Eigenvalue	% of Variance	Cumulative % of Variance
1	5.855	36.60	36.60
2	3.420	21.38	57.97
3	1.122	7.01	64.98
4	1.116	6.97	71.95
5	0.982	6.14	78.09
6	0.725	4.53	82.62
7	0.563	3.52	86.14
8	0.529	3.31	89.45
9	0.476	2.98	92.42
10	0.375	2.35	94.77

*www.plantbiology.siu.edu/PLB444/**PCA**.ppt*

Baw Baw Frog - PCA of 16 Habitat Variables



*www.plantbiology.siu.edu/PLB444/**PCA**.ppt*

Interpreting Eigenvectors

- correlations between variables and the principal axes are known as loadings
- each element of the eigenvectors represents the contribution of a given variable to a component

	1	2	3
Altitude	0.3842	0.0659	-0.1177
pH	-0.1159	0.1696	-0.5578
Cond	-0.2729	-0.1200	0.3636
TempSurf	0.0538	-0.2800	0.2621
Relief	-0.0765	0.3855	-0.1462
maxERht	0.0248	0.4879	0.2426
avERht	0.0599	0.4568	0.2497
%ER	0.0789	0.4223	0.2278
%VEG	0.3305	-0.2087	-0.0276
%LIT	-0.3053	0.1226	0.1145
%LOG	-0.3144	0.0402	-0.1067
%W	-0.0886	-0.0654	-0.1171
H1Moss	0.1364	-0.1262	0.4761
DistSWH	-0.3787	0.0101	0.0042
DistSW	-0.3494	-0.1283	0.1166
DistMF	0.3899	0.0586	-0.0175

*www.plantbiology.siu.edu/PLB444/**PCA**.ppt*

How many axes are needed?

- does the $(k+1)^{th}$ principal axis represent more variance than would be expected by chance?
- several tests and rules have been proposed
- a common “rule of thumb” when PCA is based on correlations is that axes with eigenvalues > 1 are worth interpreting

*www.plantbiology.siu.edu/PLB444/**PCA**.ppt*

Example: Thermostability

- The problem is to investigate the differences in
  - Residue composition
  - Codon composition
  - Codon usage

Among thermophilic and mesophilic prokaryotes

- Montanucci, Martelli, Fariselli, Casadio J Proteome Res, 2007

Example: Thermostability

- The data set contains 116 fully sequenced genomes from prokaryotes
  - 16 thermophilic species (11 archaea and 5 bacteria), with an OGT higher than 60 °C
  - 100 mesophilic species (95 bacteria and 5 archaea) with an OGT lower than 45 °C.
  - 7 quasi-mesophilic species (7 bacteria) with OGT between 45 and 60 °C

PCA on residue composition

Red:

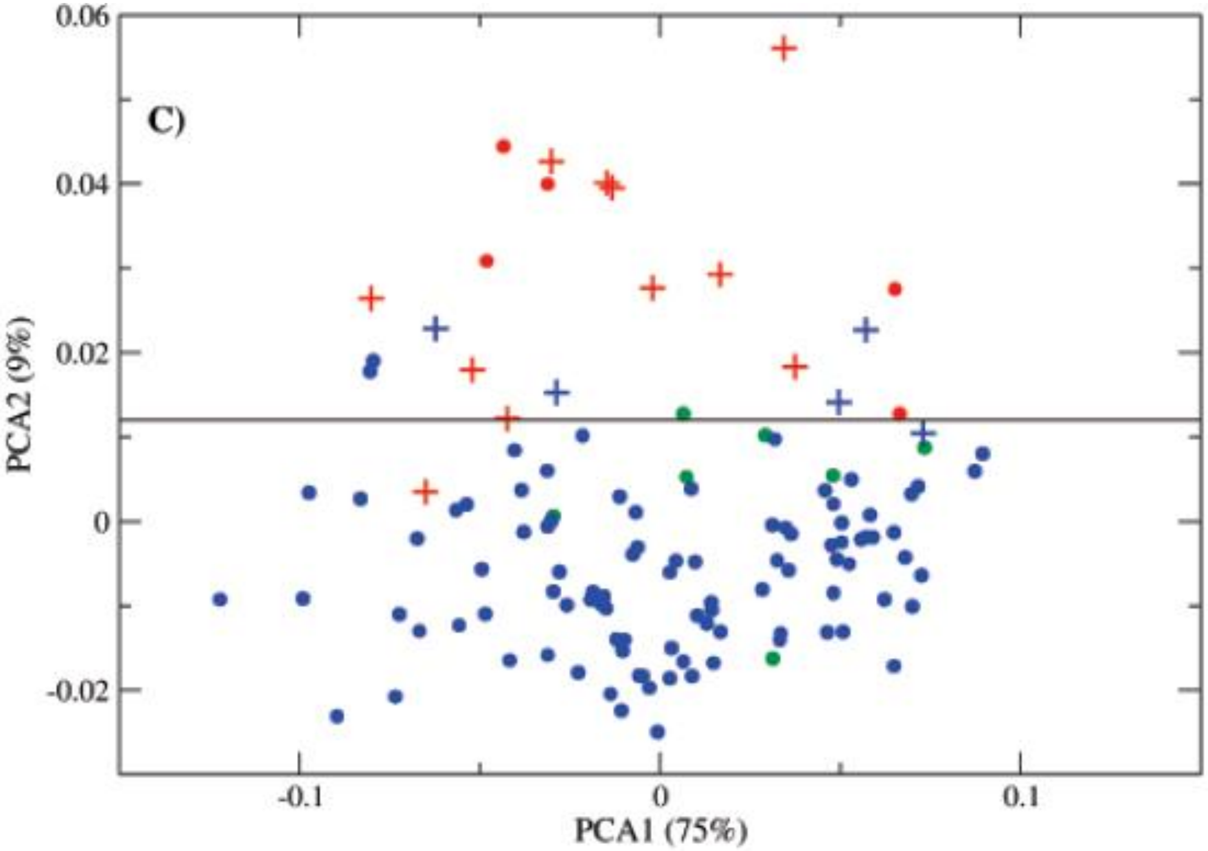
Thermophilic

Green:

Intermediate

Blue:

Mesophilic



PCA on codon usage

Red:

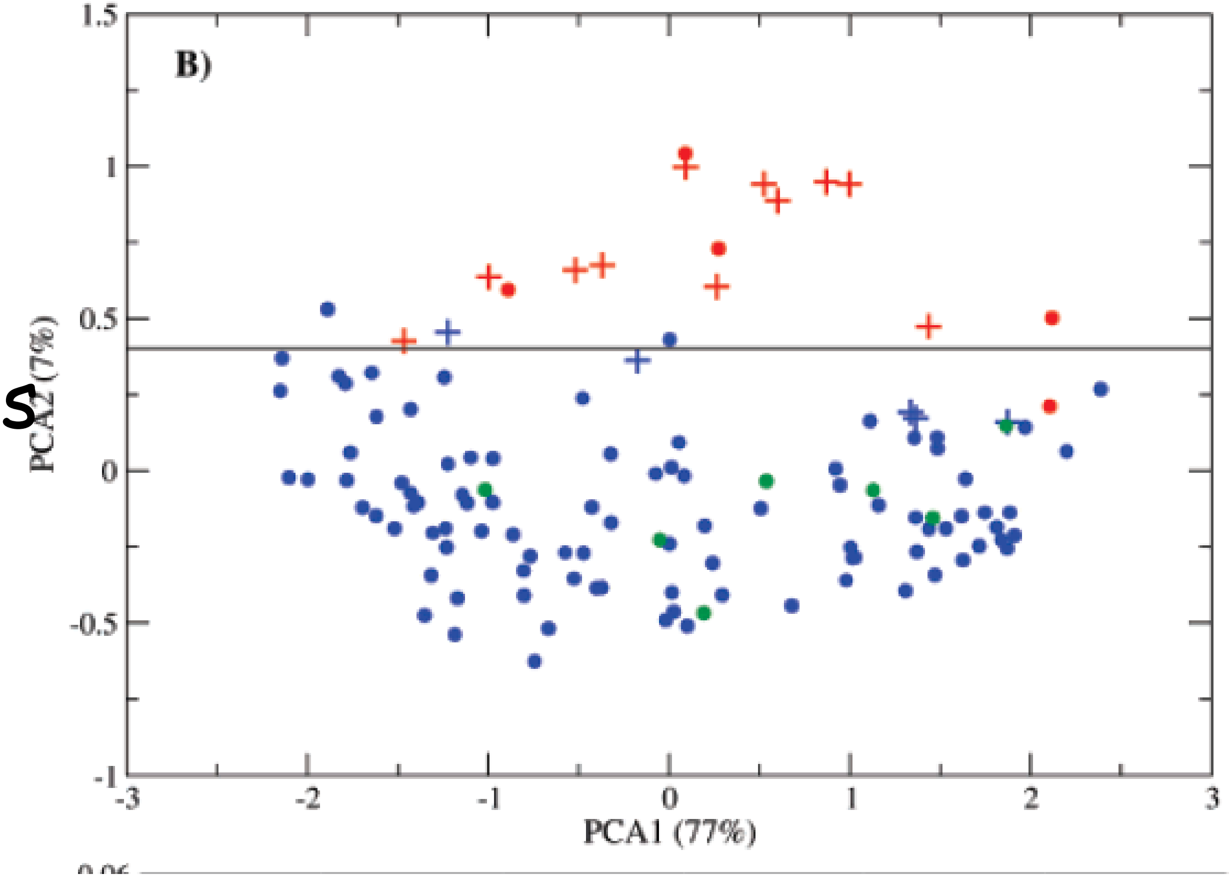
Thermophilic

Green:

Intermediates

Blue:

Mesophilic



PCA on codon composition

Red:

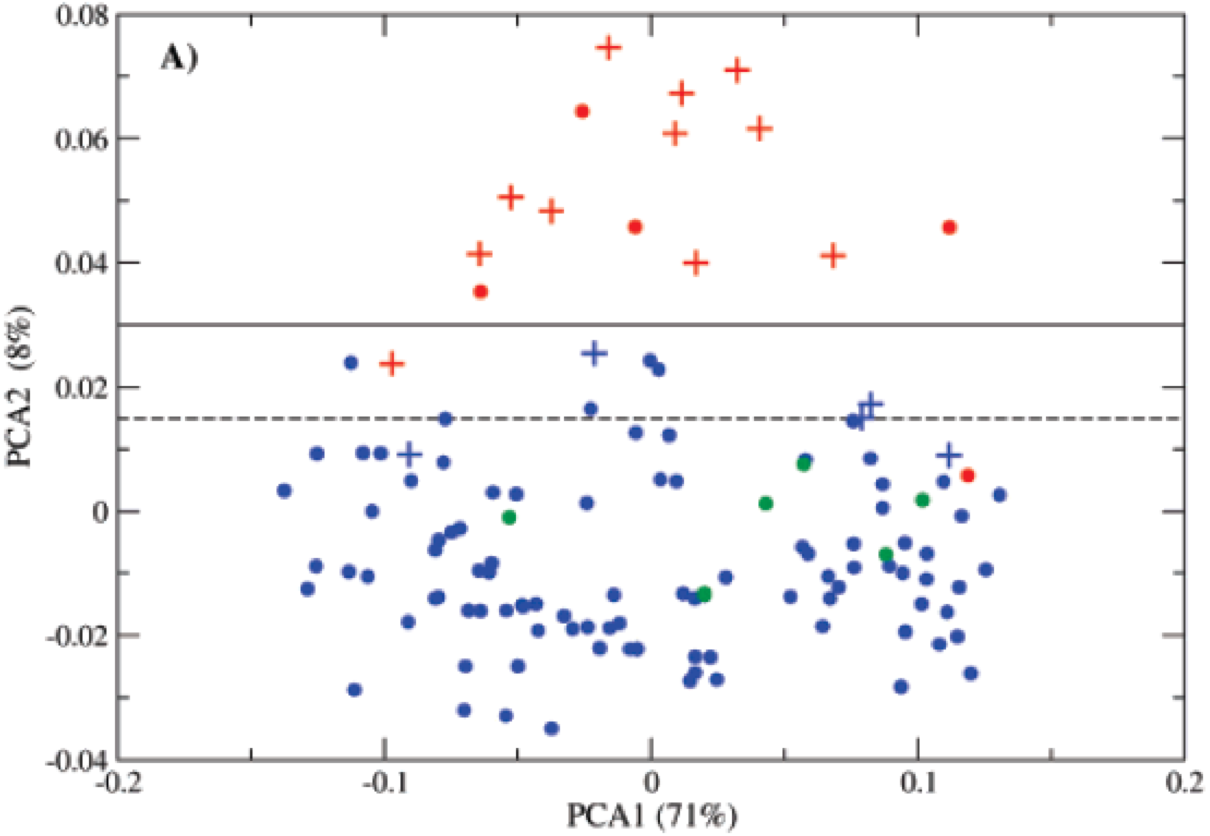
Thermophilic

Green:

Intermediates

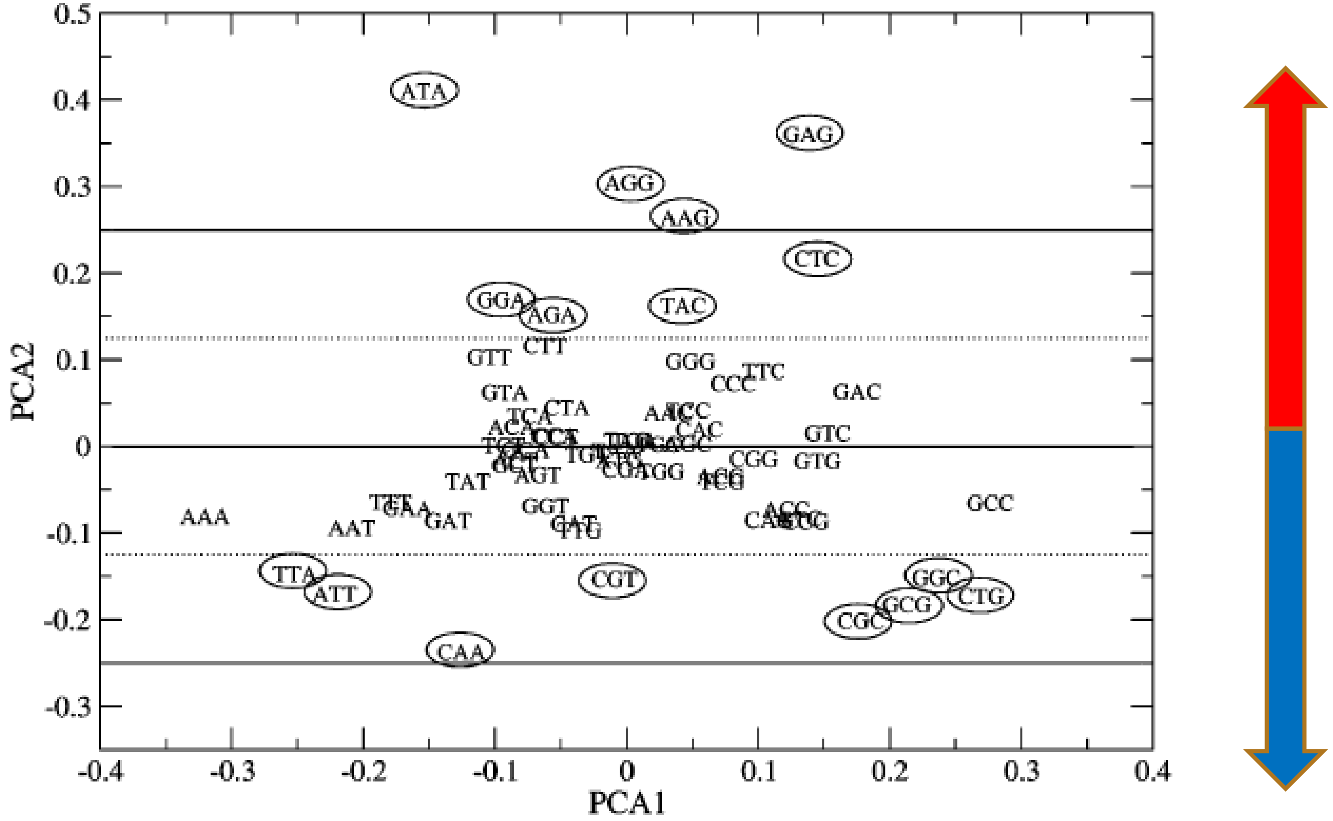
Blue:

Mesophilic

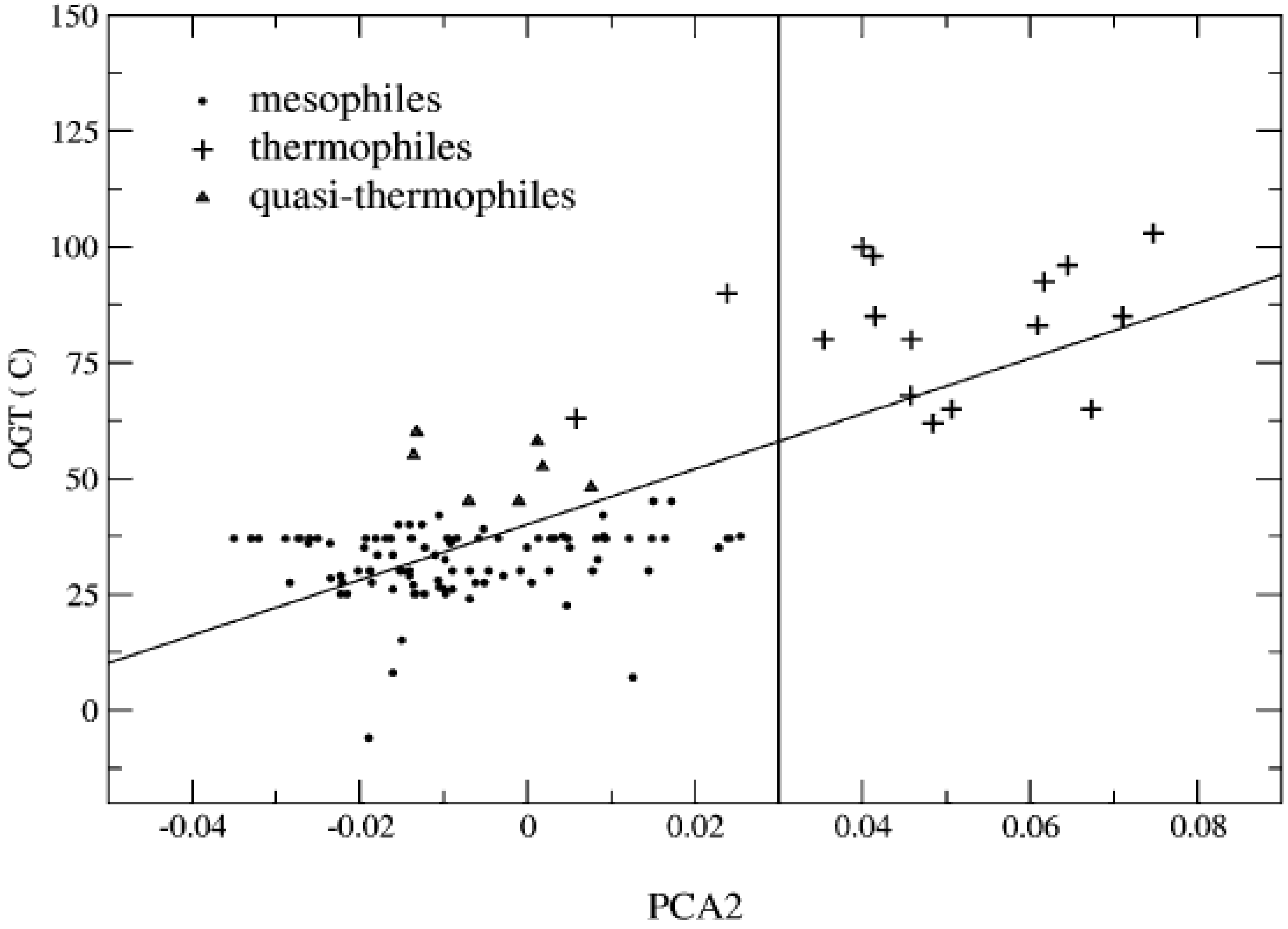


Components of 1st and 2nd PC (codon composition)

- First two components expressed as a function of the codons



Relation between 2nd component and OGT



Pitfalls of PCA

- A method such as PCA **assumes** that there are **linear relationships** between the derived components and the original variables. This is apparent from the role of the covariance or correlation matrix. If the relationships are non linear a Pearson correlation coefficient, which measures the strength of linear relationships, would underestimate the strength of the non-linear relationship.