CSE 332
Introduction to Visualization

Dimension reduction

Klaus Mueller

Computer Science Department
Stony Brook University

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Intro continued	
3	Applications of visual analytics, data, and basic tasks	
4	Data preparation and reduction	
5	Data reduction and similarity metrics	Project 1 out
6	Dimension reduction	
7	Introduction to D3	Project 2 out
8	Bias in visualization	
9	Perception and cognition	
10	Visual design and aesthetics	
11	Cluster and pattern analysis	
12	High-Dimensional data visualization: linear methods	
13	High-D data vis.: non-linear methods, categorical data	Project 3 out
14	Principles of interaction	
15	Visual analytics and the visual sense making process	
16	VA design and evaluation	
17	Visualization of graphs and hierarchies	
18	Visualization of time-varying and time-series data	Project 4 out
19	Midterm	
20	Maps and geo-vis	
21	Computer graphics and volume rendering	
22	Techniques to visualize spatial (3D) data	Project 4 halfway report due
23	Scientific and medical visualization	
24	Scientific and medical visualization	
25	Non-photorealistic rendering	
26	Memorable visualizations, visual embellishments	Project 5 out
27	Infographics design	
28	Projects Hall of Fame demos	

Last Lecture's Theme

Data Reduction

This Lecture's Theme

3D

2D

Dimension Reduction

Measure of Attribute Similarity

Are there attributes that "go together"?

Can you name a few?

Feature Vector (1)

Physical attributes

- color
- number of doors
- number of wheels
- retractable roof
- height
- length
- frames around side windows

Which attributes are useful to distinguish SUVs from convertibles?

- number of doors (4 vs. 2) --> numerical, two levels
- retractable roof (no vs. yes) --> categorical, two levels
- frames around side windows (yes vs. no) --> categorical, two levels
- height (higher vs. lower) --> numerical, many levels

Feature Vector (2)

Which attributes are not so useful?

- number of wheels (constant 4) --> no discriminative power
- length (short and long SUVs, convertibles) --> confounding
- color (colors are seemingly random, or are they?)

Is color useful?

- the convertibles seem to have more vibrant colors (red, yellow, …)
- so maybe we made a discovery

Attribute Space

Need to consider more than two attributes

- *height* attribute would have distinguished the Range Rover from the convertibles and caused it to be an outlier

Attribute Space

New classes are constantly evolving over time

- this is known as *cluster evolution*
- measuring more features will increase the chance of discovery

How Many Data Do We Need?

The more data (examples) the better

- increases the chances to discover the rare specimen

- but some attributes are useless
- we can cull them away
- perform attribute reduction or *dimension reduction*

Dimensionality Reduction

By axis rotation (linear methods)

- determine a more efficient basis
- Principal Component Analysis (PCA)
- Singular value decomposition (SVD)
- Latent semantic analysis (LSA)

By transformation (non-linear methods)

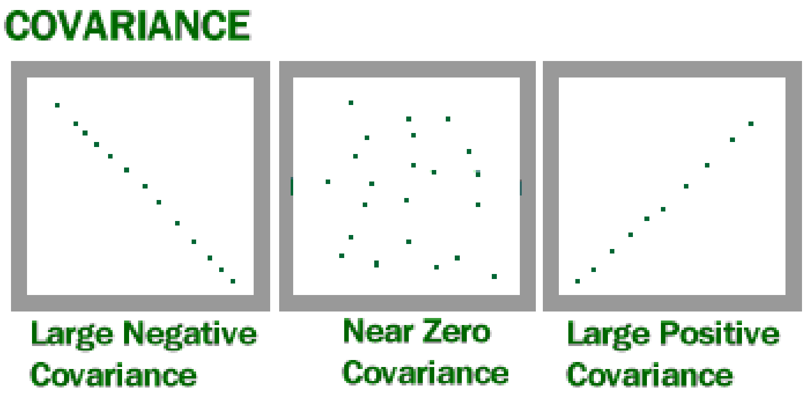
- determine a more efficient data type
- Fourier analysis and Wavelets for grids
- Multidimensional scaling (MDS) for graphs
- Locally Linear Embedding
- Isomap
- Self Organizing Maps (SOM)
- Linear Discriminant Analysis (LDA)

Principal Component Analysis (PCA)

Some Theory is Needed

Covariance

- measures how much two random variables change together



For N variable we have N^2 variable pairs

- we can write them in a matrix of size N^2 → the *covariance matrix*
- for two variables X_1 and X_2

$$\mathrm{Var}[X] = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Var}[X_2] \end{bmatrix}$$

Formulae

Covariance cov(X,Y)

mean of all data item values x_i and y_i for attributes X and Y, resp.

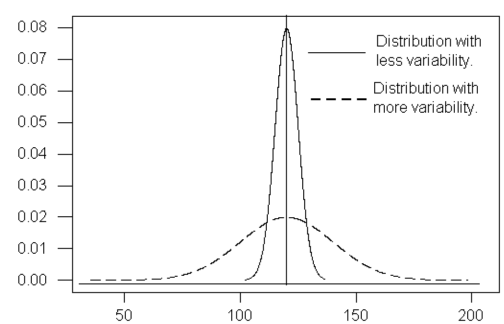
$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

Pearson's correlation r

- is covariance normalized by the individual variances for X and Y

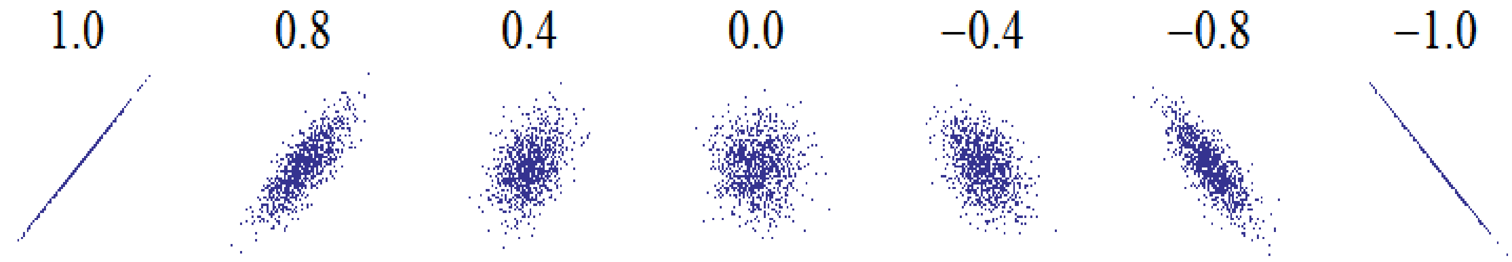
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{x})^2}}$$

individual variances for attributes X and Y



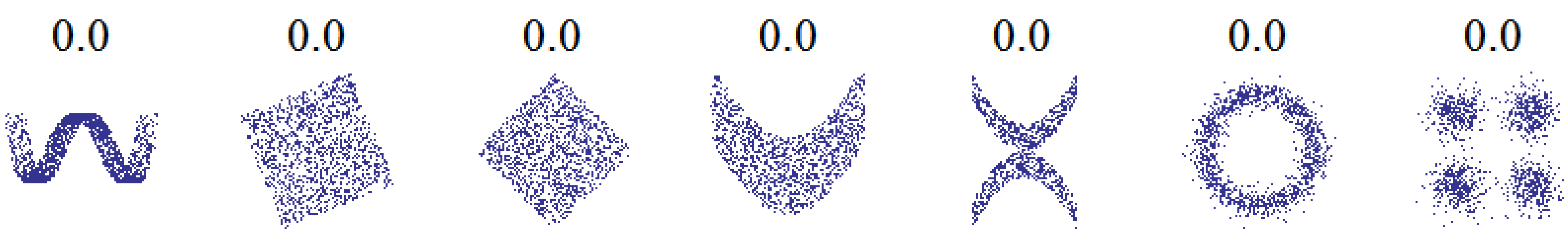
Correlation Patterns

Correlation rates between -1 and 1:



Important to note:

- correlation is defined for linear relationships
- visualization can help
- none of these point distributions have correlations:



Covariance Matrix

Analytical: $$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$

Samples: $$\sigma_{xy} = \mathrm{cov}_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

An n-D dataset has *n* variables $x_1, x_2, \ldots x_n$

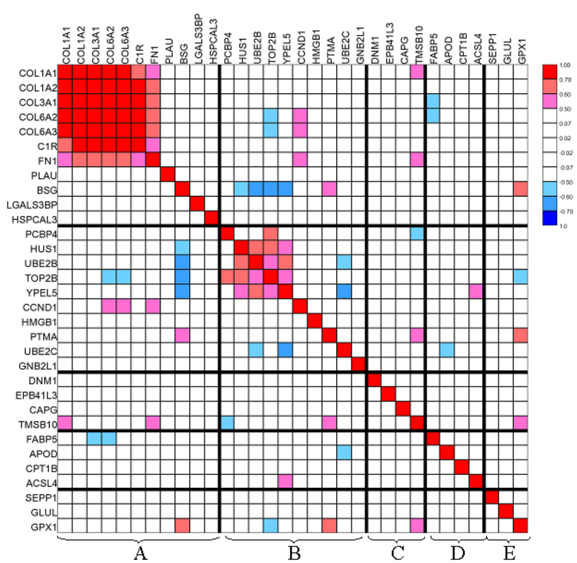
- define pairwise covariance among all of these variables
- construct a covariance matrix

$$\Sigma = \mathrm{Cov}(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

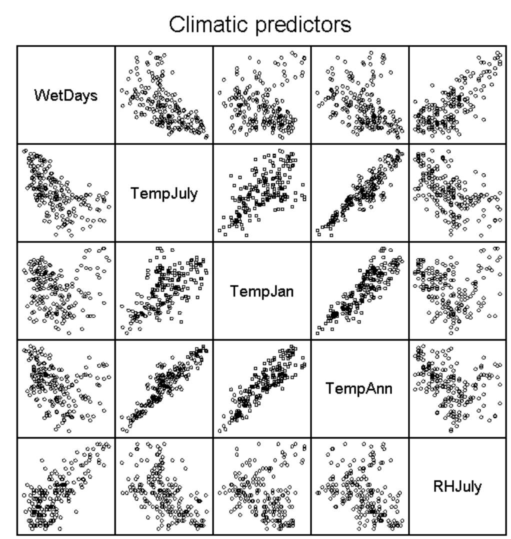
- a correlation matrix would just list the correlations instead

Correlation Matrix

	MO	FP	MP	IM	IC	FM	FE	FI	SPC	DSC	DST
MO	1.00										
FP	0.31[a]	1.00									
MP	0.32[a]	0.71[a]	1.00								
IM	0.36[a]	0.12[c]	0.14[c]	1.00							
IC	0.39[a]	0.18[b]	0.21[a]	0.62[a]	1.00						
FM	0.26[a]	0.21[a]	0.14[c]	0.30[a]	0.27[a]	1.00					
FE	0.47[a]	0.21[a]	0.18[b]	0.38[a]	0.28[a]	0.24[a]	1.00				
FI	0.53[a]	0.26[a]	0.22[a]	0.36[a]	0.37[a]	0.29[a]	0.47[a]	1.00			
SPC	0.32[a]	0.22[a]	0.31[a]	0.51[a]	0.47[a]	0.32[a]	0.37[a]	0.35[a]	1.00		
DSC	− 0.12[c]	0.03[c]	0.05[c]	0.17[b]	0.08[c]	0.18[b]	− 0.05[c]	0.06[c]	0.01[c]	1.00	
DST	− 0.02[c]	− 0.01[c]	0.05[c]	0.24[a]	0.14[c]	0.05[c]	− 0.05[c]	0.05[c]	0.05[c]	0.56[a]	1.00
DM	0.05[c]	0.144	0.136[c]	0.199[a]	0.169[b]	0.247[a]	0.08[c]	0.11[c]	0.14[c]	0.46[a]	0.71[a]



just value

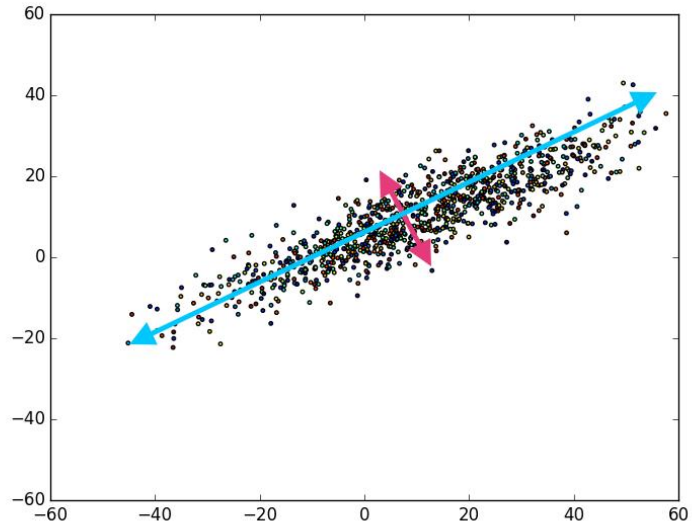


distribution (scatterplot matrix)

Principal Component Analysis

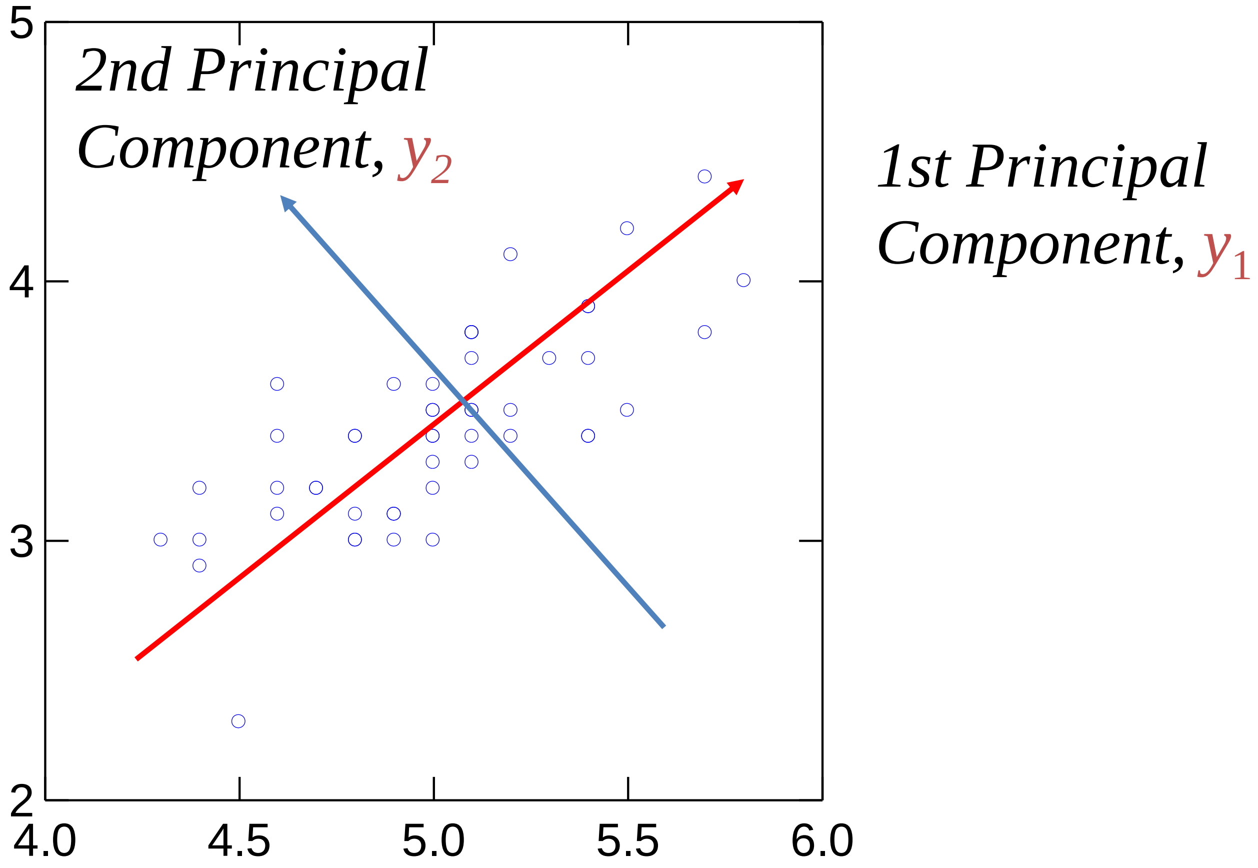
Ultimate goal:

- find a coordinate system that can represent the variance in the data with as few axes as possible



- rank these axes by the amount of variance (blue, red)
- drop the axes that have the least variance (red)

Principal Components



PCA – How To Do

Find the principal components (factors) of a distribution

First characterize the distribution by

- covariance matrix Cov
- correlation matrix Corr
- lets call it C

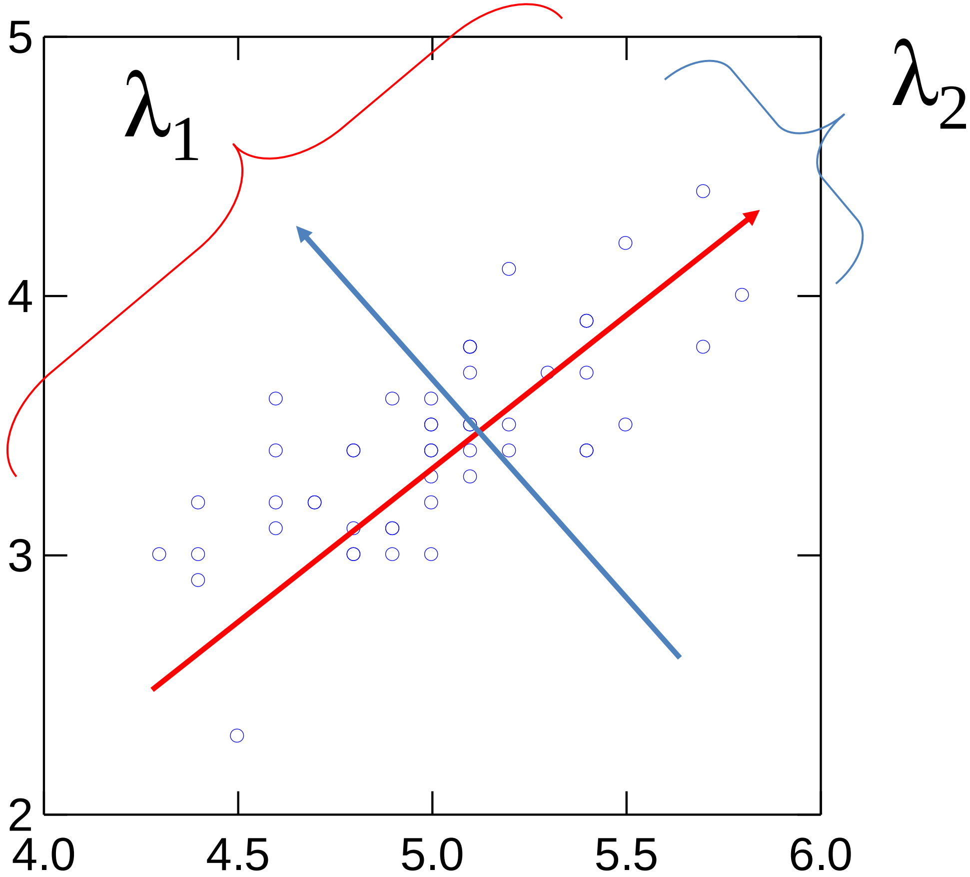
- perform QR factorization or LU decomposition to get

$$\boldsymbol{C} = Q \Lambda Q^{-1}$$

Q: matrix with Eigenvectors
Λ: diagonal matrix with Eigenvalues λ

- now order the Eigenvectors in terms of their Eigenvalues λ

Eigenvectors and values



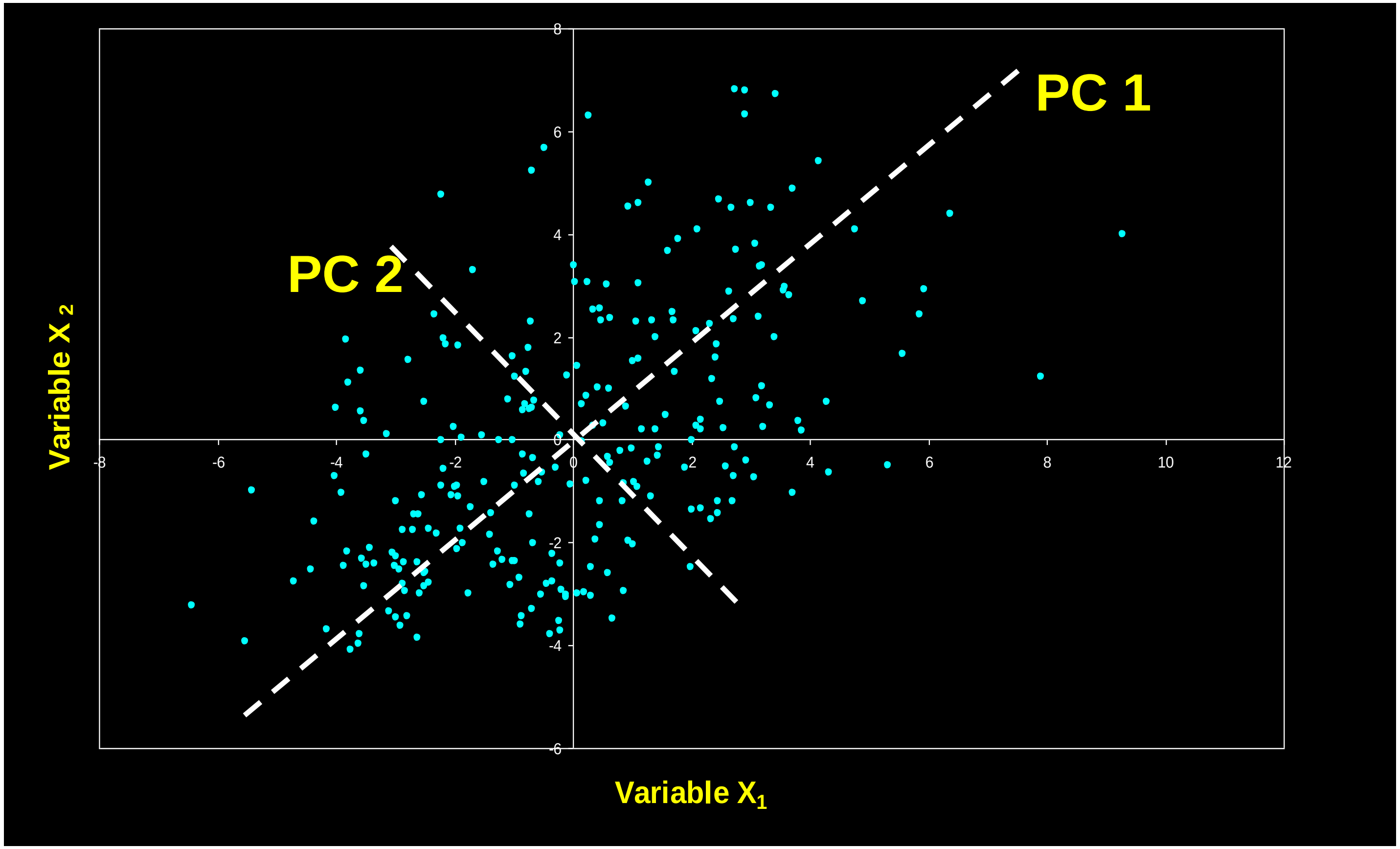
Covariance vs. Correlation

When to use what?

- use the covariance matrix when the variable scales are similar
- use the correlation matrix when the variables are on different scales
- the correlation matrix *standardizes* the data
- in general they give different results, especially when the scales are different

Example

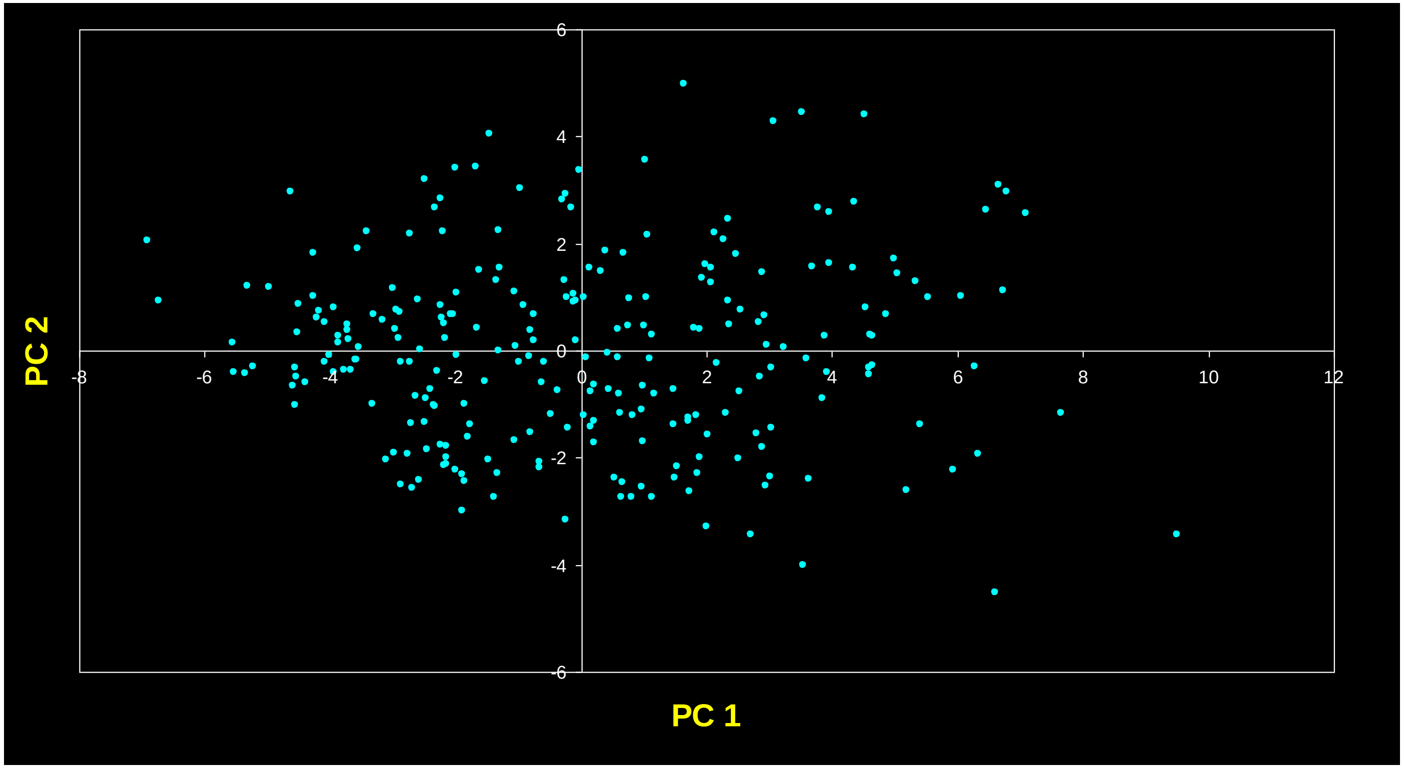
Before PCA



Example

After PCA

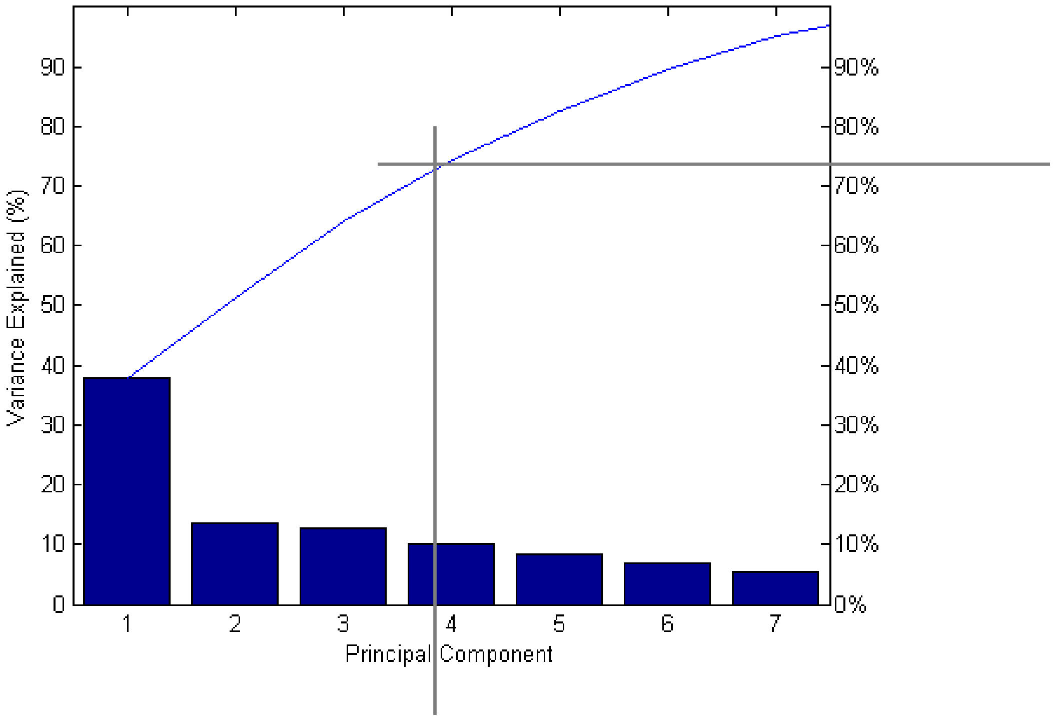
- $\lambda_1 = 9.8783$ $\lambda_2 = 3.0308$ Trace = 12.9091
- PC 1 displays ("explains") 9.8783/12.9091 = 76.5% of total variance



How Many Dimensions to Keep?

Create a *scree plot*

- plots a histogram of the Eigenvalues ordered by magnitude
- plots the explained variance as a curve



possible threshold (explain 75% of data variance)

keep top 3 principal components → reduce dimensions by a factor of 4/7 = 57%

PCA Applied To Faces

Take a set of faces:

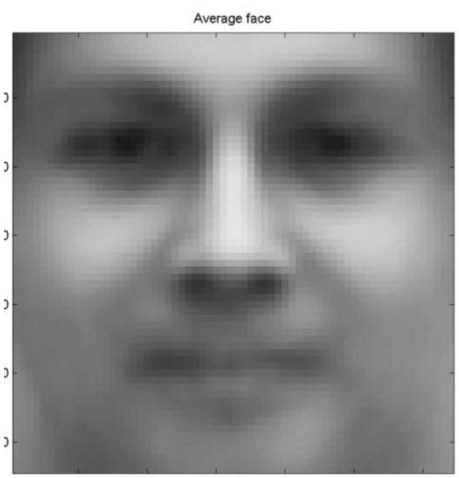
- each image has 60x60 pixels
- can write it as a 60x60 D = 3,600 D vector
- space of images is therefore 3600 D
- each image is a point in that space

Perform PCA

- will yield 3,600 Eigenvectors in 3,600 D space
- each is a face
- called "Eigenfaces"

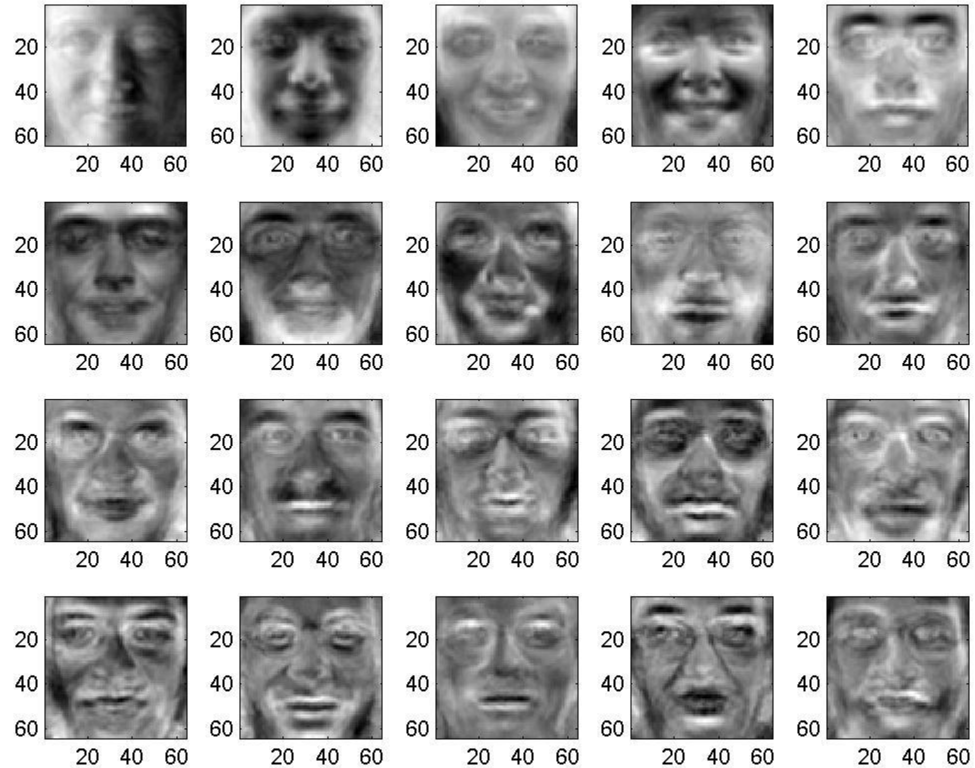
PCA Applied To Faces

We can reconstruct a face as a linear combination of these Eigenfaces [M. Turk and A. Pentland (1991)]



Average Face

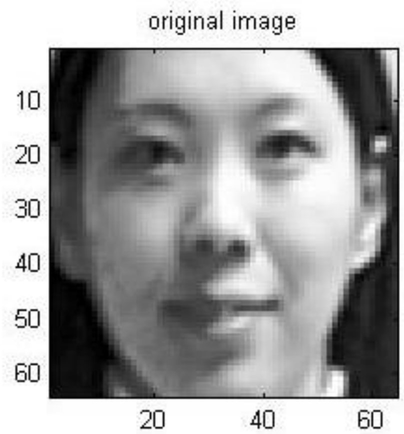
+

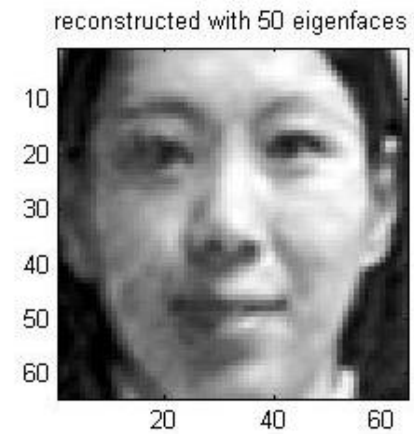


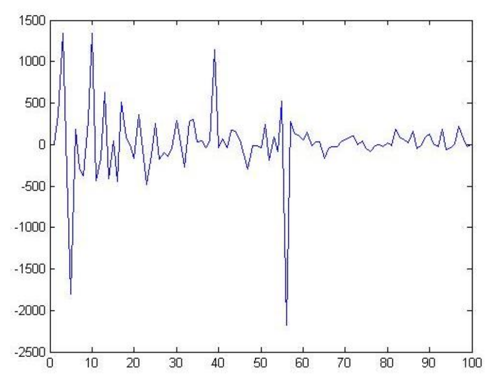
Eigenfaces

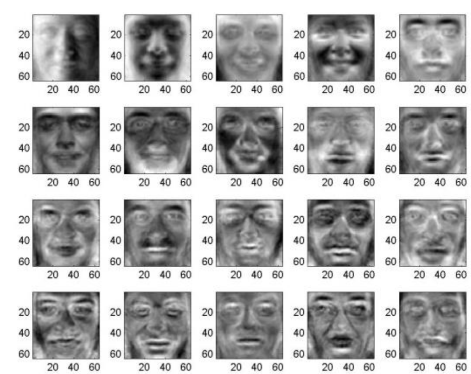
Reconstruction Using PCA

- 90% variance is captured by the first 50 eigenvectors
- Reconstruct existing faces using only 50 basis images
- We can also generate new faces by combining eigenvectors with different weights









Problem With PCA

The axes of the space generated by PCA do not mean much semantically

- the Eigenvectors are combinations of the actual data dimensions
- can we use these to determine the most important data dimensions which would be more meaningful?
- we shall explain it via an example
- see next slides

A More Challenging Example

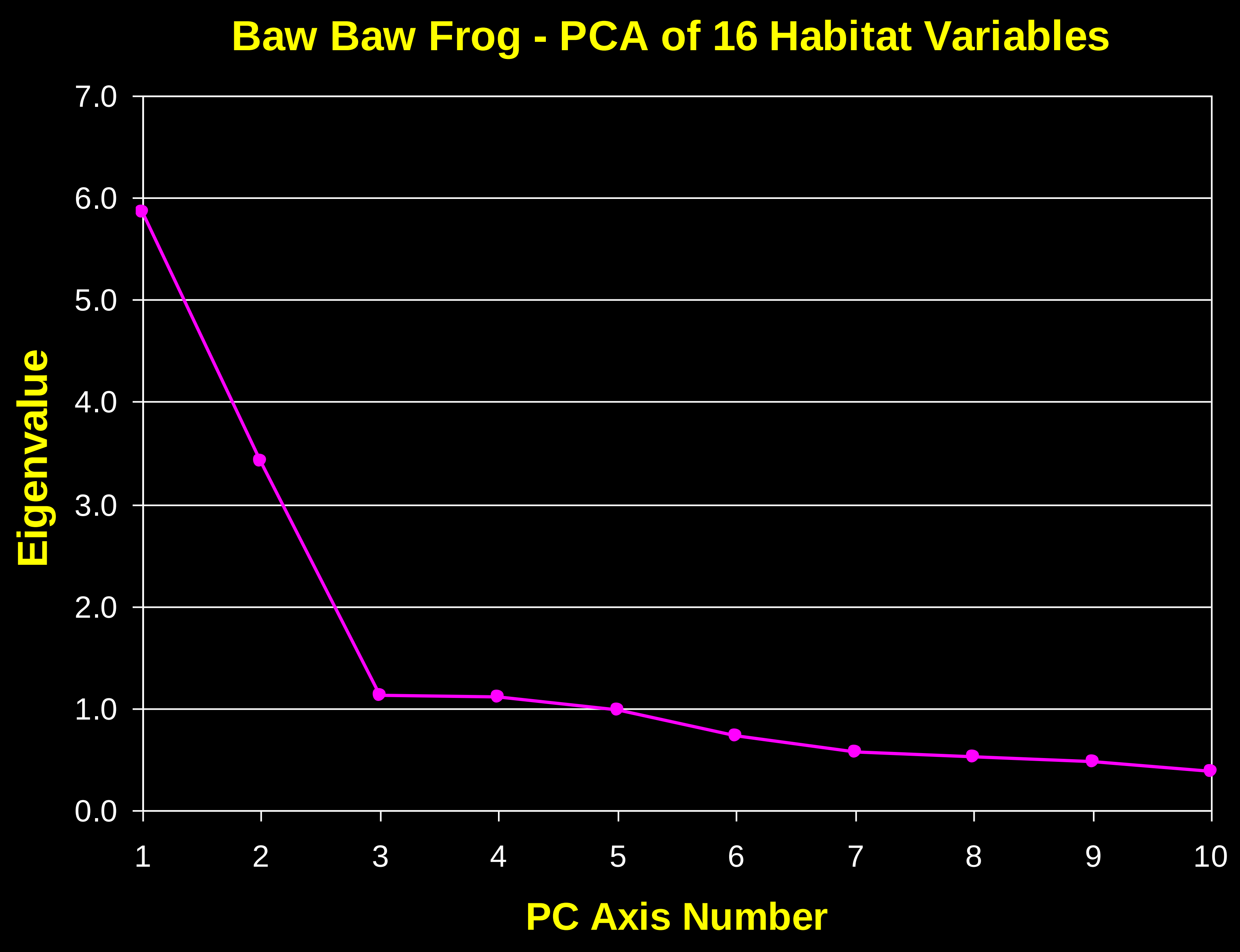
- Data from research on habitat definition of the endangered Baw Baw frog
- 16 environmental and structural variables measured at each of 124 sites
- Correlation matrix used because variables have different units

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

How Many Axes Are Needed?

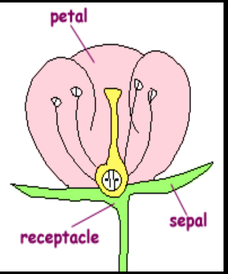
- Does the $(k+1)^{th}$ principal axis represent more variance than would be expected by chance?
- Several tests and rules have been proposed
- A common "rule of thumb" when PCA is based on correlations is that axes with eigenvalues > 1 are worth interpreting
- In our example 4 Eigenvectors fit this criterion (we shall keep 3 for simplicity)

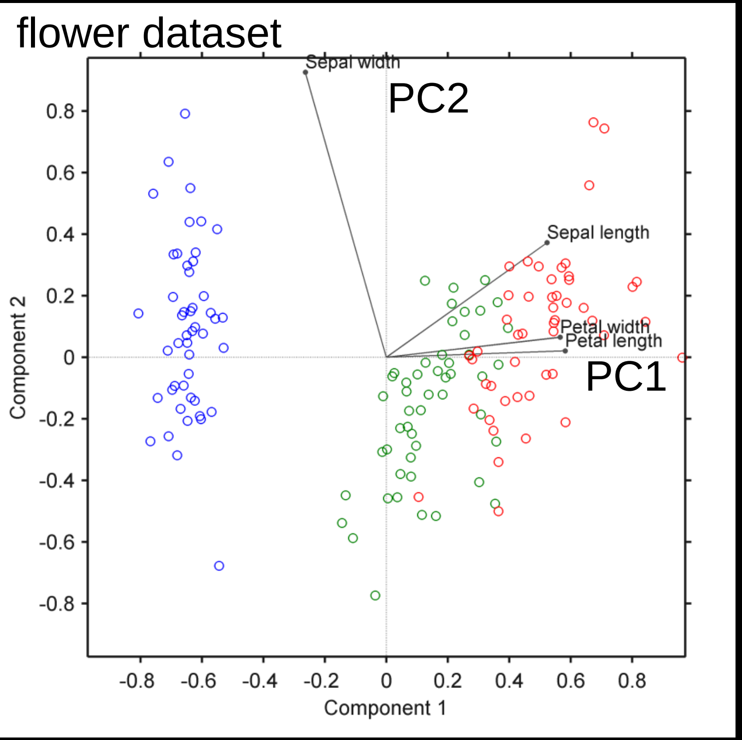

Baw Baw Frog - PCA of 16 Habitat Variables
Eigenvalue
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
1
2
3
4
5
6
7
8
9
10
PC Axis Number

Interlude – What's a "Loading"?

- The amount of weight a data dimension has on a principal component
  - petal length/width have a high loading on PC1
  - sepal width has a high loading on PC2

- Another observation
  - projection into PC basis can also bring out clusters better
  - since spread is maximized





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
- The loadings of variables on the first three PCs are shown here

	PC 1	PC 2	PC 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

Significance of Variables

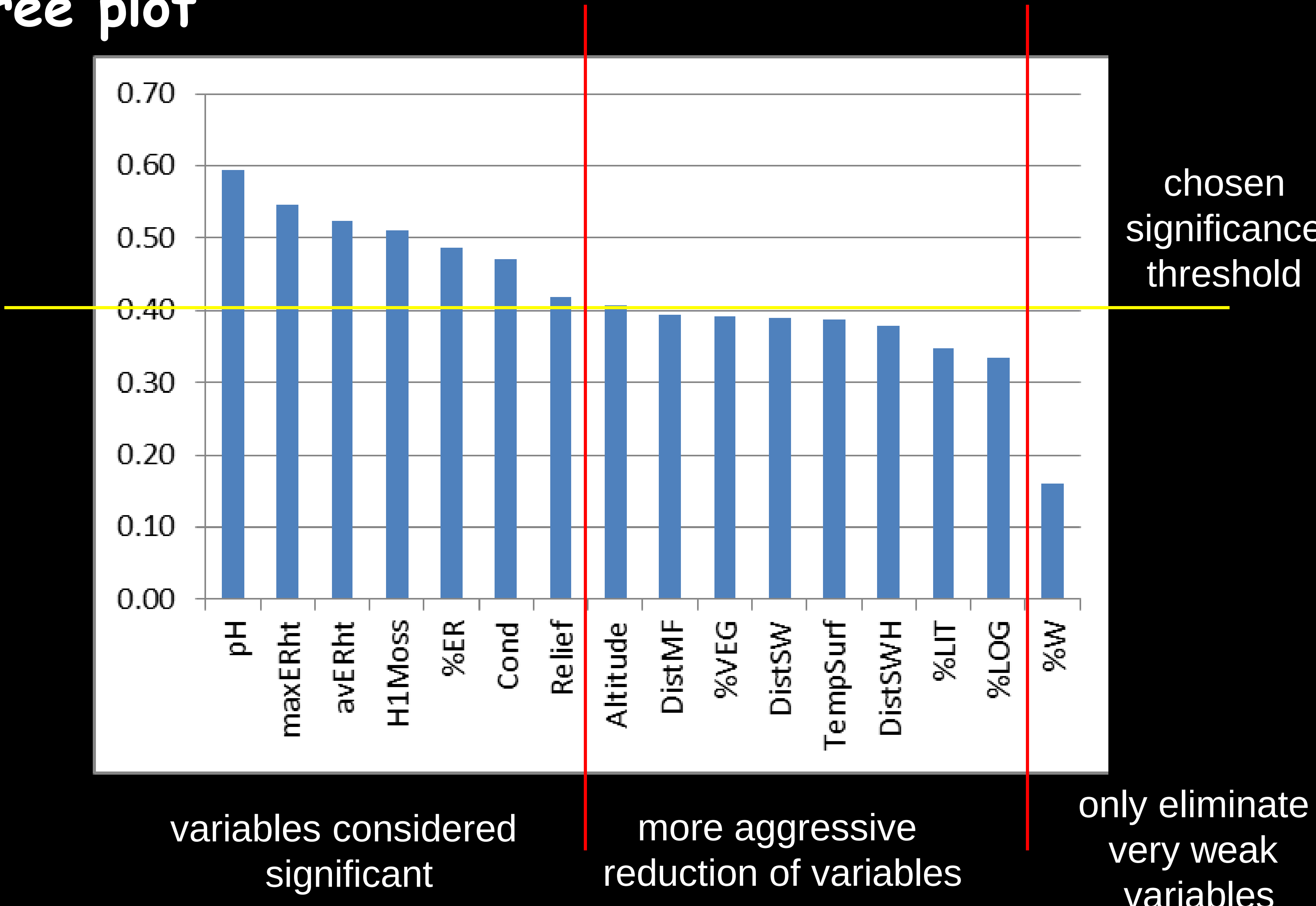
- We can compute the significance of the variables as the sum of squared loadings on to the most significant Eigenvectors we selected (3 in our example)
- The next slide shows the table of the last slide expanded with these squared loadings
- We can then sort the table by the squared loadings and make a scree plot
- The most significant variables are those above some chosen cutoff, for example 0.4 (marked in yellow in the table)

Significance of Variables

	PC 1	PC 2	PC 3	sum of squared loadings (sqrt)
Altitude	0.3842	0.0659	-0.1177	0.41
pH	-0.1159	0.1696	-0.5578	0.59
Cond	-0.2729	-0.1200	0.3636	0.47
TempSurf	0.0538	-0.2800	0.2621	0.39
Relief	-0.0765	0.3855	-0.1462	0.42
maxERht	0.0248	0.4879	0.2426	0.55
avERht	0.0599	0.4568	0.2497	0.52
%ER	0.0789	0.4223	0.2278	0.49
%VEG	0.3305	-0.2087	-0.0276	0.39
%LIT	-0.3053	0.1226	0.1145	0.35
%LOG	-0.3144	0.0402	-0.1067	0.33
%W	-0.0886	-0.0654	-0.1171	0.16
H1Moss	0.1364	-0.1262	0.4761	0.51
DistSWH	-0.3787	0.0101	0.0042	0.38
DistSW	-0.3494	-0.1283	0.1166	0.39
DistMF	0.3899	0.0586	-0.0175	0.39

Significance of Variables

- **Scree plot**



SUMMARY

Learned about:

- feature vectors, each feature is a data attribute, dimension
- distinguish useful from not so useful features with regards to data discrimination → dimension reduction
- plot data into feature space and observe clusters

- correlation vs. covariance
- algorithmic dimension reduction, summary of popular dimension reduction schemes – linear vs. non-linear

- basic linear scheme: Principal Component Analysis (PCA)
- application of PCA to face detection and generation
- scree plot to visualize and select the most important PCA axes
- use of PCA loading analysis to determine the most significant data features