Correlations, Principal Component Analysis
Correlation is “normalized covariance”

• Also called:
  Pearson correlation coefficient

\[ \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \]

is the covariance normalized to be \[-1 \leq \rho_{XY} \leq 1\]

Karl Pearson (1852–1936)
English mathematician and biostatistician
Spearman rank correlation

• **Pearson correlation** tests for **linear relationship** between X and Y

• **Unlikely** for variables with **broad distributions** → **non-linear effects** dominate

• **Spearman correlation** tests for any **monotonic relationship** between X and Y

• Calculate ranks (1 to n), \( r_X(i) \) and \( r_Y(i) \) of variables in both samples. Calculate Pearson correlation between ranks: \( \text{Spearman}(X,Y) = \text{Pearson}(r_X, r_Y) \)

• **Ties**: convert to fractions, e.g. tie for 6s and 7s place both get 6.5. This can lead to artefacts.

• If lots of ties: use **Kendall rank correlation** (Kendall tau)
Matlab exercise: Correlation/Covariation

• Generate a sample with \texttt{Stats=100,000} of two Gaussian random variables \( r_1 \) and \( r_2 \) which have mean 0 and standard deviation 2 and are:
  
  – Uncorrelated
  – Correlated with correlation coefficient 0.9
  – Correlated with correlation coefficient -0.5
  – Trick: first make uncorrelated \( r_1 \) and \( r_2 \). Then make anew variable: \( r_{1\text{mix}} = \text{mix} \cdot r_2 + (1-\text{mix}^2)^{0.5} \cdot r_1 \);
    where \( \text{mix} = \text{corr. coeff.} \)

• For each value of \( \text{mix} \) calculate covariance and correlation coefficient between \( r_{1\text{mix}} \) and \( r_2 \)

• In each case make a scatter plot: \texttt{plot(r1mix,r2,’k.’);}
Matlab exercise: Correlation/Covariation

1. Stats=100000;
2. r1=2.*randn(Stats,1);
3. r2=2.*randn(Stats,1);
4. disp('Covariance matrix='); disp(cov(r1,r2));
5. disp('Correlation='); disp(corr(r1,r2));
6. figure; plot(r1,r2,'k.');
7. mix=0.9; %Mixes r2 to r1 but keeps same variance
8. r1mix=mix.*r2+sqrt(1-mix.^2).*r1;
9. disp('Covariance matrix='); disp(cov(r1mix,r2));
10. disp('Correlation='); disp(corr(r1mix,r2));
11. figure; plot(r1mix,r2,'k.');
12. mix=-0.5; %REDO LINES 8-11
Linear Functions of Random Variables

• A function of multiple random variables is itself a random variable.

• A function of random variables can be formed by either linear or nonlinear relationships. We will only work with linear functions.

• Given random variables $X_1, X_2, \ldots, X_p$ and constants $c_1, c_2, \ldots, c_p$

  $Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$  \hspace{1cm} (5-24)

  is a linear combination of $X_1, X_2, \ldots, X_p$. 
Mean & Variance of a Linear Function

\[ Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p \]

\[ E(Y) = c_1E(X_1) + c_2E(X_2) + \ldots + c_pE(X_p) \quad (5-25) \]

\[ V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) + 2 \sum_{i<j} c_ic_j \text{cov}(X_iX_j) \quad (5-26) \]

If \( X_1, X_2, \ldots, X_p \) are independent, then \( \text{cov}(X_iX_j) = 0 \),

\[ V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) \quad (5-27) \]
Example 5-31: Error Propagation

A semiconductor product consists of three layers. The variances of the thickness of each layer is 25, 40 and 30 nm$^2$. What is the variance of the finished product?

Answer:

$X = X_1 + X_2 + X_3$

$V(X) = \sum_{i=1}^{3} V(X_i) = 25 + 40 + 30 = 95$ nm$^2$

$SD(X) = \sqrt{95} = 9.7$ nm

If adding SDs one would get $\sqrt{25}$ nm + $\sqrt{40}$ nm + $\sqrt{30}$ nm = 16.08 nm
IMPORTANT:

\( p \) independent identically distributed (i.i.d) variables

Average \( \bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_p}{p} \)

\[
E(\bar{X}) = \frac{p \cdot E(X)}{p} = \frac{p \cdot \mu}{p} = \mu
\]

\[
V(\bar{X}) = \frac{p \cdot V(X)}{p^2} = \frac{p \cdot \sigma^2}{p^2} = \frac{\sigma^2}{p}
\]

Standard deviation \( \bar{X} = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{p}} \)
Mean & Variance of an Average

If \( \bar{X} = \left( \frac{X_1 + X_2 + \ldots + X_p}{p} \right) \) and \( E(X_i) = \mu \)

Then \( E(\bar{X}) = \frac{p \cdot \mu}{p} = \mu \) \hspace{1cm} (5-28a)

If the \( X_i \) are independent with \( V(X_i) = \sigma^2 \)

Then \( V(\bar{X}) = \frac{p \cdot \sigma^2}{p^2} = \frac{\sigma^2}{p} \) \hspace{1cm} (5-28b)
Principal Component Analysis (PCA)
Suppose we have a population measured on $p$ random variables $X_1, \ldots, X_p$. Note that these random variables represent the $p$-axes of the Cartesian coordinate system in which the population resides. Our goal is to develop a new set of $p$ axes (linear combinations of the original $p$ axes) in the directions of greatest variability:

This is accomplished by rotating the axes.

Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU
PCA Scores

Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU
PCA Eigenvalues and Eigenvectors

Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU
PCA: General

From $p$ original variables: $x_1, x_2, ..., x_p$: I need to produce $p$ new variables: $y_1, y_2, ..., y_p$: 

$y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$
$y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p$
...
$y_p = a_{p1}x_1 + a_{p2}x_2 + ... + a_{pp}x_p$

such that:

$y_k$'s are uncorrelated (orthogonal)
$y_1$ explains as much as possible of original variance in data set
$y_2$ explains as much as possible of remaining variance etc.

Answer: PCA diagonalize the $p \times p$ symmetric matrix of corr. coefficients 

$$r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU
Choosing the Dimension $K$

- How many eigenvectors to use?
- Look at the decay of the eigenvalues
  - the eigenvalue tells you the amount of variance “in the direction” of that eigenvector
  - ignore eigenvectors with low variance

Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU
Applications of PCA

- **Uses:**
  - Data Visualization
  - Dimensionality Reduction
  - Data Classification

- **Examples:**
  - How to best present what is “interesting”?
  - How many unique subsets (clusters, modules) are there in the sample?
  - How are they similar / different
  - What are the underlying factors that most influence the samples?
  - Which measurements are best to differentiate between samples?
  - Which subset does this new sample rightfully belong?

Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU
Let’s work with real cancer data!

- Data from Wolberg, Street, and Mangasarian (1994)
- Fine-needle aspirates = biopsy for breast cancer
- Black dots – cell nuclei. Irregular shapes/sizes may mean cancer
- 212 cancer patients and 357 healthy individuals (column 1)
- 30 other properties (see table)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Col 2</th>
<th>Col 12</th>
<th>Col 22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radius (average distance from the center)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texture (standard deviation of gray-scale values)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perimeter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoothness (local variation in radius lengths)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compactness (perimeter² / area - 1.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concavity (severity of concave portions of the contour)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concave points (number of concave portions of the contour)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetry</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fractal dimension (“coastline approximation” - 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Matlab exercise 1

• Download cancer data in cancer_wdbc.mat

• Data in the table X (569x30). First 357 patients are healthy. The remaining 569-357=212 patients have cancer.

• Calculate the correlation matrix of all-against-all variables: 30*29/2=435 correlations. Hint: look at the help page for corr

• Visualize 30x30 table of correlations using pcolor

• Plot the histogram of these 435 correlation coefficients