Joint Probability Distributions, Correlations
What we learned so far...

- **Random Events:**
  - Working with **events as sets:** union, intersection, etc.
    - Some events are simple: Head vs Tails, Cancer vs Healthy
    - Some are more complex: $10<\text{Gene expression}<100$
    - Some are even more complex: Series of dice rolls: $1,3,5,3,2$
  - Conditional probability: $P(A|B)=P(A \cap B)/P(B)$
  - Independent events: $P(A|B)=P(A)$ or $P(A \cap B)=P(A)*P(B)$
  - Bayes theorem: relates $P(A|B)$ to $P(B|A)$
- **Random variables:**
  - Mean, Variance, Standard deviation. How to work with $E(g(X))$
  - Discrete (Uniform, Bernoulli, Binomial, Poisson, Geometric, Negative binomial, Hypergeometric, Power law);
    - PMF: $f(x)=\text{Prob}(X=x)$; CDF: $F(x)=\text{Prob}(X\leq x)$;
  - Continuous (Uniform, Exponential, Erlang, Gamma, Normal, Lognormal);
    - PDF: $f(x)$ such that $\text{Prob}(X \text{ inside } A) = \int_A f(x)dx$; CDF: $F(x)=\text{Prob}(X\leq x)$
- **Next step:** work with **multiple random variables** measured together in the same series of random experiments
Concept of Joint Probabilities

• Biological systems are usually described not by a single random variable but by many random variables

• Example: The expression state of a human cell: 20,000 random variables $X_i$ for each of its genes

• A joint probability distribution describes the behavior of several random variables

• We will start with just two random variables $X$ and $Y$ and generalize when necessary
Joint Probability Mass Function Defined

The joint probability mass function of the discrete random variables $X$ and $Y$, denoted as $f_{XY}(x, y)$, satisfies:

1. $f_{XY}(x, y) \geq 0$  
   All probabilities are non-negative

2. $\sum_{x} \sum_{y} f_{XY}(x, y) = 1$  
   The sum of all probabilities is 1

3. $f_{XY}(x, y) = P(X = x, Y = y)$  
   (5-1)
Example 5-1: # Repeats vs. Signal Bars

You use your cell phone to check your airline reservation. It asks you to speak the name of your departure city to the voice recognition system.

• Let $Y$ denote the number of times you have to state your departure city.
• Let $X$ denote the number of bars of signal strength on your cell phone.

<table>
<thead>
<tr>
<th>$y$ = number of times city name is stated</th>
<th>$x$ = number of bars of signal strength</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**Figure 5-1** Joint probability distribution of $X$ and $Y$. The table cells are the probabilities. Observe that more bars relate to less repeating.
Marginal Probability Distributions (discrete)

For a discrete joint PDF, there are marginal distributions for each random variable, formed by summing the joint PMF over the other variable.

\[
f_X(x) = \sum_y f_{XY}(x, y)
\]

\[
f_Y(y) = \sum_x f_{XY}(x, y)
\]

Called marginal because they are written in the margins.

<table>
<thead>
<tr>
<th>y = number of times city name is stated</th>
<th>x = number of bars of signal strength</th>
<th>( f_Y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
<td>0.10</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>0.10</td>
</tr>
</tbody>
</table>

\[f_X(x) = 0.20 \quad 0.25 \quad 0.55 \quad 1.00\]

Figure 5-6 From the prior example, the joint PMF is shown in green while the two marginal PMFs are shown in purple.
Mean & Variance of X and Y are calculated using marginal distributions

<table>
<thead>
<tr>
<th>y = number of times city name is stated</th>
<th>x = number of bars of signal strength</th>
<th>f(y) = y*f(y) =</th>
<th>y^2*f(y) =</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01 0.02 0.25</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>0.02 0.03 0.20</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>0.02 0.10 0.05</td>
<td>0.17</td>
<td>0.51</td>
</tr>
<tr>
<td>4</td>
<td>0.15 0.10 0.05</td>
<td>0.30</td>
<td>1.20</td>
</tr>
</tbody>
</table>

| f(x) = 0.20 0.25 0.55                  | 1.00 2.49 7.61                     |
| x*f(x) = 0.20 0.50 1.65               | 2.35                               |
| x^2*f(x) = 0.20 1.00 4.95             | 6.15                               |

\[ \mu_X = E(X) = 2.35; \quad \sigma_X^2 = V(X) = 6.15 - 2.35^2 = 6.15 - 5.52 = 0.6275 \]

\[ \mu_Y = E(Y) = 2.49; \quad \sigma_Y^2 = V(Y) = 7.61 - 2.49^2 = 7.61 - 16.20 = 1.4099 \]
Conditional Probability Distributions

Recall that \( P(B|A) = \frac{P(A \cap B)}{P(A)} \)

\[
P(Y=y|X=x) = \frac{P(X=x,Y=y)}{P(X=x)} = \frac{f(x,y)}{f_X(x)}
\]

From Example 5-1

\[
P(Y=1|X=3) = \frac{0.25}{0.55} = 0.455
\]

\[
P(Y=2|X=3) = \frac{0.20}{0.55} = 0.364
\]

\[
P(Y=3|X=3) = \frac{0.05}{0.55} = 0.091
\]

\[
P(Y=4|X=3) = \frac{0.05}{0.55} = 0.091
\]

Sum = 1.00

Note that there are 12 probabilities conditional on \( X \), and 12 more probabilities conditional upon \( Y \).
Joint Random Variable Independence

• Random variable independence means that knowledge of the value of $X$ does not change any of the probabilities associated with the values of $Y$.

• Opposite: Dependence implies that the values of $X$ are influenced by the values of $Y$
Independence for Discrete Random Variables

• Remember independence of events (slide 13 lecture 4): Events are independent if any one of the three conditions are met:
  1) \( P(A | B) = P(A \cap B) / P(B) = P(A) \) or
  2) \( P(B | A) = P(A \cap B) / P(A) = P(B) \) or
  3) \( P(A \cap B) = P(A) \cdot P(B) \)

• Random variables independent if all events \( A \) that \( Y=y \) and \( B \) that \( X=x \) are independent if any one of these conditions is met:
  1) \( P(Y=y | X=x) = P(Y=y) \) for any \( x \) or
  2) \( P(X=x | Y=y) = P(X=x) \) for any \( y \) or
  3) \( P(X=x, Y=y) = P(X=x) \cdot P(Y=y) \)
    for every pair \( x \) and \( y \)
Joint Probability Density Function Defined

The **joint probability density function** for the continuous random variables $X$ and $Y$, denoted as $f_{XY}(x,y)$, satisfies the following properties:

1. $f_{XY}(x,y) \geq 0$ for all $x, y$

2. \[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) \, dx \, dy = 1 \]

3. \[ P((X,Y) \subseteq R) = \int_{R} f_{XY}(x,y) \, dx \, dy \quad \text{(5-2)} \]

**Figure 5-2** Joint probability density function for the random variables $X$ and $Y$. Probability that $(X, Y)$ is in the region $R$ is determined by the **volume** of $f_{XY}(x,y)$ over the region $R$. 

Sec 5-1.1 Joint Probability Distributions
Figure 5-3  Joint probability density function for the continuous random variables $X$ and $Y$ of expression levels of two different genes. Note the asymmetric, narrow ridge shape of the PDF – indicating that small values in the $X$ dimension are more likely to occur when small values in the $Y$ dimension occur.
Marginal Probability Distributions (continuous)

• Rather than summing a discrete joint PMF, we integrate a continuous joint PDF.
• The marginal PDFs are used to make probability statements about one variable.
• If the joint probability density function of random variables $X$ and $Y$ is $f_{XY}(x,y)$, the marginal probability density functions of $X$ and $Y$ are:

$$f_X(x) = \int_{y} f_{XY}(x,y) \, dy$$
$$f_Y(y) = \int_{x} f_{XY}(x,y) \, dx$$

$$f_X(x) = \sum_{y} f_{XY}(x,y)$$
$$f_Y(y) = \sum_{x} f_{XY}(x,y)$$

(5-3)
Conditional Probability Density Function Defined

Given continuous random variables $X$ and $Y$ with joint probability density function $f_{XY}(x,y)$, the conditional probability density function of $Y$ given $X=x$ is

$$f_{Y|X}(y) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int_y f_{XY}(x,y) \, dy} \quad \text{if} \quad f_X(x) > 0 \quad (5-4)$$

Compare to discrete: $P(Y=y|X=x) = f_{XY}(x,y)/f_X(x)$

which satisfies the following properties:

1. $f_{Y|X}(y) \geq 0$
2. $\int f_{Y|X}(y) \, dy = 1$
3. $P(Y \subseteq B|X = x) = \int f_{Y|X}(y) \, dy$ for any set $B$ in the range of $Y$
Conditional Probability Distributions

- Conditional probability distributions can be developed for multiple random variables by extension of the ideas used for two random variables.

- Suppose \( p = 5 \) and we wish to find the distribution of \( X_1, X_2 \) and \( X_3 \) conditional on \( X_4 = x_4 \) and \( X_5 = x_5 \).

\[
f_{X_1X_2X_3|X_4X_5}(x_1, x_2, x_3) = \frac{f_{X_1X_2X_3X_4X_5}(x_1, x_2, x_3, x_4, x_5)}{f_{X_4X_5}(x_4, x_5)}
\]

for \( f_{X_4X_5}(x_4, x_5) > 0 \).
Independence for Continuous Random Variables

For random variables $X$ and $Y$, if any one of the following properties is true, the others are also true. Then $X$ and $Y$ are independent. Then $X$ and $Y$ are independent.

\[
P(Y=y \mid X=x) = P(Y=y) \quad \text{for any } x \quad \text{or} \quad P(X=x \mid Y=y) = P(X=x) \quad \text{for any } y \quad \text{or} \quad P(X=x, Y=y) = P(X=x) \cdot P(Y=y) \quad \text{for any } x \text{ and } y
\]

(1) $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$

(2) $f_{Y\mid X}(y) = f_Y(y)$ for all $x$ and $y$ with $f_X(x) > 0$

(3) $f_{X\mid Y}(y) = f_X(x)$ for all $x$ and $y$ with $f_Y(y) > 0$

(4) $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$ for any sets $A$ and $B$ in the range of $X$ and $Y$, respectively.
Example 1:

Uniform distribution in the square $-1 < x < 1$, $-1 < y < 1$

\[
\begin{align*}
f_{x,y}(x, y) &= c \quad \text{if} \quad -1 < x < 1 \quad \text{and} \quad -1 < y < 1 \\
0 &\quad \text{otherwise}
\end{align*}
\]

\[
1 = \int_{-1}^{1} \int_{-1}^{1} f_{x,y}(x, y) \, dx \, dy = c \cdot \text{Area} = c \cdot 4 \quad \Rightarrow \quad c = \frac{1}{4}
\]
Are $X$ and $Y$ independent? Yes they are.

Let's test if \( f_{XY}(x, y) = f_X(x) \cdot f_Y(y) \)

\[
\int_{-\infty}^{\infty} \int_{-1}^{1} \frac{1}{4} \, dy = \frac{1}{2} \quad \text{if} \quad -1 < x < 1
\]

Same for \( f_Y(y) = \frac{1}{2} \quad \text{if} \quad -1 < y < 1 \)

\[
\frac{1}{4} = \int_{-\infty}^{\infty} \frac{1}{2} \cdot \frac{1}{2} = f_X(x) \cdot f_Y(y)
\]

0 otherwise if both $x$ and $y$ are in $[-1, 1]$
X and Y are uniformly distributed in the disc $x^2 + y^2 \leq 1$

Are they independent?

A. yes
B. no
C. I could not figure it out

Get your i-clickers
Covariation, Correlations
Covariance - A number to measure dependence between random variables

\[ \text{Cov}(X, Y) \text{ or } \sigma_{xy} \]

\[ \sigma_{xy} = E \left[ (X - \mu_x) \cdot (Y - \mu_y) \right] = \]

\[ = E(X,Y) - \mu_x \cdot \mu_y \]

- \( \text{Var}(X) = \text{Cov}(X, X) \)

- If \( X \& Y \) are independent
  \[ \text{Cov}(X, Y) = E[X - \mu_X] \cdot E[Y - \mu_Y] = 0 \]

- \( -\infty < \text{Cov}(X, Y) < +\infty \)
  
  Can be negative!
Covariance Defined

Covariance is a number quantifying average dependence between two random variables.

The covariance between the random variables $X$ and $Y$, denoted as $\text{cov}(X,Y)$ or $\sigma_{XY}$ is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y \quad (5-14)$$

The units of $\sigma_{XY}$ are units of $X$ times units of $Y$.

Unlike the range of variance, $-\infty < \sigma_{XY} < \infty$. 
Covariance and PMF tables

<table>
<thead>
<tr>
<th>$y =$ number of times city name is stated</th>
<th>$x =$ number of bars of signal strength</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
</tr>
</tbody>
</table>

The probability distribution of Example 5-1 is shown.

By inspection, note that the larger probabilities occur as $X$ and $Y$ move in opposite directions. This indicates a negative covariance.
Covariance and Scatter Patterns

Figure 5-13  Joint probability distributions and the sign of $\text{cov}(X, Y)$. Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.
Independence Implies $\sigma = \rho = 0$ but **not vice versa**

- If $X$ and $Y$ are independent random variables, 
  
  \[
  \sigma_{XY} = \rho_{XY} = 0 \quad (5-17)
  \]

- $\rho_{XY} = 0$ is necessary, but **not a sufficient condition** for independence.
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
Prove that $\rho_{xy}$ is in $[-1, 1]$

$Z_x = \frac{x - \mu_x}{\sigma_x}$ ; $Z_y = \frac{y - \mu_y}{\sigma_y}$

$0 \leq E((Z_x - Z_y)^2) = E(Z_x^2) + E(Z_y^2) - 2E(Z_x \cdot Z_y) = 2 - 2 \frac{1}{\sigma_x \sigma_y} E((X - \mu_x)(Y - \mu_y)) = 2 - 2 \rho_{xy} \implies \rho_{xy} \leq 1$

$0 \leq E\left((Z_x + Z_y)^2\right) = E(Z_x^2) + E(Z_y^2) + 2E(Z_x \cdot Z_y) = 2 + 2 \rho_{xy} \implies \rho_{xy} \geq -1$
Spearman rank correlation

- **Pearson correlation** tests for linear relationship between X and Y
- Unlikely for variables with broad distributions $\rightarrow$ 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 Stats=100,000 of two Gaussian random variables r1 and r2 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 r1 and r2. Then make anew variable: r1mix=mix.*r2+(1-mix.^2)^0.5.*r1; where mix= corr. coeff.
• For each value of mix calculate covariance and correlation coefficient between r1mix and r2
• In each case make a scatter plot: plot(r1mix,r2,’k.’);
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 $X_1, X_2, ..., X_p$ and constants $c_1, c_2, ..., c_p$

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

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

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

\[ E(Y) = c_1E(X_1) + c_2E(X_2) + \ldots + c_pE(X_p) \]  \hspace{1cm} (5-25)

\[ V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) + 2\sum_{i<j}c_ic_j\text{cov}(X_iX_j) \]  \hspace{1cm} (5-26)

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

\[ V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) \]  \hspace{1cm} (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 $\text{nm}^2$. What is the variance of the finished product?

Answer:

\[
X = X_1 + X_2 + X_3
\]

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

\[
SD(X) = \sqrt{95} = 9.7 \text{ nm}
\]

If adding SDs one would get \( \sqrt{25 \text{ nm}^2} + \sqrt{40 \text{ nm}^2} + \sqrt{30 \text{ nm}^2} = 16.08 \text{ nm} \)
IMPORTANT:

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

\[
\text{Average } \bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_p}{p}
\]

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

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

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

If \( \bar{X} = \frac{\left(X_1 + X_2 + \ldots + X_p\right)}{p} \) 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.
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: \( x_1, x_2, \ldots, x_p \):
I need to produce \( p \) new variables:
\( y_1, y_2, \ldots, y_p \):

\[
y_1 = a_{11} x_1 + a_{12} x_2 + \ldots + a_{1p} x_p \\
y_2 = a_{21} x_1 + a_{22} x_2 + \ldots + a_{2p} x_p \\
\vdots \\
y_p = a_{p1} x_1 + a_{p2} x_2 + \ldots + a_{pp} x_p
\]

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

\[
\sigma_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}
\]

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 (perimeter^2 / 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
Matlab exercise 2

• Carry out PCA of the cancer data
   In the template I use eigs. Matlab also has a dedicated pca commands (read the manual)

• Which variables give the strongest positive or negative contributions to the 1\textsuperscript{st}, 2\textsuperscript{nd}, and 3\textsuperscript{rd} largest eigenvalues?

• Plot the scores (Score=Z*V) of the 1\textsuperscript{st} vs 2\textsuperscript{nd} eigenvalues for normal and cancer patients separately. Can these PCA scores be used to separate cancer from normal patients?
Multivariable statistics and Principal Component Analysis (PCA)

• A table of \( n \) observations in which \( p \) variables were measured

\[
p \times p \text{ symmetric matrix } R \text{ of corr. coefficients } \\
r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \\
\text{PCA: Diagonalize matrix } R
\]
Principle Component Analysis (PCA)

- $p \times p$ symmetric matrix $R$ of corr. coefficients $r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$
- $R=n^{-1}Z^*Z$ is a “square” of the matrix $Z$ of standardized r.v.:
  $z_{\alpha k} = \frac{\chi_{\alpha k} - \mu_k}{\sigma_k}$ → all eigenvalues of $R$ are non-negative
- Diagonal elements=1 → $tr(R)=p$
- Can be diagonalized: $R=V*D*V'$ where $D$ is the diagonal matrix
- $d(1,1)$ –largest eig. value, $d(p,p)$ – the smallest one
- The meaning of $V(i,k)$ – contribution of the data type $i$ to the $k$-th eigenvector
- $tr(D)=p$, the largest eigenvalue $d(1,1)$ absorbs a fraction $=d(1,1)/p$ of all correlations can be $\sim 100%$
- Scores: $Y=Z*V$: $n \times p$ matrix. Meaning of $Y(\alpha,k)$ – participation of the sample # $\alpha$ in the $k$-th eigenvector
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