Noise & Data Reduction

- Paired Sample t Test
- Data Transformation - Overview
- From Covariance Matrix to PCA and Dimension Reduction
- Fourier Analysis - Spectrum
- Dimension Reduction
Remember: Central Limit Theorem

The sampling distribution of the mean of samples of size N approaches a normal (Gaussian) distribution as N approaches infinity.

If the samples are drawn from a population with mean \( \mu \) and standard deviation \( \sigma \) then the mean of the sampling distribution is \( \mu \) and its standard deviation is \( \sigma \sqrt{\frac{1}{N}} \) as N increases.

These statements hold irrespective of the shape of the original distribution.

Z Test

- \[ Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} \]
- \[ \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2} \]

\( \bar{x} \) is the sample mean, population mean \( \mu \), standard deviation \( \sigma \)

\( s \) is the sample standard deviation

\[ t = \frac{\bar{x} - \mu}{s / \sqrt{N}} \]

\[ s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} \]

- when population standard deviation is unknown, samples are small
- population mean \( \mu \), sample mean \( \bar{x} \)
**p Values**

- Commonly we reject the H0 when the probability of obtaining a *sample statistic* given the null hypothesis is low, say < .05
- The null hypothesis is rejected but might be true
- We find the probabilities by looking them up in tables, or statistics packages provide them
  - The probability of obtaining a particular sample given the null hypothesis is called the *p* value
  - By convention, one usually does not reject the null hypothesis unless *p* < 0.05 (statistically significant)

**Example**

- Five cars parked, mean price of the cars is 20.270 € and the standard deviation of the *sample* is 5.811 €
- The mean costs of cars in town is 12.000 € (population)
- H0 hypothesis: parked cars are as expensive as the cars in town
  \[ t = \frac{20270 - 12000}{5811/\sqrt{5}} = 3.18 \]
  - For *N-1* (degrees of freedom) *t*=3.18 has a value less than 0.025, reject H0!
Paired Sample t Test

- Given a set of paired observations
- *(from two normal populations)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>( \delta = A - B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>y1</td>
<td>x1-x2</td>
</tr>
<tr>
<td>x2</td>
<td>y2</td>
<td>x2-y2</td>
</tr>
<tr>
<td>x3</td>
<td>y3</td>
<td>x3-y3</td>
</tr>
<tr>
<td>x4</td>
<td>y4</td>
<td>x4-y4</td>
</tr>
<tr>
<td>x5</td>
<td>y5</td>
<td>x5-y5</td>
</tr>
</tbody>
</table>

- Calculate the mean \( \bar{x}_\delta \) and the standard deviation \( s_\delta \) of the differences \( \delta \)
- H0: \( \mu_\delta = 0 \) *(no difference)*
- H0: \( \mu_\delta = k \) *(difference is a constant)*

\[
t_\delta = \frac{\bar{x}_\delta - \mu_\delta}{\hat{\sigma}_\delta} = \frac{\delta}{\sqrt{N_\delta}}
\]
Confidence Intervals ($\sigma$ known)

- Standard error from the standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma_{\text{Population}}}{\sqrt{N}}$$

- 95 Percent confidence interval for normal distribution is about the mean

$$\bar{x} \pm 1.96 \cdot \sigma_{\bar{x}}$$

Confidence interval when ($\sigma$ unknown)

- Standard error from the sample standard deviation

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{N}}$$

- 95 Percent confidence interval for t distribution ($t_{0.025}$ from a table) is

$$\bar{x} \pm t_{0.025} \cdot \hat{\sigma}_{\bar{x}}$$

Previous Example:
Overview Data Transformation

- Reduce Noise
- Reduce Data

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones
Data Transformation: Normalization

- Min-max normalization: to \([\text{new}_{\text{min}}, \text{new}_{\text{max}}]\)
  \[ v' = \frac{v - \text{min}}{\text{max} - \text{min}} (\text{new}_{\text{max}} - \text{new}_{\text{min}}) + \text{new}_{\text{min}}. \]

  - Ex. Let income range $12,000 to $98,000 normalized to \([0.0, 1.0]\). Then $73,000 is mapped to
  \[ \frac{73,000 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716. \]

- Z-score normalization (\(\mu\): mean, \(\sigma\): standard deviation):
  \[ v' = \frac{v - \mu}{\sigma}. \]

  - Ex. Let \(\mu = 54,000\), \(\sigma = 16,000\). Then
  \[ \frac{73,000 - 54,000}{16,000} = 1.225. \]

- Normalization by decimal scaling

How to Handle Noisy Data? (How to Reduce Features?)

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

- Regression
  - smooth by fitting the data into regression functions

- Clustering
  - detect and remove outliers

- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)
Data Reduction Strategies

- A data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation
  - Dimensionality reduction—remove unimportant attributes
  - Data Compression
  - Numerosity reduction—fit data into models
  - Discretization and concept hierarchy generation

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - Divides the range into $N$ intervals of equal size: uniform grid
  - if $A$ and $B$ are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well.
- Equal-depth (frequency) partitioning:
  - Divides the range into $N$ intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.
Binning Methods for Data Smoothing

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries (min and max are identified, bin value replaced by the closest boundary value):
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34
Regression

\[ y = x + 1 \]

Feature space

- Sample \( \{ \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}, \ldots, \mathbf{x}^{(n)} \} \)

\[
\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix}, \quad ||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}
\]
Scaling

- A well-known scaling method consists of performing some scaling operations
  - subtracting the mean and dividing the standard deviation
    \[ y_i = \frac{(x_i - m_i)}{s_i} \]
  - \( m_i \) sample mean
  - \( s_i \) sample standard deviation

- According to the scaled metric the **scaled** feature vector is expressed as
  \[ \| \bar{y} \|_s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - m_i)^2}{s_i^2}} \]
  - shrinking **large** variance values
    - \( s_i > 1 \)
  - stretching **low** variance values
    - \( s_i < 1 \)

- Fails to preserve distances when **general linear transformation** is applied!
- **Covariance**
  - Measuring the tendency two features $x_i$ and $x_j$ varying in the **same direction**
  - The covariance between features $x_i$ and $x_j$ is estimated for $n$ patterns

\[
C_{ij} = \frac{\sum_{k=1}^{n}(x_i^{(k)} - m_i)(x_j^{(k)} - m_j)}{n - 1}
\]

\[
C = \begin{bmatrix}
  c_{11} & c_{12} & \cdots & c_{1d} \\
  c_{21} & c_{22} & \cdots & c_{2d} \\
  \vdots & \vdots & \ddots & \vdots \\
  c_{d1} & c_{d2} & \cdots & c_{dd}
\end{bmatrix}
\]
Correlation

- Covariances are symmetric \( c_{ij} = c_{ji} \)
- Covariance is related to correlation

\[
 r_{ij} = \frac{\sum_{k=1}^{n} (x_i^{(k)} - m_i)(x_j^{(k)} - m_j)}{(n-1)s_i s_j} = \frac{c_{ij}}{s_i s_j} \in [-1, 1]
\]

Principal Component Analysis

- Intuition: find the axis that shows the greatest variation, and project all points into this axis
**Karhunen-Loève Transformation**

- Covariance matrix $C$ of a $d \times d$ matrix
  - Symmetric and positive definite
  
  $$U^T C U = \Lambda = \text{diag} (\lambda_1, \lambda_2, ..., \lambda_d)$$
  
  $$(\lambda I - C) u = 0$$
  
  - There are $d$ eigenvalues and eigenvectors
    
    $$C \vec{u}_i = \lambda_i \vec{u}_i$$
    
    - is the $\lambda_i$ ith eigenvalue of $C$ and $u_i$ the $i$th column of $U$, the $i$th eigenvectors

- Eigenvectors are always **orthogonal**
  - $U$ is an orthonormal matrix $UU^T = U^T U = I$
  - $U$ defines the K-L transformation
  - The transformed features by the K-L transformation are given by
    
    $$\vec{y} = U^T \vec{x}$$
    
    (linear Transformation)
    
    - K-L transformation rotates the feature space into alignment with **uncorrelated** features
Example

\[
C = \begin{bmatrix}
2 & 1 \\
1 & 1
\end{bmatrix}
\quad |\lambda I - C| = 0
\quad \lambda^2 - 3\lambda + 1 = 0
\]

- \(\lambda_1 = 2.618\)  \(\lambda_2 = 0.382\)

\[
\begin{bmatrix}
-0.618 & -1 \\
-1 & 1.618
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2
\end{bmatrix} = 0
\]

- \(u^{(1)} = [1 \ 0.618]\)  \(u^{(2)} = [-1 \ 1.618]\)

- \(U = [u^{(1)}, u^{(2)}]\)
PCA (Principal Components Analysis)

- New features $y$ are uncorrelated with the covariance Matrix
- Each eigenvector $u_i$ is associated with some variance associated by $\lambda_i$
- Uncorrelated features with higher variance (represented by $\lambda_i$) contain more information
- Idea:
  - Retain only the significant eigenvectors $u_i$
  - Example
    - $U=[u^{(1)}, u^{(2)}]$ $\lambda_1=2.618 \ \lambda_2=0.382$
    - $U^*=[u^{(1)}]$ $\bar{y} = U^{*T}\bar{x}$

Dimension Reduction

- How many eigenvectors (and corresponding eigenvector) to retain
- Kaiser criterion
  - Discards eigenvectors whose eigenvalues are below 1
Problems

- Principal components are linear transformation of the original features
- It is difficult to attach any semantic meaning to principal components
- For new data which is added to the dataset, the PCA has to be recomputed

Suppose we have a covariance matrix:

\[ C = \begin{pmatrix} 3 & 1 \\ 1 & 21 \end{pmatrix} \]

- What is the matrix of the K-L transformation?
First we have to compute the eigenvalues
The system has to become linear dependent (singular)

\[ \det(\lambda I - C) = 0 \]

The determinant has to become zero

\[ \lambda^2 - 24\lambda + 62 = 0 \]

Solving it we get

- \( \lambda_1 = 2.94461 \)
- \( \lambda_2 = 21.05538 \)
Now, let's compute the two eigenvectors.

To do it you have to solve two singular, dependent systems.

For the first eigenvalue $\lambda_1 = 2.94461$

$$ (\lambda_1 I - C) \vec{u}_1 = 0 $$

Or if you prefer more ...

$$ \lambda_1 \vec{u}_1 = C \vec{u}_1 $$
$$ C \vec{u}_1 = \lambda_1 \vec{u}_1 $$

And for the second eigenvalue $\lambda_2 = 21.05538$

$$ (\lambda_2 I - C) \vec{u}_2 = 0 $$

Or if you prefer more ...

$$ \lambda_2 \vec{u}_2 = C \vec{u}_2 $$
$$ C \vec{u}_2 = \lambda_2 \vec{u}_2 $$
For $\lambda_1 = 2.94461$

$\mathbf{u}_1 = (u_1, u_2)$

\[
\begin{bmatrix}
2.94461 & 0 \\
0 & 2.94461
\end{bmatrix}
- 
\begin{bmatrix}
3 & 1 \\
1 & 21
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2
\end{bmatrix}
= 0
\]

- We have to find a nontrivial solution!
  - Trivial solution is $u = [0, 0]$

\[
\begin{bmatrix}
-0.05539 & -1 \\
-1 & -18.055
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2
\end{bmatrix}
= 0
\]

- Because the system is linear dependable, the left column is multiple value of the right column
- There are infinity many solution!!!!

- We have only to determine the direction of the eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$
- But be careful, the normalized vectors have to be orthogonal to each other
- $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = 0$
Let be \( u_1=1 \) then we have to determine \( u_2 \)
\[
\begin{bmatrix}
-0.05539 & -1 \\
-1 & -18.055
\end{bmatrix}
\begin{bmatrix}
1 \\
u_2
\end{bmatrix} = 0
\]
\[
\begin{bmatrix}
-0.05539 \\
-1
\end{bmatrix}
\begin{bmatrix}
1 \\
u_2
\end{bmatrix} = \begin{bmatrix}
18.055
\end{bmatrix}
\]

\( u_1=[u_1,u_2]=[1,-0.05539] \)

For \( \lambda_2 = 21.05538 \)

\( u_1=(u_1,u_2) \)
\[
\begin{bmatrix}
21.05538 & 0 \\
0 & 21.05538
\end{bmatrix}
- \begin{bmatrix}
3 & 1 \\
1 & 21
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2
\end{bmatrix} = 0
\]

We have to find a nontrivial solution!

- Trivial solution is \( u=[0,0] \)... D\'\c{e}j\`a vu?

\[
\begin{bmatrix}
18.055 & -1 \\
-1 & 0.05538
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2
\end{bmatrix} = 0
\]
Let be \( u_1 = 1 \) then we have to determine \( u_2 \):

\[
\begin{bmatrix}
18,055 \\
-1
\end{bmatrix} = \begin{bmatrix} 1 \\ -0.05538 \end{bmatrix} u_2
\]

\( u_2 = [u_1, u_2] = [1, 18,055] \)

\( u_1 = [u_1, u_2] = [1, -0.05539] \)

- for \( \lambda_1 = 2,94461 \)

\( u_2 = [u_1, u_2] = [1, 18,055] \)

- \( \lambda_2 = 21,05538 \)

Orthogonal? Yes \( <u_1, u_2> = 0 \)

Which of the two eigenvectors is more significant?

- \( u_2 \), because \( \lambda_1 = 2,94461 < \lambda_2 = 21,05538 \)
Fourier Analysis

- It is always possible to analyze “complex” periodic waveforms into a set of sinusoidal waveforms.
- Any periodic waveform can be approximated by adding together a number of sinusoidal waveforms.
- Fourier analysis tells us what particular set of sinusoids go together to make up a particular complex waveform.

Spectrum

- In the Fourier analysis of a complex waveform the **amplitude** of each sinusoidal component depends on the shape of particular complex wave:
  - Amplitude of a wave: maximum or minimum deviation from zero line
  - $T$ duration of a period
  - $f = \frac{1}{T}$
Noise reduction or Dimension Reduction

- It is difficult to identify the frequency components by looking at the original signal
- Converting to the frequency domain
- If dimension reduction, store only a fraction of frequencies (with high amplitude)
- If noise reduction
  - (remove high frequencies, fast change, smoothing)
  - (remove low frequencies, slow change, remove global trends)
  - Inverse discrete Fourier transform
Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
- The attribute with the most distinct values is placed at the lowest level of the hierarchy
- Exceptions, e.g., weekday, month, quarter, year

- country: 15 distinct values
- province or state: 365 distinct values
- city: 3567 distinct values
- street: 674,339 distinct values
- Paired Sample t Test
- Data Transformation - Overview
- From Covariance Matrix to PCA and Dimension Reduction
- Fourier Analysis - Spectrum
- Dimension Reduction

- Mining Association rules
- Apriori Algorithm (Chapter 6, Han and Kamber)