SBE 4032: Advanced Topics in Medical Informatics (2)

موضوعات متقدمة في المعلوماتية الطبية (2)

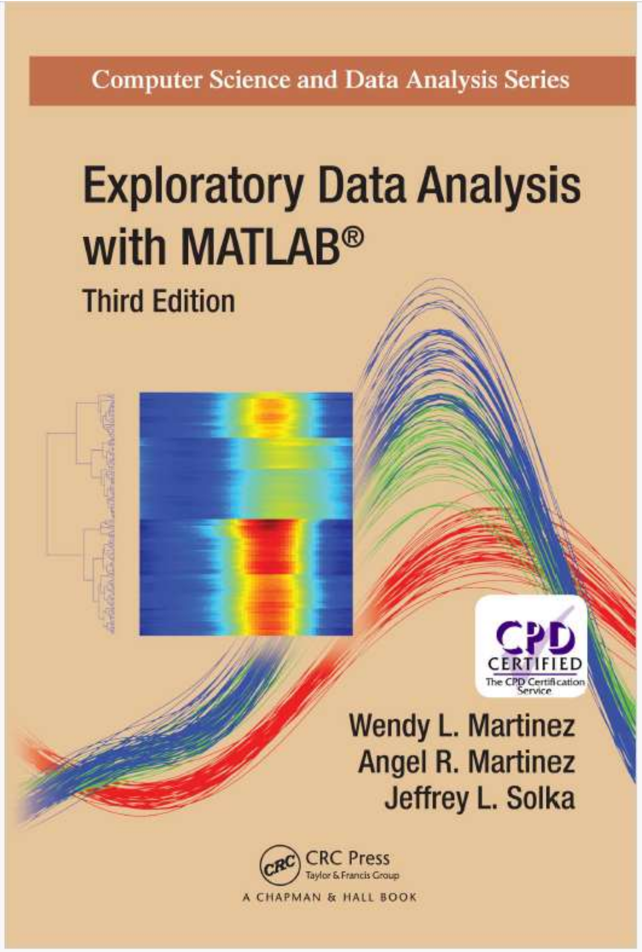
Lecture #1: Introduction to Exploratory Data Analysis

Hisham Abdeltawab, Ph.D.
Assistant Professor,
Cairo University

Syllabus and Grading

- **Tentative Topics:**
  - Introduction to Exploratory Data Analysis
  - Dimensionality Reduction
  - Data Tours
  - Finding Clusters
  - Graphical Methods for EDA
    - Visualizing Clusters
    - Distribution Shapes
    - Multivariate Visualization
    - Visualizing Categorical Data

- 40 Marks (Class work) and 60 Marks (Final Exam).



Exploratory Data Analysis vs. Confirmatory Data Analysis

❑ Exploratory Data Analysis (EDA):

- “detective work – numerical detective work – or counting detective work – or graphical detective work” [Tukey, 1977, page 1].
- It is mostly a philosophy of data analysis where the researcher examines the data without any pre-conceived ideas in order to discover what the data can tell him or her about the phenomena being studied.
- Time spent playing with the data and looking at it from different angles.

❑ Confirmatory Data Analysis (CDA):

- It is mostly concerned with hypothesis testing.
- CDA methods typically involve the process of making inferences about or estimates of some population characteristic and then trying to evaluate the precision associated with the results.

❑ EDA and CDA should be used in a complementary way. The analyst explores the data looking for patterns and structure that leads to hypotheses and models.

Exploratory Data Analysis vs. Confirmatory Data Analysis

❑ Hartwig and Dearing [1979] described the CDA mode as one that answers questions such as “Do the data confirm hypothesis XYZ?” Whereas, EDA tends to ask “What can the data tell me about relationship XYZ?

❑ Hartwig and Dearing specify two principles for EDA: skepticism and openness.

- ✓ This might involve visualization of the data to look for anomalies or patterns, the use of resistant (or robust) statistics to summarize the data, openness to the transformation of the data to gain better insights, and the generation of models.
- ✓ Question: What are the robust statistics?

Transforming Data

- In many real-world applications, the data analyst will have to deal with raw data that are not in the most convenient form.

- The data might need to be re-expressed to produce effective visualization or an easier, more informative analysis.

Transforming Data: Power Transformation

- Some common transformations include taking roots (square root, cube root, etc.), finding reciprocals, calculating logarithms, and raising variables to positive integral powers.
- These transformations provide adequate flexibility for most situations in data analysis.

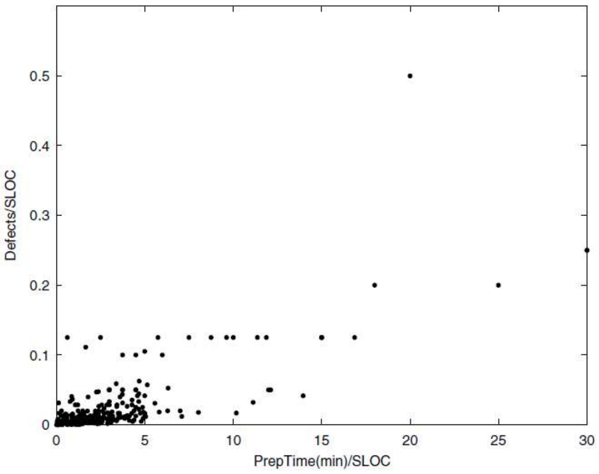
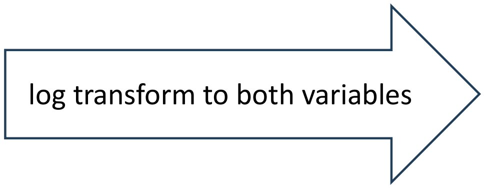


FIGURE 1.2
This is a scatterplot of the software inspection data. The relationship between the variables is difficult to see.



log transform to both variables

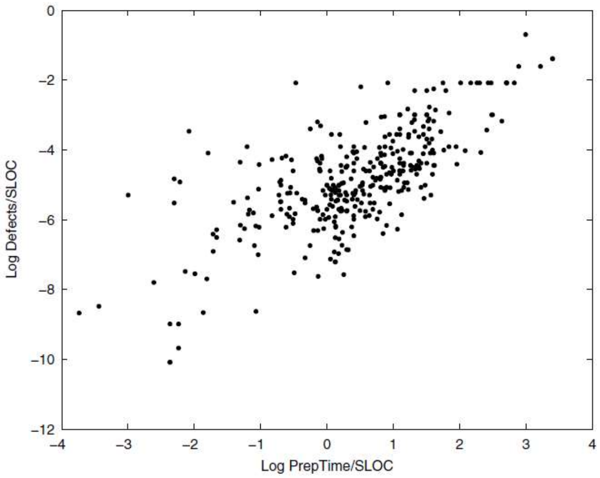


FIGURE 1.3
Each variate was transformed using the logarithm. The relationship between preparation time per SLOC and number of defects found per SLOC is now easier to see.

- the data are skewed, and the relationship between the variables is difficult to understand.

- We now have a better idea of the relationship between these two variables

Standardization: Transformation Using the Standard Deviation

- If the variables are measurements along a different scale or if the standard deviations for the variables are different from one another, then one variable might dominate the distance used in the analysis.

- The first standardization we discuss is called the sample z-score. The transformed variates are found using:

$$z = \frac{(x - \bar{x})}{s},$$

where x is the original observed data value, $\bar{x}$ is the sample mean, and s is the sample standard deviation. In this standardization, the new variate z will have a mean of zero and a variance of one.

- to make sure that all the variables have the same scale and same importance in the model (machine learning model)

- Gradient Descent Convergence (faster).

Standardization: Transformation Using the Range

Instead of dividing by the standard deviation, as above, we can use the range of the variable as the divisor. This yields the following two forms of standardization:

$$z = \frac{x}{\max(x) - \min(x)},$$

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

The standardization in this equation is bounded by zero and one

- to make all the elements lie between 0 and 1 thus bringing all the values of numeric columns in the dataset to a common scale.

Sphering the Data

- This type of standardization called ***sphering*** pertains to multivariate data.
- It serves a similar purpose as the 1–D standardization methods given above.
- The transformed variables will have a p-dimensional mean of **0** and a covariance matrix given by the identity matrix.
- We start off with the p-dimensional sample mean given by:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i .$$

- We then find the sample covariance matrix given by the following:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T ,$$

Sphering the Data

We sphere the data using the following transformation:

$$\mathbf{Z}_i = \Lambda^{-1/2}\mathbf{Q}^T(\mathbf{x}_i - \bar{\mathbf{x}}) \qquad i = 1, \ldots, n,$$

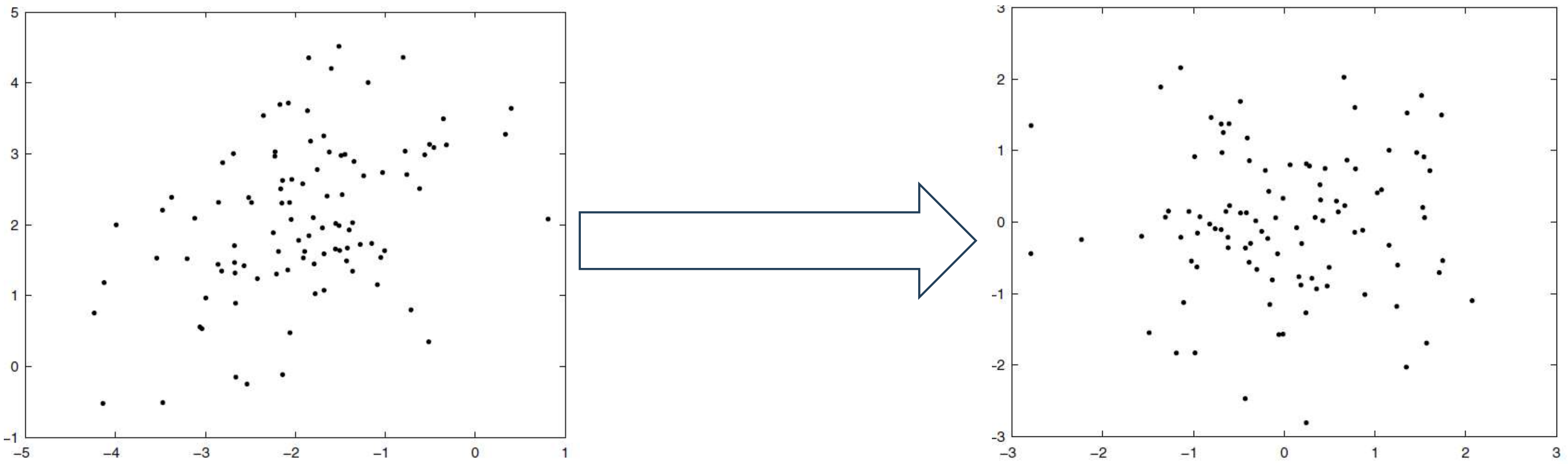
where the columns of **Q** are the eigenvectors obtained from **S**, Λ is a diagonal matrix of corresponding eigenvalues, and x_i is the i-th observation.

- Whitening or Sphering is a data pre-processing step. It can be used to remove correlation or dependencies between features in a dataset. This may help to better train a machine learning model.

Sphering the data: Example

```
% First generate some 2-D multivariate normal
% random variables, with mean MU and
% covariance SIGMA.
n = 100;
mu = [-2, 2];
sigma = [1,.5;.5,1];
X = mvnrnd(mu,sigma,n);
plot(X(:,1),X(:,2),'.')
```

```
% Now sphere the data.
xbar = mean(X);
% Get the eigenvectors and eigenvalues of the
% covariance matrix.
[V,D] = eig(cov(X));
% Center the data.
Xc = X - ones(n,1)*xbar;
% Sphere the data.
Z = ((D)^(-1/2)*V'*Xc')';
plot(Z(:,1),Z(:,2),'.')
```



This is similar to the z-score standardization in 1–D.

Thank You
&
Questions