

MICHIGAN STATE UNIVERSITY

Statistical Analysis of Experimental Data

Yuzhen Lu
 Email: luyuzhen@msu.edu
<https://github.com/jingweimo/BE815-Ch6-SAED>

1

MICHIGAN STATE UNIVERSITY

Overview

- General Concepts
- Outlier Detection
- Least-square Fitting
- Multivariate Analysis: PCA

2

MICHIGAN STATE UNIVERSITY

- General Concepts
 - Measure of central tendency
 - Mean
 - Median
 - Mode
 - Measure of dispersion
 - Variance
 - **Standard deviation (SD)**
 - Coefficient of variation (CV, SD/Mean)

3

MICHIGAN STATE UNIVERSITY

Standard deviation (SD) vs. Standard error (SE)

SD, a **descriptive** term, is a measure of the dispersion of measurements.

SE, also called the standard error of mean, is a measure of how precisely the sample mean estimates the underlying population mean, and **SE = SD/sqrt(N)** where N is the number of samples

SE is mainly used as **an inferential tool**. It is common to see the mean is reported as **Mean ± SE**

SE explains the reason for collecting a large set of sample for reliable mean estimation (high precision)

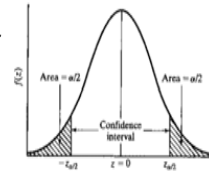
4

Interval Estimation of the Population Mean μ

- When the sample size $N > 30$, the sample mean is normally distributed:

$$\mu = \bar{x} \pm z_{\alpha/2} SE$$

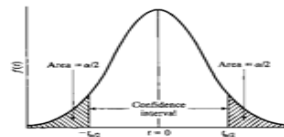
For a 95% confidence interval, the uncertainty is equal to $1.96SE$



- If $N \leq 30$, μ is calculated based on the Student's t distribution

$$\mu = \bar{x} \pm t_{\alpha/2} SE$$

$$t_{\alpha/2} > z_{\alpha/2}$$



5

Example

TABLE 6.1 Results of 60 Temperature Measurements in a Duct

Number of readings	Temperature (°C)
1	1089
1	1092
2	1094
4	1095
8	1098
9	1100
12	1104
4	1105
5	1107
5	1108
4	1110
3	1112
2	1115

Calculate statistics using R

```
mean(x)
getmode(x)
median(x)
var(x)
sd(x)
sd(x)/mean(x) #coefficient of variation
sd(x)/sqrt(length(x)) #standard error
```

```
#user function: getmode()
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v,
    uniqv)))]
}
```

```
#uncertainty of the mean
qnorm(p=0.975)*sd(x)/sqrt(length(x))
```

6

Overview

- General Concepts
- Outlier Detection
- Least-square Fitting
- Multivariate Analysis: PCA

7

Outlier Detection

- Modified Thompson τ technique

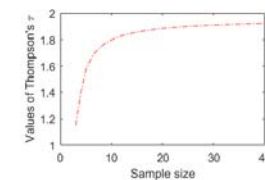
$$\tau = \frac{t_{\alpha/2} \cdot (n-1)}{\sqrt{n} \sqrt{n-2 + t_{\alpha/2}^2}}$$

TABLE 6.8 Values of Thompson's τ

Sample size	τ	Sample size	τ
3	1.150	22	1.893
4	1.393	23	1.896
5	1.572	24	1.899
6	1.656	25	1.902
7	1.711	26	1.904
8	1.749	27	1.906
9	1.777	28	1.908
10	1.798	29	1.910
11	1.815	30	1.911
12	1.829	31	1.913
13	1.840	32	1.914
14	1.849	33	1.916
15	1.858	34	1.917
16	1.865	35	1.919
17	1.871	36	1.920
18	1.876	37	1.921
19	1.881	38	1.922
20	1.885	39	1.923
21	1.889	40	1.924

Source: ASME (1998).

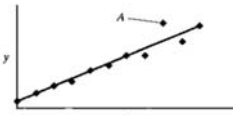
- Compute the mean, standard deviation (SD) and the deviation $\delta_i = |x_i - \bar{x}|$
- If $\delta_i > \tau SD$, the sample is rejected as an outlier
- Repeat 1-2 for the remaining samples until no outliers are identified



8

MICHIGAN STATE UNIVERSITY


- Outliers in x-y data pairs



Fit the dataset and the point has a large deviation from the fitting line or residual likely to be an outlier.

- Outliers in multivariate data

Each sample has more than one measurand, each of which corresponds to a variable space.



9

MICHIGAN STATE UNIVERSITY

Euclidean distance (ED) vs Mahalanobis distance (MD)

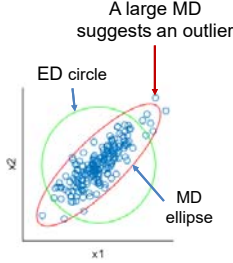
Let two samples be defined by the vectors of x_A and x_B whose entries are the values of different variables, and then ED and MD can be calculated as follows:

$$ED = [(x_A - x_B)^T (x_A - x_B)]^{1/2}$$

$$MD = [(x_A - x_B)^T C^{-1} (x_A - x_B)]^{1/2}$$

↑
Sample covariance matrix

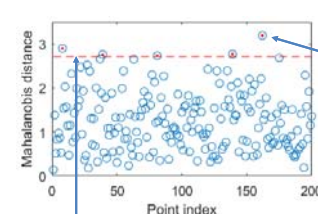
- ✓ Compared to ED, MD accounts for the **covariance structure** of data
- ✓ MD can be used as guidance in outlier detection, based on the calculation of MD from the data center and the fact the **squared MD** follows a **chi-square distribution** with degree of freedom equal to the number of variables involved



10

Johnson, RA & Wicheren, DW (2007). Applied Multivariate Statistical Analysis. Pearson Prentice Hall

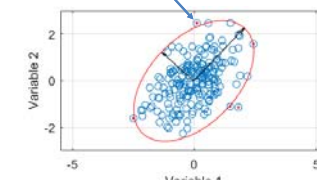
MICHIGAN STATE UNIVERSITY



#R command
`mahalanobis(data, data_center, cov)`

Identified outliers

%cut-off distance
`sqrt(chi2inv(0.975,2))`



%Matlab command

```
temp = (data - repmat(data_center,[n,1]))/cov(data)*(data - repmat(data_center,[n,1]));
distMaha = sqrt(diag(temp));
```

11

MICHIGAN STATE UNIVERSITY

Overview

- General Concepts
- Outlier Detection
- Least-square Fitting
- Multivariate Analysis: PCA

12

Least-square (LS) fitting

- Linear system

$$\begin{cases} x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1n}\beta_n = y_1 \\ x_{21}\beta_1 + x_{22}\beta_2 + \dots + x_{2n}\beta_n = y_2 \\ \vdots \\ x_{m1}\beta_1 + x_{m2}\beta_2 + \dots + x_{mn}\beta_n = y_m \end{cases}$$

m samples

n variables (i.e., parameter β)

Linear regression

LS (line or curve) fitting is to solve for a LS solution (i.e., β_{LS}) for an **over-determined** linear system that has $m > n$.

LS solution

Matrix notation: $X\beta = Y$

The LS solution occurs when the projection error is minimized, which is satisfied when the projection error is **perpendicular to** the variable space:

$$(X\beta - Y) \perp \{X\beta \mid \beta \in R^n\} \implies \beta_{LS} = (X^T X)^{-1} X^T Y$$

%Matlab commands or using cftool

```
L = [0, 0.5, 1, 1.5, 2, 2.5]';
V = [0.05, 0.52, 1.03, 1.50, 2.00, 2.56]';
figure;
plot(L,V,'b+');
xlabel('L'); ylabel('V');
```

```
A = [L, ones(length(L),1)];
x = (A'*A)\A'*V; %LS solution
hold on;
plot(L,A*x,'r-');
```

Example 6.20

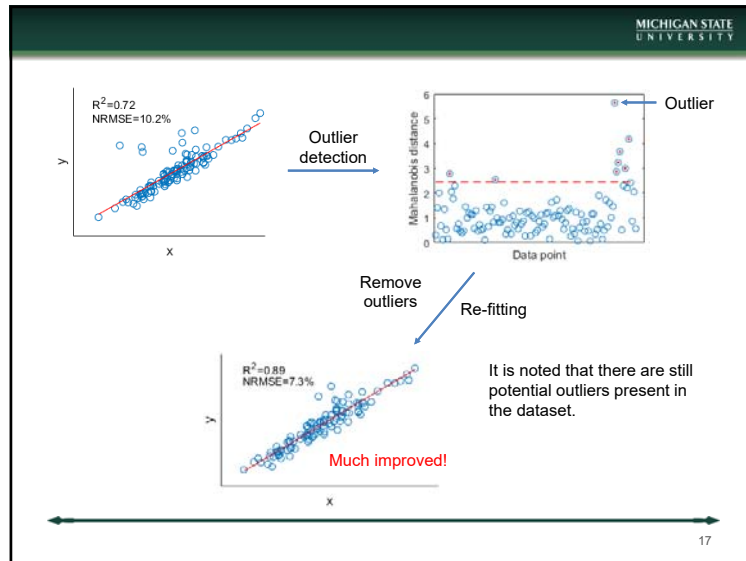
The following table represents the output (volts) of a linear variable differential transformer (LVDT; an electric output device used for measuring displacement) for five length inputs:

L(cm)	0.00	0.50	1.00	1.50	2.00	2.50
V(V)	0.05	0.52	1.03	1.50	2.00	2.56

Performance criteria

- Determination coefficient (R^2) $R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$ y_i : measured values \hat{y}_i : predicted values
- Normalized root mean square of error (NRMSE)

$$NRMSE = \frac{1}{\bar{y}} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \text{ or } \frac{1}{y_{\max} - y_{\min}} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$



MICHIGAN STATE UNIVERSITY

Overview

- General Concepts
- Outlier Detection
- Least-square Fitting
- Multivariate Analysis: PCA

18

MICHIGAN STATE UNIVERSITY

■ Multivariate Analysis: PCA

- Principal component analysis (PCA) is considered the mother of all multivariate analysis methods
- Functions of PCA:
 - Data visualization
 - Data decorrelation
 - Diagnostic plots
 - Dimension reduction
 - Feature extraction
 -

Most often $n \gg k$

Koch, I. (2014). Analysis of multivariate and high-dimensional data. Cambridge University Press.

19

MICHIGAN STATE UNIVERSITY

The axis or direction that captures the largest variance of data

- PCA Matrix notation:

$$T = XP$$

Score matrix Data matrix loading matrix
- PCA is to find the loading matrix that transforms the original data into a score matrix.
- Column vectors of P are eigenvectors orthogonal to each other
- Column vectors of T are principal components (PCs) orthogonal to each other as well, and PCs are referred to as the extracted features

20

- Implementation of PCA

Singular value decomposition (SVD) is the most widely used tool for PCA.

$$\text{SVD: } X = USV^T \longrightarrow \text{PC Scores: } T = US$$

For mean-centered X , the matrix U is the normalized PCA score matrix, and V is the PCA loading matrix, and S is a diagonal matrix containing singular values that are the standard deviations of PC scores.

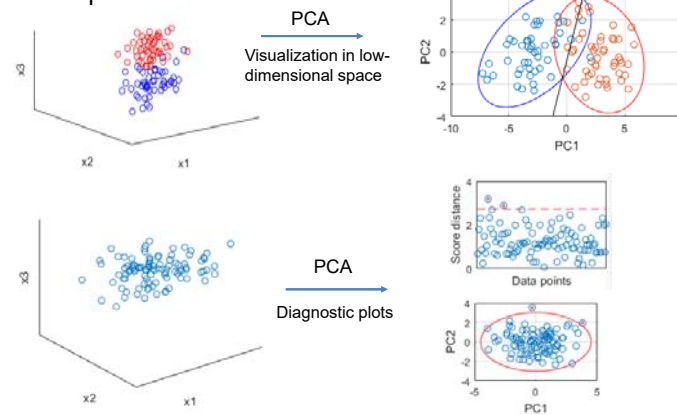
Alternatively, **eigen-decomposition**, i.e., decomposition of $X^T X$ to obtain its eigenvector matrix that is equal to P , or XX^T to obtain the eigenvector matrix that is equal to T .

#Matlab command

<code>%svd.m</code>	<code>% or eig.m</code>	<code>% or simply by pca.m</code>
<code>[U,S,V] = svd(X);</code>	<code>eigenvec = eig(cov(X));</code>	<code>[P,T] = pca(X);</code>
<code>T = U*S;</code>	<code>T = X*eigenvec;</code>	

21

- Example



22

- Extensions to PCA

- Principal component regression (PCR)
- Partial least-squares regression (PLSR)
- Linear discriminant analysis (LDA)
- Partial least-squares discriminant analysis (PLS-DA)
- Independent component analysis (ICA)
-

- Recommended R package for Multivariate analysis

Garcia, H & Filzmoser (2017). Multivariate statistical analysis using the R package chemometrics.
<http://cran.ms.unimelb.edu.au/web/packages/chemometrics/>

23

Lessons & Tools Learned

- Basic statistics: mean, variance, SD, SE, etc., and use SE for confidence interval estimation of mean
- Outlier detection by the Thompson's rule for single-variable measurements, and the MD for more than one variable
- LS line fitting, evaluation, and improvement through outlier detection & removal
- PCA for multivariate data visualization, dimension reduction and diagnostics
- Software
 - R (for statistics) <https://www.r-project.org/>
 - Matlab
 - Lecture codes: <https://github.com/jingweimo/BE815-Ch6-SAEED>

24