STA437/2005 - Methods for Multivariate Data

Lecture 1

Gun Ho Jang

September 8, 2014

Introduction

Interest of this course

This course is concerned with "statistical methods designated to elicit information from data sets."

- data include simultaneous measurements on many aspects so called random variables
- *multivariate analysis* is concerned about many random variables
- The underlying relationships between variables are one of interests.
- Inference and prediction are also most important part of multivariate analysis.

Some Aspects of Multivariate Analysis

- *Data reduction or structural simplification:* Make data as simple as possible. Cf. minimal sufficient statistic.
- *Sorting and grouping:* Classification. As a result, accuracies of estimation and prediction might increased.
- *Investigation of the dependence among variables:* Independence assessment. Recognition of dependence structure.
- *Prediction:* Forcast the value of interest using the other variables. One of the most important topics in statistics.
- *Hypothesis construction and testing:* One of the most important topics in statistics.

Data Format

- p: the number of variables
- n: the number of subjects
- x_{ij}: the measurement of jth variable on ith subject.

Random Variable/Vectors convention

- small characters are designated for single random variables
- capital characters are designated for random vectors
- boldfaces are designated for aggregation of p variables

Data Format

$$
\begin{array}{lcccccc}
 & \text{Variable 1} & \text{Variable 2} & \cdots & \text{Variable } j & \cdots & \text{Variable } p \\
\text{Subject 1:} & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\
\text{Subject 2:} & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\text{Subject } i\text{:} & x_{i1} & x_{ii} & \cdots & x_{ij} & \cdots & x_{ip} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\text{Subject } n\text{:} & x_{n1} & x_{nn} & \cdots & x_{nj} & \cdots & x_{np}
\end{array}
$$

Or simply

$$
\mathbf{X} = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\
\cdots & \cdots & \cdots & \ddots & \ddots & \cdots \\
x_{i1} & x_{ii} & \cdots & x_{ij} & \cdots & x_{ip} \\
\cdots & \cdots & \cdots & \ddots & \ddots & \cdots \\
x_{n1} & x_{nn} & \cdots & x_{nj} & \cdots & x_{np}
\end{pmatrix}
$$

Example: Book Sales Record

Variables of Interest

- Varialbe 1 (sales amount in dollars)
- Variable 2 (number of books sold)

Data record

Varialbe 1 (sales amount in dollars)	42	52	48	58
Variable 2 (number of books sold)	4	5	4	3

Form of data

$$x_{11} = 42, x_{21} = 52, x_{31} = 48, x_{41} = 58,$$

$$x_{12} = 4, x_{22} = 5, x_{32} = 4, x_{42} = 3$$

$$\mathbf{X} = \begin{pmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{pmatrix}$$

Descriptive Statistics

Sample Means

$$\bar{x}_j = \frac{1}{n} \sum_{k=1}^{n} x_{kj}$$

Sample Variance

$$s_j^2 = \frac{1}{n} \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2$$

Sample Covariance

$$s_{jl} = \frac{1}{n} \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)(x_{kl} - \bar{x}_l)$$

Note that $s_j^2 = s_{jj}$ for all $j = 1, \ldots, p$.

Descriptive Statistics

Sample correlation

$$r_{jl} = \frac{s_{jl}}{\sqrt{s_{jj}s_{ll}}} = \frac{\sum_{k=1}^{n}(x_{kj} - \bar{x}_j)(x_{kl} - \bar{x}_l)}{\sqrt{\sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2 \sum_{k=1}^{n}(x_{kl} - \bar{x}_l)^2}}$$

Descriptive Statistics

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

- Correlations are always between -1 and 1 inclusively.
- $\mathbf{S}, \mathbf{R}$ are symmetric and non-negative definite, in general, positive definite.

Example

$$
\begin{aligned}
\bar{x}_1 &= (42 + 52 + 48 + 58)/4 = 50, \\
\bar{x}_2 &= (4 + 5 + 4 + 3)/4 = 4, \\
s_1^2 = s_{11} &= \sum_{k=1}^{n} (x_{k1} - \bar{x}_1)^2 / n = 34, \\
s_2^2 = s_{22} &= 0.5, \\
s_{12} = s_{21} &= -1.5.
\end{aligned}
$$

$$
\bar{\mathbf{x}} = \begin{pmatrix} 50 \\ 4 \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} 1 & -0.36 \\ -0.36 & 1 \end{pmatrix}
$$

Example: Paper Strength

Random variables

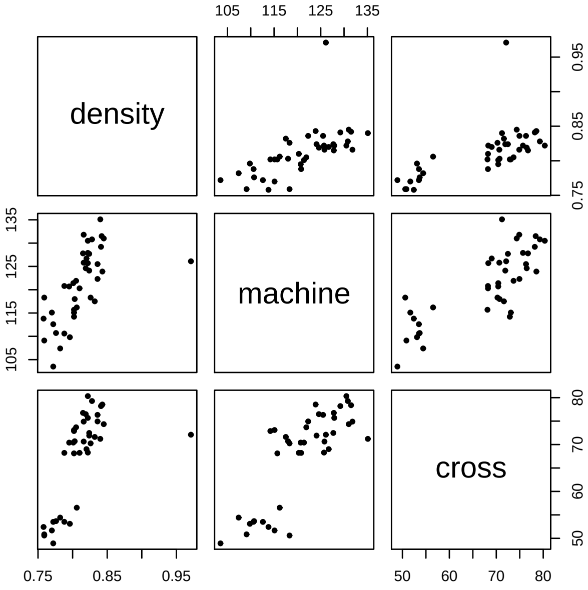
- x_1: density (grams per cubic centimer)
- x_2: strength (pounds) in the machine direction
- x_3: strength (pounds) in the cross direction

Data

density	machine	cross
0.801	121.41	70.42
0.824	127.7	72.47
0.841	129.2	78.2
0.816	131.8	74.89
0.84	135.1	71.21
0.842	131.5	78.39
0.82	126.7	69.02
⋮	⋮	⋮
0.758	113.8	52.42

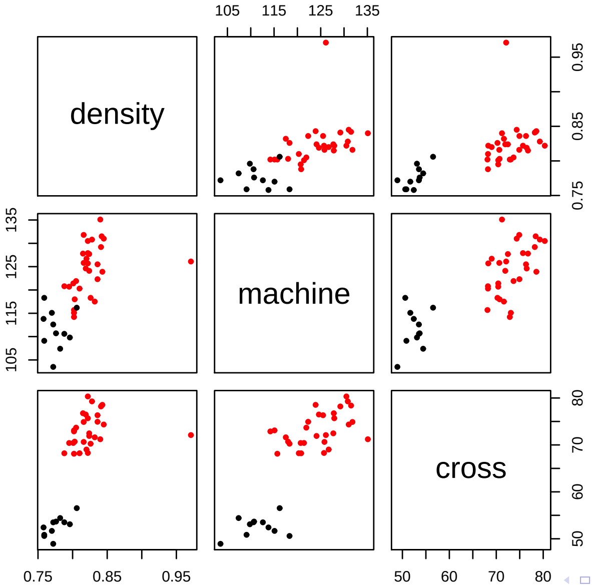
$n = 41.$

Example: Paper Strength



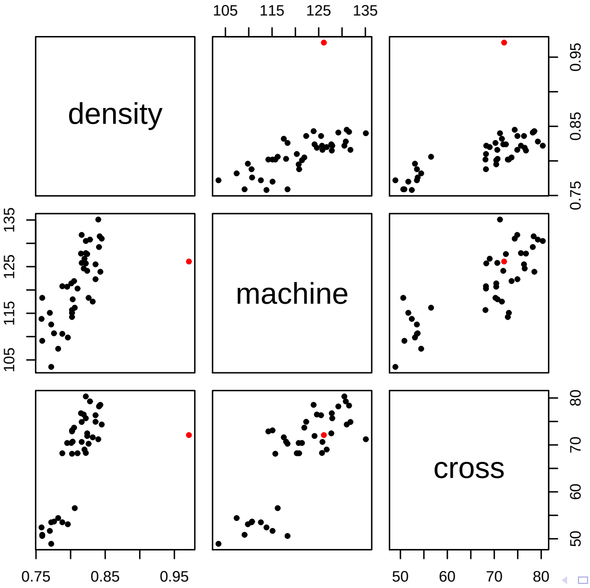
`density` and `machine` seem to have positive correlation.

Example: Paper Strength



The data can be separated into two groups according to strength cross.

Example: Paper Strength



A data point seems to be an outlier.

Distance/Metric

The Euclidean distance between two same dimensional random vectors $\mathbf{x} = (x_1, \ldots, x_k)$, $\mathbf{y} = (y_1, \ldots, y_k)$ is

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} = ||\mathbf{x} - \mathbf{y}||.$$

If each coordinate have different weight, then

$$d\mathbf{w}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k} w_i (x_i - y_i)^2}$$

becomes a weighted distance where $w_i \geq 0$ and $\sum w_i > 0$.

Distance/Metric

In general distance is a function between two points satisfying

(a) $d(\mathbf{x}, \mathbf{y}) \geq 0$ and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$. (*nonnegative*)

(b) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (*symmetric*)

(c) $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ (*triangle inequality*)

Example

$d(\mathbf{x}, \mathbf{y}) = \max_{i=1,\ldots,k} |x_i - y_i|$ is a distrance. But
$d(\mathbf{x}, \mathbf{y}) = \min_{i=1,\ldots,k} |x_i - y_i|$ is not a distrance.
For a positive definite matrixk A, define $d(\mathbf{x}, \mathbf{y}) = [(\mathbf{x} - \mathbf{y})^\top A(\mathbf{x} - \mathbf{y})]^{1/2}$ which is a distrance.

R demonstration

- R is a free statistical computing software
- R package can be download from http://www.r-project.org/
- Current stable version is 3.1.1
- There are many free cutting edge packages

R demonstration I

```
# win loss data
dt <- read.table("data/T1-1.DAT");
names(dt) <- c("payroll","winloss");
plot(dt$payroll/1e6,dt$winloss,pch=20);
plot(dt$payroll/1e6,dt$winloss,pch=20,xlim=c(0,4),ylim=c(0,.7

# sample mean, standard deviation
colMeans(dt);
mean(dt[,1]);

# unbiased sample standard deviation
sd(dt[,1]);
sd(dt[,2]);

# unbiased variance-covariance matrix
var(dt);
cor(dt);
```

R demonstration II

```
# paper strength data
dt <- read.table("data/T1-2.DAT");
names(dt) <- c("density","machine","cross");
colMeans(dt);
plot(dt,pch=20);
plot(dt,pch=20,col=1+(dt$cross > 60));
plot(dt,pch=20,col=1+1*(dt$density > .9));
```