STA437/2005 - Methods for Multivariate Data

Lecture 1–5

Gun Ho Jang

Fall 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:* Forecast 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

	Variable 1	Variable 2	$\cdots$	Variable j	$\cdots$	Variable p
Subject 1:	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1p}
Subject 2:	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Subject i:	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Subject n:	x_{n1}	x_{nn}	$\cdots$	x_{nj}	$\cdots$	x_{np}

Or simply

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \vdots & \ddots & \cdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \vdots & \ddots & \cdots \\ x_{n1} & x_{nn} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Example: Book Sales Record

Variables of Interest

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

Data record

Variable 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

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

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

Example: Paper Strength

Random variables

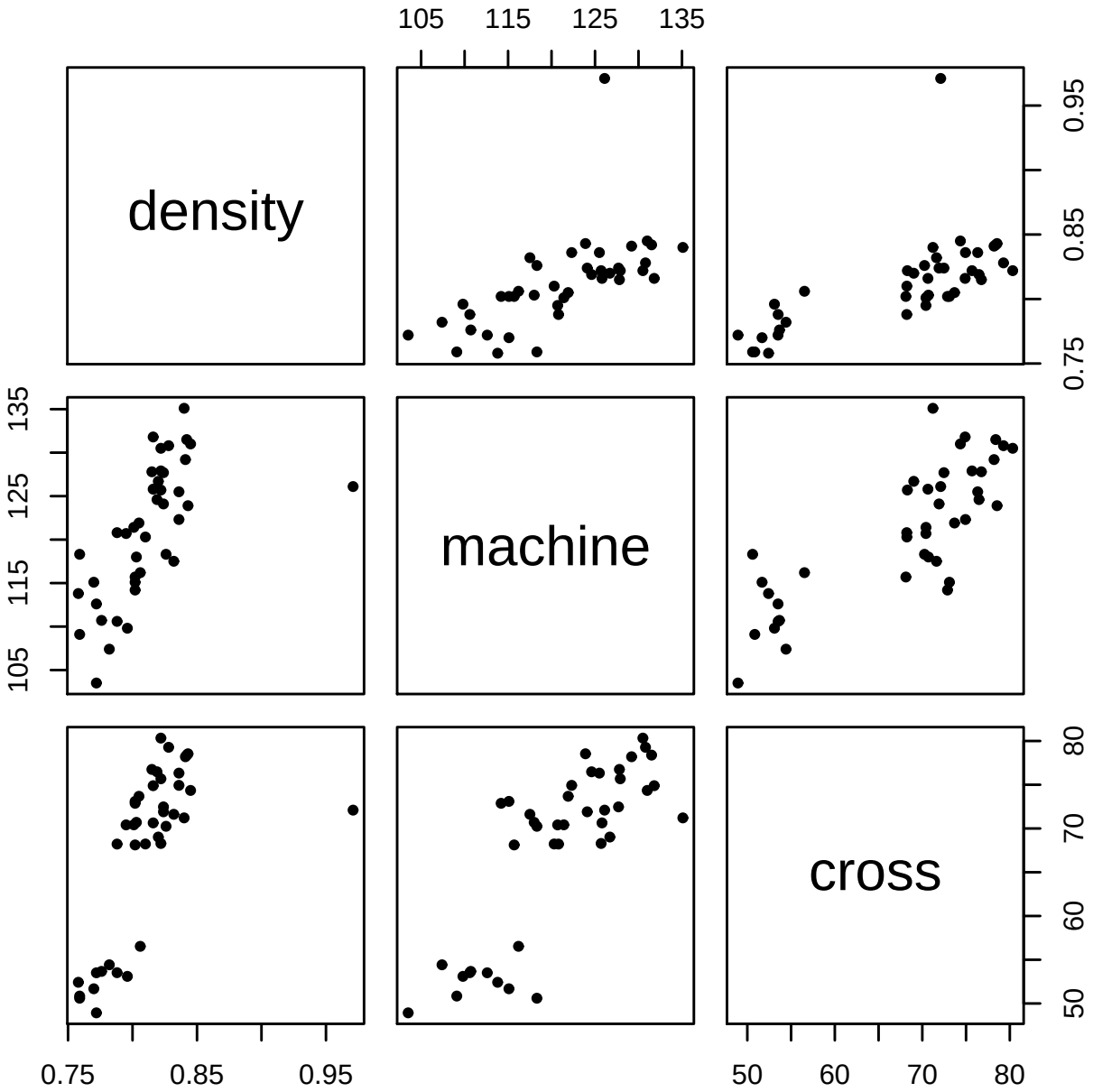
- x_1: density (grams per cubic centimetre)
- x_2: strength (pounds) in the machine direction
- x_3: strength (pounds) in the cross direction

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

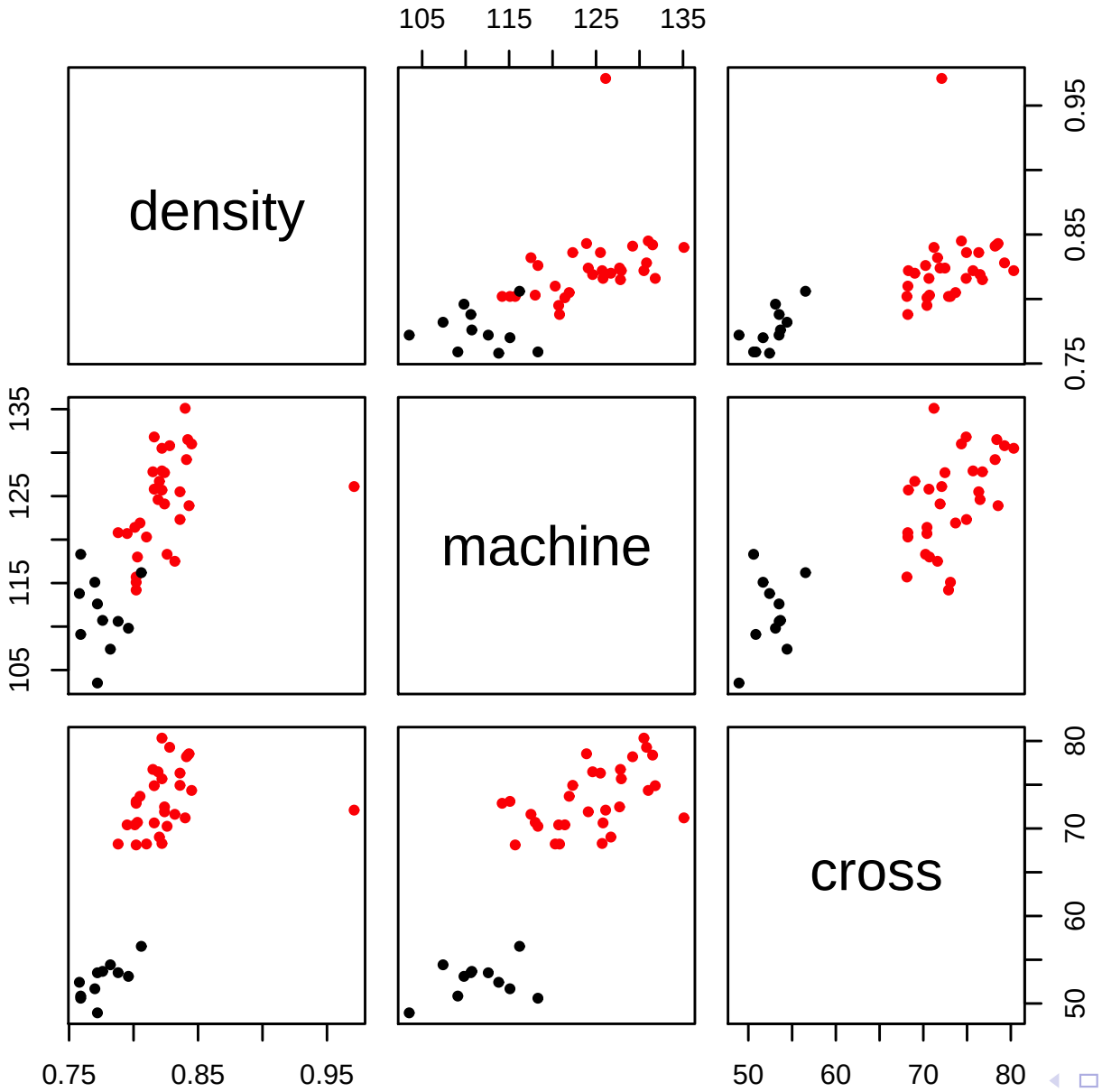
$n = 41.$

Example: Paper Strength



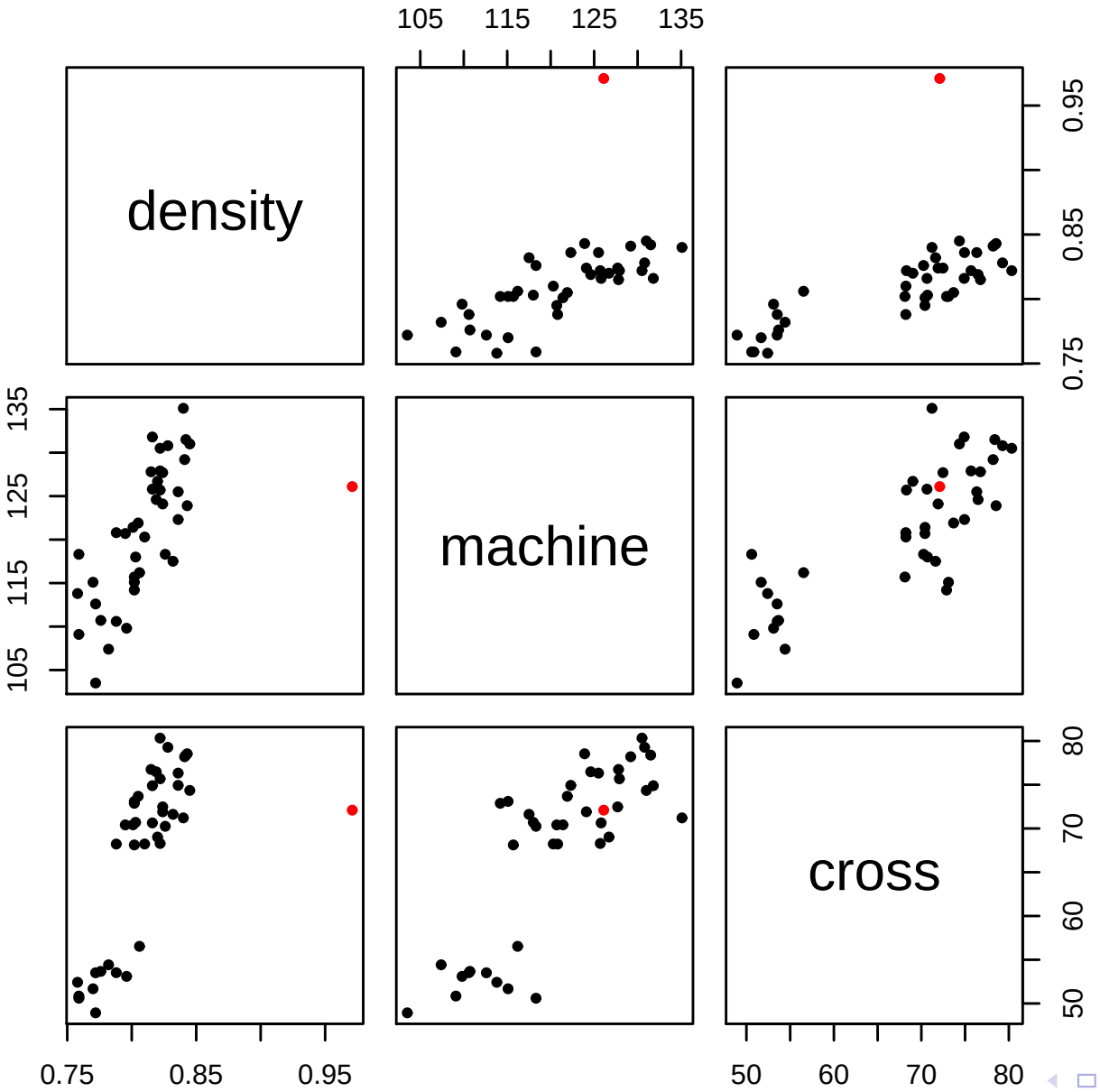
`density` and `machine` seem to have positive correlation.

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

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

```
1   # win loss data
2   dt <- read.table("data/T1-1.DAT");
3   names(dt) <- c("payroll","winloss");
4   plot(dt$payroll/1e6,dt$winloss,pch=20);
5   plot(dt$payroll/1e6,dt$winloss,pch=20,xlim=c(0,4),ylim=c(0,.7
6
7   # sample mean, standard deviation
8   colMeans(dt);
9   mean(dt[,1]);
10
11  # unbiased sample standard deviation
12  sd(dt[,1]);
13  sd(dt[,2]);
14
15  # unbiased variance-covariance matrix
16  var(dt);
17  cor(dt);
```

R demonstration II

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

Lecture 2

Matrix Algebra

Definition

A $k \times k$ matrix A is *non-negative definite* if and only if $\mathbf{v}^\top A\mathbf{v} \geq 0$ for any $\mathbf{v} \in \mathbb{R}^k$.
A $k \times k$ matrix A is *positive definite* if and only if $\mathbf{v}^\top A\mathbf{v} > 0$ for any $\mathbf{v} \in \mathbb{R}^k \backslash \{0\}$.

Definition

Eigen values are solution to $|A - \lambda I| = 0$. For any eigen value λ, there exists an *eigen vector* $\mathbf{v} \neq 0$ such that $A\mathbf{v} = \lambda\mathbf{v}$.

It is possible to have many eigen vectors for a eigen value.

Theorem (Spectral decomposition)

Let A be a symmetric $k \times k$ matrix. Then there exist k orthonormal eigen vectors $e_1, \ldots, e_k$ and corresponding eigen values $\lambda_1, \ldots, \lambda_k$ so that

$$A = \lambda_1 e_1 e_1^\top + \cdots + \lambda_k e^k e_k^\top .$$

A sketch proof.

Let $\mathbf{e} = (e_1 \ \ldots \ e_k)$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$. From the orthonormality, $\mathbf{e}^\top \mathbf{e} = (e_i^\top e_j) = I_k$. The uniqueness of inverse implies $\mathbf{e}\mathbf{e}^\top = I_k$. Since e_i's are eigen vectors, $A\mathbf{e} = (\lambda_1 e_1 \ \ldots \ \lambda_k e_k) = \mathbf{e}\Lambda$. Then $A = (\mathbf{e}\Lambda)\mathbf{e}^{-1} = \mathbf{e}\Lambda\mathbf{e}^\top = \lambda_1 e_1 e_1^\top + \cdots + \lambda_k e_k e_k^\top$. □

Expectation of Random Matrix

For a $n \times p$ random matrix $\mathbf{X} = (x_{ij})$, the expectation is defined by

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(x_{ij})) = \begin{pmatrix} \mathbb{E}(x_{11}) & \mathbb{E}(x_{12}) & \cdots & \mathbb{E}(x_{1p}) \\ \mathbb{E}(x_{21}) & \mathbb{E}(x_{22}) & \cdots & \mathbb{E}(x_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(x_{n1}) & \mathbb{E}(x_{n2}) & \cdots & \mathbb{E}(x_{np}) \end{pmatrix}.$$

Expectation of Random Matrix

Proposition

Let $\mathbf{X}, \mathbf{Y}$ *be two* $n \times p$ *random matrices and* A, B *be two conformable matrices. Then,*
(a) $\mathbb{E}(\mathbf{X} + \mathbf{Y}) = \mathbb{E}(\mathbf{X}) + \mathbb{E}(\mathbf{Y})$
(b) $\mathbb{E}(A\mathbf{X}B) = A\mathbb{E}(\mathbf{X})B$.

Proof.

(a) Let $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$ so that $Z_{ij} = X_{ij} + Y_{ij}$. Then
$\mathbb{E}(\mathbf{X} + \mathbf{Y}) = \mathbb{E}(\mathbf{Z}) = (\mathbb{E}(Z_{ij})) = (\mathbb{E}(X_{ij} + Y_{ij})) = (\mathbb{E}(X_{ij}) + \mathbb{E}(Y_{ij})) = (\mathbb{E}(X_{ij})) + (\mathbb{E}(Y_{ij})) = \mathbb{E}(\mathbf{X}) + \mathbb{E}(\mathbf{Y})$.
(b) Let $\mathbf{W} = A\mathbf{X}B$. Then $W_{kl} = \sum_{i=1}^{n} \sum_{j=1}^{p} A_{ki} X_{ij} B_{jl}$ and $\mathbb{E}(\mathbf{W}) = (\mathbb{E}(W_{kl})) = (\sum_{i=1}^{n} \sum_{j=1}^{p} A_{ki} \mathbb{E}(X_{ij}) B_{jl}) = ([A\mathbb{E}(\mathbf{X})B]_{kl}) = A\mathbb{E}(\mathbf{X})B$. □

Random Vector

Random vector $X = (x_1, \ldots, x_n)^\top$ is a $n \times 1$ random matrix. Hence the mean of X is defined by

$$\mathbb{E}(X) = (\mathbb{E}(x_i)) = \begin{pmatrix} \mathbb{E}(x_1) \\ \mathbb{E}(x_2) \\ \vdots \\ \mathbb{E}(x_n) \end{pmatrix}.$$

The *variance* of X is defined by the variance-covariance matrix, that is,

$$\begin{aligned} \mathbb{V}ar(X) &= (\mathrm{Cov}(x_i, x_j)) = (\mathbb{E}((x_i - \mathbb{E}(x_i))(x_j - \mathbb{E}(x_j)))) \\ &= \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top]. \end{aligned}$$

In general the *covariance* of two random vectors X and Y is defined by

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^\top].$$

Random Vector

Proposition

Let X, Y be two $n \times 1$ random vectors. Then,
(a) for $a, b \in \mathbb{R}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, $Cov(aX + \mathbf{v}, bY + \mathbf{w}) = abCov(X, Y)$,
(b) The mean and variance of $Z = AX$ for a matrix $A \in \mathbb{R}^{k \times n}$ are $\mathbb{E}(Z) = A\mathbb{E}(X)$ and $\mathbb{V}ar(Z) = A\mathbb{V}ar(X)A^\top$.

Proof.

(a)
$\text{Cov}(aX+\mathbf{v}, bY+\mathbf{w}) = \mathbb{E}((aX+\mathbf{v}-\mathbb{E}(aX+\mathbf{v}))(bY+\mathbf{w}-\mathbb{E}(bY+\mathbf{w}))) =$
$ab\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^\top) = ab\text{Cov}(X, Y)$.
(b) $\mathbb{E}(Z) = \mathbb{E}(AX) = A\mathbb{E}(X)$ and $\mathbb{V}ar(Z) = \mathbb{E}(ZZ^\top) - \mathbb{E}(Z)\mathbb{E}(Z)^\top =$
$\mathbb{E}(AXX^\top A^\top) - A\mathbb{E}(X)\mathbb{E}(X)^\top A^\top = A\mathbb{V}ar(X)A^\top$. □

Partition

Let X be a $n \times 1$ random vector. Consider a partition $X = ({X^{(1)}}^\top, {X^{(2)}}^\top)^\top)$, that is, for some $k, l > 0$ with $k + l = n$, $X^{(1)} = (X_1, \ldots, X_k)^\top$ and $X^{(2)} = (X_{k+1}, \ldots, X_{k+l})^\top$.

Proposition

The mean and variance of the partition becomes

$$\mathbb{E}(X) = \mathbb{E}\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X^{(1)}) \\ \mathbb{E}(X^{(2)}) \end{pmatrix}$$

$$\mathbb{V}ar(X) = \begin{pmatrix} Cov(X^{(1)}, X^{(1)}) & Cov(X^{(1)}, X^{(2)}) \\ Cov(X^{(2)}, X^{(1)}) & Cov(X^{(2)}, X^{(2)}) \end{pmatrix}$$

Hence the mean and variance can be computed blockwise.

Partition

Proof. Definitions and partition gives

$$\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(x_1) \\ \vdots \\ \mathbb{E}(x_k) \\ \mathbb{E}(x_{k+1}) \\ \vdots \\ \mathbb{E}(x_{k+l}) \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X^{(1)}) \\ \mathbb{E}(X^{(2)}) \end{pmatrix}$$

Similarly

$$\mathbb{V}ar(X) = (\mathrm{Cov}(x_i, x_j)) = \begin{pmatrix} \mathrm{Cov}(x_1, x_1) & \mathrm{Cov}(x_1, x_2) & \cdots & \mathrm{Cov}(x_1, x_n) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Cov}(x_2, x_2) & \cdots & \mathrm{Cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_n, x_1) & \mathrm{Cov}(x_n, x_2) & \cdots & \mathrm{Cov}(x_n, x_n) \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{Cov}(X^{(1)}, X^{(1)}) & \mathrm{Cov}(X^{(1)}, X^{(2)}) \\ \mathrm{Cov}(X^{(2)}, X^{(1)}) & \mathrm{Cov}(X^{(2)}, X^{(2)}) \end{pmatrix}$$

Partition

Exercise

Let Σ be a symmetric positive definite matrix with partition $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. *Express Σ^{-1} using Σ_{ij}'s.*

Random Sample

- Vectors of p measurements are supposed to be independent.
- $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is i.i.d.
- sample mean $\bar{\mathbf{x}} = (\mathbf{x}_1 + \cdots + \mathbf{x}_n)/n$ is unbiased estimator of $\mu = \mathbb{E}(\mathbf{x}_i)$
- covariance matrix of $\bar{\mathbf{x}}$ is Σ/n where $\Sigma = \mathbb{V}ar(\mathbf{x}_i)$.
- sample covariance matrix $S_n = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top/n$ is a consistent estimator of Σ with bias $-\Sigma/n$.
- $S = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})/(n-1)$ so that S is an unbiased estimator for Σ.

Random Sample

$$\mathbb{E}(\bar{\mathbf{x}}) = \mathbb{E}(\sum_{i=1}^{n} \mathbf{x}_i/n) = \sum_{i=1}^{n} \mathbb{E}(\mathbf{x}_i)/n = n\mathbb{E}(\mathbf{x}_1)/n = \mathbb{E}(\mathbf{x}_1) = \mu.$$

$$\mathbb{V}ar(\bar{x}) = \mathbb{V}ar(\sum_{i=1}^{n} x_i/n) = n^{-2} \sum_{i=1}^{n} \mathbb{V}ar(\mathbf{x}_i) = \mathbb{V}ar(\mathbf{x}_1)/n = \Sigma/n.$$

$$\begin{aligned}
\mathbb{E}(S_n) &= \frac{1}{n}\mathbb{E}(\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top) = \mathbb{E}((\mathbf{x}_1 - \bar{\mathbf{x}})(\mathbf{x}_1 - \bar{\mathbf{x}})^\top) \\
&= \mathrm{Cov}(\mathbf{x}_1 - \bar{\mathbf{x}}, \mathbf{x}_1 - \bar{\mathbf{x}}) = \mathbb{V}ar(\mathbf{x}_1) - \mathrm{Cov}(\mathbf{x}_1, \bar{\mathbf{x}}) - \mathrm{Cov}(\bar{\mathbf{x}}, \mathbf{x}_1) + \mathbb{V}ar(\bar{\mathbf{x}}) \\
&= \Sigma - \Sigma/n - \Sigma/n + \Sigma/n = \Sigma(1 - 1/n).
\end{aligned}$$

$$\mathbb{E}(S) = \mathbb{E}[\frac{n}{n-1} S_n] = \frac{n}{n-1}\Sigma\frac{n-1}{n} = \Sigma.$$

Generalized Variance

- Variance-covariance matrix contains $p \times p$ elements which is big to consider simultaneously.
- Simplified variance might be useful in interpretation.
- Suggestion: $|S|$, the determinant of unbiased variance-covariance matrix.
- $|S| = (n-1)^{-p}(\text{volume})^2$
- If $|S| = 0$, then there exists a linear relationship between variables.

Matrix form

Sample mean and variance can be expressed as a matrix and vector form.

$$\bar{\mathbf{x}} = \mathbf{X}^\top(\frac{1}{n}\mathbf{1}), \quad S = \frac{1}{n-1}\mathbf{X}^\top(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{X}.$$

Similar formulation is useful in multiple regression such as, for a regression model $Y = \beta^\top \mathbf{X} + \text{error}$,

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$
$$Y - \hat{Y} = Y - \hat{\beta}^\top X = Y(I - X(X^\top X)^{-1} X^\top)$$

Multivariate Normal Distribution

- Univariate normal density: $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$
- Let $z_1, \ldots, z_k \sim i.i.d.\ N(0,1)$.
- joint density of $Z = (z_1, \ldots, z_k)$ is

$$\text{pdf}_Z(\mathbf{z}) = \prod_{i=1}^{k} (2\pi)^{-1/2} \exp(-z_i^2/2) = |2\pi I_k|^{-1/2} \exp(-\frac{1}{2}\mathbf{z}^\top \mathbf{z}).$$

- Let $X = \mu + \Sigma^{1/2} Z$
- Density of X can be obtained using the change of variable formula

$$\begin{aligned}
\text{pdf}_X(\mathbf{x}) &= \text{pdf}_Z(\Sigma^{-1/2}(\mathbf{x}-\mu)) \cdot \left|\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right| \\
&= |2\pi I_k|^{-1/2} \exp(-\frac{1}{2}(\Sigma^{-1/2}(\mathbf{x}-\mu))^\top (\Sigma^{-1/2}(\mathbf{x}-\mu))) \times |\Sigma^{-1/2}| \\
&= |2\pi I_k|^{-1/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)) \\
&= |2\pi\Sigma|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)).
\end{aligned}$$

Multivariate Normal Distribution

Proposition

If $X \sim N_p(\mu, \Sigma)$, then for a conformable matrix A, $AX \sim N(A\mu, A\Sigma A^\top)$.

Proof.

Note that $X = \mu + \Sigma^{1/2} Z$ implies $AX = A(\mu + \Sigma^{1/2} Z) = A\mu + A\Sigma^{1/2} Z$. Hence AX is a normal distribution with mean $A\mu$ and variance $A\Sigma^{1/2}(A\Sigma^{1/2})^\top = A\Sigma A^\top$. □

Proposition

If $X \sim N_k(\mu, \Sigma)$ with $|\Sigma| > 0$ and $k = rank(\Sigma)$, then $(X-\mu)^\top \Sigma^{-1}(X-\mu) \sim \chi^2(k)$.

Proof.

Note $X = \mu + \Sigma^{1/2} Z$ for $Z \sim N_k(O, I_k)$. Then $(X-\mu)^\top \Sigma^{-1}(X-\mu) = (\mu + \Sigma^{1/2} Z - \mu)^\top \Sigma^{-1/2}\Sigma^{-1/2}(\mu + \Sigma^{1/2} Z - \mu) = Z^\top Z = Z_1^2 + \cdots + Z_p^2 \sim \chi^2(p)$. □

Multivariate Normal Distribution

- the ellipsoid $C_\gamma = \{\mathbf{x} : (\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) \leq \chi^2_\gamma(p)\}$ has probability γ from $N_p(\mu, \Sigma)$ for $0 < \gamma < 1$
- the structure of high probable regions have of the C_γ forms.
- similar form of confidence regions will be studied.
- It will reappear in Hotelling's T-statistic

Multivariate Normal Distribution

Proposition

The moment generating function of $X \sim N_k(\mu, \Sigma)$ *is* $mgf_X(\mathbf{t}) = \exp(\mathbf{t}^\top \mu + \mathbf{t}^\top \Sigma \mathbf{t}/2)$.

Proof.

$$\begin{aligned} \text{mgf}_X(\mathbf{t}) &= \mathbb{E}[\exp(\mathbf{t}^\top X)] = \mathbb{E}[\exp(\mathbf{t}^\top(\mu + \Sigma^{1/2} Z))] \\ &= \exp(\mathbf{t}^\top \mu)\text{mgf}_Z((\mathbf{t}^\top \Sigma^{1/2})^\top). \end{aligned}$$

Since $z_1, \ldots, z_k$ are independent, for $\mathbf{u} = (\mathbf{t}^\top \Sigma^{1/2})^\top = \Sigma^{1/2}\mathbf{t}$,

$$\begin{aligned} \text{mgf}_Z(\mathbf{u}) &= \prod_{i=1}^{k} \exp(u_i^2/2) = \exp(\mathbf{u}^\top \mathbf{u}/2) = \exp((\Sigma^{1/2}\mathbf{t})^\top(\Sigma^{1/2}\mathbf{t})/2) \\ &= \exp(\mathbf{t}^\top \Sigma \mathbf{t}/2). \end{aligned}$$

Hence $\text{mgf}_X(\mathbf{t}) = \exp(\mathbf{t}^\top \mu + \mathbf{t}^\top \Sigma \mathbf{t}/2)$. □

Multivariate Normal Distribution

Exercise

Show that if $x_j \sim N(\mu_j, \Sigma_j)$ *are independent, then* $x_1 + \cdots + x_k \sim N(\mu_1 + \cdots + \mu_k, \Sigma_1 + \cdots + \Sigma_k)$.

Multivariate Normal Distribution

Proposition

Suppose that $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \sim N(0, \Sigma)$ *with* $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ *and* $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. *Then,*

(a) $X^{(i)} \sim N(\mu_i, \Sigma_{ii})$.

(b) If $\Sigma_{12} = O$, *then* $X^{(1)}$ *and* $X^{(2)}$ *are independent.*

(c) $X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}X^{(2)}$ *and* $X^{(2)}$ *are independent.*

(d) $X^{(1)} \mid X^{(2)} = x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11\cdot 2})$ *where* $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Multivariate Normal Distribution

Proof. (a) Let $\mathbf{t} = (\mathbf{t}_1^\top, O_{1\times l})^\top$. Then
$\text{mgf}_{X^{(1)}}(\mathbf{t}_1) = \mathbb{E}[\exp(\mathbf{t}_1^\top X^{(1)})] = \mathbb{E}[\exp(\mathbf{t}^\top X)] = \text{mgf}_X(\mathbf{t}) =$
$\exp(\mathbf{t}^\top \mu + \mathbf{t}^\top \Sigma \mathbf{t}/2) = \exp(\mathbf{t}_1^\top \mu_1 + \mathbf{t}_1^\top \Sigma_{11} \mathbf{t}_1/2)$. Thus $X^{(1)} \sim N(\mu_1, \Sigma_{11})$.
Similarly, $X^{(2)} \sim N(\mu_2, \Sigma_{22})$.
(b) Let $\mathbf{t} = (\mathbf{t}_1^\top, \mathbf{t}_2^\top)^\top$. Then

$$\text{mgf}_{X^{(1)},X^{(2)}}(\mathbf{t}_1, \mathbf{t}_2) = \text{mgf}_X(\mathbf{t}) = \exp(\mathbf{t}^\top \mu + \mathbf{t}^\top \Sigma \mathbf{t}/2)$$
$$= \exp(\mathbf{t}_1^\top \mu_1 + \mathbf{t}_2^\top \mu_2 + \mathbf{t}_1^\top \Sigma_{11} \mathbf{t}_1/2 + \mathbf{t}_1^\top \Sigma_{12} \mathbf{t}_2/2 + \mathbf{t}_2^\top \Sigma_{21} \mathbf{t}_1/2 + \mathbf{t}_2^\top \Sigma_{22} \mathbf{t}_2/2)$$
$$= \exp(\mathbf{t}_1^\top \mu_1 + \mathbf{t}_1^\top \Sigma_{11} \mathbf{t}_1/2) \times \exp(\mathbf{t}_2^\top \mu_2 + \mathbf{t}_2^\top \Sigma_{22} \mathbf{t}_2/2) = \text{mgf}_{X^{(1)}}(\mathbf{t}_1) \times \text{mgf}_{X^{(2)}}(\mathbf{t}_2$$

Hence $X^{(1)}$ and $X^{(2)}$ are independent if and only if $\Sigma_{12} = O$.
(c) The covariance between $X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1} X^{(2)}$ and $X^{(2)}$ is

$$\text{Cov}(X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1} X^{(2)}, X^{(2)}) = \text{Cov}(X^{(1)}, X^{(2)}) - \Sigma_{12}\Sigma_{22}^{-1} \text{Cov}(X^{(2)}, X^{(2)})$$
$$= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = \Sigma_{12} - \Sigma_{12} = O.$$

Hence $X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1} X^{(2)}$ and $X^{(2)}$ are independent.

Multivariate Normal Distribution

(d) From (c), $X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}X^{(2)}$ is independent from $X^{(2)}$ and normally distributed with mean $\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$ and variance $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. If $X^{(2)} = \mathbf{x}_2$ is given,

$$\begin{aligned} X^{(1)} \mid X^{(2)} = \mathbf{x}_2 &\equiv^d X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}X^{(2)} + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{x}_2 \\ &\sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11\cdot 2}) + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{x}_2 \\ &\sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2), \Sigma_{11\cdot 2}). \end{aligned}$$

Lecture 3

Matrix Algebra I

Definition

Let A be a $k \times k$ matrix. The *trace* of A is $A_{11} + \cdots + A_{kk} = \sum_{i=1}^{k} A_{ii}$.

Theorem

Let A and B be two $k \times k$ matrices. Then
(a) $tr(A^\top) = tr(A)$
(b) $tr(A + B) = tr(A) + tr(B)$
(c) $tr(AB) = tr(BA)$
(d) For $C, D^\top \in \mathbb{R}^{l \times k}$, $tr(CAD) = tr(ADC)$.

Proof.

(a) $\text{tr}(A^\top) = \sum_{i=1}^k (A^\top)_{ii} = \sum_{i=1}^k A_{ii} = \text{tr}(A)$.
(b) $\text{tr}(A+B) = \sum_{i=1}^k (A+B)_{ii} = \sum_{i=1}^k [A_{ii} + B_{ii}] =$
$\sum_{i=1}^k A_{ii} + \sum_{i=1}^k B_{ii} = \text{tr}(A) + \text{tr}(B)$.
(c) $\text{tr}(AB) = \sum_{i=1}^k (AB)_{ii} = \sum_{i=1}^k \sum_{j=1}^k A_{ij} B_{ji} = \sum_{j=1}^k \sum_{i=1}^k B_{ji} A_{ij} =$
$\sum_{j=1}^k (BA)_{jj} = \text{tr}(BA)$.
(d) $\text{tr}(CAD) = \sum_{m=1}^l (CAD)_{mm} = \sum_{m=1}^l \sum_{i=1}^k \sum_{j=1}^l C_{mi} A_{ij} (D)_{jm} =$
$\sum_{i=1}^k \sum_{j=1}^l \sum_{m=1}^l A_{ij} D_{jm} C_{mi} = \sum_{i=1}^k (ADC)_{ii} = \text{tr}(ADC)$. □

Matrix Algebra III

Definition

Let A be a $k \times k$ matrix. The *determinant* of A is A_{11} if $k = 1$ and for $k > 1$ and any j

$$|A| = \sum_{i=1}^{k} A_{ij}(-1)^{i+j}|A_{-i,-j}|$$

where $A_{-i,-j}$ is the minor matrix of A removed ith row and jth column.

Proposition

For $A, B \in \mathbb{R}^{k\times k}$, $|A^\top| = |A|$, $|AB| = |A| \times |B|$ *and* $A^{-1} = ((-1)^{i+j}|A_{-j,-i}|/|A|)$.

Maximal Likelihood Estimation I

The density of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times p}$ is the joint density of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ given by

$$\begin{aligned} \text{pdf}_{\mathbf{X}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= \prod_{i=1}^{n} |2\pi\Sigma|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}_i - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \mu)) \\ &= |2\pi\Sigma|^{-n/2} \exp\Big(-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \mu) \Big) \end{aligned}$$

The sum of quadratic form in the exponent can be simplified using

$$\begin{aligned} &(\mathbf{x}_i - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \mu) = (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \mu) \\ &= (\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + (\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu) \\ &\quad + (\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu). \end{aligned}$$

Maximal Likelihood Estimation II

Similarly the sum of quadratic form separated into four parts as follows

$$\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu)$$

$$+ \sum_{i=1}^{n}(\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + n(\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu)$$

$$= \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + n(\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu)$$

The first term is sum of traces, that is,
$(\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) = \mathrm{tr}[(\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})] = \mathrm{tr}[\Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top]$.

Maximal Likelihood Estimation III

Hence, the sum becomes

$$\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) = \sum_{i=1}^{n} \mathrm{tr}[\Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top]$$
$$= \mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\right] = \mathrm{tr}(AB)$$

where $A = \Sigma^{-1}$ and $B = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$. For the same A and B, the density function becomes

$$(2\pi)^{-np/2}|A|^{n/2}\exp(-\mathrm{tr}(AB)/2 - n(\bar{\mathbf{x}} - \mu)^\top A(\bar{\mathbf{x}} - \mu))$$
$$\leq (2\pi)^{-np/2}|A|^{n/2}\exp(-\mathrm{tr}(AB)/2).$$

The equality holds if and only if $\mu = \bar{\mathbf{x}}$. Which implies the maximum likelihood estimator of μ is $\hat{\mu}_{\mathrm{MLE}} = \bar{\mathbf{x}}$.

Maximal Likelihood Estimation IV

The maximum likelihood estimator Σ can be obtained by maximizing

$$n \log |A| - \text{tr}(AB) = n \log \left(\sum_{k=1}^{p} A_{ik} (-1)^{i+k} |A_{-k,-i}| \right) - \sum_{k=1}^{p} \sum_{l=1}^{p} A_{kl} B_{lk}.$$

Since the partial derivative of $|A|$ with respect to A_{ij} is

$$\frac{\partial |A|}{\partial A_{ij}} = \frac{\partial}{\partial A_{ij}} \sum_{k=1}^{p} (-1)^{i+k} A_{ik} |A_{-k,-i}| = (-1)^{i+j} |A_{-j,-i}|.$$

Then the first and second partial derivatives with respect to A_{ij} are

$$n \frac{1}{|A|} \frac{\partial |A|}{\partial A_{ij}} - B_{ji} = n \frac{(-1)^{i+j} |A_{-j,-i}|}{|A|} - B_{ji} = n[A^{-1}]_{ji} - B_{ji},$$

$$- n(-1)^{i+j} \frac{|A_{-j,-i}|}{|A|^2} \frac{\partial |A|}{\partial A_{ij}} = -n(-1)^{i+j} \frac{|A_{-j,-i}|}{|A|^2} \times (-1)^{i+j} |A_{-j,-i}|$$

$$= -n \frac{|A_{-j,-i}|^2}{|A|^2} \leq 0$$

Maximal Likelihood Estimation V

Hence the maximum is obtained at $n[A^{-1}]_{ji} = B_{ji}$. In other words, $A^{-1} = B/n$. Thus the maximum likelihood estimator $\widehat{\Sigma}_{\text{MLE}} = \widehat{A}^{-1} = B/n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$.

In sum the maximum likelihood estimators are

Maximum likelihood estimator

$\widehat{\mu} = \bar{\mathbf{x}}$ and $\widehat{\Sigma} = S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$.

Method of Moment Estimator

The first and second moments are

$$\mathbb{E}[\mathbf{x}_i] = \mu \quad \text{and} \quad \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \mathbb{V}ar(\mathbf{x}_i) + \mathbb{E}(\mathbf{x}_i)\mathbb{E}(\mathbf{x}_i)^\top = \Sigma + \mu\mu^\top.$$

The corresponding sample moments solve the method of moment estimator (MME), that is,

$$\widehat{\mu}_{\text{MME}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \bar{\mathbf{x}}, \quad \widehat{\Sigma}_{\text{MME}} + \widehat{\mu}_{\text{MME}}\widehat{\mu}_{\text{MME}}^\top = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$$

Hence the solutions are

$$\widehat{\mu}_{\text{MME}} = \bar{\mathbf{x}}, \quad \widehat{\Sigma}_{\text{MME}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

Comparison

Unbiased Estimator

The (minimum variance) unbiased estimators are $\widehat{\mu}_{\text{UE}} = \bar{\mathbf{x}}$ and $\widehat{\Sigma}_{\text{UE}} = S = \frac{n}{n-1} S_n = \frac{n}{n-1} \widehat{\Sigma}_{\text{MLE}}$.

Exercise

Show that MLE and MME are the same for univariate normal distribution.

Note

Even for multivariate normal distribution, MLE and MME are the same.

Note

Since the joint density function is a function of $\bar{\mathbf{x}}$ and S, the pair $(\bar{\mathbf{x}}, S)$ is a sufficient statistic.

The Distribution of $\bar{\mathbf{x}}$ and S

- The distribution of the sample mean $\bar{\mathbf{x}}$ is a multivariate normal because it is a weighted sum of independent multivariate normal random variables.
- $\mathbb{E}(\bar{\mathbf{x}}) = \mu$, $\mathbb{V}ar(\bar{\mathbf{x}}) = n^{-2}\mathbb{V}ar(\mathbf{x}_1 + \cdots + \mathbf{x}_n) = \Sigma/n$
- $\bar{\mathbf{x}} \sim N(\mu, \Sigma/n)$.
- The distribution of S is a bit complicated
- Recall that for the univariate case $(n-1)s^2/\sigma^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1) \sim \text{Gamma}((n-1)/2, 1/2)$
- $s^2 \sim \text{Gamma}((n-1)/2, (n-1)/(2\sigma^2))$.

Wishart Distribution

Definition

Let $Z_i \sim i.i.d.\ N_p(O, \Sigma)$ for $i = 1, \ldots, m$. The distribution of the quadratic sum

$$W_p(\Sigma, m) \equiv^d \sum_{i=1}^{m} Z_i Z_i^\top$$

is called the *Wishart* distribution with m degree of freedom and parameter Σ where p is the rank of Σ.

Note

Wishart distribution is a multivariate version of χ^2 distribution.

Proposition

The moment generating function of $\mathbf{A} \sim W_p(\Sigma, m)$ is $mgf_{\mathbf{A}}(U) = |I_p - 2U\Sigma|^{-m/2}$ for $U \in \mathbb{R}^{p \times p}$.

Wishart Distribution

Proof.

The matrix version of moment generating function is

$$\mathbb{E}[\exp(\sum_{i=1}^{p}\sum_{j=1}^{p} U_{ij}A_{ij})] = \mathbb{E}[\exp(\mathrm{tr}(U^\top \mathbf{A}))] = \mathbb{E}[\exp(\sum_{i=1}^{m} \mathrm{tr}(U^\top Z_i Z_i^\top))]$$

$$= \prod_{i=1}^{m} \mathbb{E}[\exp(\mathrm{tr}(U^\top Z_i Z_i^\top))]. \quad \text{Then,}$$

$$\mathbb{E}[\exp(\mathrm{tr}(U^\top Z_i Z_i^\top))] = \mathbb{E}[\exp(\mathrm{tr}(Z_i^\top U^\top Z_i))] = \mathbb{E}[\exp(\mathrm{tr}(Z_i^\top U Z_i))]$$

$$= \mathbb{E}[\exp(Z_i^\top U Z_i)] = \int \exp(\mathbf{x}^\top U\mathbf{x}) \times |2\pi\Sigma|^{-1/2} \exp(-\mathbf{x}^\top \Sigma^{-1}\mathbf{x}/2)\, d\mathbf{x}$$

$$= |2\pi\Sigma|^{-1/2} \int \exp(-\mathbf{x}^\top(\Sigma^{-1} - 2U)\mathbf{x}/2)\, d\mathbf{x}$$

$$= |2\pi\Sigma|^{-1/2} |2\pi(\Sigma^{-1} - 2U)^{-1}|^{1/2} = |I_p - 2U\Sigma|^{-1/2}.$$

Hence $\mathrm{mgf}_{\mathbf{A}}(U) = |I_p - 2U\Sigma|^{-m/2}$ for some $U \in \mathbb{R}^{p\times p}$ around O. □

Wishart Distribution

Proposition

(a) If $\mathbf{A} \sim W_p(\Sigma, m)$ *and* $\mathbf{B} \sim W_p(\Sigma, n)$ *are independent, then* $\mathbf{A} + \mathbf{B} \sim W_p(\Sigma, m+n)$.
(b) If $\mathbf{A} \sim W_p(\Sigma, m)$ *and* $C \in \mathbb{R}^{k \times p}$, *then* $C\mathbf{A}C^\top \sim W_k(C\Sigma C^\top, m)$.

Proof.

(a) $\text{mgf}_{\mathbf{A}+\mathbf{B}}(U) = \mathbb{E}[\exp(\text{tr}(U^\top(\mathbf{A}+\mathbf{B})))] =$
$\mathbb{E}[\exp(\text{tr}(U^\top \mathbf{A}))]\mathbb{E}[\exp(\text{tr}(U^\top \mathbf{B}))] = |I_p - 2U\Sigma|^{-m/2}|I_p - 2U\Sigma|^{-n/2} =$
$|I_p - 2U\Sigma|^{-(m+n)/2} \sim W_p(\Sigma, m+n)$.
(b) There exists $Z_1, \ldots, Z_m \sim i.i.d.\ N(O, \Sigma)$ such that
$\mathbf{A} = Z_1 Z_1^\top + \cdots + Z_m Z_m^\top$. Then $C\mathbf{A}C^\top = C(Z_1 Z_1^\top + \cdots + Z_m Z_m^\top)C^\top =$
$(CZ_1)(CZ_1)^\top + \cdots + (CZ_m)(CZ_m)^\top \sim W_k(C\Sigma C^\top, m)$ because
$Y_i = CZ_i \sim i.i.d.\ N_k(O, C\Sigma C^\top)$ □

Wishart Distribution

Proposition

The density function of $\mathbf{A} \sim W_p(\Sigma, m)$ *is*

$$pdf_{\mathbf{A}}(\mathbf{A}) = |\mathbf{A}|^{(m-p-1)/2} \exp(-tr(\Sigma^{-1}\mathbf{A})/2)/[2^{mp/2}|\Sigma|^{m/2}\Gamma_p(m/2)]$$

where $\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(x + (1-j)/2)]$.

A proof can be found in "Muirhead (2005). Aspects of Multivariate Statistical Theory."

Proposition

If $\mathbf{x}_i \sim i.i.d.\ N_p(\mu, \Sigma)$, *then* $\bar{\mathbf{x}} \sim N_p(\mu, \Sigma/n)$ *and* $(n-1)S \sim W_p(\Sigma, n-1)$ *are independent.*

Proof of Proposition I

Note that $\mathrm{Cov}(\bar{\mathbf{x}}, \mathbf{x}_i - \bar{\mathbf{x}}) = \mathrm{Cov}(\bar{\mathbf{x}}, \mathbf{x}_i) - \mathbb{V}ar(\mathbf{x}) = \Sigma/n - \Sigma/n = O$.
Hence $\bar{\mathbf{x}}$ and $\{\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}\}$ are independent. So are $\bar{\mathbf{x}}$ and $(n-1)S = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$.
There exists an orthonormal matrix $U = (u_{ij}) = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ such that $UU^\top = I_n$ and $\mathbf{u}_n = \mathbf{1}_n/\sqrt{n}$. Then
$\mathbf{u}_j^\top \mathbf{u}_n = \sum_{i=1}^n u_{ij} u_{in} = n^{-1/2} \sum_{i=1}^n u_{ij} = 0$ and $\sum_{i=1}^n u_{ij}^2 = 1$. For $\mathbf{x}_j \sim i.i.d.\ N_p(\mu, \Sigma)$, define $Y_j = \sum_{i=1}^n u_{ij} \mathbf{x}_i$. Then, for $j = 1, \ldots, n-1$, $Y_j \sim N_p(\sum_{i=1}^n u_{ij}\mu, \sum_{i=1}^n u_{ij}^2 \Sigma) \sim N_p(O, \Sigma)$ and
$\mathrm{Cov}(Y_j, Y_k) = \sum_{i=1}^n \mathrm{Cov}(u_{ij}\mathbf{x}_i, u_{ik}\mathbf{x}_i) = \sum_{i=1}^n u_{ij} u_{ik} \Sigma = \mathbf{u}_j^\top \mathbf{u}_k \Sigma = O$ if $j \neq k$. Hence $Y_1, \ldots, Y_{n-1} \sim i.i.d.\ N_p(O, \Sigma)$. By the definition, $\sum_{i=1}^{n-1} Y_i Y_i^\top \sim W_p(\Sigma, n-1)$ and

$$\sum_{j=1}^{n-1} Y_j Y_j^\top = \sum_{j=1}^{n-1} \sum_{i=1}^{n} u_{ij} \mathbf{x}_i \sum_{k=1}^{n} u_{kj} \mathbf{x}_k^\top = \sum_{i=1}^{n} \sum_{k=1}^{n} \mathbf{x}_i \mathbf{x}_k^\top \sum_{j=1}^{n-1} u_{ij} u_{kj}$$

Proof of Proposition II

The assumption $UU^\top = I_p = U^\top U$ implies
$\sum_{j=1}^{n-1} u_{ij}u_{kj} = \sum_{j=1}^{n} u_{ij}u_{kj} - u_{in}u_{kn} = I(i=k) - 1/n.$

$$= \sum_{i=1}^{n}\sum_{k=1}^{n} \mathbf{x}_i\mathbf{x}_k^\top (I(i=k) - 1/n) = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top - n\bar{\mathbf{x}}\bar{\mathbf{x}}^\top$$

$$= \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top = (n-1)S.$$

Comparison

Univariate	item	Multivariate
$X_i \sim N(\mu, \sigma^2)$	sample	$\mathbf{x}_i \sim N_p(\mu, \Sigma)$
$\exp(\mu t + t^2\sigma^2/2)$	mgf	$\exp(\mathbf{t}^\top \mu + \mathbf{t}^\top \Sigma \mathbf{t}/2)$
$\bar{X} \sim N(\mu, \sigma^2/n)$	distribution	$\bar{\mathbf{x}} \sim N_p(\mu, \Sigma/n)$
$(n-1)S/\sigma^2 \sim \chi^2(n-1)$		$(n-1)S \sim W_p(\Sigma, n-1)$

Large Sample Property I

Univariate law of large numbers

For an i.i.d. $X_1, X_2, \ldots$ with $\mathbb{E}(X_i) = \mu$, the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$ converges to μ almost surely.

Proposition

Let $Y_1, Y_2, \ldots$ be i.i.d. with mean $\mathbb{E}(Y_i) = \mu \in \mathbb{R}^p$. Then $\bar{Y} = (Y_1 + \cdots + Y_n)/n \to \mu$ in probability.

Proof.

The law of large numbers can be applicable for each coordinate, that is, $U_{jn} = (Y_{1j} + \cdots + Y_{nj})/n \to \mathbb{E}(Y_{ij}) = \mu_j$ almost surely. Then $P(\lim_{n\to\infty} \bar{Y}_n \neq \mu) \leq \sum_{j=1}^{p} P(\lim_{n\to\infty} U_{jn} \neq \mu_j) = 0$. □

Large Sample Property II

Example

$S_n \to \Sigma$ and $S \to \Sigma$ almost surely. It is shown that
$S_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top$. Hence
$[S_n]_{ij} = \frac{1}{n}\sum_{k=1}^{n} x_{ni}x_{nj} - (\bar{\mathbf{x}})_i(\bar{\mathbf{x}})_j \to \mathbb{E}(x_{1i}x_{1j}) - \mu_i\mu_j = \mathrm{Cov}(x_{1i}, x_{1j}) = \Sigma_{ij}$
almost surely. Hence $S = S_n n/(n-1) \to \Sigma$ almost surely.

Proposition

Let $Y_1, Y_2, \ldots$ be i.i.d. with mean $\mathbb{E}(Y_j) = \mu \in \mathbb{R}^p$ and $\mathbb{V}ar(Y_j) = \Sigma$. Then $\sqrt{n}(\bar{Y} - \mu) \to N(O, \Sigma)$ in distribution.

Large Sample Property III

Proof.

The theorem can be proven using the convergence of characteristic functions. Fix $\mathbf{t} \in \mathbb{R}^p$. Define $Z_j = \mathbf{t}^\top (Y_j - \mu)$ so that $\mathbb{E}(Z_j) = \mathbf{t}^\top (\mathbb{E}(Y_j) - \mu) = 0$ and $\mathbb{V}ar(Z_j) = \mathbf{t}^\top \mathbb{V}ar(Y_j)\mathbf{t} = \mathbf{t}^\top \Sigma \mathbf{t}$. Using the central limit theorem for univariate random variables,

$$\text{chf}_{\sqrt{n}\bar{Z}}(u) = \mathbb{E}[\exp(iu\sqrt{n}\bar{Z})] \to \exp(-u^2 \mathbf{t}^\top \Sigma \mathbf{t}/2).$$

Then the characteristic function of $\sqrt{n}(\bar{Y} - \mu)$ at $\mathbf{t}$ is

$$\begin{aligned}\text{chf}_{\sqrt{n}(\bar{Y}-\mu)}(\mathbf{t}) &= \mathbb{E}[\exp(i\mathbf{t}^\top \sqrt{n}(\bar{Y} - \mu))] = \mathbb{E}[\exp(i\sqrt{n}\bar{Z})] = \text{chf}_{\sqrt{n}\bar{Z}}(1) \\ &\to \exp(-\mathbf{t}^\top \Sigma \mathbf{t}/2).\end{aligned}$$

Hence $\sqrt{n}(\bar{Y} - \mu)$ converges to $N(O, \Sigma)$ in distribution. □

Large Sample Property IV

Example

Using the continuous mapping theorem and the central limit theorem,

$$n(\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu) = [\sqrt{n}(\bar{\mathbf{x}} - \mu)]^\top \Sigma^{-1}[\sqrt{n}(\bar{\mathbf{x}} - \mu)] \rightarrow Z^\top \Sigma^{-1} Z \sim \chi^2(p)$$

in distribution where $Z \sim N_p(O, \Sigma)$.

Exercise

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be i.i.d. with mean μ and variance Σ. Show that $n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \rightarrow \chi^2(p)$.

Lecture 4

Assessment of Normality

Most of data analysis methods were developed under normality assumption of observed data. Hence it should be checked whether or not observed data is normally distributed. Simple check-ups.

- Whether or not marginal distributions follow normal distribution
- Whether or not pair-wise distributions follow normal distribution
- Whether or not there are wild observations which are quite different from normal distributions.

Univariate normal check-ups

- Let $Z \sim N(\mu, \sigma^2)$.
- $P(Z \in (\mu - k\sigma, \mu + k\sigma)) = \Phi(k) - \Phi(-k) = 2(\Phi(k) - 1/2) = 2\Phi(k) - 1$.
- The probability is 0.6827 if $k = 1$ and 0.9545 if $k = 2$.
- Normality is violated if the proportions of samples contained in $\widehat{\mu} \pm k\widehat{\sigma}$ is quite different from normal distribution.

- Let $Z_1, \ldots, Z_n$ be i.i.d. $N(\mu, \sigma^2)$.
- Let $C_j = \sum_{i=1}^{n} I(Z_i \in \widehat{\mu} \pm k\widehat{\sigma}) \sim$ Binomial(n, p_k) where $p_k = \Phi(k) - \Phi(-k)$.
- Approximately, $\sqrt{n}(C_k - np_k) \approx N(0, p_k(1 - p_k))$.
- Big $|\sqrt{n}(C_k - np_k)|$ indicates departure from normal distribution.
- If $|\sqrt{n}(C_k - np_k)| > 3$, then departure from normality with confidence more than 99.7%.

Q-Q plot

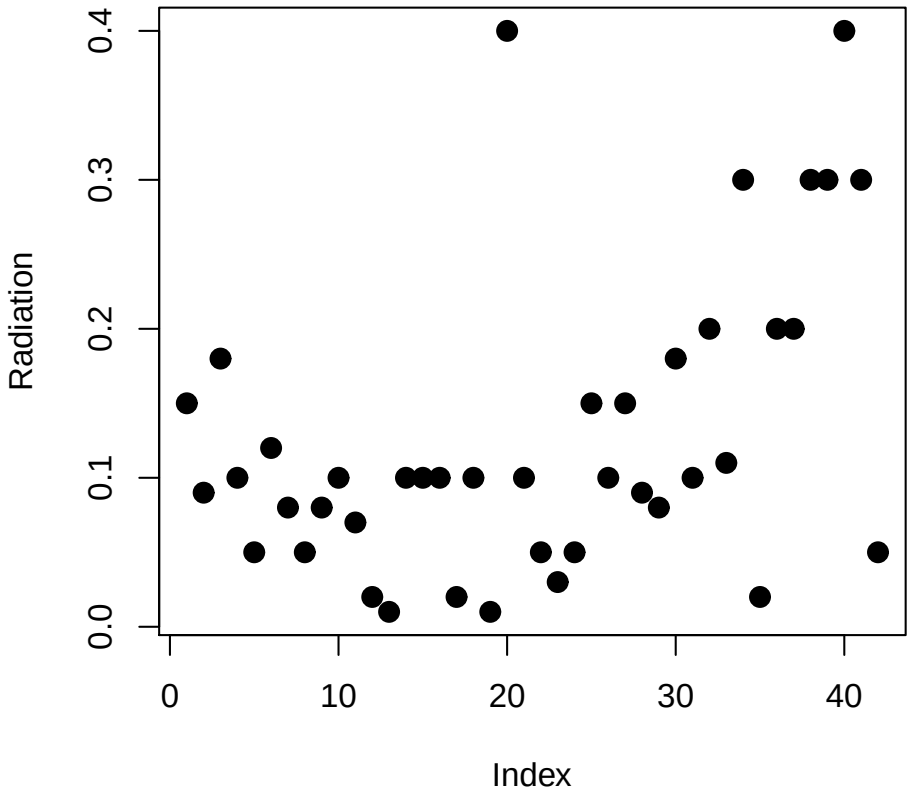
- Let $Z_1, \ldots, Z_n \sim i.i.d.\ N(0,1)$.
- Order statistics $Z_{(1)}, \ldots, Z_{(n)}$
- Each $Z_{(k)}$ is expected to be centred around $q_{(k)}$.
- Such $q_{(k)}$ is approximated by the $(k-1/2)/n$th (or $k/(n+1)$th or $(k-3/4)/(n+1/4)$th) quantile of the standard normal.
- If the data is normally distributed, then the plot of $(q_{(k)}, z_{(k)})$ should be closed to a line.
- The correlation coefficient of $(q_{(k)}, z_{(k)})$ is given by

$$r_Q = \frac{\sum_{i=1}^n (z_{(i)} - \bar{z})(q_{(i)} - \bar{q})}{\left[\sum_{i=1}^n (z_{(i)} - \bar{z})^2 \sum_{i=1}^n (q_{(i)} - \bar{q})^2\right]^{1/2}}$$

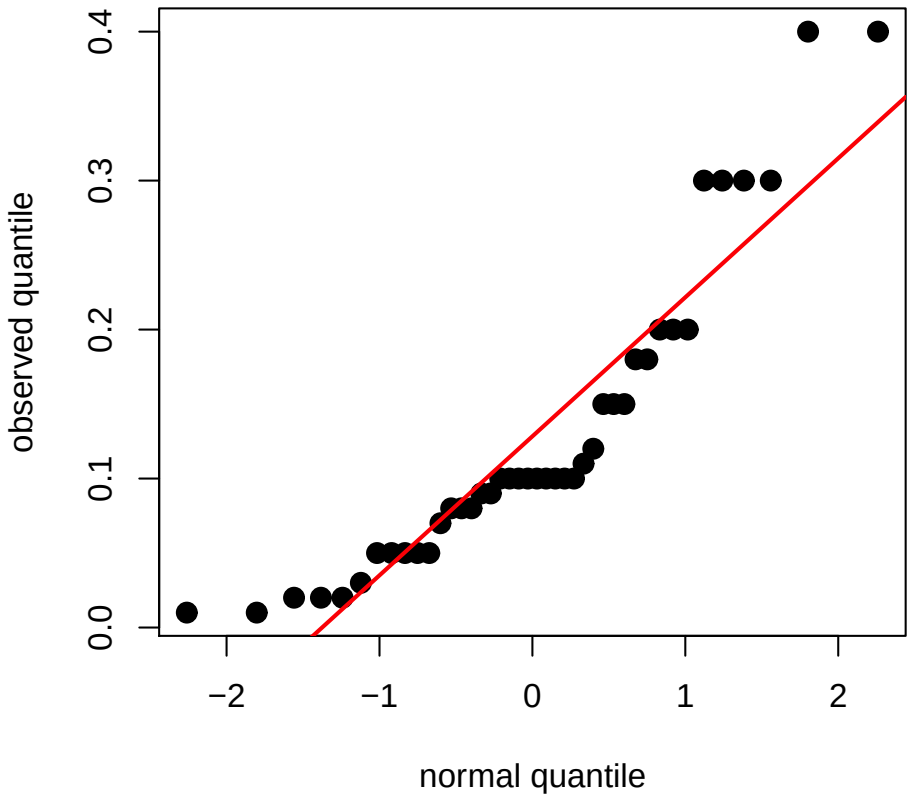
- This coefficient is closed to 1 if the observed data is normally distributed.
- A few tests were developed on ordered numbers.
- Shapiro-Wilks test is very popular which is implemented in R as `shapiro.test`.
- Jarque-Bera test is popular in time series data analysis which is based on the limit behaviour of third and fourth moment.

Example

Radiation data in textbook

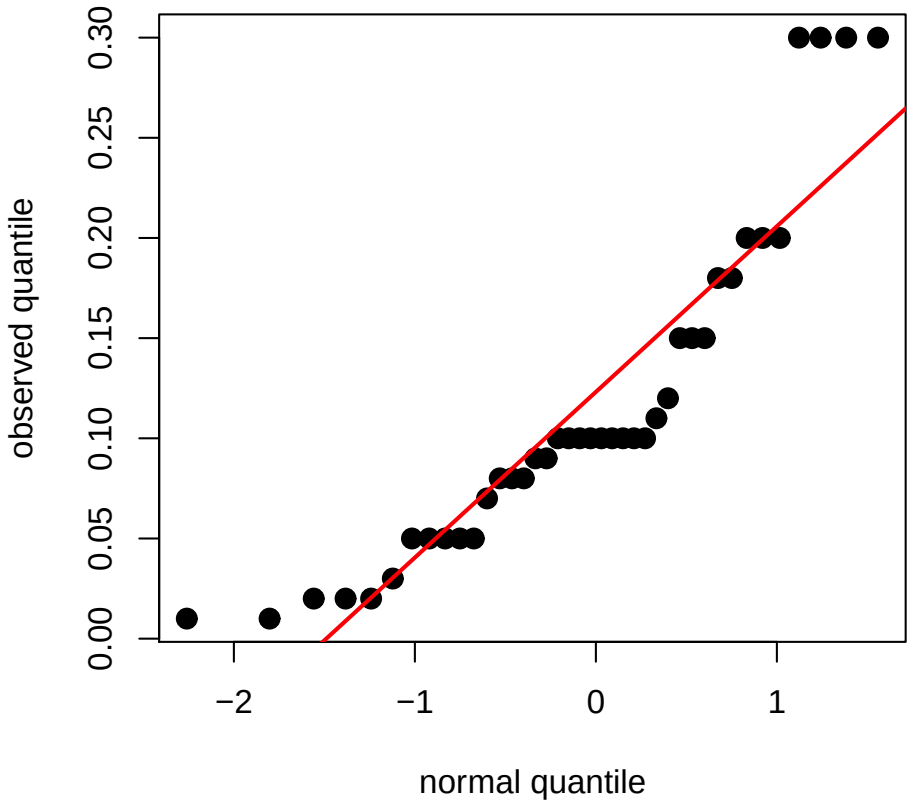


Example



Shapiro-Wilks test p-value is $9.902e-5$.

Example



Shapiro-Wilks test p-value is $4.956e-4$.

Jarque-Bera Test

- One of the most popular normality test is Shapiro-Wilks test which assesses linearity of sample quantile and normal quantile
- Most statistics packages implement Shapiro-Wilks test. In `R`, `shapiro.test` tests normality up to of size 5,000
- One of the most common normality test in time series is Jarque-Bera test which is asymptotic test based on third and fourth moments, that is,

$$JB = \frac{n}{6} Skewness^2 + \frac{n}{24}(Kurtosis - 3)^2 \quad \xrightarrow{d} \chi^2(2)$$

  where $Skewness = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3/\hat{\sigma}^3$ and $Kurtosis = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4/\hat{\sigma}^4$, and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
- The asymptotic distribution might be a bit different from the sample distribution if the sample size is small.
- In `R`, Jarque-Bera test is implemented in `jarque.bera.test` in `tseries` package.

Normality of Multivariate

- Normality of a multivariate random vector can be assessed through χ^2 distribution.
- $(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) \leq \chi^2_\gamma(k)$ can be a basis for normality test.
- For example, $\#\{\mathbf{x}_i : (\mathbf{x}_i-\bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) \leq \chi^2_\gamma(k)\}$ can be approximated by Binomial(n,γ).

Assessment of Multivariate Normality

- Compute $d_j^2 = (\mathbf{x}_i-\bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})$
- Sort and get order statistics $d^2_{(1)} \leq d^2_{(2)} \leq \cdots \leq d^2_{(n)}$
- Plot $(\chi^2_{(j-1/2)/n}(p), d^2_{(j)})$
- Assess whether the plot is linear or not.

Outlier Detection

- Many data sets contain a few unusual data point which may not belong to the pattern of the most observation.
- Such *outliers* may hinder recognition of the pattern of the most observed data.
- With appropriate justifications, outliers can be removed and the pattern of most observed data can be efficiently recognized.

Outlier Detection

- In most cases, outliers can be visually detectable

Steps for Detecting Outliers

(1) Make a univariate data plot such as histogram or density.

(2) Make a scatter plot for each pair of variables

(3) Compute standardized value $z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$. *Large* values of $|z_{jk}|$ could be resulted in outliers. Interpretations of "large" are depend on the dimension of data. A recommended cutoff is 3.5 for moderate sample size.

(4) Compute χ^2 statistics $(\mathbf{x}_j - \bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$. If the value is large, then the data point could be an outlier. A cutoff should be extremely high quantile of $\chi^2(p)$.

Variance Stabilization Transformation

Estimators of certain types show different characteristics in variance. For example, the variance of count data like Poisson random variables is proportional to the mean. Which can be stabilized using a square-root transformation.

Variance Stabilization

count data, y	$\sqrt{y}$
proportion, $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log(\frac{\hat{p}}{1-\hat{p}})$ **parameter dependent**
correlation, $\hat{r}$	Fisher's $z(\hat{r}) = \frac{1}{2}\log(\frac{1+\hat{r}}{1-\hat{r}}) \approx N(\frac{1}{2}\log(\frac{1+r}{1-r}), \frac{1}{n-3})$

Box-Cox Transformation

A generalized power transformation given by

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(x) & \lambda = 0 \end{cases}$$

for positive $x > 0$. The tuning parameter λ can be chosen to maximize

$$\ell(\lambda) = -\frac{n}{2} \log \left(\frac{1}{n} \sum_{j=1}^{n} (x_j^{(\lambda)} - \overline{x^{(\lambda)}})^2 \right) + (\lambda - 1) \sum_{j=1}^{n} \log x_j$$

where $\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)}$.
The function `boxcox` can be used in `MASS` package of `R`.

Inference about Mean Vector

Consider univariate case, that is, $x_1, \ldots, x_n \sim i.i.d.\ N(\mu, \sigma^2)$. Then $\sqrt{n}(\bar{x} - \mu)/\sqrt{s^2} \sim t(n-1)$. Hence a hypothesis assessment $H_0 : \mu = \mu_0$ can base on the statistic $\sqrt{n}(\bar{x} - \mu)/\sqrt{s^2} \sim t(n-1)$.
A natural generalization of univariate t-statistic for multivariate is

$$T^2 = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu)$$

which is called *Hotelling's T^2-statistic* since it is square of t-statistic for univariate case.
It is known that $T^2 \sim \frac{(n-1)p}{n-p} F(p, n-p)$. A long proof is given by Harold Hotelling.

Example

Note that $t(k) \sim N(0,1)/[\chi^2(k)/k]^{1/2}$. Then $t(k)^2 \sim \chi^2(1)/(\chi^2(k)/k) \sim kF(1,k)$.

Note that $n(\bar{\mathbf{x}} - \mu)^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mu) \sim \chi^2(p)$. Roughly speaking, $\Sigma^{-1/2} S \Sigma^{-1/2} \sim (n-1)^{-1} W_p(I_p, n-1)$ contributes $[(n-p)/(p(n-1))]\chi^2(n-p)$ along with $\Sigma^{-1/2}(\bar{\mathbf{x}} - \mu)$.

Likelihood Ratio Region

Definition (Likelihood ratio test)

The likelihood ratio test for H_0 vs H_1 is

$$\Lambda = \max_{\theta \in H_0} L(\theta) \Big/ \max_{\theta \in H_0 \cup H_1} L(\theta)$$

which has a limit distribution

$$-2 \log \Lambda \xrightarrow{d} \chi^2(p)$$

where p is the number of free parameters in H_1 but not in H_0.

Proposition

The likelihood ratio test for $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ is a function of Hotelling's T^2-statistic.

Proof of Proposition I

Under H_0, the parameter value of μ is μ_0 and the maximum likelihood estimator for Σ is $\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mu_0)(\mathbf{x}_i - \mu_0)^\top$. While under H_1, the maximum likelihood estimators for μ and Σ are $\bar{\mathbf{x}}$ and $\hat{\Sigma}_1 = S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$. Hence the corresponding likelihood functions are

$$
\begin{aligned}
L(\hat{\Sigma}_0) &= |2\pi\hat{\Sigma}_0|^{-n/2}\exp(-\frac{1}{2}\mathrm{tr}(\hat{\Sigma}_0^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mu_0)(\mathbf{x}_i - \mu_0)^\top)) \\
&= |2\pi\hat{\Sigma}_0|^{-n/2}\exp(-0.5\mathrm{tr}(nI_p)) = |2\pi\hat{\Sigma}_0|^{-n/2}\exp(-np/2).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
L(\bar{\mathbf{x}}, \hat{\Sigma}_1) &= |2\pi\hat{\Sigma}_1|^{-n/2}\exp(-\frac{1}{2}\mathrm{tr}(\hat{\Sigma}_1^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top)) \\
&= |2\pi\hat{\Sigma}_1|^{-n/2}\exp(-0.5\mathrm{tr}(nI_p)) = |2\pi\hat{\Sigma}_1|^{-n/2}\exp(-np/2).
\end{aligned}
$$

Proof of Proposition II

Thus the likelihood ratio statistic becomes

$$\Lambda = |2\pi\hat{\Sigma}_0|^{-n/2}\exp(-np/2)\Big/|2\pi\hat{\Sigma}_1|^{-n/2}\exp(-np/2) = (|\hat{\Sigma}_0|/|\hat{\Sigma}_1|)^{-n/2}.$$

It is not hard to see that

$$\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\mu_0)(\mathbf{x}_i-\mu_0)^\top = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^\top + (\bar{\mathbf{x}}-\mu_0)(\bar{\mathbf{x}}-\mu_0)^\top.$$

This implies that

$$\begin{aligned}
|\hat{\Sigma}_0| &= |\hat{\Sigma}_1|\cdot|I_p + \hat{\Sigma}_1^{-1/2}(\bar{\mathbf{x}}-\mu_0)(\bar{\mathbf{x}}-\mu_0)^\top\hat{\Sigma}_1^{-1/2}| \\
&= |\hat{\Sigma}_1|\cdot(1+(\bar{\mathbf{x}}-\mu_0)^\top\hat{\Sigma}_1^{-1}(\bar{\mathbf{x}}-\mu_0)) = |\hat{\Sigma}_1|\cdot(1+(\bar{\mathbf{x}}-\mu_0)^\top\hat{\Sigma}_1^{-1}(\bar{\mathbf{x}}-\mu_0)) \\
&= |\hat{\Sigma}_1|\cdot(1+T^2/(n-1))
\end{aligned}$$

Therefore the likelihood ratio statistic becomes

$$\Lambda = (|\hat{\Sigma}_0|/|\hat{\Sigma}_1|)^{-n/2} = (1+T^2/(n-1))^{-n/2}.$$

Which is a function of Hotelling's T^2-statistic.

Lecture 5

Confidence Region

Definition

A random region $R(\mathbf{X})$ is called a γ-*confidence region* of a parameter θ if

$$P_\theta(\theta \in R(\mathbf{X})) \geq \gamma$$

for any $\theta \in \Theta$.

Example

If $\mathbf{x}_j \sim N_p(\mu, \Sigma)$, then Hotelling's T^2 statistic satisfies $T^2 = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \sim \frac{(n-1)p}{n-p} F(p, n-p)$. Thus $P(T^2 = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \leq \frac{(n-1)p}{n-p} F_\gamma(p, n-p)) = \gamma$ regardless of μ and Σ. Then

$$R(\bar{\mathbf{x}}, S) = \{\mu : n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \leq \frac{(n-1)p}{n-p} F_\gamma(p, n-p)\}$$

is a γ-confidence region for μ.

Confidence Region

Example

- Radiation example in text book.
- Let y_{i1} and y_{i2} be measure radiation with door closed and open, respectively.
- Both Y_1 and Y_2 are not normally distributed.
- Box-Cox transformation is applied with $\lambda = 1/4$ for both of them. Let $x_{ij} = (y_{ij})^{1/4}$.
- sample mean and variance are

$$\bar{\mathbf{x}} = \begin{pmatrix} 0.5643 \\ 0.6030 \end{pmatrix} \quad S = \begin{pmatrix} 0.0144 & 0.0117 \\ 0.0117 & 0.0146 \end{pmatrix}$$

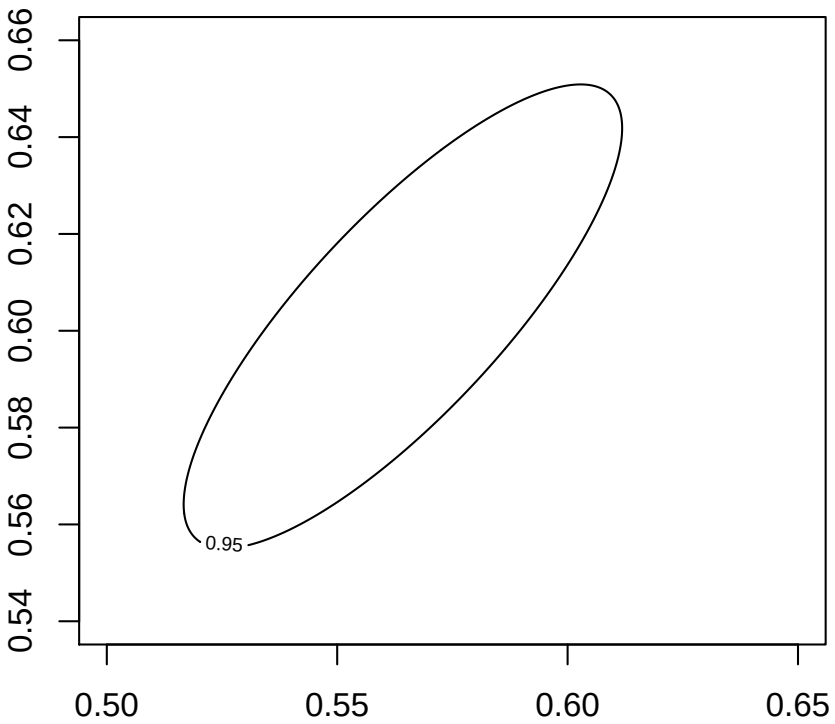
Then 95% confidence region of μ is

$$R_{0.95} = R_{0.95}(\bar{\mathbf{x}}, S)$$

$$= \{\mu = (\mu_1, \mu_2)^\top : n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \leq F_{0.95}(p, n-p)\frac{(n-1)p}{n-p}\}.$$

Confidence Region

The corresponding 95%-confidence region is



Confidence Regions of Marginal Parameters I

- a linear combination of mean vector is of interest rather than the full parameter, that is, parameter of interest is

$$\psi = a_1\mu_1 + \cdots + a_p\mu_p$$

- Let $\mathbf{x}_j \sim N_p(\mu, \Sigma)$ and $\mathbf{a} = (a_1, \ldots, a_p)^\top$.
- Define $z_j = \mathbf{a}^\top \mathbf{x}_j$ so that $z_j \sim N(\mathbf{a}^\top\mu, \mathbf{a}^\top\Sigma\mathbf{a}) \sim N(\psi, \zeta)$ where $\zeta = \mathbf{a}^\top\Sigma\mathbf{a}$.
- sample mean and unbiased variance are

$$\bar{z} = \frac{1}{n}\sum_{j=1}^{n} z_j = \frac{1}{n}\sum_{j=1}^{n} \mathbf{a}^\top\mathbf{x}_j = \mathbf{a}^\top\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j = \mathbf{a}^\top\bar{\mathbf{x}}$$

$$s_z^2 = \frac{1}{n-1}\sum_{j=1}^{n}(z_j - \bar{z})^2 = \frac{1}{n-1}\mathbf{a}^\top(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^\top\mathbf{a} = \mathbf{a}^\top S\mathbf{a}.$$

Confidence Regions of Marginal Parameters II

γ-confidence region

Note the t statistic

$$\frac{\bar{z}-\psi}{s_z/\sqrt{n}} = \frac{\sqrt{n}\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)}{\sqrt{\mathbf{a}^\top S\mathbf{a}}} \sim t(n-1).$$

Then a γ-confidence region (or interval) for ψ is

$$\mathbf{a}^\top\bar{\mathbf{x}} \pm t_{(1+\gamma)/2}(n-1)\sqrt{\mathbf{a}^\top S\mathbf{a}}/\sqrt{n}.$$

Confidence Region for vector parameter

- Parameter of interest: $\psi = A\mu \in \mathbb{R}^k$
- $\mathbf{z}_j = A\mathbf{x}_j \sim N_k(A\mu, A\Sigma A^\top)$
- γ-confidence region becomes

$$\{\psi : n(A\bar{\mathbf{x}}-\psi)^\top(ASA^\top)^{-1}(A\bar{\mathbf{x}}-\psi) \leq \frac{(n-1)k}{n-k}F_\gamma(k,n-k)\}.$$

Confidence Regions of Marginal Parameters III

- Confidence intervals vary as linear combinations changes
- because of the correlations.
- "Is it possible to have a simple form of simultaneous γ-confidence intervals?"
- Note CIs are $n\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)^\top(\bar{\mathbf{x}}-\mu)\mathbf{a}/\mathbf{a}^\top S\mathbf{a} \le c^2$.
- For $\mathbf{a}_1, \ldots, \mathbf{a}_k$,

$$
\begin{aligned}
&P(n\mathbf{a}_j^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}_j/(\mathbf{a}_j^\top S\mathbf{a}_j) \le c, j=1,\ldots,k)\\
&\ge P(\max_{\mathbf{a}} n\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}/(\mathbf{a}^\top S\mathbf{a}) \le c)\\
&= P(n(\bar{\mathbf{x}}-\mu)^\top S^{-1}(\bar{\mathbf{x}}-\mu) \le c)
\end{aligned}
$$

where the last equality can be obtained when $\mathbf{a}$ is proportional to $S^{-1}(\mathbf{x}-\mu)$, that is,

Confidence Regions of Marginal Parameters IV

- Let $\mathbf{b} = S^{1/2}\mathbf{a}$ or $\mathbf{a} = S^{-1/2}\mathbf{b}$

$$\max_{\mathbf{a}} \frac{n\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}}{\mathbf{a}^\top S\mathbf{a}} = \max_{\mathbf{b}} \frac{n\mathbf{b}^\top S^{-1/2}(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top S^{-1/2}\mathbf{b}}{\mathbf{b}^\top\mathbf{b}}$$

$$= n \max_{\mathbf{b}} \frac{||(\bar{\mathbf{x}}-\mu)^\top S^{-1/2}\mathbf{b}||^2}{||\mathbf{b}||^2}$$

Maximum is when $\mathbf{b} \propto S^{-1/2}(\bar{\mathbf{x}}-\mu)$ and $\mathbf{a} \propto S^{-1/2}S^{-1/2}(\bar{\mathbf{x}}-\mu) \propto S^{-1}(\bar{\mathbf{x}}-\mu)$.

- The simultaneous confidence interval is a γ-confidence region,

$$P\left(\frac{n\mathbf{a}_j^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}_j}{\mathbf{a}_j^\top S\mathbf{a}_j} \leq \frac{(n-1)p}{n-p}F_\gamma(p,n-p), j=1,\ldots,k\right)$$

$$\geq P\left(n(\bar{\mathbf{x}}-\mu)^\top S^{-1}(\bar{\mathbf{x}}-\mu) \leq \frac{(n-1)p}{n-p}F_\gamma(p,n-p)\right) = \gamma.$$

Confidence Regions of Marginal Parameters V

Simultaneous confidence interval

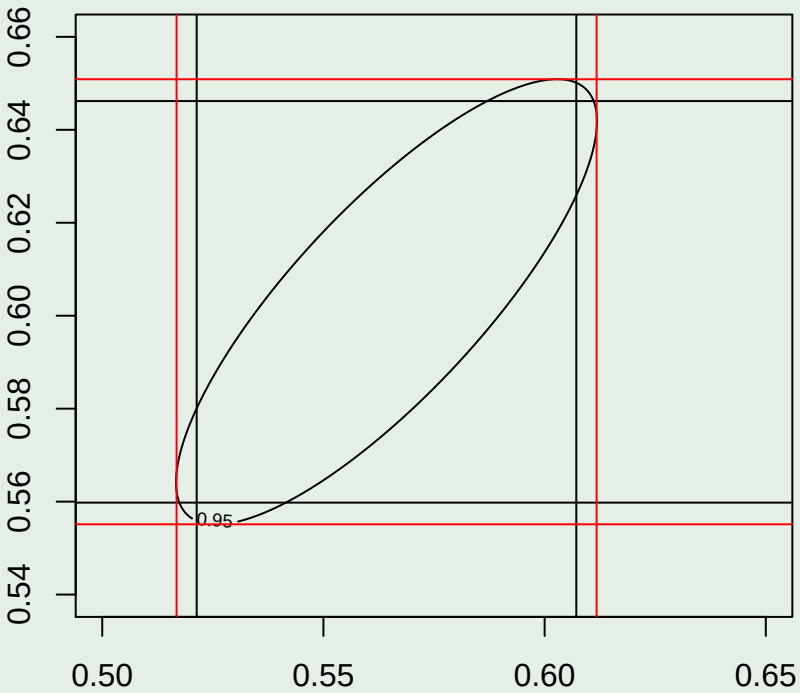
The simultaneous confidence intervals for any $\mathbf{a}$ is

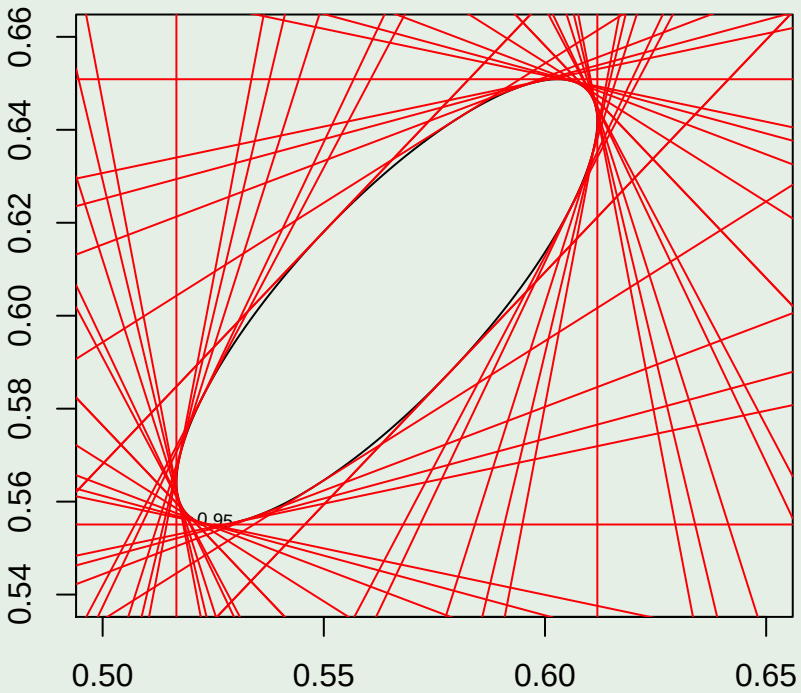
$$\mu_j \in \bar{x}_j \pm \sqrt{\frac{(n-1)p}{n-p} F_\gamma(p, n-p)} \sqrt{\frac{S_{jj}}{n}} \quad \text{for} \quad j = 1, \ldots, p$$

has confidence at least γ.

Confidence Regions of Marginal Parameters VI

Example (Radition Example)





Bonferroni Correction

If all coordinates are independent, simultaneous marginal confidence regions have confidence

$$P(\mu_j \in \bar{\mathbf{x}}_j \pm t_{(1+\gamma)/2}(n-1)\sqrt{S_{jj}/n}, j = 1, \ldots, p) = \gamma^p \leq \gamma.$$

It becomes very conservative. To make the confidence close to nominate confidence take γ^* a bit bigger, that is,

$$P(\mu_j \in \bar{\mathbf{x}}_j \pm t_{(1+\gamma^*)/2}(n-1)\sqrt{S_{jj}/n}, j = 1, \ldots, p) = 1 - P(\mu_j \notin \bar{\mathbf{x}}_j \pm t_{(1+\gamma^*)/2}(n$$

$$\geq 1 - \sum_{j=1}^{p} P(\mu_j \notin \bar{\mathbf{x}}_j \pm t_{(1+\gamma^*)/2}(n-1)\sqrt{S_{jj}/n}) = 1 - p(1-\gamma^*) \approx \gamma$$

Which gives $\gamma^* = 1 - (1-\gamma)/p \geq \gamma$.

Large Sample Confidence Intervals

When the sample size is large, Hotelling's T^2 statistic follows approximately a $\chi^2(p)$ distribution using the central limit theorem and the continuous mapping theorem. Hence the region

$$\{\mu : n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \leq \chi^2_\gamma(p)\}$$

has confidence approximately γ.
Similarly, for any vector $\mathbf{a}$, the confidence of the interval

$$\mathbf{a}^\top \bar{\mathbf{x}} \pm \sqrt{\chi^2_\gamma(p)}\sqrt{\mathbf{a}^\top S\mathbf{a}/n}$$

is approximately γ.

Inference with Missing Observations I

- Often there are missing values in practice.
- If the proportion of missing data is not big, then mean and variance matrix can be estimated very efficiently using expectation-maximization (EM) algorithm.

- complete data set $Y_c = (Y_o, Y_m)$ with parameter θ
- Y_o, Y_m are sets of observed/missed data.
- MLE $\hat{\theta}$ can be obtained using the following steps.

Initial step Set an initial parameter $\theta^{(0)}$

E-step Compute the conditional log likelihood

$$Q(\theta \mid \theta^{(l)}) = \mathbb{E}[\log \text{pdf}_{Y_c}(y_o, y_m \mid \theta) \mid y_o, \theta^{(l)}]$$

given observed data and current parameter value $\theta^{(l)}$.

M-step Find new estimator $\theta^{(l+1)}$ maximizing $Q(\theta \mid \theta^{(l)})$.

Repeat Repeat E-step and M-step until the parameter converges.

Inference with Missing Observations II

Example (Multivariate Normal)

- $\mathbf{x}_j \sim N_p(\mu, \Sigma)$ with some missing.
- $Q(\mu, \Sigma \mid \hat{\mu}, \hat{\Sigma})$ function is the log likelihood function of complete data with missing values replaced by the conditional expectation given $\hat{\mu}, \hat{\Sigma}$.
- For example, if x_{i4}, x_{i5} are missing while x_{i1}, x_{i2}, x_{i3} are observed, the Q function is the likelihood function of $(x_{i1}, x_{i2}, x_{i3}, \mathbb{E}((x_{i4}, x_{i5}) \mid \hat{\mu}, \hat{\Sigma}))$.
- Plug in Y_m by $\mathbb{E}(Y_m \mid Y_o, \mu^{(l)}, \Sigma^{(l)})$
- Compute new sample mean and variance of (Y_o, Y_m).