THE UNIVERSITY OF CHICAGO
Booth School of Business
Business 41912, Spring Quarter 2020, Mr. Ruey S. Tsay

Lecture 1 of Multivariate Data Analysis

1 Preliminary

Data frame: n data points on p variables. We follow the notation commonly used in statistics.

		Variable 1	Variable 2	$\cdots$	Variable k	$\cdots$	Variable p
Item 1	Obs 1	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Item 2	Obs 2	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item 2	Obs j	x_{i1}	x_{i2}	$\cdots$	x_{ik}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n	Obs n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Notation: x_{ik} = measurement of the kth variable on the ith item. Typically, column denotes variables and row denotes subjects (or sample).
In matrix notation: The jth observation is $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jk})'$, and

$$\boldsymbol{X} = \left[\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{array}\right]_{n\times p} \equiv [x_{ik}] \quad \text{or} \quad \boldsymbol{X} = \left[\begin{array}{c} \boldsymbol{x}'_1 \\ \boldsymbol{x}'_2 \\ \vdots \\ \boldsymbol{x}'_n \end{array}\right].$$

Vectors and matrices are used extensively in this class. If needed, students should review the basic concepts of vector and matrix. $\boldsymbol{A}'$ denotes the transpose of the matrix (or vector) $\boldsymbol{A}$. For instance,

$$\boldsymbol{b} = (1.2, 0.3), \quad \boldsymbol{A} = \left[\begin{array}{cc} 0.2 & 0.3 \\ -0.6 & 1.1 \end{array}\right] \Rightarrow \boldsymbol{b}' = \left[\begin{array}{c} 1.2 \\ 0.3 \end{array}\right], \quad \boldsymbol{A}' = \left[\begin{array}{cc} 0.2 & -0.6 \\ 0.3 & 1.1 \end{array}\right].$$

Descriptive statistics (over the observations)

1. Sample means: $\bar{x}_k = \frac{1}{n}\sum_{j=1}^n x_{jk}$, $k = 1, \ldots, p$.
2. Sample variance: $s_k^2 = \frac{1}{n-1}\sum_{j=1}^n (x_{jk} - \bar{x}_k)^2$, $k = 1, \ldots, p$.

   Alternatively, $s_k^2 = \frac{1}{n}\sum_{j=1}^n (x_{jk} - \bar{x}_k)^2$.

3. Sample covariance: $s_{ik} = \frac{1}{n-1}\sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k), \quad i, k = 1, \ldots, p.$

4. Sample correlation coefficients

$$r_{ik} = \frac{s_{ik}}{s_i \times s_k} = \frac{\sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2 \times \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2}}.$$

   Remarks on correlation:

   - $-1 \leq r_{ik} \leq 1$.
   - r_{ik} measures the strength of linear association.
   - r_{ik} is scale invariant.
   - r_{ik} is the referred to as the **Pearson's correlation coefficient**. For measurement of general dependence (including nonlinear dependence), one can use Kandall's tau or Spearman's rho.

Matrix representation:

(a) Sample mean: $\bar{\boldsymbol{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)' = \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j$.

(b) Sample covariance matrix: $\boldsymbol{S}_n = [s_{ik}] = \frac{1}{n-1}\sum_{j=1}^n (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$.

(c) Sample correlation matrix: $\boldsymbol{R}_n = [r_{ik}]$ with $r_{ii} = 1$.

Example: The paper-quality measurements of Table 1.2, page 15.

R demonstration:

```
> setwd("C:/teaching/ama")
> y=read.table("T1-2.dat")
> colnames(y) <- c("density","mach-dir","cross-dir")
> colMeans(y)  % Same as apply(y,2,mean)
    density    mach-dir   cross-dir
  0.8118537 120.9534146  67.7231707
> var(y)         %  The command ‘‘cov’’ works too.
              density   mach-dir  cross-dir
density   0.001264578  0.1684468  0.2252480
mach-dir  0.168446762 59.3211480 60.9925314
cross-dir 0.225247976 60.9925314 95.8566672
> apply(y,2,var)
     density     mach-dir    cross-dir
 0.001264578 59.321148049 95.856667195
> cor(y)
            density  mach-dir cross-dir
```

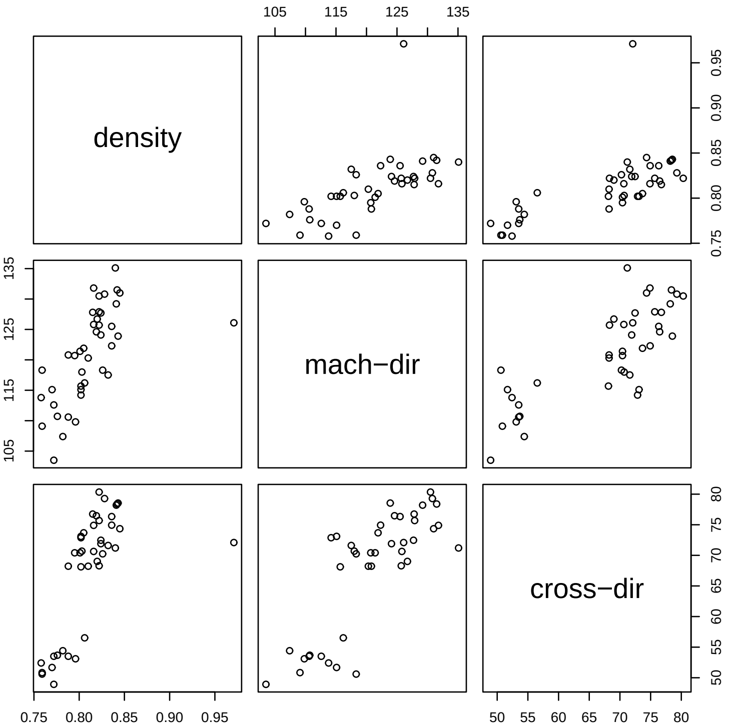

Figure 1: Pairwise scatter-plot or matrix plot

```
density   1.0000000 0.6150141 0.6469592
mach-dir  0.6150141 1.0000000 0.8088365
cross-dir 0.6469592 0.8088365 1.0000000

> plot(y) % Pairwise-scatter plot. See Figure 1 for demonstration.
% 3-dimensional plot using the package "rgl".
> require(rgl)  % Install the package ``rgl'' before using it.
> plot3d(y[,1],y[,2],y[,3],xlab="density",ylab='mach-dir',zlab='cross-dir')
> q() % exit R
```

Kendall's tau: Let F be a continuous bivariate cumulative distribution function (CDF) of random varibe $\boldsymbol{x} = (x_1, x_2)'$. Let (X_1, X_2) and (Y_1, Y_2) be independent random pairs with distribution F. Then, Kendall's tau is

$$\begin{aligned}\tau &= Pr[(X_1 - Y_1)(X_2 - Y_2) > 0] - Pr[(X_1 - Y_1)(X_2 - Y_2) < 0] \\ &= 2Pr[(X_1 - Y_1)(X_2 - Y_2) > 0] - 1.\end{aligned}$$

Kendall-τ is also known as Kendall rank correlation coefficient. It is a measure of the similarity of the orderings of the data when ranked by each of the quantities.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)},$$

where n_c is the number of *concordant pairs* and n_d the number of *discordant pairs* in the data set. The denominator is the total number of pairs of the data. A high τ indicates that most paris are concordant, indicating that the two rankings are consistent.

For ease in notation, rewrite the pair of random vectors as (X_1, Y_1) and (X_2, Y_2). By a concordant pair, we mean $\text{sign}(X_2 - X_1) = \text{sign}(Y_2 - Y_1)$, where $\text{sign}(d)$ = -1, 0, 1 for $d < 0, = 0, > 0$, respectively. A pair is discordant if $\text{sign}(X_2 - X_1) = -\text{sign}(Y_2 - Y_1)$. Based on the sign function, Kendall-τ can be rewritten as

$$\tau = \frac{\sum_{i<j} \text{sign}(X_j - X_i) * \text{sign}(Y_j - Y_i)}{n(n-1)/2}.$$

Spearman's rho: Let F be a continuous bivariate cumulative distribution function (CDF) of random varibe $\boldsymbol{x} = (x_1, x_2)'$. Let F_1 and F_2 be the two marginal CDF. Assume $(X_1, X_2) \sim F$. Then, Spearman's rho is the correlation of $F_1(X_1)$ and $F_2(X_2)$. Since $F_1(X_1)$ and $F_2(X_2)$ are uniform $U(0,1)$, their means are 1/2 and their variances are 1/12. Thus, Spearman's rho is

$$\rho_S = 12 \int\int F_1(x_1) F_2(x_2) dF(x_1, x_2) - 3.$$

Spearman's rho is also a rank correlation coefficient. For data $\{(X_i, Y_i)\}_{i=1}^n$, let $\{(x_i, y_i)\}_{i=1}^n$ be the corresponding rank pairs and $d_i = x_i - y_i$ be the difference between the ranks. Then, Spearman's rho is

$$\rho_s = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)},$$

assuming no ties in the data.

In R, the command `cor` can be used to compute Kendall's tau and Spearman's rho. See the subcommand method with `help(cor)` in R.

R Demonstration

```
> x=rnorm(100)
> cor(x,x)
[1] 1
> cor(x,x*10)
[1] 1
> cor(exp(x),exp(x))
```

```
[1] 1
> cor(exp(x),exp(x*10))
[1] 0.7977132
> cor(exp(x),exp(x),method="kendall")
[1] 1
> cor(exp(x),exp(x*10),method="kendall")
[1] 1
> cor(exp(x),exp(x*10),method="spearman")
[1] 1
```

Discussion: Pearson's correlation coefficient measures the linear relation. Nothing more! Kendall's tau and Spearman's rho, on the other hand, show more general dependence between the variables. Rank correlations are also robust to outliers.

2 Some useful results of matrix theory

Eigenvalue and Eigenvector:
A p-dimensional vector $\boldsymbol{x}$ is an eigenvector with eigenvalue λ of the $m \times n$ matrix $\boldsymbol{A}$ if $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$.
Since $\boldsymbol{A}(c\boldsymbol{x}) = \lambda(c\boldsymbol{x})$ for any non-zero constant c, eigenvector is not unique. Normalization is often used such as $\|\boldsymbol{x}\| = 1$, where $\|.\|$ denotes the Euclidean norm, i.e. $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{p} x_i^2}$ if $\boldsymbol{x} = (x_1, \dots, x_p)'$.
Positive definite matrix:
A square matrix $\boldsymbol{A}$ is positive definite if (a) $\boldsymbol{A} = \boldsymbol{A}'$, i.e. symmetric, and (b) $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ for any non-zero vector vector $\boldsymbol{x}$.
If $\boldsymbol{A}$ is positive definite matrix, then all eigenvalues of $\boldsymbol{A}$ are positive.
Spectral decomposition: If $\boldsymbol{A}$ is a $p \times p$ symmetric matrix, then

$$\boldsymbol{A} = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1' + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2' + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p',$$

where $\boldsymbol{e}_i'\boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i'\boldsymbol{e}_j = 0$ for $i \neq j$. The λ_i are eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{e}_i$ are the corresponding eigenvectors.
If $\lambda_i > 0$ for all i, then $\boldsymbol{A}$ is positive definite.
If $\lambda_i = 0$ for some i, then $\boldsymbol{A}$ is non-negative definite.
For a positive definite matrix $\boldsymbol{A}_{p\times p}$, we have

1. $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i' = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, where $\boldsymbol{P} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p]$ is the matrix of eigenvectors and $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_p\}$. Often we also assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

2. $\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}' = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i'$.

3. The square-root matrix of $\boldsymbol{A}$ is $\boldsymbol{A}^{1/2} = \sum_{i=1}^{p} \sqrt{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i' = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}'$. The matrix $\boldsymbol{A}^{1/2}$ is also positive definite. Similarly, $\boldsymbol{A}^{-1/2} = \sum_{i=1}^{p} \frac{1}{\sqrt{\lambda_i}} \boldsymbol{e}_i \boldsymbol{e}_i'$.

Cauchy-Schwarz inequality: For any two p-dimensional vectors $\boldsymbol{b}$ and $\boldsymbol{d}$, we have

$$(\boldsymbol{b}'\boldsymbol{d})^2 \leq (\boldsymbol{b}'\boldsymbol{b})(\boldsymbol{d}'\boldsymbol{d})$$

with equality if and only if $\boldsymbol{b} = c\boldsymbol{d}$ for some constant c.

Extension: $(\boldsymbol{b}'\boldsymbol{d})^2 \leq (\boldsymbol{b}'\boldsymbol{B}\boldsymbol{b})(\boldsymbol{d}'\boldsymbol{B}^{-1}\boldsymbol{d})$ for a positive definite matrix $\boldsymbol{B}_{p\times p}$.

Maximization Lemma: Let $\boldsymbol{B}_{p\times p}$ be positive definite and $\boldsymbol{d}$ a given p-dimensional vector. Then,

$$\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{(\boldsymbol{x}'\boldsymbol{d})^2}{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}} = \boldsymbol{d}'\boldsymbol{B}^{-1}\boldsymbol{d}$$

with the maximum attained when $\boldsymbol{x} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ for some non-zero constant c.
Proof: By the extended Cauchy-Schwarz inequality.

More properties of positive definite matrix $\boldsymbol{B}_{p\times p}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ be the eigenvalues and $\boldsymbol{e}_i$ are the corresponding eigenvectors so that $\boldsymbol{B} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i'$. Then,

1. $\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_1$ (attained when $\boldsymbol{x} = \boldsymbol{e}_1$).

2. $\min_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_p$ (attained when $\boldsymbol{x} = \boldsymbol{e}_p$).

3. $\max_{\boldsymbol{x}\perp\boldsymbol{e}_1,\ldots,\boldsymbol{e}_\ell;\neq\boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_{\ell+1}$ (attained when $\boldsymbol{x} = \boldsymbol{e}_{\ell+1}$).

Example: Compute eigenvalues and eigenvectors in R

```
> x=matrix(c(2,1,1,4),2,2)  % a simple 2-by-2 matrix.

> m1=eigen(x)
> names(m1)
[1] "values"  "vectors"
> m1$values
[1] 4.414214 1.585786
> m1$vectors
          [,1]       [,2]
[1,] 0.3826834  0.9238795
[2,] 0.9238795 -0.3826834
%%% Verification
> ev=m1$values
> vec=m1$vectors
> y=x%*%vec
> y
         [,1]       [,2]
```

```
[1,] 1.689246  1.4650756
[2,] 4.078202 -0.6068542
> y1=vec%*%diag(ev)
> y1
         [,1]       [,2]
[1,] 1.689246  1.4650756
[2,] 4.078202 -0.6068542
```

3 Random vectors and matrices

Definition: $\boldsymbol{X} = [\boldsymbol{X}_{ij}]_{n\times p}$ is a random matrix if X_{ij} is a random variable.
Expectation: $E(\boldsymbol{X}) = [E(X_{ij})]$.
There are two basic properties: (1) $E(\boldsymbol{X}+\boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y})$ and (2) $E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant matrices.
If $p = 1$, the $\boldsymbol{X}$ is a n-dimensional random vector. Let $E(\boldsymbol{X}) = \boldsymbol{\mu}_x$ and $\text{cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_x$. Then, for any constant n-dimensional vector $\boldsymbol{c}$, $E(\boldsymbol{c}'\boldsymbol{X}) = \boldsymbol{c}'\boldsymbol{\mu}_x$ and $\text{Var}(\boldsymbol{c}'\boldsymbol{X}) = \boldsymbol{c}'\boldsymbol{\Sigma}_x\boldsymbol{c}$.
For a random vector $\boldsymbol{X}$ with mean $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_x$, $E(\boldsymbol{X}\boldsymbol{X}') = \boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x\boldsymbol{\mu}_x'$.
The generalized variance of a random vector $\boldsymbol{X}$ is $|\boldsymbol{\Sigma}_x|$, the determinant of its covariance matrix.

Linear combination of random variables. A linear combination of random variables in $\boldsymbol{X}$ can be written as $Y = \boldsymbol{c}'\boldsymbol{X}$, where $\boldsymbol{c}$ is a p-dimensional constant vector. The mean and variance of Y are

- $E(Y) = \boldsymbol{c}'\boldsymbol{\mu}_x$.
- $\text{Var}(Y) = \boldsymbol{c}'\boldsymbol{\Sigma}_x\boldsymbol{c}$.

4 Random Samples

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from a joint distribution that has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_x$. Let

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i, \quad \boldsymbol{S}_n = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'.$$

Then, $E(\bar{\boldsymbol{X}}) = \boldsymbol{\mu}$, $\text{cov}(\bar{\boldsymbol{X}}) = \frac{1}{n}\boldsymbol{\Sigma}_x$, and $E(\boldsymbol{S}_n) = \frac{n-1}{n}\boldsymbol{\Sigma}_x$.

Proof: $E(\bar{\boldsymbol{X}}) = \boldsymbol{\mu}$ is straightforward. For $\text{cov}(\bar{\boldsymbol{X}})$, consider

$$(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \;=\; \left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu})\right]\left[\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{X}_j - \boldsymbol{\mu})\right]'$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_j-\boldsymbol{\mu})'.$$

Therefore,

$$\begin{aligned}
\text{cov}(\bar{\boldsymbol{X}}) &= E[(\bar{\boldsymbol{X}}-\boldsymbol{\mu})(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E[(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_j-\boldsymbol{\mu})'] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}E[(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})'] \quad \text{(because of independence)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\boldsymbol{\Sigma}_x = \frac{1}{n}\boldsymbol{\Sigma}_x.
\end{aligned}$$

Finally,

$$\begin{aligned}
E(\boldsymbol{S}_n) &= \frac{1}{n}\sum_{i=1}^{n}E(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})' \\
&= \frac{1}{n}\sum_{i=1}^{n}E[\boldsymbol{X}_i\boldsymbol{X}_i'-\boldsymbol{X}_i\bar{\boldsymbol{X}}'-\bar{\boldsymbol{X}}\boldsymbol{X}_i+\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}'] \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n}E(\boldsymbol{X}_i\boldsymbol{X}_i')+E\left[-\sum_{i=1}^{n}\boldsymbol{X}_i\bar{\boldsymbol{X}}'-\sum_{i=1}^{n}\bar{\boldsymbol{X}}\boldsymbol{X}_i+\sum_{i=1}^{n}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}'\right]\right\} \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}E(\boldsymbol{X}_i\boldsymbol{X}_i')-nE(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}')\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}E(\boldsymbol{X}_i\boldsymbol{X}_i')-E(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}') \\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\Sigma}_x+\boldsymbol{\mu}\boldsymbol{\mu}')-[\frac{1}{n}\boldsymbol{\Sigma}_x+\boldsymbol{\mu}\boldsymbol{\mu}'] \\
&= \boldsymbol{\Sigma}_x+\boldsymbol{\mu}\boldsymbol{\mu}'-\frac{1}{n}\boldsymbol{\Sigma}_x-\boldsymbol{\mu}\boldsymbol{\mu}' \\
&= \frac{n-1}{n}\boldsymbol{\Sigma}_x.
\end{aligned}$$

From the result, $\boldsymbol{S}_n$ is a biased estimate of $\boldsymbol{\Sigma}_x$, but $\boldsymbol{S} = \frac{n}{n-1}\boldsymbol{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})'$ is an unbiased estimate of $\boldsymbol{\Sigma}_x$.

5 Multivariate Normal Distribution

Definition: Let $\boldsymbol{X} = (X_1, \ldots, X_p)'$ be a p-dimensional random vector. We say that $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if the probability density function of $\boldsymbol{X}$ is

$$f(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[\frac{-1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right],$$

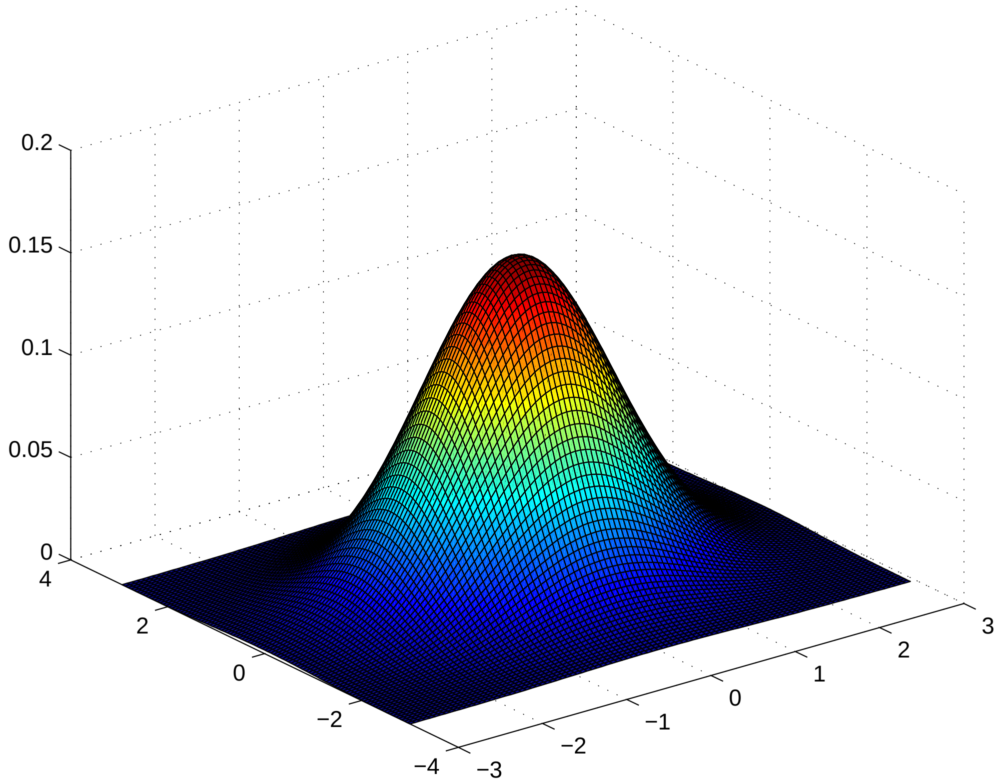

Figure 2: Bivariate standard normal

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and $\boldsymbol{\Sigma}$ is a positive definite $p \times p$ matrix.
Basic properties

- $E(\boldsymbol{X}) = \boldsymbol{\mu}$
- $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$
- Characteristic function: $\phi(\boldsymbol{t}) = E(e^{i\boldsymbol{t}'\boldsymbol{X}}) = \exp[i\boldsymbol{t}'\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}],$, where $\boldsymbol{t} = (t_1, \ldots, t_p)'$.
- Moment generating function: $\Phi(\boldsymbol{t}) = \exp[\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}]$.

Alternative definition: $\boldsymbol{X}$ has a p-dimensional normal distribution if and only if $\boldsymbol{c}'\boldsymbol{X}$ is univariate normal for all fixed p-dimensional (nonzero) vector $\boldsymbol{c}$.

Some plots of bivariate normal density function.

Key properties of $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

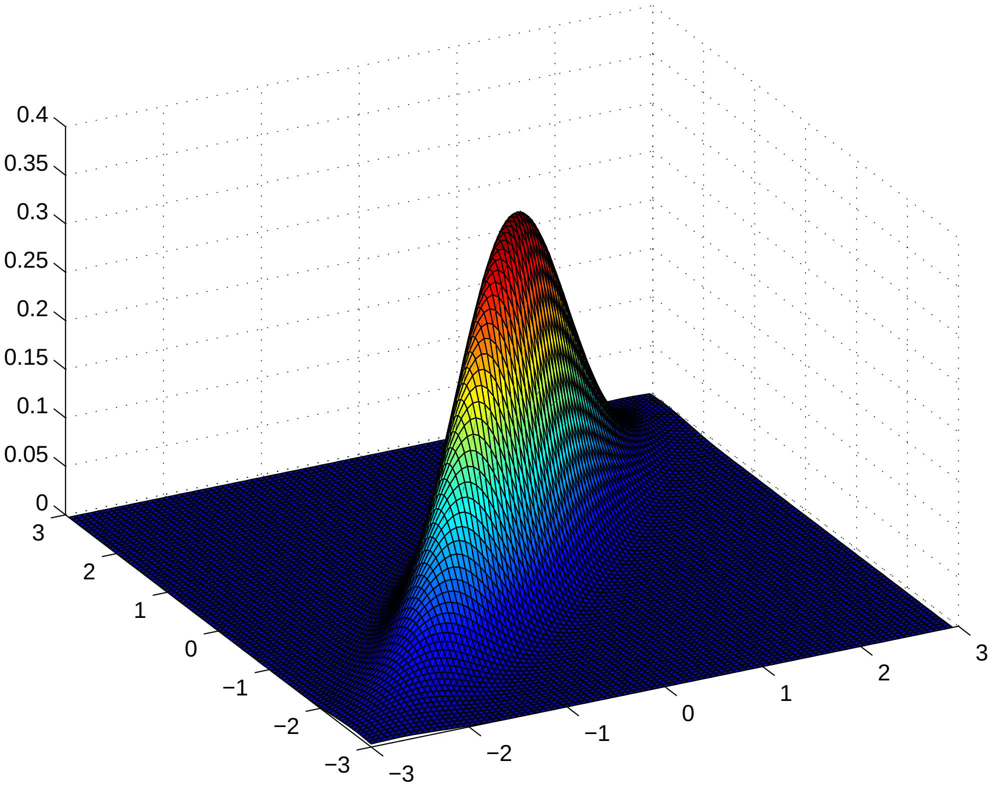

Figure 3: Bivariate normal with correlation 0.9

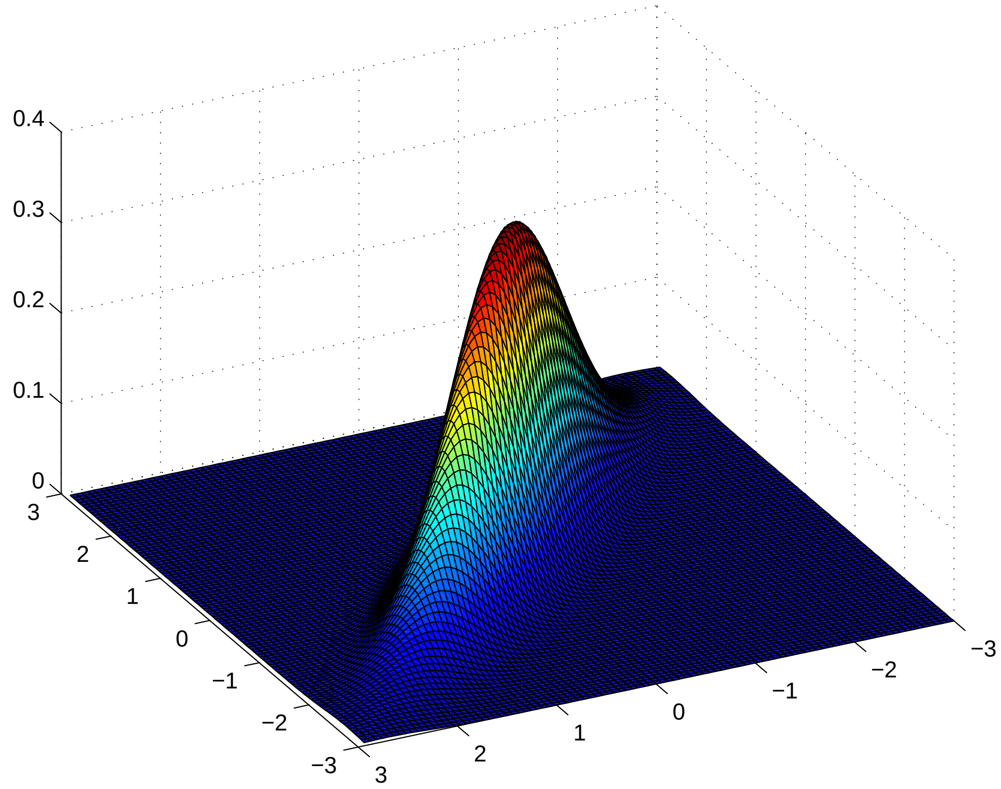

Figure 4: Bivariate normal with correlation −0.9

1. Linear combinations of the components of $\boldsymbol{X}$ are normally distributed. Let $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}$, where $\boldsymbol{C}$ is a $k \times p$ matrix with Rank($\boldsymbol{C}$) $= k \leq p$. Then, $\boldsymbol{Y} \sim N_k(\boldsymbol{C}\boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}')$.

   [Can easily be shown by using the characteristic function.]

2. All subsets of the components of $\boldsymbol{X}$ have a (multivariate) normal distribution. Let $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ where $\boldsymbol{X}_1 = (X_1, \ldots, X_k)'$ and $\boldsymbol{X}_2 = (X_{k+1}, \ldots, X_p)'$ with $1 \leq k \leq p$. Partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly as

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')', \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

   Then, $\boldsymbol{X}_1 \sim N_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. In particular, each component $X_i \sim N(\mu_i, \sigma_{ii})$, where σ_{ii} is the (i,i)th element of $\boldsymbol{\Sigma}$.

   The converse is not necessarily true. Specifically, if $\boldsymbol{X}_i \sim N(\boldsymbol{\mu}_i, \sigma_{ii})$, then $\boldsymbol{X} = (X_1, \ldots, X_p)'$ may not be normally distributed.

   Example. Suppose $X_1 \sim N(0,1)$ and X_2 is defined as

$$X_2 = \begin{cases} -X_1 & \text{if } -1 \leq X_1 \leq 1 \\ X_1 & \text{otherwise.} \end{cases}$$

   Then, X_2 is also $N(0,1)$, but $\boldsymbol{X} = (X_1, X_2)'$ is not bivariate normal.

   Reason: The linear combination $X_1 - X_2$ is not normally distributed. More precisely, $Pr(X_1 - X_2 = 0) = Pr(|X_1| > 1) = 1 - Pr(|X_1| \leq 1) = 0.3174$.

3. If $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')' \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent if and only if cov$(\boldsymbol{X}_1, \boldsymbol{X}_2) = \mathbf{0}$.

   In general, if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, then cov$(\boldsymbol{X}_1, \boldsymbol{X}_2) = \mathbf{0}$. On the other hand, if $\boldsymbol{X}_i \sim N_{k_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$ and are independent, where k_i is the dimension of $\boldsymbol{X}_i$, then $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ is jointly multivariate normal with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ and cov$(\boldsymbol{X}) =$ diag$\{\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}\}$.

4. The conditional distributions of the components are (multivariate) normal. Again, consider the partition mentioned before. We have

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_{k_1}[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}],$$

   where $\boldsymbol{\Sigma}_{22}$ is positive definite.

   This property is very important and widely used. For example, it is the theory behind the well known **Kalman Filter** in time series analysis and the forward filtering backward sampling (FFBS) algorithm in Markov chain Monte Carlo methods.

5. $(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_p^2$, chi-square distribution with p degrees of freedom.

Maximum likelihood estimation: Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be *iid* random samples from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The joint probability density function (pdf) is

$$\begin{aligned} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp[-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})] \\ &= \frac{1}{(2\pi)^{np/2}} \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right]. \end{aligned}$$

Treating $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as given, $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. It is called the likelihood function of the random sample.
The values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximize $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are called the maximum likelihood estimate (MLE) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.
It turns out that the MLEs are

$$\begin{aligned} \widehat{\boldsymbol{\mu}} &= \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i, \\ \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'. \end{aligned}$$

To derive the results, we make use of the following two properties of a $p \times p$ symmetric matrix $\boldsymbol{A}$:

1. $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = tr(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}')$.

2. $tr(\boldsymbol{A})$ = sum of the eigenvalues of $\boldsymbol{A}$.

Derivation of MLE for Gaussian distribution:
First, focus on the exponent of the likelihood function

$$(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) = tr[(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})] = tr[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})'].$$

Therefore,

$$\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) = \sum_{i=1}^{n} tr[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})'] = tr\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})'\right)\right].$$

Next, adding and subtracting $\bar{\boldsymbol{x}}$, we have

$$\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \boldsymbol{\mu})(\boldsymbol{x}_i - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \boldsymbol{\mu})' = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})',$$

because the cross-product terms are zero.
Denoting the likelihood function by $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}tr\left\{\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\right)\right\}\right] \\ &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}tr\left\{\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'\right)\right\} - \frac{n}{2}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\right]. \end{aligned}$$

From the prior equation, $\boldsymbol{\mu}$ only appears on the 2nd term so that the likelihood function is maximized with respect to $\boldsymbol{\mu}$ when the 2nd term is 0. Furthermore, since $\boldsymbol{\Sigma}$ is positive definite, $\boldsymbol{\Sigma}^{-1}$ is also positive definite. Thus, $(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})=0$ if and only if $\bar{\boldsymbol{x}}=\boldsymbol{\mu}$. Consequently, $\widehat{\boldsymbol{\mu}}=\bar{\boldsymbol{x}}$.
To obtain $\widehat{\boldsymbol{\Sigma}}$, we make use of the following inequality:
Result 4.10 of the textbook: Given a $p\times p$ positive definite matrix $\boldsymbol{B}$ and a scalar $b>0$, it follows that

$$\frac{1}{|\boldsymbol{\Sigma}|^b}e^{-tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{B})/2}\le\frac{1}{|\boldsymbol{B}|^b}(2b)^{pb}e^{-bp},$$

for all positive definite matrix $\boldsymbol{\Sigma}_{p\times p}$, with equality holding only for $\boldsymbol{\Sigma}=(1/2b)\boldsymbol{B}$.
Proof of Result 4.10 can be found on pages 170-171 of the textbook.
Return to MLE. Plugging in the estimate $\widehat{\boldsymbol{\mu}}=\bar{\boldsymbol{x}}$, we obtain the likelihood function as

$$L(\widehat{\boldsymbol{\mu}},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}}\exp\left[-\frac{1}{2}tr\left\{\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'\right)\right\}\right],$$

which is a function of $\boldsymbol{\Sigma}$. Let $b=n/2$ and $\boldsymbol{B}=\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'$. Apply Result 4.10, the maximum of the likelihood function occurs at

$$\widehat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'.$$

The maximized likelihood function value is

$$L(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}})=\frac{1}{(2\pi)^{np/2}}e^{-(np)/2}\frac{1}{|\widehat{\boldsymbol{\Sigma}}|^{n/2}}.$$

Furthermore, since $|\widehat{\boldsymbol{\Sigma}}|=[(n-1)/n]^p|\boldsymbol{S}|$, where $\boldsymbol{S}$ is the unbiased estimate of the covariance matrix, the maximized likelihood function value is

$$L(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}})=\text{constant}\times(\text{generalized variance})^{-n/2}.$$

The maximized likelihood function plays an important role in model selection using information criteria, e.g. AIC and BIC. It is the measure of goodness-of-fit of the model to the data.

Remark: MLEs possess an invariance property. Let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$ and $h(\boldsymbol{\theta})$ be a function of $\boldsymbol{\theta}$. Then the MLE of $h(\boldsymbol{\theta})$ is $h(\widehat{\boldsymbol{\theta}})$. For instance, the MLE of the individual variances of the components of $\boldsymbol{X}$ are $\hat{\sigma}_{ii}=\frac{1}{n}\sum_{i=j}^{n}(x_{ij}-\bar{x}_i)^2$, for $i=1,\ldots,p$.
Sufficient statistics: Since the likelihood function of a random sample of normal distribution only depends on the sample mean $\bar{\boldsymbol{x}}$ and the sum-of-squares-and-cross-product matrix $\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'=(n-1)\boldsymbol{S}$, we say that $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ are the sufficient statistics of the multivariate normal distribution.

5.1 The sampling distribution of $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$

Based on the prior discussion, for a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the maximum likelihood estimators are $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\widehat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\boldsymbol{S}$. What are the sampling distributions of $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$?
For $\bar{\boldsymbol{X}}$, the answer is simple because $\bar{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$.
If $p = 1$, then $(n-1)S = \sum_{i=1}^n (X_i - \bar{X})^2$ is distributed as σ^2 times a chi-square random variable with $(n-1)$ degrees of freedom. What is the generalization? It turns out that $(n-1)\boldsymbol{S}$ follows a *Wishart random matrix* with $(n-1)$ degrees of freedom.
Finally, similar to the univariate case, $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are independent.

Wishart distribution:
Definition: Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$ be a random sample from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then, the distribution of $\sum_{i=1}^m \boldsymbol{Z}_i\boldsymbol{Z}_i'$ is called a Wishart distribution with m degrees of freedom and is denoted by $\boldsymbol{W}_m(\cdot|\boldsymbol{\Sigma})$.
If $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, then we have a standard Wishart distribution with m degrees of freedom. Note that the dimension of Wishart distribution is implicitly given by $\boldsymbol{\Sigma}$.
When $m > p$, the probability density function (pdf) of a Wishart distribution $\boldsymbol{W}_m(\cdot|\boldsymbol{\Sigma})$ is

$$\boldsymbol{w}_m(\boldsymbol{A}|\boldsymbol{\Sigma}) = \frac{|\boldsymbol{A}|^{(m-p-1)/2} e^{-tr[\boldsymbol{A}\boldsymbol{\Sigma}^{-1}]/2}}{2^{pm/2}\pi^{p(p-1)/4}|\boldsymbol{\Sigma}|^{m/2}\prod_{i=1}^p \Gamma(\frac{1}{2}(m+1-i))}$$

where $\Gamma(\cdot)$ is the gamma function and $\boldsymbol{A}$ is positive definite. See Anderson's book.

Some properties of the Wishart distribution:

1. If $\boldsymbol{X} \sim W_m(\cdot|\boldsymbol{\Sigma})$ with $m > p$, then $E(\boldsymbol{X}) = m\boldsymbol{\Sigma}$ and $\text{Var}[X_{ij}] = m(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})$.
2. If $\boldsymbol{M} \sim \boldsymbol{W}_m(\cdot|\boldsymbol{\Sigma})$ and $\boldsymbol{B}$ is a $p \times q$ matrix, then $\boldsymbol{B}'\boldsymbol{M}\boldsymbol{B} \sim \boldsymbol{W}_m(\cdot, \boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})$.
3. If $\boldsymbol{M}_i \sim \boldsymbol{W}_{m_i}(\cdot, \boldsymbol{\Sigma})$ for $i = 1$ and 2. In addition, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are independent, then $\boldsymbol{M}_1 + \boldsymbol{M}_2 \sim \boldsymbol{W}_{m_1+m_2}(\cdot, \boldsymbol{\Sigma})$.
4. If $\boldsymbol{M} \sim \boldsymbol{W}_m(\cdot, \boldsymbol{\Sigma})$ and $\boldsymbol{a}$ is any fixed p-dimensional vector such that $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \neq 0$, then $\boldsymbol{a}'\boldsymbol{M}\boldsymbol{a}/(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}) \sim \chi_m^2$.

Remark: Given a positive definite covariance matrix $\boldsymbol{\Sigma}$, one can generate random samples of Wishart distribution in R. The command is `x <- rWishart(n,df,Sigma)`, where n is the sample size, df is the degrees of freedom, Sigma is $\boldsymbol{\Sigma}$.

6 Large sample properties

Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are random samples from a population distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. If the population is normally distributed, then the finite-sample behavior of $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ is stated earlier. Here we consider the case in which the population distribution is not normal.

Law of Large Number:

1. $\bar{\boldsymbol{X}}$ converges in probability to $\boldsymbol{\mu}$ as $n \to \infty$.
2. $\boldsymbol{S}$ converges in probability to $\boldsymbol{\Sigma}$ as $n \to \infty$.

Central Limit Theorem:

1. $\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \to_d N_p(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$, where $\to_d$ denotes convergence in distribution.
2. $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \to_d \chi^2_p$ as $n \to \infty$.

7 Assessing the normality assumption

Recall that in the univariate case a simple approach to assess the normality assumption if the *normal probability plot*, i.e. the normal quantile to quantile (QQ) plot.

Univariate QQ-plot: Let $x_1, \ldots, x_n$ be a random sample from a population distribution. Let $\{x_{(i)}\}$ be the ordered sample, i.e. $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. When x_i are distinct, then there are exactly j observations that are smaller than or equal to $x_{(j)}$. That is, $x_{(j)}$ is the j/nth quantile of the sample. Let $q_{(j)}$ be the $(j - 0.5)/n$ quantile of a standard normal distribution. The plot of $(q_{(j)}, x_{(j)})$ is called normal probability plot or the normal QQ-plot of the random sample. If x_is are indeed from a normal distribution, then the QQ-plot should show a straight line. Note that the subtraction of 0.5 from j is referred to as the *continuity* correction.

The following R commands create the normal QQ-plot shown in Figure 5.

```
> x=rnorm(100)
> par(mfcol=c(2,1))
> hist(x)
> qqnorm(x)
> prob=(c(1:100)-0.5)/100
> z=qnorm(prob)
> y=sort(x)
> cor(y,z)
```

A simple statistic to check the straight line in a QQ-plot is the correlation coefficient between $\{(q_{(j)}, x_{(j)})\}$. Specifically,

$$r_q = \frac{\sum_{j=1}^n (x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^n (x_{(j)} - \bar{x})^2}\sqrt{\sum_{j=1}^n (q_{(j)} - \bar{q})^2}}.$$

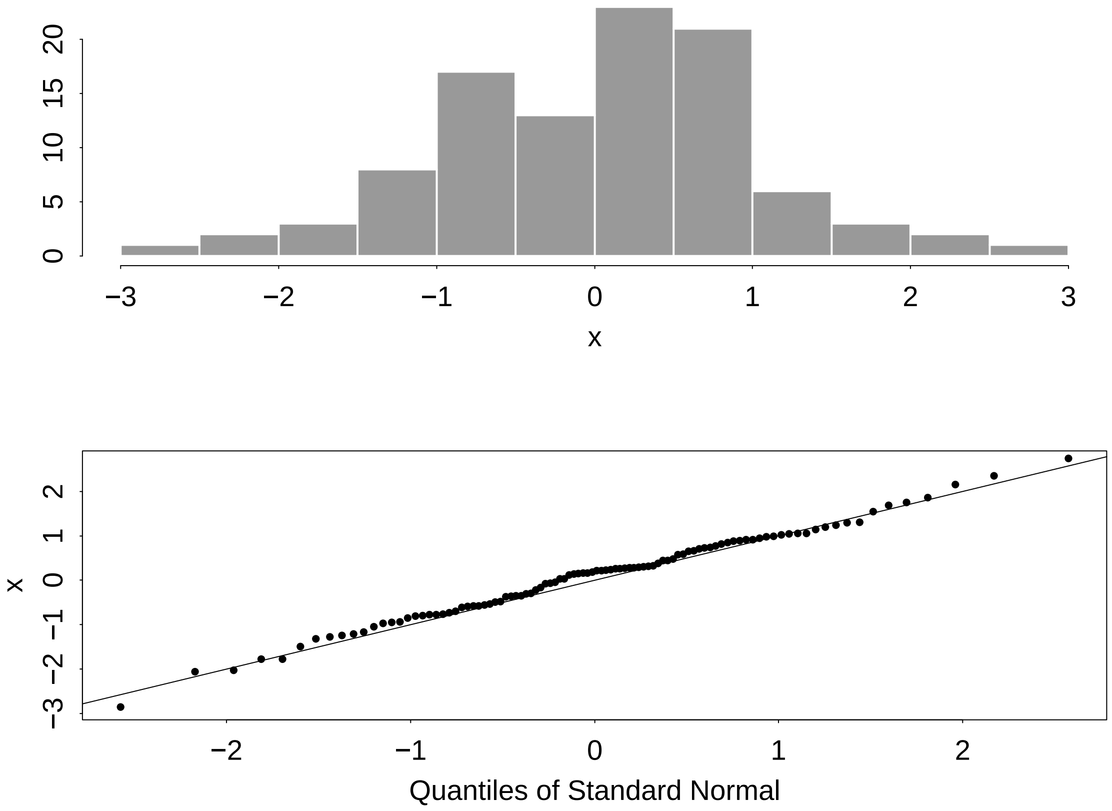

Figure 5: A random sample of 100 points from N(0,1): (a) histogram, (b) QQ-plot with a straight line imposed.

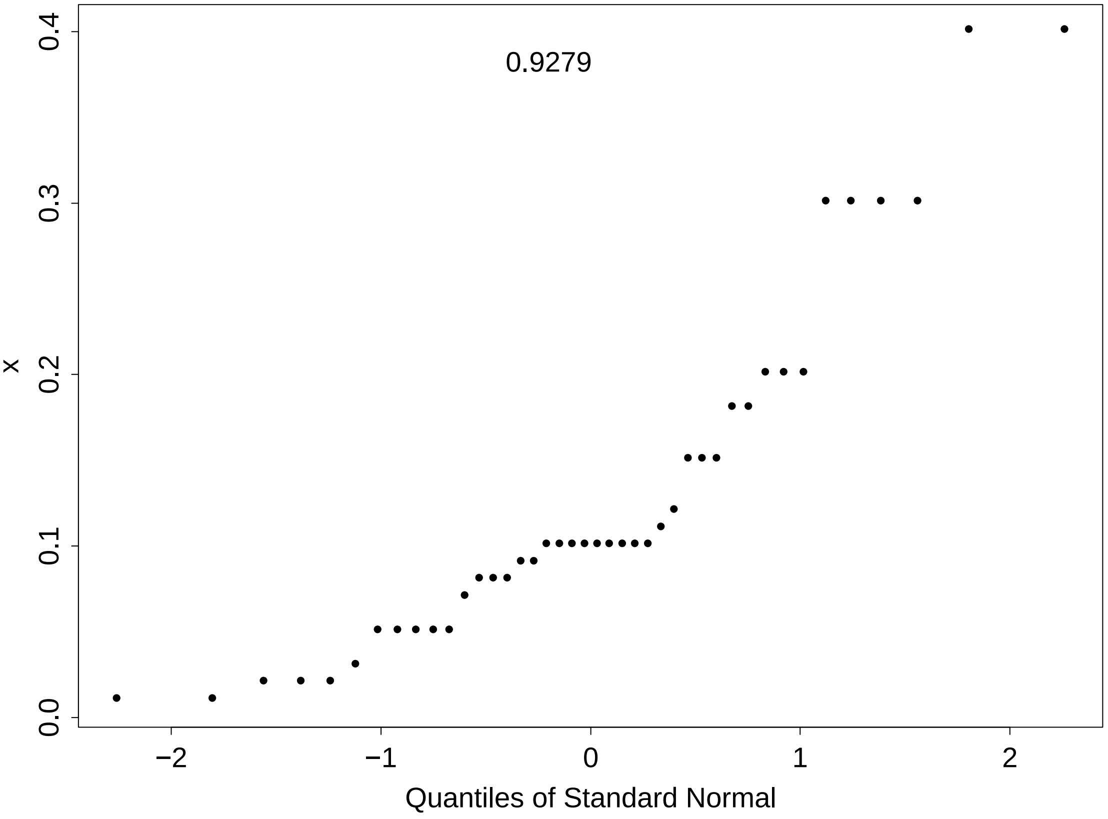

Figure 6: Normal QQ-plot for the Radiation Data of Table 4.1

A table of critical values for r_q is given in Table 4.2 of the textbook, page 181. For the simulated example, $r_Q = 0.9961$ so that, as expected, we cannot reject the normality assumption. [For $n = 100$, the 5% critical value of the correlation coefficient is 0.9873.]
Next, consider the Radiation data on Table 4.1 of the text with 42 observations. The normal QQ-plot is given in Figure 6. The correlation coefficient is 0.9279, which is smaller than the approximate 5% critical value 0.9749. Thus, the normality assumption is rejected.
Multivariate case: Here we make use of the properties of multivariate normal distribution. In particular,

$$d_j^2 = (\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}}) \sim \chi_p^2.$$

Thus, we may use the chi-square QQ-plot of $\{d_j\}$ to check the multivariate normality assumption.
Remark: Programs to construct chi-square QQ-plot for R, Matlab and S-Plus are available from the course web page. For illustration, Figure 7 shows the chi-square QQ-plot for the Stiffness Data on Table 4.3 of the textbook. The R commands used are given below:

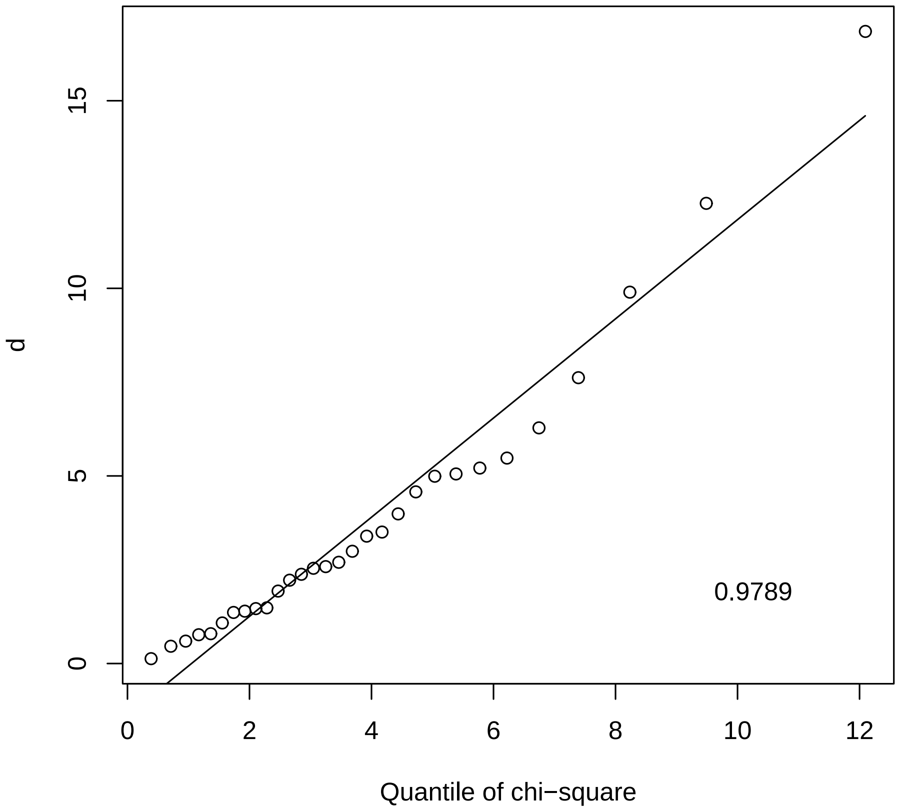

Figure 7: Chi-square QQ-plot for the Stiffness Data of Table 4.3

```
> x=read.table(‘‘T4-3.DAT’’) % Load the data into R workspace
> x=x[,1:4]
> source(‘‘ama.R’’)  % Compile a collection of multivariate analysis commands.
> qqchi2(x)
```

In practice, one can also assess the normality of any (non-zero) linear combination of $\boldsymbol{X}$. In particular, the marginal distributions are often checked.

Remark: Some researchers show that the chi-square approximation of d_j^2 can be poor. An alternative is to use the result of Gnanadesikan and Ketternring (1972), who show that

$$u_i = \frac{nd_j^2}{(n-1)^2},$$

has a beta distribution with parameter $\alpha = p/2$ and $\beta = (n-p-1)/2$ and the corresponding quantile v_i for the ordered $u_{(i)}$ is given by the probability $(i-a)/(n-a-b+1)$ with $a=(p-2)/(2p)$ and $b=(n-p-3)/[2(n-p-1)]$. The R script for this alternative is `qqbeta.R` on the course web.

8 Outliers

Often the random sample contains a few unusual observations that do not seem to belong to the pattern of variability produced by the other observations. These unusual observations are called *outliers* of the data set. Note that not all outliers are wrong numbers. They are just different from the majority of the data.
In multivariate case, some steps can be taken to identify possible outliers.

1. Examine the dot plot of each variable
2. Study the scatter plot for each pair of the variables.
3. Calculate the generalized squared distance $d_j^2=(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})$. Examine the unusually large values of d_j^2. A chi-square QQ-plot might be helpful here.

9 Transformation to near normality

Some transformations are known to move the data close to being normal.

1. Counts, y: square-root transformation $z=\sqrt{y}$.
2. Proportion p: $\text{Logit}(p)=\frac{1}{2}\ln\left(\frac{p}{1-p}\right)$.
3. Correlation r: $z(r)=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$.

For other cases, the Box-Cox transformation can be used. In the univariate case, the power transformation is

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

And λ is obtained by maximizing

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

where $x^{(\lambda)}$ is defined earlier and

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

is the arithmetic average of the transformed data.
In the multivariate case, one can transform the components either individually or jointly. Let λ_i be the power of the transformation of the ith variable. These λ_is can be estimated jointly by maximizing the function $\ell(\lambda_1, \ldots, \lambda_p)$ given in Eq. (4.40) of the textbook.

R packages: The packages `quantmod` and `quantdl` allow user to download many economic and financial data via Internet. The sources include FRED (Federal Reserve Economic Data) of the Federal Reserve Bank of St. Louis, Yahoo, Google, etc.