TMA4267 Linear Statistical Models V2017 [L5]

Part 1: Multivariate RVs, and the multivariate normal distribution

Properties of the multivariate normal distribution [H:2.6,4.4,5.1]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: January 24, 2017

Last lecture: derived the MGF and pdf of the multivariate normal distribution

1. $Z \sim N_1(0,1)$
   - MGF: $M_Z(t) = \mathrm{E}(e^{tz}) = e^{\frac{1}{2}t^2}$
2. $Z_1, Z_2, \ldots, Z_p$ iid $N_1(0,1) \rightarrow \boldsymbol{Z}_{p\times 1} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$
   - MGF: $M_{\boldsymbol{Z}}(\boldsymbol{t}) = \mathrm{E}(e^{\boldsymbol{t}^T\boldsymbol{z}}) = e^{\frac{1}{2}\boldsymbol{t}^T\boldsymbol{t}}$
3. $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$, $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$ gives $\boldsymbol{X}_{p\times 1} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
   - MGF: $M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}(e^{\boldsymbol{t}^T\boldsymbol{x}}) = e^{\boldsymbol{t}^T\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{t}^T\boldsymbol{t}}$
   - pdf (Σ invertible):

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

Why is the mulitivariate normal distribution so important in statistics?

- Many natural phenomena may be modelled using this distribution (just as in the univariate case).
- Multivariate version of the central limit theorem- the sample mean will be approximately multivariate normal for large samples.
- Good interpretability of the covariance.
- Mathematically tractable.
- Building block in many models and methods.

Today: six properties of the mvN

Let $\boldsymbol{X}_{(p\times 1)}$ be a random vector from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. The grapical contours of the mvN are ellipsoids (shown using spectral decomposition).
2. Linear combinations of components of $\boldsymbol{X}$ are (multivariate) normal (proof using MGF).
3. All subsets of the components of $\boldsymbol{X}$ are (multivariate) normal (special case of the above).
4. Zero covariance implies that the corresponding components are independently distributed (proof using MGF).
5. $\boldsymbol{A\Sigma B}^T = \boldsymbol{0} \Leftrightarrow \boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent (will be very important in Part 2)
6. The conditional distributions of the components are (multivariate) normal. $\boldsymbol{X}_2 \mid (\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim N_{p2}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$.

Diabetes data

We will study a data set on diabetes in Part 2. The data set has measurements on $n = 442$ diabetes patients, and $p = 11$ different measurements are taken for each patients. These measurements are:

- age
- sex
- body mass index (bmi)
- mean arterial blood pressure (map)
- six blood serum measurements: total cholesterol (tc), ldl cholesterol (ldl), hdl cholesterol (hdl), tch, ltg, glu.
- a quantitative measurement of disease progression one year after baseline (prog)

We will look at the four variables bmi, map, tc and ldl. Can we assume that these follow a multivariate normal distribution?

Contours of multivariate normal distribution

- Contours of constant density for the p-dimensional normal distribution are ellipsoids defined by $\boldsymbol{x}$ such that

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = b$$

  where $b > 0$ is a constant.
  These ellipsoids are centered at $\boldsymbol{\mu}$ and have axes $\pm\sqrt{b\lambda_i}\boldsymbol{e}_i$, where $\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i$, for $i = 1, ..., p$.
- $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is distributed as χ^2_p.
- The volume inside the ellipsoid of $\boldsymbol{x}$ values satisfying

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \leq \chi^2_p(\alpha)$$

  has probability $1-\alpha$.

Example: Slightly modified version of Exam K2014 1b

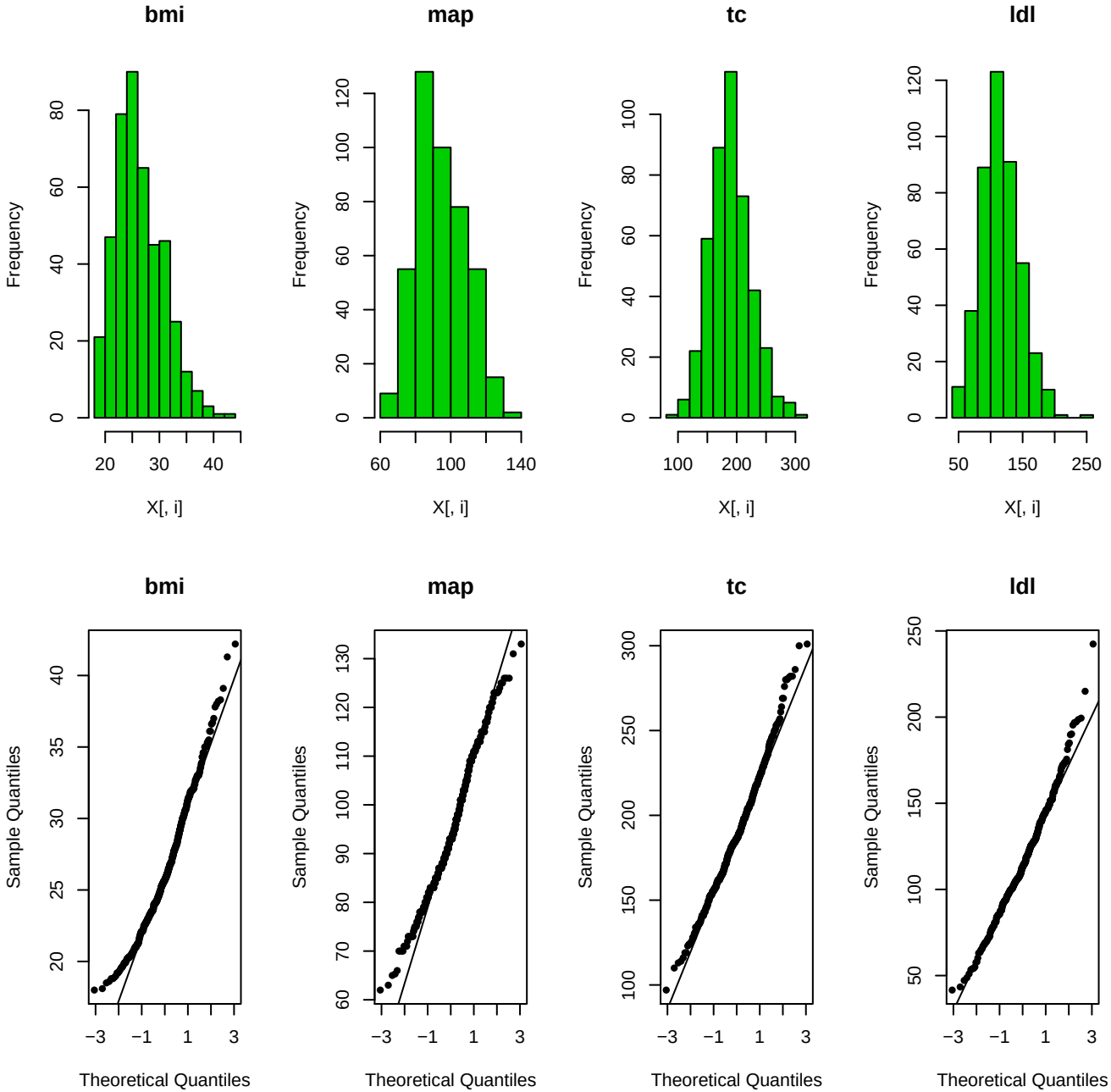
Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be a bivariate normal random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$.

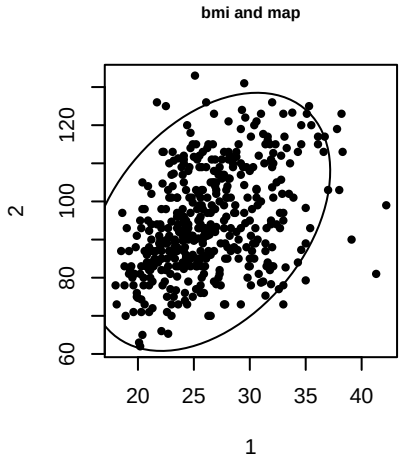
You find the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ on the next slide.

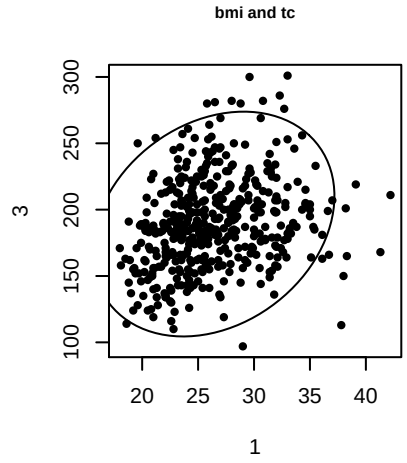
Describe the graph of the equation $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = b$ where $b > 0$ is a constant.
Make a drawing of the graph, for $b = 1$ found above.
What is the probability that a random sample from this distribution will be inside this graph?

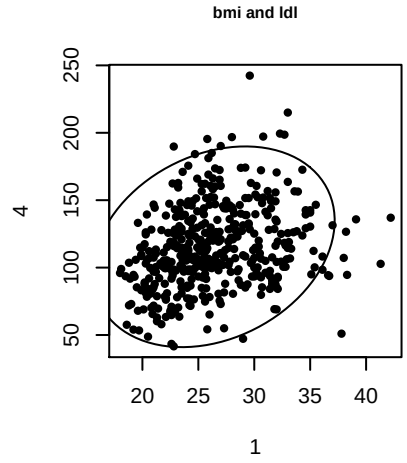
Example: Exam K2014 1b

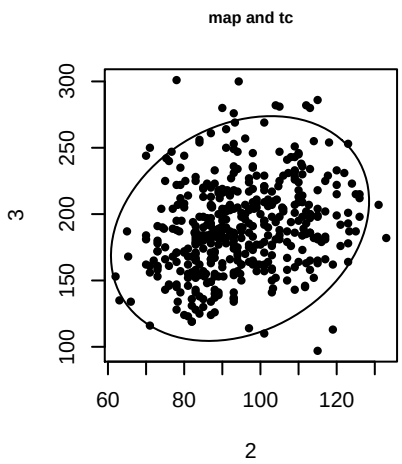
```
> sigma <- matrix(c(1,0.5,0.5,2),ncol=2)
> eigen(sigma)
$values
[1] 2.2071068 0.7928932

$vectors
           [,1]       [,2]
[1,] 0.3826834 -0.9238795
[2,] 0.9238795  0.3826834
```

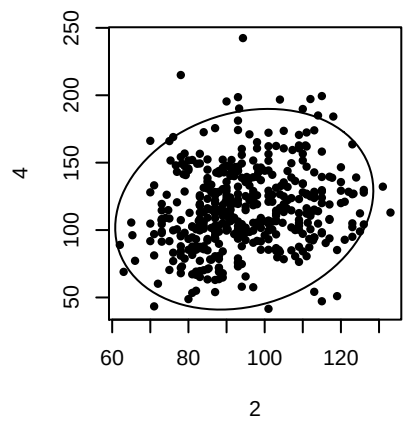

bmi
map
tc
ldl
Frequency
X[, i]
Sample Quantiles
Theoretical Quantiles

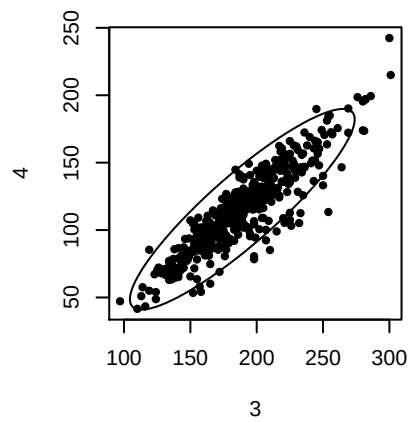

bmi and map
2
1
60
80
100
120
20
25
30
35
40


bmi and tc
3
1
100
150
200
250
300
20
25
30
35
40


bmi and ldl
4
1
50
100
150
200
250
20
25
30
35
40


map and tc
3
2
100
150
200
250
300
60
80
100
120


map and ldl
4
2
50
100
150
200
250
60
80
100
120


tc and ldl
4
3
50
100
150
200
250
100
150
200
250
300

Multivariate distributions - in 3D: task for the intermission!

Let $\mathbf{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$.

The following four 3D-printed figures have been made:

- A: $\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0.3$
- B: $\sigma_x = 1$, $\sigma_y = 1$, $\rho = 0$
- C: $\sigma_x = 1$, $\sigma_y = 1$, $\rho = 0.5$
- D: $\sigma_x = 1$, $\sigma_y = 2$, $\rho = 0$

The figures have the following colours:

- white
- purple
- red
- black

Task: match letter and colour by writing the correct letter after the name of the colour on the available sheets and take the sheet with you. We report on the solution after the intermission.

Today: six properties of the mvN

Let $\boldsymbol{X}_{(p\times 1)}$ be a random vector from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. The grapical contours of the mvN are ellipsoids (shown using spectral decomposition).
2. Linear combinations of components of $\boldsymbol{X}$ are (multivariate) normal (proof using MGF).
3. All subsets of the components of $\boldsymbol{X}$ are (multivariate) normal (special case of the above).
4. Zero covariance implies that the corresponding components are independently distributed (proof using MGF).
5. $\boldsymbol{A\Sigma B}^T = \boldsymbol{0} \Leftrightarrow \boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent (will be very important in Part 2)
6. The conditional distributions of the components are (multivariate) normal. $\boldsymbol{X}_2 \mid (\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim N_{p2}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$.

Example: Exam K2014 1a

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be a bivariate normal random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$.

Let $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $Y_1 = 3X_1 - 2X_2$ and $Y_2 = X_1 + X_2$.

What is the distribution of $\boldsymbol{Y}$?

What is the distribution of Y_1?

Let $Z = X_1 + aX_2$. How can you choose a so that Z and Y_2 are independent?

Example: Exam K2014 1a (slightly modified)

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be a bivariate normal random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$. Let $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $Y_1 = 3X_1 - 2X_2$ and $Y_2 = X_1 + X_2$.
What is the distribution of $\boldsymbol{Y}$?

What is the distribution of Y_1?

Let $Z = X_1 + aX_2$. How can you choose a so that Z and Y_2 are independent?

Example: Exam K2014 1a (slightly modified)

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be a bivariate normal random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$. Let $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $Y_1 = 3X_1 - 2X_2$ and $Y_2 = X_1 + X_2$.
What is the distribution of $\boldsymbol{Y}$?

What is the distribution of Y_1?

Let $Z = X_1 + aX_2$. How can you choose a so that Z and Y_2 are independent?

Today: six properties of the mvN

Let $\boldsymbol{X}_{(p\times 1)}$ be a random vector from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. The grapical contours of the mvN are ellipsoids (shown using spectral decomposition).
2. Linear combinations of components of $\boldsymbol{X}$ are (multivariate) normal (proof using MGF).
3. All subsets of the components of $\boldsymbol{X}$ are (multivariate) normal (special case of the above).
4. Zero covariance implies that the corresponding components are independently distributed (proof using MGF).
5. $\boldsymbol{A\Sigma B}^T = \boldsymbol{0} \Leftrightarrow \boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent (will be very important in Part 2)
6. The conditional distributions of the components are (multivariate) normal. $\boldsymbol{X}_2 \mid (\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim N_{p2}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$.

Independent variables?

Let $\boldsymbol{X}_{p\times 1} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

- List the pairs of variables that are independent.

Example: Exam K2014 1a - cont.

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be a bivariate normal random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$.

Let $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $Y_1 = 3X_1 - 2X_2$ and $Y_2 = X_1 + X_2$.

Let $Z = X_1 + aX_2$. How can you choose a so that Z and Y_2 are independent?

Today: six properties of the mvN

Let $\boldsymbol{X}_{(p\times 1)}$ be a random vector from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. The grapical contours of the mvN are ellipsoids (shown using spectral decomposition).
2. Linear combinations of components of $\boldsymbol{X}$ are (multivariate) normal (proof using MGF).
3. All subsets of the components of $\boldsymbol{X}$ are (multivariate) normal (special case of the above).
4. Zero covariance implies that the corresponding components are independently distributed (proof using MGF).
5. $\boldsymbol{A\Sigma B}^T = \boldsymbol{0} \Leftrightarrow \boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent (will be very important in Part 2)
6. The conditional distributions of the components are (multivariate) normal. $\boldsymbol{X}_2 \mid (\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim N_{p2}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$.

Example: Exam V2010, Problem 1

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \begin{pmatrix} 4 \\ -3 \\ 1 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & -1.5 \\ 0 & -1.5 & 5 \end{pmatrix}$.

a) Find the distribution of $X_1 + X_2 + X_3$ and of X_2 given $X_1 = x_1$ and $X_3 = x_3$.

Help: for $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right)$ we have

$\boldsymbol{X}_2 \mid (\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim N(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$

Today: six properties of the mvN

Let $\boldsymbol{X}_{(p\times 1)}$ be a random vector from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. The grapical contours of the mvN are ellipsoids (shown using spectral decomposition). [CompEx1.1b]
2. Linear combinations of components of $\boldsymbol{X}$ are (multivariate) normal (proof using MGF). [CompEx1.1a]
3. All subsets of the components of $\boldsymbol{X}$ are (multivariate) normal (special case of the above).
4. Zero covariance implies that the corresponding components are independently distributed (proof using MGF). [CompEx1.1a]
5. $\boldsymbol{A\Sigma B}^T = \boldsymbol{0} \Leftrightarrow \boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent (will be very important in Part 2). [CompEx1.2b]
6. The conditional distributions of the components are (multivariate) normal. $\boldsymbol{X}_2 \mid (\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim N_{p2}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$.