TMA4267 Linear Statistical Models V2017 [L1]

Introduction to the course
Part 1: Multivariate random variables, and the multivariate normal distribution

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: January 10, 2017

TMA4267 Linear Statistical Models

- Statistics, linear statistical models and movie recommender systems.
- Learning outcome.
- TMA4267 core and parts.
- Background knowledge in probability and statistical inference.
- TMA4267 course information.
- Voting and questionnaire.
- Part 1: Multivariate RVs and the multivariate normal distribution.

What is Statistics?

What is Statistics?

- The true foundation of theology is to ascertain the character of God.
- It is by the aid of *Statistics* that law in the social sphere can be ascertained and codified,
- and, certain aspects of the character of God hereby revealed.
- The study of statistics is thus a *religious service*.

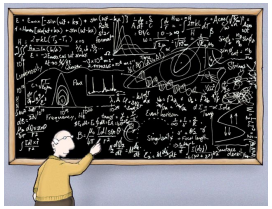
What is Statistics?

- The true foundation of theology is to ascertain the character of God.
- It is by the aid of *Statistics* that law in the social sphere can be ascertained and codified,
- and, certain aspects of the character of God hereby revealed.
- The study of statistics is thus a *religious service*.

Florence Nightingale (1820-1910). Quotation from "Games, Gods and Gambling: A History of Probability and Statistical Ideas" by F. N. David.

What is Statistics?

- The goal of Statistics is to expand our knowledge based on collection and analysis of empirical data.
- Two branches:
  - Probability: the mathematical study of the probability of random events.
  - Statistical Inference: models and methods for collecting, describing, analysing and interpreting numerical data.



Drawing taken from http://www.nearingzero.net - now at http://www.lab-initio.com/

Word cloud: Probability

Word cloud: Statistical Inference

Linear Statistical Models

Simple linear regression (height of child explained by mid-parent height):

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Linear Statistical Models

Simple linear regression (height of child explained by mid-parent height):

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Multiple linear regression (also include other explanatory variables):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

The multiple linear regression model is our linear statistical model!

Linear Statistical Models

Simple linear regression (height of child explained by mid-parent height):

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Multiple linear regression (also include other explanatory variables):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

The multiple linear regression model is our linear statistical model!
So, why is this course not called "Regression"?

Linear Statistical Models

Simple linear regression (height of child explained by mid-parent height):

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Multiple linear regression (also include other explanatory variables):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

The multiple linear regression model is our linear statistical model!
So, why is this course not called "Regression"?
We include theory that focus on mathematical understanding: multivariate random variables, the multivariate normal distribution, projections, idempotent matrices, hypothesis tests, design of experiments, ...

Recommender systems

- Recommender systems are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item.

Recommender systems

- Recommender systems are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item.
- Recommender systems have become extremely common in recent years, and are applied in a variety of applications. The most popular ones are probably movies, music, news, books, research articles, search queries, social tags, and products in general. However, there are also recommender systems for experts, collaborators, jokes, restaurants, financial services,life insurance, persons (online dating), and Twitter followers.

Recommender systems

- Recommender systems are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item.
- Recommender systems have become extremely common in recent years, and are applied in a variety of applications. The most popular ones are probably movies, music, news, books, research articles, search queries, social tags, and products in general. However, there are also recommender systems for experts, collaborators, jokes, restaurants, financial services,life insurance, persons (online dating), and Twitter followers.

Recommender systems

- Recommender systems are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item.
- Recommender systems have become extremely common in recent years, and are applied in a variety of applications. The most popular ones are probably movies, music, news, books, research articles, search queries, social tags, and products in general. However, there are also recommender systems for experts, collaborators, jokes, restaurants, financial services,life insurance, persons (online dating), and Twitter followers.

Source: Wikipedia: Recommender systems

The Netflix Price: 2006

Text from `http:\www.netflixprice.com`.

- To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch.

The Netflix Price: 2006

Text from `http:\www.netflixprice.com`.

- To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch.
- Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.

The Netflix Price: 2006

Text from `http:\www.netflixprice.com`.

- To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch.
- Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.
- We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.)

The Netflix Price: 2006

Text from `http:\www.netflixprice.com`.

- To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch.
- Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.
- We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.)
- Remark: At this point in time DVDs were sent to customers by mail - this was before the age of online streaming.

The Netflix Price: 2006

Text from `http:\www.netflixprice.com`.

- To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch.
- Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.
- We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.)
- Remark: At this point in time DVDs were sent to customers by mail - this was before the age of online streaming.

Cinematch

The Cinematch recommender system: *use statistical linear models with a lot of data conditioning*. I have not found any other information on the algorithm online.

"Simple" linear suggestion:
(predicted score on movie for person)=
(some overall score for this movie)+
(some overall score used by this person)+
(similarity of this movie with other movie this person has seen)*
(how much this person liked that movie)+
the same for all the movies this person has rated+
error term.

The Netflix Price: 2006

Text from `http:\www.netflixprice.com`.

- To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch.
- Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.
- We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.)
- Remark: At this point in time DVDs were sent to customers by mail - this was before the age of online streaming.

The prize was awarded the team *BellKor's Pragmatic Chaos* in 2009.

The Netflix Price: Training data

- The training data set consists of more than 100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18 thousand movie titles.
- The ratings are on a scale from 1 to 5 (integral) stars. The date of each rating and the title and year of release for each movie are provided.
- No other customer or movie information is provided. No other data were employed to compute Cinematch's accuracy values used in this Contest.

Text from `http:\www.netflixprice.com`.

The Netflix Price: Test data

- A qualifying test set is provided containing over 2.8 million customer/movie id pairs with rating dates but with the ratings withheld.
- Eligible algorithms must provide predictions for all the withheld ratings for each customer/movie id pair in the qualifying set.
- The qualifying set is divided into two disjoint subsets containing randomly selected pairs from the qualifying set. The assignment of pairs to these subsets is not disclosed.
  - The Site will score each subset by computing the square root of the averaged squared difference between each prediction and the actual rating (the root mean squared error or "RMSE") in the subset, rounded to the nearest .0001.
  - The RMSE for the first "quiz" subset will be reported publicly on the Site,
  - the RMSE for the second "test" subset will not be reported publicly but will be employed to qualify a submission as described below.

Text from `http:\www.netflixprice.com`.

The winning algorithm: lessons to learn

Bell, Koren and Volinsky (2010): *All Together Now: A Perspective on the Netflix Price*, Chance, 23, p. 24-29.

- Most entries into the competition looked at the problem as a set of algorithms – focus on prediction rather than on understanding what *drives* the preditions.

The winning algorithm: lessons to learn

Bell, Koren and Volinsky (2010): *All Together Now: A Perspective on the Netflix Price*, Chance, 23, p. 24-29.

- Most entries into the competition looked at the problem as a set of algorithms – focus on prediction rather than on understanding what *drives* the preditions.

The winning algorithm: lessons to learn

Bell, Koren and Volinsky (2010): *All Together Now: A Perspective on the Netflix Price*, Chance, 23, p. 24-29.

- Most entries into the competition looked at the problem as a set of algorithms – focus on prediction rather than on understanding what *drives* the preditions.
- Complex models are prone to over fitting or matching small details rather than the big picture, especially where data are scarce (importance of cross-validation).

The winning algorithm: lessons to learn

Bell, Koren and Volinsky (2010): *All Together Now: A Perspective on the Netflix Price*, Chance, 23, p. 24-29.

- Most entries into the competition looked at the problem as a set of algorithms – focus on prediction rather than on understanding what *drives* the preditions.
- Complex models are prone to over fitting or matching small details rather than the big picture, especially where data are scarce (importance of cross-validation).
- The final model is an *ensemble model* combining many different prediction models (at least more than 100), including nearest neighbour methods, latent factor models, neural networks, weighting determined by ridge regression.

The winning algorithm: lessons to learn

Bell, Koren and Volinsky (2010): *All Together Now: A Perspective on the Netflix Price*, Chance, 23, p. 24-29.

- Most entries into the competition looked at the problem as a set of algorithms – focus on prediction rather than on understanding what *drives* the preditions.
- Complex models are prone to over fitting or matching small details rather than the big picture, especially where data are scarce (importance of cross-validation).
- The final model is an *ensemble model* combining many different prediction models (at least more than 100), including nearest neighbour methods, latent factor models, neural networks, weighting determined by ridge regression.
- The winning model was never implemented by Netflix, partly due to implementation issues - but also due to the increase of available data after "sending DVDs by mail" was replaced by online streaming.

The winning algorithm: lessons to learn

Bell, Koren and Volinsky (2010): *All Together Now: A Perspective on the Netflix Price*, Chance, 23, p. 24-29.

- Most entries into the competition looked at the problem as a set of algorithms – focus on prediction rather than on understanding what *drives* the preditions.
- Complex models are prone to over fitting or matching small details rather than the big picture, especially where data are scarce (importance of cross-validation).
- The final model is an *ensemble model* combining many different prediction models (at least more than 100), including nearest neighbour methods, latent factor models, neural networks, weighting determined by ridge regression.
- The winning model was never implemented by Netflix, partly due to implementation issues - but also due to the increase of available data after "sending DVDs by mail" was replaced by online streaming.

The winning algorithm: lessons to learn

Bell, Koren and Volinsky (2010): *All Together Now: A Perspective on the Netflix Price*, Chance, 23, p. 24-29.

- Most entries into the competition looked at the problem as a set of algorithms – focus on prediction rather than on understanding what *drives* the preditions.
- Complex models are prone to over fitting or matching small details rather than the big picture, especially where data are scarce (importance of cross-validation).
- The final model is an *ensemble model* combining many different prediction models (at least more than 100), including nearest neighbour methods, latent factor models, neural networks, weighting determined by ridge regression.
- The winning model was never implemented by Netflix, partly due to implementation issues - but also due to the increase of available data after "sending DVDs by mail" was replaced by online streaming.

Read more: Link to talk with interesting points raised.

TMA4267 Linear statistical methods
Learning outcome, Knowledge

- The student has strong theoretical knowledge about the most popular statistical models and methods that are used in science and technology, with emphasis on regression-type statistical models.

TMA4267 Linear statistical methods
Learning outcome, Skills

- The student knows how to design an experiment and

TMA4267 Linear statistical methods
Learning outcome, Knowledge

- The student has strong theoretical knowledge about the most popular statistical models and methods that are used in science and technology, with emphasis on regression-type statistical models.
- The statistical properties of the multivariate normal distribution are well known to the student, and the student is familiar with the role of the multivariate normal distribution within linear statistical models.

13 / 29

TMA4267 Linear statistical methods
Learning outcome, Skills

- The student knows how to design an experiment and
- how to collect informative data of high quality to study a phenomenon of interest.

14 / 29

TMA4267 Linear statistical methods
Learning outcome, Skills

- The student knows how to design an experiment and
- how to collect informative data of high quality to study a phenomenon of interest.
- Subsequently, the student is able to choose a suitable statistical model,

TMA4267 Linear statistical methods
Learning outcome, Skills

- The student knows how to design an experiment and
- how to collect informative data of high quality to study a phenomenon of interest.
- Subsequently, the student is able to choose a suitable statistical model,
- apply sound statistical methods, and

TMA4267 Linear statistical methods
Learning outcome, Skills

- The student knows how to design an experiment and
- how to collect informative data of high quality to study a phenomenon of interest.
- Subsequently, the student is able to choose a suitable statistical model,
- apply sound statistical methods, and
- perform the analyses using statistical software.

TMA4267 Linear statistical methods
Learning outcome, Skills

- The student knows how to design an experiment and
- how to collect informative data of high quality to study a phenomenon of interest.
- Subsequently, the student is able to choose a suitable statistical model,
- apply sound statistical methods, and
- perform the analyses using statistical software.
- The student knows how to present the results from the statistical analyses, and how to draw conclusions about the phenomenon under study.

TMA4267: Parts

- Part 1: Multivariate RVs and the multivariate normal distribution [week 2-5].
  - Data consists of simultaneous measurements on many variables: we work with *random vectors and random matrices*.
  - There is a strong connection between the *multivariate normal distribution* and the classical linear model.
- Part 2: The classical linear model [week 6-9]
  - We want to understand the relationship between many variables: with focus on linear relationships through the *classical linear model* (multiple linear regression).
- Part 3: Hypothesis tests and analysis of variance [week 9-11]
  - Is there and association between a response and an explanatory variable? Does a response vary between treatment groups?
- Part 4: Design of Experiments [week 12-13+project]
  - If we want to collect data, we need to do know how to *design an experiment*.

Do you know this?

Recommended background: TMA4240/TMA4245 Statistics.

- Probability: (continuous) random variables (RV), probability distribution function (pdf), cumulative distribution function (cdf), mean E, variance Var, covariance Cov, correlation Corr, transformation formula, momentgenerating function (MFG), normal, chi-square and t-distributions.
- Inference: population and sample philosophy, parameter estimation, confidence interval, hypothesis test, p-value, simple linear regression.

Do you know this?

Recommended background: TMA4240/TMA4245 Statistics.

- Probability: (continuous) random variables (RV), probability distribution function (pdf), cumulative distribution function (cdf), mean E, variance Var, covariance Cov, correlation Corr, transformation formula, momentgenerating function (MFG), normal, chi-square and t-distributions.
- Inference: population and sample philosophy, parameter estimation, confidence interval, hypothesis test, p-value, simple linear regression.

Do you know this?

Recommended background: TMA4240/TMA4245 Statistics.

- Probability: (continuous) random variables (RV), probability distribution function (pdf), cumulative distribution function (cdf), mean E, variance Var, covariance Cov, correlation Corr, transformation formula, momentgenerating function (MFG), normal, chi-square and t-distributions.
- Inference: population and sample philosophy, parameter estimation, confidence interval, hypothesis test, p-value, simple linear regression.
- Linear methods: vector and matrix algebra (trace, determinant, eigenvalues/vectors), real vector spaces, orthogonality, spectral decomposition.

TMA4267 Linear Statistical Models
Course information

`https://innsida.ntnu.no/bb`

- Course information.
- Course material.
- Lectures (and handouts).
- Statistical software.
- Exercises (6 recommended and 4 compulsory).
- Exam (80% of portfolio assessment).

Is this the correct course for you?

Are you afraid that this course have a too strong focus on theory and to little on the practical aspects of statistics? You may also look at at the following similar courses (that is, a second course in statistics, with focus on inference)

- ST2304 Statistical modelling for biology/biotechnology: `https://wiki.math.ntnu.no/st2304/`
- TMA4255 Applied statistics, for all siv.ing. studiprograms (except IndMat): `https://wiki.math.ntnu.no/tma4255/`
- KLMED Medical statistics II: `https://www.ntnu.no/studier/emner/KLMED8005`

Is this the correct course for you?

Are you afraid that this course have a too strong focus on theory and to little on the practical aspects of statistics?

Electronic voting

– more than an anonymous show of hands?

For student: check that topics are understood, compare to class, focus on the question asked, while preserving anonymity.

For lecturer: collect data to design sessions that are more contingent.

Software: clicker. math.ntnu.no (single questions), Kahoot! (end-of-lecture sum-up), quiz in Blackboard.

Future studies?

What is your current plan of topic for future studies?

- A: Statistics
- B: Mathematics
- C: Numerics
- D: Other
- E: Don't know

Electronic voting

Use your smart phone, or other devise with internet access and go to **http://clicker.math.ntnu.no/**, and then select TMA4267 as classroom.
Answers

- A: Statistics
- B: Mathematics
- C: Numerics
- D: Other
- E: Don't know

Start voting now!

Part 1: Multivariate random vectors and the multivariate normal distribution

- Härdle and Simar (2015): *Applied Multivariate Statistical Analysis*. Springer.
  - Chapter 2 (p. 53-76): A Short Excursion into Matrix Algebra (partly lectured, manly assumed known).
  - Chapter 3.3 (p. 89-93): Summary statistics.
  - Chapter 4.1-4.5 (p. 117-149): Multivariate Distributions.
  - Chapter 5.1 (p. 183-190): Elementary Properties of the Multinormal.
- Fahrmair, Kneib, Lang and Marx (2013): *Regression*. Springer.
  - Appendix B: Def B.11 (chis q), B.13 (t) , B14 (F), Theorem B.2 and B3.3 (distribution of quadratic forms).

A merged pdf named TMA4267Part1.pdf is available from Bb. Both eBooks and can be downloaded without charge for NTNU students.

The Cork deposit data

- Classical data set from Rao (1948).
- Weigth of bark deposits of $n = 28$ cork trees in $p = 4$ directions (N, E, S, W).

Tree	N	E	S	W
1	72	66	76	77
2	60	53	66	63
3	56	57	64	58
⋮	⋮	⋮	⋮	⋮
28	48	54	57	43

How may we define a random vector in connection to the cork deposit data set?

Hands-on

Let $\boldsymbol{X}_{(2\times 1)}$ have joint pdf (see the 3D-printed figure)

$$f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} \text{ for } -\infty < x_1, x_2 < \infty$$

Find:

1. the marginal distributions $f_1(x_1)$ and $f_2(x_2)$,
2. the conditional distributions $f(x_1 \mid x_2)$ and $f(x_2 \mid x_1)$.
3. What about $F(x_1, x_2)$?

Hint (why?):

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 1$$

Sklar's Theorem [H4.1, p121-122]

Let F be a joint (cumulative) distribution function with marginal distribution functions F_1 and F_2. Then a copula C exists with

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

for every $x_1, x_2 \in \mathbb{R}$. If F_1 and F_2 are continuous, then C is unique. On the other hand, if C is a copula and F_1 and F_2 are (cumulative) distribution functions, then the function F defined above, is a joint distribution function with marginals F_1 and F_2.

Remark: this is not part of the core of the course (not suitable as an exam question), but it is a nice concept and you should have heard about it.

Copula [H4.1, p120]

A two-dimensional copula is a function $C : [0,1]^2 \to [0,1]$ with the following properties:

- For every $u \in [0,1] : C(0,u) = C(u,0) = 0$.
- For every $u \in [0,1] : C(u,1) = u$ and $C(1,u) = u$.
- For every $(u_1, u_2), (v_1, v_2) \in [0,1] \times [0,1]$ with $u_1 \le v_1$ and $u_2 \le v_2$:
$$C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \ge 0$$

(The last property is called "2-increasing".)

Remark: this is not part of the core of the course (not suitable as an exam question), but it is a nice concept and you should have heard about it.

Bivariate Copulas

Farlie-Gumbel-Morgenstern family

$$C(u,v) = uv + \theta uv(1-u)(1-v), \ \theta \in [-1,1]$$

The only copulas that are polynomial quadratic in u and v, symmetric.

Normal (Gaussian) copulas [H.p141]

$$f(x_1, x_2) = \frac{1}{2\pi} \frac{1}{\sigma_1 \sigma_2 \sqrt{1-\rho^2}} e^{-\frac{1}{2}Q(x_1, x_2)}$$

$$Q(x_1, x_2) = \frac{1}{1-\rho^2} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) \right]$$

$$C(u,v) = \int_{-\infty}^{\Phi_1^{-1}(u)} \int_{-\infty}^{\Phi_2^{-1}(v)} f(x_1, x_2) dx_1 dx_2$$

Read more? Properties and applications of copulas: A brief survey, Roger B. Nelsen (And same remark as before.)

More on copulas?

We will later in Part 1 (recommended exercise 2) look at data and contour plots from Gaussian copulas, and other copulas poplar in finance - using R and the `copula` library in R.

What have we worked with today?

- A random vector is ... a vector of random variables.
- Joint distribution function.

What have we worked with today?

- A random vector is ... a vector of random variables.

What have we worked with today?

- A random vector is ... a vector of random variables.
- Joint distribution function.
- From joint distribution function to marginal and conditional distributions.

What have we worked with today?

- A random vector is ... a vector of random variables.
- Joint distribution function.
- From joint distribution function to marginal and conditional distributions.
- Cumulative distribution.

What have we worked with today?

- A random vector is ... a vector of random variables.
- Joint distribution function.
- From joint distribution function to marginal and conditional distributions.
- Cumulative distribution.
- Independence.

29 / 29

What have we worked with today?

- A random vector is ... a vector of random variables.
- Joint distribution function.
- From joint distribution function to marginal and conditional distributions.
- Cumulative distribution.
- Independence.
- From marginal cumulative distribution functions to joint using copula.

29 / 29

What have we worked with today?

- A random vector is ... a vector of random variables.
- Joint distribution function.
- From joint distribution function to marginal and conditional distributions.
- Cumulative distribution.
- Independence.
- From marginal cumulative distribution functions to joint using copula.

29 / 29

What have we worked with today?

- A random vector is ... a vector of random variables.
- Joint distribution function.
- From joint distribution function to marginal and conditional distributions.
- Cumulative distribution.
- Independence.
- From marginal cumulative distribution functions to joint using copula.

Next lecture: Mean vector and covariance matrix. You may want to look into how to define a positive definite matrix, how to define eigenvalues/vectors and results for symmetric matrices.

Part 1 : multivariate random vectors and the multivariate normal distribution [L1] 10.01.2017

Random vector, $f(x)$, $F(x)$

Random vector = vector of random variables (RV)

$\underset{p\times 1}{\mathbf{X}} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$ Ex: height, weight, physical, diet

We will focus on continuous random variables.

1) $\mathbf{X}$ has a joint distribution (density) function (pdf)

$$f(x) = f(x_1, x_2, \dots, x_p)$$

Remember $f(x) \geq 0 \quad \forall x$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_p)\, dx_1 \cdots dx_p = 1$$

2) $\mathbf{X}$ has a joint cumulative distribution function (cdf) $F(x) = P(\mathbf{X} \leq x) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_p \leq x_p)$

$$\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_p} f(t_1, t_2, \dots, t_p)\, dt_1 \cdots dt_p$$

$f(x) \rightarrow F(x) = F(x_1, x_2, \dots, x_p)$

$f(x_1, \dots, x_p) \leftarrow \frac{\partial^p F(x_1, \dots, x_p)}{\partial x_1 \partial x_2 \cdots \partial x_p}$

Let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_A \\ \mathbf{X}_B \end{bmatrix}$

$\mathbf{X}_A \in \mathbb{R}^k$, $\mathbf{X}_B \in \mathbb{R}^{p-k}$ eg. $\mathbf{X}_A = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}$ $\mathbf{X}_B = \begin{bmatrix} X_{k+1} \\ \vdots \\ X_p \end{bmatrix}$

3) $\mathbf{X}_A$ has marginal distribution function

$$f_A(x_A) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x)\, dx_B$$

"integrating out the variables not of interest"

4) and conditional distribution of $\mathbf{X}_B$ given $\mathbf{X}_A$

$$f(x_B \mid x_A) = \frac{f(x_A, x_B)}{f_A(x_A)} \quad \text{for } f_A(x_A) > 0$$

Ex: $f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)}$ $-\infty < x_1, x_2 < \infty$

Hint: $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2}\, dx_1 = 1$

1) Find $f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} dx_2$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2}\, dx_2$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} \cdot \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2} dx_2}_{1}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2}$$

$f_2(x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2}$

2) $f(x_2 \mid x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{\frac{1}{2\pi} e^{-\frac{1}{2}x_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2}}$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2} = f_2(x_2)$$

3) $F(x) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(t_1, t_2)\, dt_1\, dt_2 = \dots$

$F(x_1, x_2)$ not on closed form, use R software

Independence

DEF: X_A and X_B are independent

$$\Updownarrow$$

$$f(x_A, x_B) = f_A(x_A) \cdot f_B(x_B)$$

Also: $f(x_A \mid x_B) = f_A(x_A)$, $f(x_B \mid x_A) = f_B(x_B)$

Ex: $f(x_1, x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2} = f_1(x_1) \cdot f_2(x_2)$

What about $F(x)$, $F_A(x_A)$, $F_B(x_B)$ when X_A and X_B are independent?

$$F(x) = \int_{-\infty}^{x_A} \int_{-\infty}^{x_B} f(t) \, dt_A \, dt_B$$

$$\uparrow$$

$$f_A(t_A) \cdot f_B(t_B)$$

$$= \int_{-\infty}^{x_A} f_A(t_A) \, dt_A \cdot \int_{-\infty}^{x_B} f_B(t_B) \, dt_B = F_A(x_A) \cdot F_B(x_B)$$

From $f(x_1, x_2) \rightarrow \frac{f_2(x_2)}{f_1(x_1)}$ by integrating

How can we get from $f_1(x_1)$ and $f_2(x_2)$ to $f(x_1, x_2)$?

4

$f_1(x_1)$ and $f_2(x_2) \Rightarrow f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$

x_1 and x_2 independent

But what if x_1 and x_2 are not independent?

↓

solution is Copula ⇒ see slides!

5

TMA4267 Linear Statistical Models V2017 [L2]

Part 1: Multivariate RVs, and the multivariate normal distribution
Moments: mean and covariance [H:4.2]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: January 13, 2017

Last lecture

- A random vector $\boldsymbol{X}_{(p\times 1)}$ is ... a p-dimensional vector of random variables.
  - Weight of cork deposits in $p = 4$ directions (N, E, S, W).
  - Rent index in Munich: rent, area, year of construction, location, bath condition, kitchen condition, central heating, district.
- Joint distribution function: $f(\boldsymbol{x})$.
- From joint distribution function to marginal (and conditional distributions).

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_p) dx_2 \cdots dx_p$$

- Cumulative distribution (definite integrals!) used to calculate probabilites.
- Independence: $f(x_1, x_2) = f_1(x_1) \cdot f(x_2)$ and $f(x_1 \mid x_2) = f_1(x_1)$.
- From marginal cumulative distribution functions to joint using copula.

Word cloud: Probability

Today

- Moments: important properties about the distribution of $\boldsymbol{X}$.
- E: Mean of random vector and random matrices.
- Cov: Covariance matrix.
- Corr: Correlation matrix.
- E and Cov of multiple linear combinations.

The Cork deposit data

- Classical data set from Rao (1948).
- Weigth of bark deposits of $n = 28$ cork trees in $p = 4$ directions (N, E, S, W).

Tree	N	E	S	W
1	72	66	76	77
2	60	53	66	63
3	56	57	64	58
⋮	⋮	⋮	⋮	⋮
28	48	54	57	43

How may we define a random vectors and random matrices for cork trees?

The Cork deposit data

Draw a random sample of size $n = 28$ from the population of cork treed and observe a $p = 4$ dimensional random vector for each tree.

$$\boldsymbol{X}_{(28\times 4)} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ \vdots & \vdots & \ddots & \vdots \\ X_{28,1} & X_{28,2} & X_{28,3} & X_{28,4} \end{bmatrix}$$

and $\mathrm{E}(\boldsymbol{X}) = \{\mathrm{E}(X_{ij})\}$.

Random vectors and matrices: rules for means

- Random vector $\boldsymbol{X}_{(p\times 1)}$ with mean vector $\boldsymbol{\mu}_{(p\times 1)}$:

$$\boldsymbol{X}_{(p\times 1)} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \qquad \boldsymbol{\mu}_{(p\times 1)} = \mathrm{E}(\boldsymbol{X}) = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_p) \end{bmatrix}$$

- 1) Random matrix $\boldsymbol{X}_{(n\times p)}$ and random matrix $\boldsymbol{Y}_{(n\times p)}$:

$$\mathrm{E}(\boldsymbol{X}+\boldsymbol{Y}) = \mathrm{E}(\boldsymbol{X}) + \mathrm{E}(\boldsymbol{Y})$$

- 2) Random matrix $\boldsymbol{X}_{(n\times p)}$ and conformable constant matrices $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\mathrm{E}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}\mathrm{E}(\boldsymbol{X})\boldsymbol{B}$$

Hands-on

Let $\boldsymbol{X}_{4\times 1}$ have variance-covariance matrix

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

Explain to your neighbour what this means.

Variance-covariance matrix

- Random vector $\boldsymbol{X}_{(p\times 1)}$ with mean vector $\boldsymbol{\mu}_{(p\times 1)}$:

$$\boldsymbol{X}_{(p\times 1)} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \quad \boldsymbol{\mu}_{(p\times 1)} = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

- Variance-covariance matrix $\boldsymbol{\Sigma}$ (real and symmetric)

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}$$

- $\sigma_{ij} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)]$

Correlation matrix

Correlation matrix $\boldsymbol{\rho}$ (real and symmetric)

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}\sigma_{pp}}} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$$

$$\boldsymbol{\rho} = (\boldsymbol{V}^{\frac{1}{2}})^{-1}\boldsymbol{\Sigma}(\boldsymbol{V}^{\frac{1}{2}})^{-1}, \text{ where } \boldsymbol{V}^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Hands-on

Let $\boldsymbol{X}_{4\times 1}$ have variance-covariance matrix

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

Find the correlation matrix.

Hands-on: Focus on $\boldsymbol{C}$?

$$\boldsymbol{X} = \begin{bmatrix} X_N \\ X_E \\ X_S \\ X_W \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_N \\ \mu_E \\ \mu_S \\ \mu_W \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{NN} & \sigma_{NE} & \sigma_{NS} & \sigma_{NW} \\ \sigma_{NE} & \sigma_{EE} & \sigma_{ES} & \sigma_{EW} \\ \sigma_{NS} & \sigma_{EE} & \sigma_{SS} & \sigma_{SW} \\ \sigma_{NW} & \sigma_{EW} & \sigma_{SW} & \sigma_{WW} \end{bmatrix}$$

- Scientists would like to compare the following three contrasts: N-S, E-W and (E+W)-(N+S),
- and define a new random vector $\boldsymbol{Y}_{(3\times 1)} = \boldsymbol{C}_{(3\times 4)}\boldsymbol{X}_{(4\times 1)}$ giving the three contrasts.
- Write down $\boldsymbol{C}$.
- Use the formulas we just developed and *explain* how to find $\mathrm{E}(Y_1)$ and $\mathrm{Cov}(Y_1, Y_3)$.

Linear combinations

- Random vector $\boldsymbol{X}_{(p\times 1)}$ with mean vector $\boldsymbol{\mu}_{\boldsymbol{X}} = \mathrm{E}(\boldsymbol{X})$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathrm{Cov}(\boldsymbol{X})$.
- The linear combinations $\boldsymbol{Z} = \boldsymbol{C}\boldsymbol{X}$ have

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{Z}} &= \mathrm{E}(\boldsymbol{Z}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{X}) = \boldsymbol{C}\boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Z}} &= \mathrm{Cov}(\boldsymbol{Z}) = \mathrm{Cov}(\boldsymbol{C}\boldsymbol{X}) = \boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{C}^T \end{aligned}$$

Exam V2014: Problem 1a

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ be a random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Further, let $\boldsymbol{A} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$ be a matrix of constants.

Define $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \boldsymbol{A}\boldsymbol{X}$.

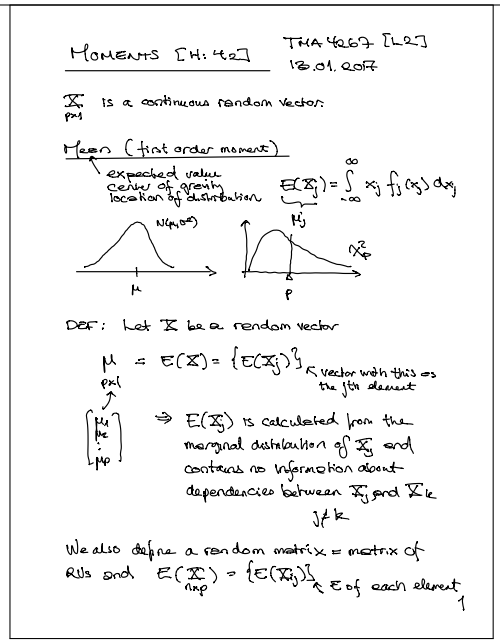
Find $\mathrm{E}(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{Y})$.
Are X_1 and X_2 independent?
Are Y_1 and Y_2 independent? Justify your answers.
Find the mean of $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$.

The covariance matrix

Random vector $\boldsymbol{X}_{(p\times 1)}$ with mean vector $\boldsymbol{\mu}_{(p\times 1)}$ and covariance matrix

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}$$

The covariance matrix is by construction symmetric, and we would only consider covariance matrices that are positive definite (PD).
Why would we only consider PD matrices?
Homework for next lecture: Read H.Chapter 2.1-2.2 to remind yourself of spectral decomposition (diagonalization), positive definite matrix, eigenvalues and eigenvectors.



MOMENTS [H: 4.2] TMA4267 [L2] 18.01.2017

$\underset{p\times 1}{\boldsymbol{X}}$ is a continuous random vector:

Mean (first order moment)

expected value, center of gravity, location of distribution

$E(X_j) = \int_{-\infty}^{\infty} x_j f_j(x_j)\,dx_j$

$N(\mu,\sigma^2)$ μ_j χ^2_p μ p

DEF: Let $\boldsymbol{X}$ be a random vector

$\underset{p\times 1}{\boldsymbol{\mu}} = E(\boldsymbol{X}) = \{E(X_j)\}_p$ ← vector with this as the jth element

$\Rightarrow$ $E(X_j)$ is calculated from the marginal distribution of X_j and contains no information about dependencies between X_j and X_k $j \neq k$

We also define a random matrix = matrix of RVs and $E(\underset{n\times p}{\boldsymbol{X}}) = \{E(X_{ij})\}$ ← E of each element

1

What have we worked with today?

- Mean: $\boldsymbol{\mu}_X = E(\boldsymbol{X}) = \mathrm{E}(X_j)$
- Covariance: $\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{E}((\boldsymbol{X}-\boldsymbol{\mu}_X)(\boldsymbol{Y}-\boldsymbol{\mu}_Y)^T)$.
- Variance-covariance: $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}((\boldsymbol{X}-\boldsymbol{\mu}_X)(\boldsymbol{X}-\boldsymbol{\mu}_X)^T)$, also sometimes denoted $\mathrm{Var}(\boldsymbol{X})$.
- Correlation: $\mathrm{Corr}(\boldsymbol{X}) = \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{V}^{-\frac{1}{2}}$.
- $\boldsymbol{CX}$: $\mathrm{E}(\boldsymbol{CX}) = \boldsymbol{C}\boldsymbol{\mu}_X$ and $\mathrm{Cov}(\boldsymbol{CX}) = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T$.

Next lecture: First work with the covariance matrix and positive definiteness, then start with the multivariate normal distribution (where we use moment generating functions and a multivariate version of the transformation formula).

Remember: X scalar $\Rightarrow$ $E(aX+b) = \int_{-\infty}^{\infty}(ax+b)f(x)\,dx$

$= a\int_{-\infty}^{\infty} x f(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx$

$= a\,E(X) + b$

RULES FOR MEAN

$\underset{n\times p}{\boldsymbol{X}}$ and $\underset{n\times p}{\boldsymbol{Y}}$ are random matrices (n may be 1)

and $\underset{m\times n}{\boldsymbol{A}}$ and $\underset{p\times q}{\boldsymbol{B}}$ are constant matrices

1) $E(\boldsymbol{X}+\boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y})$

Proof: Look at element $Z_{ij} = X_{ij} + Y_{ij}$ and see that $E(Z_{ij}) = E(X_{ij}+Y_{ij}) = E(X_{ij}) + E(Y_{ij})$

2) $E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}\,E(\boldsymbol{X})\,\boldsymbol{B}$

Proof: look at element (i,j) of $\underbrace{\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}}_{m\times q}$

$e_{ij} = \sum_{k=1}^{n} a_{ik} \sum_{l=1}^{p} X_{kl} b_{lj}$

and see that $E(e_{ij})$ is the element (i,j) of $\boldsymbol{A}\,E(\boldsymbol{X})\,\boldsymbol{B}$.

2

Covariance matrix [H4.2]

① From TMA4240/42 & ST1101: X and Y are scalar RVs with $E(X)=\mu_X$ and $E(Y)=\mu_Y$.

a) $\mathrm{Cov}(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big) \overset{c)}{=} E(X\cdot Y) - \mu_X\mu_Y$

Remember: $\mathrm{Cov}(X,Y) > 0$: high X and high Y; < 0: high X and low Y; $= 0$: no linear pattern

b) $\mathrm{Var}(X) = \mathrm{Cov}(X,X) = E(X^2) - \mu_X^2$

d) $\mathrm{Cov}(X,Y) = 0$ if X and Y are independent, but not necessarily the opposite

② Now: random vectors

Ex: Chemical process; X = temp, pressure; Y = yield, quality

a) DEF [H4.16]

$\underset{p\times 1}{X}$ and $\underset{q\times 1}{Y}$ are two random vectors.

$E(X) = \mu_X$ and $E(Y) = \mu_Y$

$$\mathrm{Cov}(X,Y) = E\Big[\underbrace{(X-\mu_X)}_{p\times 1}\underbrace{(Y-\mu_Y)^T}_{1\times q}\Big] \quad (p\times q)$$

$$= \Sigma_{XY} \quad \text{(Sigma)}$$

and $\mathrm{Cov}(X,Y) = 0$ if X and Y are independent.

If $p=2$, $q=3$:

$$\mathrm{Cov}(X,Y) = \begin{bmatrix} \mathrm{Cov}(X_1,Y_1) & \mathrm{Cov}(X_1,Y_2) & \mathrm{Cov}(X_1,Y_3) \\ \mathrm{Cov}(X_2,Y_1) & \mathrm{Cov}(X_2,Y_2) & \mathrm{Cov}(X_2,Y_3) \end{bmatrix}$$

Not a square matrix, but $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)^T$

b) $\underset{p\times p}{\mathrm{Cov}(X,X)} = E\big[(X-\mu_X)(X-\mu_X)^T\big]$

$$= \begin{bmatrix} \underbrace{E[(X_1-\mu_1)(X_1-\mu_1)]}_{\mathrm{Var}(X_1)} & \cdots & \underbrace{E[(X_1-\mu_1)(X_p-\mu_p)]}_{\mathrm{Cov}(X_1,X_p)} \\ \vdots & \ddots & \\ \underbrace{E[(X_p-\mu_p)(X_1-\mu_1)]}_{\mathrm{Cov}(X_p,X_1)} & & \underbrace{E[(X_p-\mu_p)(X_p-\mu_p)]}_{\mathrm{Var}(X_p)} \end{bmatrix}$$

$= \Sigma = \Sigma_{XX}$

$= \underline{\mathrm{Cov}(X)}$ or $\mathrm{Var}(X)$

= "variance - covariance matrix"

Example: 4×4 matrix

Observations: $\mathrm{Cov}(X_1,X_2) = \mathrm{Cov}(X_2,X_1)$

$= E[(X_1-\mu_1)(X_2-\mu_2)]$

$= \int_{-\infty}^{\infty}\int (x_1-\mu_1)(x_2-\mu_2) f(x_1,x_2)\,dx_1\,dx_2$

Σ is a symmetric matrix.

Cont.

c) $\mathrm{Cov}(X,Y) = E(X\cdot Y^T) - \mu_X\cdot\mu_Y^T$ (↑ Rec Ex1. P3)

$$\underset{p\times p}{\mathrm{Cov}(X)} = E\big(\underset{p\times 1}{X}\underset{1\times p}{X^T}\big) - \underset{p\times 1}{\mu_X}\underset{1\times p}{\mu_X^T}$$

d) As before if X_i and Y_j are independent $\Rightarrow \mathrm{Cov}(X_i,Y_j) = 0$.

And if all pairs of Xs and Ys are independent ($\forall i,j$) $\Rightarrow \mathrm{Cov}(X,Y) = 0$

Correlation

X_1 and X_2: $\mathrm{Corr}(X_1,X_2) = \dfrac{\mathrm{Cov}(X_1,X_2)}{\sqrt{\mathrm{Var}(X_1)\cdot\mathrm{Var}(X_2)}}$

Rec Ex1. P1i do this in R

See matrix formula on slides

Ex: Given Σ find ρ:

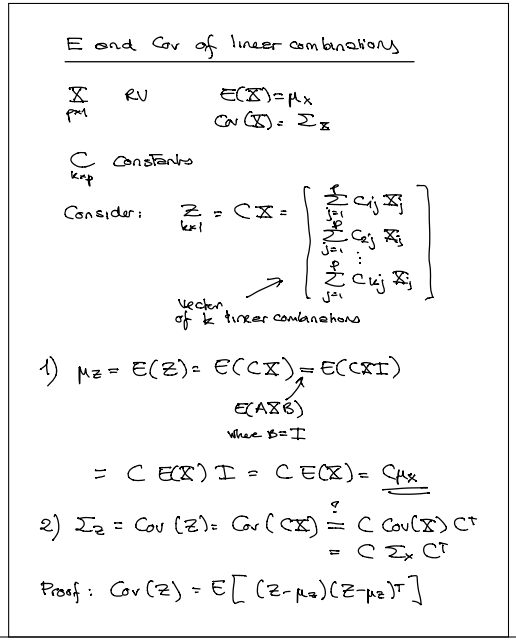
$$\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix} \to \begin{bmatrix} 1 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}$$

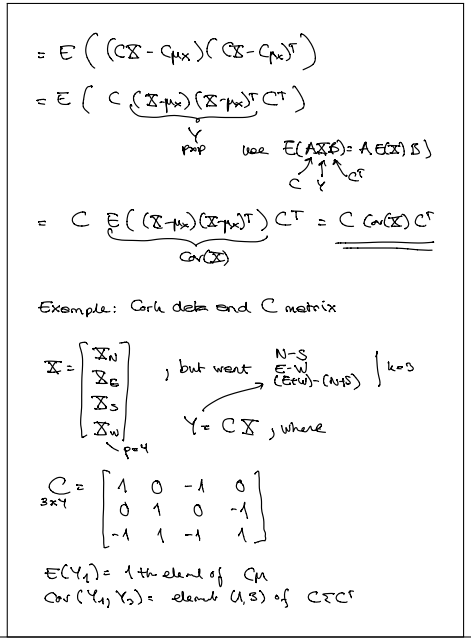
Q: can information in Σ be condensed?

$\mathrm{tr}(\Sigma) = \sum_{i=1}^{p}\mathrm{Var}(X_i)$ = "total variance"

(↑ trace = sum of diagonal elements)

$\det(\Sigma)$ = "generalized variance"





TMA4267 Linear Statistical Models V2017 [L3]

Part 1: Multivariate RVs and normal distribution (L3)
Covariance and positive definiteness [H:2.2,2.3,3.3],
Principal components [H11.1-11.3]
Quiz with Kahoot!

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: January 17, 2017

Previously

- $\boldsymbol{X}_{(p\times 1)}$: random vector, described by
- joint probability distribution function (pdf) $f(\boldsymbol{x})$ and cumulative distribution function (cdf) $F(\boldsymbol{x})$, or, as we will see (in L4), by the (multivariate) moment generating function (MGF) $M_{\boldsymbol{X}}(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{X}}$.
- Important aspects: moments.
- $\mathrm{E}(\boldsymbol{X})$: mean of a random vector (or matrix) is found as the mean of each element.
- $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}((\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T)$: $p \times p$ variance-covariance matrix, with variances on the diagonal and covariance off-diagonal, real, symmetric.
- Rules for vector of linear combinations $\boldsymbol{CX}$: $\mathrm{E}(\boldsymbol{CX}) = \boldsymbol{C\mu}$ and $\mathrm{Cov}(\boldsymbol{CX}) = \boldsymbol{C\Sigma C}^T$.

Today!

- Requirements and properties of $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$: symmetric, positive definite (SPD), via spectral decomposition (eigenvalues/eigenvectors).
- The square root matrix.
- Linear combinations with maximal variability: principal components are linear combinations made from eigenvectors.
- PCA-plots.
- Kahoot! on what we have worked with so far.
- Next lecture: move on to multivariate normal data!

Drinking habits data set

	Coffee	Tea	Cocoa	Liquer	Wine	Beer
Norway	9.800000	0.21	0.61	1.1	6.4	52.0
Danmark	10.400001	0.39	0.54	1.4	20.7	123.2
Finland	12.450000	0.17	0.03	3.1	5.4	79.0
Iceland	8.270001	0.23	0.00	2.2	5.2	23.7
Sweden	10.710000	0.32	0.16	1.8	12.3	57.4
France	5.490000	0.20	1.18	2.5	73.8	40.5
Ireland	0.550000	3.14	2.76	1.4	3.9	114.0
Italy	4.670000	0.08	0.97	1.0	67.0	22.7
Jugoslavia	3.100000	0.11	0.59	1.6	20.3	46.8
The Netherlands	10.970000	0.82	15.35	2.0	14.7	87.0
Poland	1.400000	0.54	0.56	4.3	7.6	30.8
Portugal	3.080000	0.03	0.02	0.8	51.5	60.7
Soviet Union	0.300000	0.98	0.56	1.9	6.6	19.3
Spain	4.250000	0.03	1.11	2.8	38.3	70.7
Schweitz	9.400000	0.25	3.06	1.9	49.6	69.3
Great Britain	2.060000	2.62	2.78	1.8	11.5	110.5
Chech Repl	2.200000	0.13	1.21	3.3	13.6	132.9
Germany	8.970000	0.22	3.94	2.0	26.0	143.0
Hungary	2.270000	0.07	0.85	4.6	21.5	103.9
Austria	10.220000	0.16	1.80	1.5	34.8	119.5
New Zealand	1.920000	1.46	0.03	1.4	14.6	114.2

The variance-covariance matrix and positive definiteness

$\boldsymbol{X}_{(p\times 1)}$ with symmetric $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$

- Want variance of linear combination to be positive: $\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}$ for all $\boldsymbol{c} \neq \boldsymbol{0}$,
- which means that $\boldsymbol{\Sigma}$ needs to positive definite. Write $\boldsymbol{\Sigma} > 0$.
- This is true if all eigenvalues of $\boldsymbol{\Sigma}$ are positive (eigenvalues of symmetric matrix are real).
- Spectral decompositions (diagonalization): $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$.
- Square root matrix defined from spectral decomposition: $\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^T$.

Drinking habits

	Coffee	Tea	Cocoa	Liquer	Wine	Beer
Norway	9.800000	0.21	0.61	1.1	6.4	52.0
Danmark	10.400001	0.39	0.54	1.4	20.7	123.2
Finland	12.450000	0.17	0.03	3.1	5.4	79.0
Iceland	8.270001	0.23	0.00	2.2	5.2	23.7
Sweden	10.710000	0.32	0.16	1.8	12.3	57.4
France	5.490000	0.20	1.18	2.5	73.8	40.5
Ireland	0.550000	3.14	2.76	1.4	3.9	114.0
Italy	4.670000	0.08	0.97	1.0	67.0	22.7
Jugoslavia	3.100000	0.11	0.59	1.6	20.3	46.8
The Netherlands	10.970000	0.82	15.35	2.0	14.7	87.0
Poland	1.400000	0.54	0.56	4.3	7.6	30.8
Portugal	3.080000	0.03	0.02	0.8	51.5	60.7
Soviet Union	0.300000	0.98	0.56	1.9	6.6	19.3
Spain	4.250000	0.03	1.11	2.8	38.3	70.7
Schweitz	9.400000	0.25	3.06	1.9	49.6	69.3
Great Britain	2.060000	2.62	2.78	1.8	11.5	110.5
Chech Repl	2.200000	0.13	1.21	3.3	13.6	132.9
Germany	8.970000	0.22	3.94	2.0	26.0	143.0
Hungary	2.270000	0.07	0.85	4.6	21.5	103.9
Austria	10.220000	0.16	1.80	1.5	34.8	119.5
New Zealand	1.920000	1.46	0.03	1.4	14.6	114.2

Estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ [H3.3]

$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ i.i.d $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X})$.

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j$$

$$\boldsymbol{S}^2 = \frac{1}{n-1}\sum_{j=1}^{n} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^T$$

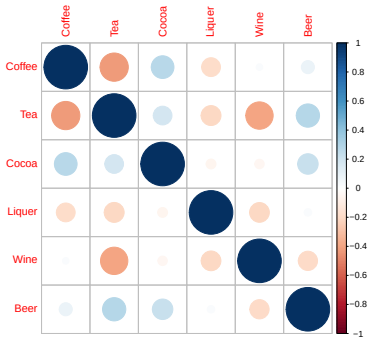
are two commonly used estimators for the mean and covariance matrix.

RecEx1.P7: we may write $\boldsymbol{S}^2$ using a centering matrix.

Drinking habits data set

```
> drinkcov
           Coffee        Tea      Cocoa     Liquer       Wine       Beer
Coffee 16.5471497 -1.4476576  3.6381233 -0.7651859   1.841446  14.386113
Tea    -1.4476576  0.7168891  0.5154248 -0.1786857  -6.939988   9.503595
Cocoa   3.6381233  0.5154248 10.8676967 -0.1700214  -3.191860  28.673374
Liquer -0.7651859 -0.1786857 -0.1700214  1.0222857  -4.307429   1.142071
Wine    1.8414458 -6.9399885 -3.1918596 -4.3074288 432.878505 -159.797675
Beer   14.3861129  9.5035954 28.6733741  1.1420707 -159.797675 1549.731430
```


Correlation plot



prcomp

```
drink <- read.csv("drikke.TXT",sep=",",header=TRUE)
drink <- na.omit(drink) # remove missing data

# now for PCA
pca <- prcomp(drink,scale=TRUE) # scale: variables are scaled, and automatically center=TRUE

> names(pca)
[1] "sdev"     "rotation" "center"   "scale"    "x"

> pca$rotation # the loadings
              PC1         PC2         PC3          PC4         PC5         PC6
Coffee -0.26029733  0.66788815 -0.22475187  0.4132467433  0.07431918  0.5092751
Tea     0.65540048 -0.09539757  0.36756357 -0.0002927055 -0.12503940  0.6407898
Cocoa   0.23510209  0.57754726 -0.06603093 -0.4200858712 -0.61199325 -0.2362164
Liquer  0.02190508 -0.32118904 -0.79997824 -0.3292322714 -0.12307455  0.3644878
Wine   -0.50599685  0.06551597  0.37109534 -0.6765579799  0.15862233  0.3450672
Beer    0.43693234  0.32219426 -0.17985159 -0.2943302832  0.75099779 -0.1493533

> s <- cor(drink) # cor, not cov, since covariates are scaled
> eigen(s) # same as pca$rotations - opposite sign of some vectors
$values
[1] 1.7204307 1.4295795 1.1408597 0.7731249 0.7354586 0.2005467
$vectors
             [,1]        [,2]        [,3]          [,4]         [,5]        [,6]
[1,]  0.26029733  0.66788815 -0.22475187  0.4132467433  0.07431918 -0.5092751
[2,] -0.65540048 -0.09539757  0.36756357 -0.0002927055 -0.12503940 -0.6407898
[3,] -0.23510209  0.57754726 -0.06603093 -0.4200858712 -0.61199325  0.2362164
[4,] -0.02190508 -0.32118904 -0.79997824 -0.3292322714 -0.12307455 -0.3644878
[5,]  0.50599685  0.06551597  0.37109534 -0.6765579799  0.15862233 -0.3450672
[6,] -0.43693234  0.32219426 -0.17985159 -0.2943302832  0.75099779  0.1493533
```

Principal components

- Let $\mathbf{\Sigma}$ be the covariance matrix associated with the random vector $\boldsymbol{X}_{p\times 1}$. The covariance matrix has the eigenvalue-vector pairs $(\lambda_j, \boldsymbol{e}_j)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
- The mth principal component is given by

$$Y_m = \boldsymbol{e}_m^T \boldsymbol{X} = e_{m1}X_1 + e_{m2}X_2 + \cdots + e_{mp}X_p$$

- and has

$$\begin{aligned} \mathrm{Var}(Y_m) &= \boldsymbol{e}_m^T \mathbf{\Sigma} \boldsymbol{e}_m = \lambda_m, \quad i = 1, 2, ..., p \\ \mathrm{Cov}(Y_i, Y_m) &= \boldsymbol{e}_i^T \mathbf{\Sigma} \boldsymbol{e}_m = 0 \quad i \neq m \end{aligned}$$

Principal components: idea

1. We choose principal component 1, $\mathrm{PC}_1 = \boldsymbol{c}_1^T \boldsymbol{X}$, to have maximal variance

$$\max_{\boldsymbol{c}_1 \neq 0, \boldsymbol{c}_1^T \boldsymbol{c}_1 = 1} \mathrm{Var}(\boldsymbol{c}_1^T \boldsymbol{X})$$

2. We choose principal component 2, $\mathrm{PC}_2 = \boldsymbol{c}_2^T \boldsymbol{X}$, to have maximal variance and to be uncorrelated with PC_1.

$$\max_{\boldsymbol{c}_2 \neq 0, \boldsymbol{c}_2^T \boldsymbol{c}_2 = 1} \mathrm{Var}(\boldsymbol{c}_2^T \boldsymbol{X}) \text{ and } \boldsymbol{c}_1^T \mathbf{\Sigma} \boldsymbol{c}_2 = 0$$

Principal components: idea

3. We choose principal component 3, $\mathrm{PC}_1 = \boldsymbol{c}_3^T \boldsymbol{X}$, to have maximal variance and be uncorrelated with PC_1 and PC_2.

$$\max_{\boldsymbol{c}_3 \neq 0, \boldsymbol{c}_3^T \boldsymbol{c}_3 = 1} \mathrm{Var}(\boldsymbol{c}_3^T \boldsymbol{X}) \text{ and } \boldsymbol{c}_i^T \mathbf{\Sigma} \boldsymbol{c}_3 = 0$$

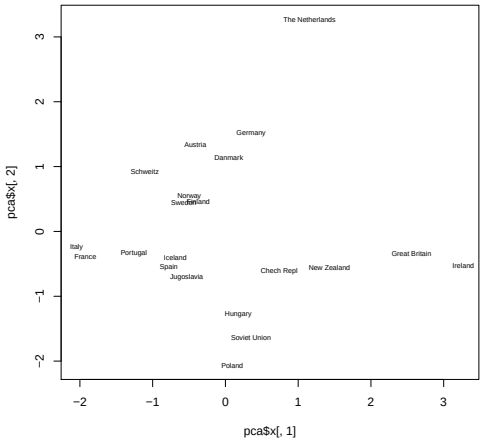
   for $i = 1, 2$.
4. and so on.

It can be shown that choosing $\boldsymbol{c}_i = \boldsymbol{e}_i$ (ith eigenvector of $\mathbf{\Sigma}$) fulfills these requirements.

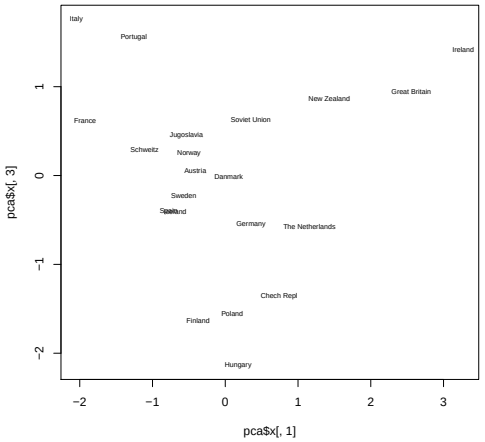
Principal component scores

```
> pca$x
                    PC1        PC2        PC3        PC4          PC5          PC6
Norway        -0.49755266  0.5423321  0.24730564  1.64969605 -0.141194597 -0.259224773
Danmark        0.04736942  1.1407629 -0.01389387  0.62447776  1.286968641  0.032325551
Finland       -0.37212624  0.4628190 -1.62846743  1.17223112  0.284805095  0.685797001
Iceland       -0.68875794 -0.4028320 -0.40970218  1.46441365 -0.741785804  0.092078151
Sweden        -0.57111547  0.4409190 -0.21949104  1.33928137  0.005514957  0.299908368
France        -1.92532968 -0.3909512  0.61668510 -1.42238114 -0.200223837  0.818931395
Ireland        3.27114811 -0.5256052  1.42206819 -0.04493512 -0.014828505  0.477585062
Italy         -2.04540738 -0.2394487  1.76095443 -0.63624103 -0.367324127  0.054512802
Jugoslavia    -0.53258860 -0.7075896  0.45055250  0.39569929 -0.299213104 -0.741778021
The Netherlands 1.15707155 3.2747502 -0.57171870 -0.93460573 -2.324855354 -0.377956367
Poland         0.09667998 -2.0690624 -1.55793948 -0.11990952 -1.118921154  0.196370232
Portugal      -1.25572594 -0.3355743  1.55429933 -0.39165369  0.418471226 -0.587655889
Soviet Union   0.35234451 -1.6341096  0.62745644  0.66818676 -1.138865302 -0.446485485
Spain         -0.77751355 -0.5475718 -0.39572230 -0.70847071  0.084179130 -0.055157063
Schweitz      -1.10755090  0.9248648  0.29621598 -0.49777226 -0.047226891  0.484657020
Great Britain  2.55842016 -0.3473694  0.94755863 -0.24512125  0.028322241  0.555177625
Chech Repl     0.74010278 -0.6054459 -1.36137719 -0.75417409  0.950808503 -0.708681733
Germany        0.35362098  1.5089269 -0.54743939 -0.46975244  0.999859480 -0.289909548
Hungary        0.17766893 -1.2757040 -2.13925810 -1.16454698  0.376532314 -0.009844322
Austria       -0.41110347  1.3402182  0.06024028 -0.01769055  1.108477110  0.029376612
New Zealand    1.43034542 -0.5543289  0.86167320  0.09326849  0.850499977 -0.250026615
```

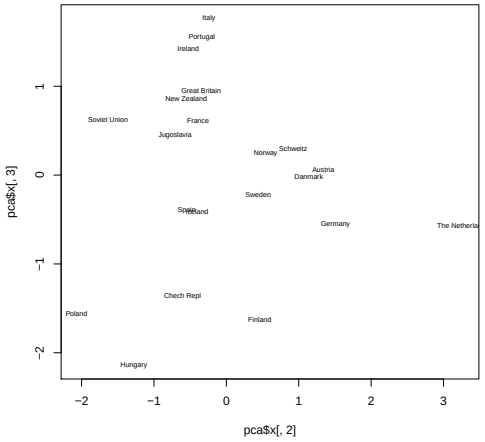
Scores PCA1 vs PCA2



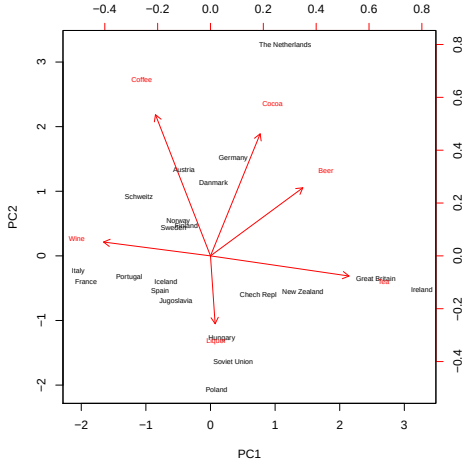
Scores PCA1 vs PCA3



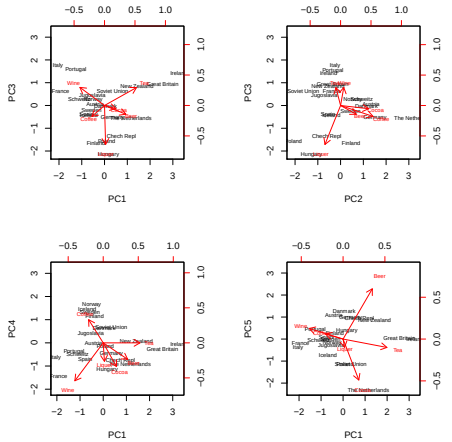
Scores PCA2 vs PCA3



Scores PCA1 vs PCA2 with biplot



Biplots



Proportion of total population variance

- Total population variance:
  $\sum_{j=1}^{p} \mathrm{Var}(X_j) = \mathrm{tr}\mathbf{\Sigma} = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \mathrm{Var}(Z_j)$.
- Proportion of total population variance explained by PC m:

$$\frac{\lambda_m}{\sum_{j=1}^{p} \lambda_j}$$

- Proportion of total population variance explained by the first m PCs:

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

How many PCs are needed?

Dependent on:

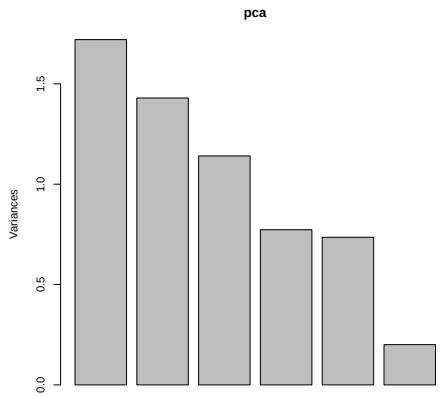
- The proportion of the total sample variance that we would like to explain. 80%? More?
- Look at the eigenvalues; small eigenvalues may be an evidence of collinearity problems.

Importance of components

```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6
Standard deviation     1.3117 1.1957 1.0681 0.8793 0.8576 0.44782
Proportion of Variance 0.2867 0.2383 0.1901 0.1288 0.1226 0.03342
Cumulative Proportion  0.2867 0.5250 0.7151 0.8440 0.9666 1.00000

> eigen(s)
$values
[1] 1.7204307 1.4295795 1.1408597 0.7731249 0.7354586 0.2005467
> sqrt(eigen(s)$values)
[1] 1.3116519 1.1956502 1.0681103 0.8792752 0.8575888 0.4478244
```

Screeplot



Principal components from singular value decomposition

The singular value decomposistion of a (data) matrix $\boldsymbol{X}_{n\times p}$ is given by:

$$\boldsymbol{X}_{n\times p} = \boldsymbol{U}_{n\times p}\boldsymbol{D}_{p\times p}\boldsymbol{V}^T_{p\times p}$$

where

- the columns of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$
- $\boldsymbol{D}$ is a diagonal matrix with singular values on the diagonal, i.e. the square root of the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$ (they have the same eigenvalues).
- the columns of $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$.

And, the principal components (scores) of the data are defined as the columns of

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}$$

PC from standardized variables

- $\boldsymbol{X}$ can be standardized to have mean $\boldsymbol{0}$ and unit variances.

$$\boldsymbol{X}^* = \boldsymbol{V}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})$$

- Principal components made from standardized variables will be based on the eigenvalues and eigenvectors of the correlation matrix $\boldsymbol{\rho} = \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{V}^{-\frac{1}{2}}$.
- Achilles heel: Since $\boldsymbol{\Sigma}$ and $\boldsymbol{\rho}$ do not have the same eigenvectors/eigenvalues, the principal components made from $\boldsymbol{\Sigma}$ and $\boldsymbol{\rho}$ will not be the same.
- Unless we have a good reason to compare the variances for the different X_js we should make PCs from the standardized variables.
- For standardized variables $\sum_{j=1}^{p} \mathrm{Var}(X_j^*) = p$, and
- Proportion of total population variance explained by PC m: $\frac{\lambda_m}{p}$.

PCR: summary

- PCA finds linear combinations $\boldsymbol{Y}$ that "best" represents the $\boldsymbol{X}$.
- The PCs are found in an unsupervised way. The "truth" is not known.
- A plot of PC1 vs PC2 is often used to see if there is separation (subgroups in the data).
- The principal component loadings are often given interpretation (overall consumption,
- PCA can be combined with linear regression.

Quiz

with Kahoot! at kahoot.it. - based on what we have gone through so far!

The covariance matrix [ch2.2, H:2.2] TMA4267 L3 17.01.2017

$\underset{p\times 1}{X}$ random vector $E(X) = \underset{p\times 1}{\mu}$

$Cov(X) = \underset{p\times p}{\Sigma} = E\left[(X-\mu)(X-\mu)^T\right]$

$$= \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_p) \\ Cov(X_1,X_2) & Var(X_2) & & \\ Cov(X_1,X_p) & \cdots & & Var(X_p) \end{bmatrix}$$

Ex: Drinking habits X_1 = coffee, X_2 = tea, X_3 = cocoa, X_4 = liquer, X_5 = wine, X_6 = beer (consumption of).

i.i.d $n = 21$ countries.

Σ is real and symmetric. Other requirements? We will need Σ^{-1} so we require that Σ is invertible, that $\det(\Sigma) \neq 0$ (or we might use generalized inverse)

But, we looked at $\underset{k\times p}{C}\underset{p\times 1}{X}$ and found $Cov(CX) = C\Sigma C^T$.

If $k=1$ then $\underbrace{C^T}_{1\times p}\underbrace{X}_{p\times 1}$ and $Cov(C^TX) = \underbrace{Var(C^TX)}_{\text{scalar}} = \underbrace{C^T\Sigma C}_{\text{scalar}}$

We want $Var(C^TX) = C^T\Sigma C$ to be positive (because we don't want 0 or negative variances).

So, what do we need?

$\underset{1\times p}{C^T}\underset{p\times p}{\Sigma}\underset{p\times 1}{C} > 0$ for all $C \neq 0$

this is the definition of a S P D matrix.

Symmetric, positive, definite

Useful result:

If Σ is SPD $\iff$ all $\lambda_i > 0$ (λ: eigenvalues of Σ)

Proof via spectral decomposition:

Let $P\Lambda P^T = \Sigma$ where $P = [e_1\ e_2\ \cdots\ e_p]$ is a matrix with the normalized eigenvectors of Σ as column vectors and $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_p)$ is a diagonal matrix with the eigenvalues of Σ on the diagonal.

Remember: (λ_i, e_i) satisfy $\Sigma e_i = \lambda_i e_i$ and use $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Aha that a (real) symmetric matrix has real eigenvalues (and eigenvectors of distinct eigenvalues are orthogonal)

$\Leftarrow$

$\underset{1\times p}{C^T}\underset{p\times p}{P}\Lambda P^T C > 0$ for all $C \neq 0$ (P is invertible)

$\underset{1\times p}{y^T}\underset{p\times p}{\Lambda}\underset{p\times 1}{y} > 0$ for all $y \neq 0$ since $P^TC = y$

$\sum_{i=1}^{p} \lambda_i \underbrace{y_i^2}_{>0} > 0$ This is true when all λ_i are positive.

$\Rightarrow$

If $y = \begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} \rightarrow \lambda_1$, same for all the eigenvalues

QED

This also implies Recall 1.P4

$\det(\Sigma) = \prod_{i=1}^{p}\lambda_i > 0$

$tr(\Sigma) = \sum_{i=1}^{p}\lambda_i > 0$

Homework: why is $\det(\Sigma) = \prod\lambda_i$ and $tr(\Sigma) = \sum\lambda_i$?

What about Σ^{-1}

If Σ is SPD then Σ^{-1} is also SPD, and the eigenvalues of Σ^{-1} are the inverses of the eigenvalues of Σ.

Proof: $\Sigma^{-1} = (P\Lambda P^T)^{-1} = (P^T)^{-1}\Lambda^{-1}P^{-1}$

$(P^T)^{-1} = P$ and $P^{-1} = P^T$ for orthogonal matrix

$$\Sigma^{-1} = P\Lambda^{-1}P^T = P\begin{bmatrix}\frac{1}{\lambda_1} & & 0\\ & \frac{1}{\lambda_2} & \\ 0 & & \ddots & \frac{1}{\lambda_p}\end{bmatrix}P^T$$

eigenvectors; eigenvalues of Σ^{-1}

Finally: $\Sigma^{-\frac{1}{2}} = P\Lambda^{-\frac{1}{2}}P^T$

$$[e_1 \cdots e_p]\begin{bmatrix}\frac{1}{\sqrt{\lambda_1}} & & 0\\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\lambda_p}}\end{bmatrix}$$

Recall: $\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}} = \Sigma$

$\Sigma^{-\frac{1}{2}} = (\Sigma^{\frac{1}{2}})^{-1} = P\Lambda^{-\frac{1}{2}}P^T$

$\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}} = I$ (show)

$$\begin{bmatrix}\frac{1}{\sqrt{\lambda_1}} & & 0\\ & \frac{1}{\sqrt{\lambda_2}} & \\ 0 & & \ddots & \frac{1}{\sqrt{\lambda_p}}\end{bmatrix}$$

The transformation $Y = \Sigma^{-\frac{1}{2}}(X-\mu)$

Mahlenobis transform

Principal Component Analysis [Wild. 11.3]

why is Σ important and what can Σ be?

$\underset{p\times 1}{X}$ RV and $E(X) = \mu$, $Cov(X) = \Sigma$

$\underset{p\times 1}{X_1}, \underset{p\times 1}{X_2}, \ldots, \underset{p\times 1}{X_n}$ i.i.d (identically distributed, independent)

$\mu = E(X_j)$ and $\Sigma = Cov(X_j)$ $j = 1, \ldots, n$.

Estimator for μ and Σ?

$\underset{\sim}{\hat{\mu}} = \frac{1}{n}\sum_{j=1}^{n} X_j = \bar{X}$ p-dim vector

$Cov(X) = E\left((X-\mu)(X-\mu)^T\right) = \Sigma$

$\underset{p\times p}{\underline{\underline{S^2}}} = \frac{1}{n-1}\sum_{j=1}^{n}\underset{p\times 1}{(X_j-\bar{X})}\underset{1\times p}{(X_j-\bar{X})^T}$ $p\times p$ matrix

We will work with scaled variables so $\Sigma = \rho$.

Rizzavalg: 6 Tiddlibom
9 Sult

Glutenfri =

Vegeler: 2 + 3

Q: When faced with a large set of correlated variables – is it possible to define a set of linear combinations of the original variables that capture a large part of the variability of the data?
- interpretability
- data reduction
- visualization

} Aims!

A: Yes. Let (λ_i, e_i) be eigenvalue/vector pairs of Σ where $\lambda_1 \geq \cdots \geq \lambda_p$. Then

Construct $Y_i = e_i^T X$

Y_i: principal components; e_i: loadings or rotations

Important property:

$\underset{p\times 1}{X}$, $Cov(X) = \underset{p\times p}{\Sigma}$ and (λ_i, e_i) eigenvalue/vector pairs of Σ, $i = 1, \ldots, p$

Let $Y_i = e_i^T X$ and $Y = \begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_p\end{bmatrix} = P^T X$

Then $Cov(Y) = Cov(P^T X)$

$= P^T Cov(X) P = P^T \Sigma P = \underbrace{P^T}_{I} \underbrace{P}_{I} \Lambda \underbrace{P^T P}_{I}$

$= \Lambda$

so $Var(e_i^T X) = \lambda_i$ and

$Cov(e_i^T X, e_j^T X) = 0$

$\Rightarrow$ the PCs Y_i have $Var(Y_i) = \lambda_i$

and are uncorrelated.

PC scores

$Y_1 = e_1^T X$
$Y_2 = e_2^T X$
$\vdots$
$Y_p = e_p^T X$

$\Big\} \; Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix}$ new random vector

I have a data set of n X's

Ex: $n = 21$

For each we may calculate Y = PC scores $y_1, \ldots, y_k$

$\Rightarrow$ sum up next time!

Ex: Norway:

Data

$x_1 = 9.8$
$x_2 = 6.21$
$x_3 = 3.61$
$x_4 = 1.1$
$x_5 = 6.4$
$x_6 = 52$

$e_1^T x \to y_1 = -0.497$
$e_2^T x \to y_2 = 0.54$
$y_3 =$
$\vdots$
$y_k =$

$y_1 = x_1 \cdot e_{11} + x_2 e_{12} + \ldots + x_6 \cdot e_{16}$
$= 9.8 \cdot -0.26 + 0.21 \cdot 0.66 + 0.61 \cdot 0.24 + \ldots = -0.497$

The same for each country

PCA plot

x-axis: PCA1 score
y-axis: PCA2 score
} n points

Biplot: - display graphically the loadings in the PC-space:

1) Coffee arrow: PC1 loading: -0.26, PC2: 0.67

Tea arrow: PC1: 0.65, PC2: -0.09

$f(x)$

Let X_1 and X_2 be two (continuous) RVs, and $f(x_1, x_2)$ be the joint pdf and $f_1(x_1)$ and $f_2(x_2)$ be the marginal pdfs, and C is a copula. What is true?

A $f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$

B $f(x_1, x_2)$ is found from $f_1(x_1)$ and $f_2(x_2)$ alone

C $f_1(x_1)$ is found from $f(x_1, x_2)$

D $f(x_1, x_2) = C(f_1(x_1), f_2(x_2))$

Mean of sum

$\boldsymbol{X}$ and $\boldsymbol{Y}$ are two bivariate random vectors with $\mathrm{E}(\boldsymbol{X}) = (1, 2)^T$ and $\mathrm{E}(\boldsymbol{Y}) = (2, 0)^T$. What is $\mathrm{E}(\boldsymbol{X} + \boldsymbol{Y})$?

A $(1.5, 1)^T$

B $(3, 2)^T$

C $(-1, 2)^T$

D $(1, -2)^T$

Mean of linear combination

$\boldsymbol{X}$ is a 2-dimensional random vector with $\mathrm{E}(\boldsymbol{X}) = (2, 5)^T$, and $\boldsymbol{b} = (0.5, 0.5)^T$ is a constant vector. What is $\mathrm{E}(\boldsymbol{b}^T\boldsymbol{X})$?

A 3.5

B 7

C 2

D 5

Covariance

$\boldsymbol{X}$ is a p-dimensional random vector with mean $\boldsymbol{\mu}$. Which of the following defines the covariance matrix?

A $E[(\boldsymbol{X} - \boldsymbol{\mu})^T(\boldsymbol{X} - \boldsymbol{\mu})]$

B $E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$

C $E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})]$

D $E[(\boldsymbol{X} - \boldsymbol{\mu})^T(\boldsymbol{X} - \boldsymbol{\mu})^T]$

Mean of linear combinations

$\boldsymbol{X}$ is a p-dimensional random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{C}$ is a constant matrix. What is then the mean of the k-dimensional random vector $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}$?

A $\boldsymbol{C}\boldsymbol{\mu}$

B $\boldsymbol{C}\boldsymbol{\Sigma}$

C $\boldsymbol{C}\boldsymbol{\mu}\boldsymbol{C}^T$

D $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T$

Covariance of linear combinations

$\boldsymbol{X}$ is a p-dimensional random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{C}$ is a constant matrix. What is then the covariance of the k-dimensional random vector $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}$?

A $\boldsymbol{C}\boldsymbol{\mu}$

B $\boldsymbol{C}\boldsymbol{\Sigma}$

C $\boldsymbol{C}\boldsymbol{\mu}\boldsymbol{C}^T$

D $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T$

Symmetric positive definite matrix

Which of the following is not correct for a symmetric positive definite matrix?

A The trace equals the rank of the matrix.

B The determinant is positive.

C The trace is the sum of the eigenvalues.

D All the eigenvalues are positive.

Correlation

$\boldsymbol{X}$ is a 2-dimensional random vector with covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Then the correlation between the two elements of $\boldsymbol{X}$ are:

A 0.10

B 0.25

C 0.40

D 0.80

PCA interpretation

Data set: student's score on a Math test, a Physics test, a Reading comprehension test, and a Vocabulary test.

First PC represents overall academic ability,
second PC represents a contrast between quantitative ability and verbal ability.

What loadings would be consistent with that interpretation?

A (0.5,0.5,0.5,0.5) and (0.71,0.71,0,0)
B (0.5,0.5,0.5,0.5) and (0.5,0.5,-0.5,-0.5)
C (0.71,0.71,0,0) and (0,0,0.71,-0.71)
D (0.71,0,-0.71,0) and (0 ,0.71,0,-0.71)

Correct?

Are you sure you want to read the correct answers? Maybe try first? The answers are explained on the next two slides.

Answers

1. C: We go from joint to marginal distribution by integration. The product of marginals equal the joint only for independent variables. We need information on the dependency structure to construct a joint from marginals, and that is what is done with the copula - but the formula is based on the cumulative distribution functions.
2. B: Mean of sum $(1,2)^T + (2,0)^T = (3,2)^T$.
3. A: Mean of linear combination $(0.5, 0.5)^T(2,5) = 3.5$.
4. B: Covariance matrix defined as $E\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\}$. This was the only formula that gave a $p \times p$ matrix. A gave a scalar and C and D did not match in dimensions.

Answers

5. A: $\boldsymbol{C\mu}$ is the mean of $\boldsymbol{Y} = \boldsymbol{CX}$.
6. D: $\boldsymbol{C\Sigma C}^T$ is the covariance matrix of $\boldsymbol{Y} = \boldsymbol{CX}$.
7. C: Correlation is 0.40 since covariance was 0.8 and variances 4 and 1.
8. A: NOT true for a symmetric positive definite matrix: the trace is in general not equal to the rank - but it is for idempotent symmetric matrices.
9. B: average means equal weight for all values, difference between quantitative and verbal means opposite signs for quantitative (maths and physics) and verbal (reading and vocabular).

TMA4267 Linear Statistical Models V2017 [L4]

Part 1: Multivariate RVs, and the multivariate normal distribution

The multivariate normal distribution (pdf and mgf) [H:4.2-4.4]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: January 20, 2017

What we know, and the plan for this lecture

What we know, and the plan for this lecture

- A random vector $\boldsymbol{X}$ can be described by the joint pdf $f(\boldsymbol{x})$.

What we know, and the plan for this lecture

- A random vector $\boldsymbol{X}$ can be described by the joint pdf $f(\boldsymbol{x})$.
- Mean: $\boldsymbol{\mu} = E(\boldsymbol{X}) = \{\mathrm{E}(X_j)\}$

What we know, and the plan for this lecture

- A random vector $\boldsymbol{X}$ can be described by the joint pdf $f(\boldsymbol{x})$.
- Mean: $\boldsymbol{\mu} = E(\boldsymbol{X}) = \{\mathrm{E}(X_j)\}$
- Covariance matrix: $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T)$, symmetric and we often require the matrix to be positive definite.

What we know, and the plan for this lecture

- A random vector $\boldsymbol{X}$ can be described by the joint pdf $f(\boldsymbol{x})$.
- Mean: $\boldsymbol{\mu} = E(\boldsymbol{X}) = \{\mathrm{E}(X_j)\}$
- Covariance matrix: $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T)$, symmetric and we often require the matrix to be positive definite.
- Linear combinations $\boldsymbol{CX}$: $\mathrm{E}(\boldsymbol{CX}) = \boldsymbol{C\mu}_X$ and $\mathrm{Cov}(\boldsymbol{CX}) = \boldsymbol{C\Sigma C}^T$.

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.

What we know, and the plan for this lecture

- A random vector $\boldsymbol{X}$ can be described by the joint pdf $f(\boldsymbol{x})$.
- Mean: $\boldsymbol{\mu} = E(\boldsymbol{X}) = \{\mathrm{E}(X_j)\}$
- Covariance matrix: $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T)$, symmetric and we often require the matrix to be positive definite.
- Linear combinations $\boldsymbol{CX}$: $\mathrm{E}(\boldsymbol{CX}) = \boldsymbol{C\mu}_X$ and $\mathrm{Cov}(\boldsymbol{CX}) = \boldsymbol{C\Sigma C}^T$.
- Now: derive the joint pdf and the moment generating function for the multivariate normal distribution.

Cramer-Wold and moment generating functions

$\boldsymbol{X}_{(p\times 1)}$ is a random vector. The distribution of $\boldsymbol{X}$ is completely determined by the set of all one-dimensional distributions of the linear combinations $Y = \boldsymbol{t}^T\boldsymbol{X} = \sum_{i=1}^{p} t_i X_i$ where $\boldsymbol{t}$ ranges over all fixed p-vectors.

- $Y = \boldsymbol{t}^T\boldsymbol{X}$ has MGF $M_Y(s) = \mathrm{E}(\exp(sY)) = \mathrm{E}(\exp(s\boldsymbol{t}^T\boldsymbol{X}))$.
- If we choose $s = 1$ $M_Y(1) = \mathrm{E}(\exp(\boldsymbol{t}^T\boldsymbol{X})) = M_{\boldsymbol{X}}(\boldsymbol{t})$, which is the MGF of $\boldsymbol{X}$ and thus determines the distribution of $\boldsymbol{X}$.

Härdle and Simes (2015) use characteristic functions, $\mathrm{E}(e^{i\boldsymbol{t}^T\boldsymbol{X}})$ but we stick with moment generating functions $\mathrm{E}(e^{\boldsymbol{t}^T\boldsymbol{X}})$. Why: we will only work with nice distributions and do not have problems with integrals not existing, and we know MGFs from previous course.

Multivariate transformation formula [H:4.3]

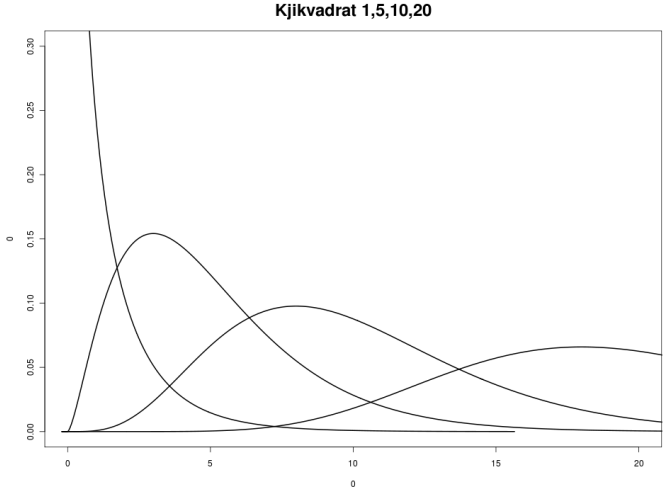
$$X = u(Y) \tag{4.43}$$

for a one-to-one transformation u: $\mathbb{R}^p \to \mathbb{R}^p$. Define the Jacobian of u as

$$\mathcal{J} = \left(\frac{\partial x_i}{\partial y_j}\right) = \left(\frac{\partial u_i(y)}{\partial y_j}\right)$$

and let abs($|\mathcal{J}|$) be the absolute value of the determinant of this Jacobian. The pdf of Y is given by

$$f_Y(y) = \text{abs}(|\mathcal{J}|) \cdot f_X\{u(y)\}. \tag{4.44}$$



The Chi-square distribution

pdf χ^2_p:

$$f(y) = \frac{1}{2^{p/2}\Gamma(p/2)} y^{p/2-1} e^{(-y/2)} \text{ for } y > 0$$

MGF χ^2_p:

$$M_Y(t) = \frac{1}{(1-2t)^{p/2}}$$

Addition property:
Let $X_1 \sim \chi^2_p$ and $X_2 \sim \chi^2_q$, and let X_1 and X_2 be independent. Then $X_1 + X_2 \sim \chi^2_{p+q}$.
Subtraction property:
Let $X = X_1 + X_2$ with $X_1 \sim \chi^2_p$ and $X \sim \chi^2_{p+q}$. Assume that X_1 and X_2 are independent. Then $X_2 \sim \chi^2_q$.

This 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 $N_1(0,1) \to \boldsymbol{Z}_{p\times 1} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$
   - MGF: $M_{\boldsymbol{Z}}(\boldsymbol{t}) = \mathrm{E}(e^{\boldsymbol{t}^T\boldsymbol{z}}) = e^{\frac{1}{2}\boldsymbol{t}^T\boldsymbol{t}}$
3. $\boldsymbol{X} = \boldsymbol{AZ} + \boldsymbol{\mu}$, $\boldsymbol{AA}^T = \boldsymbol{\Sigma}$ 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 (invertible):

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

Properties of the mvN - plan for L5

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

1. Probability density function $f(\boldsymbol{x})$ (both when $\boldsymbol{\Sigma}$ is invertible and not).
2. Moment generating function: $M_X(\boldsymbol{t})=\exp(\boldsymbol{t}^T\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$
3. Graphical display, contours (ellipsoids), and chisq-distributed $(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$.
4. Linear combinations of components of $\boldsymbol{X}$ are (multivariate) normal.
5. All subsets of the components of $\boldsymbol{X}$ are (multivariate) normal.
6. Zero covariance implies that the corresponding components are independently distributed.
7. $\boldsymbol{A\Sigma B}^T=\boldsymbol{0}\Leftrightarrow \boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent.
8. The conditional distributions of the components are (multivariate) normal. $\boldsymbol{X}_2\mid(\boldsymbol{X}_1=\boldsymbol{x}_1)\sim N_{p2}(\boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1),\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$

And then remains estimators for parameters and properties of quadratic forms in L6.

Characterizing X:

- $f(x)$ (pdf), $F(x)$ (cdf)
- $M_X(t)=E(e^{t^TX})=E(e^{t_1X_1+t_2X_2+\dots+t_pX_p})$

  (t: p×1, X: p×1)

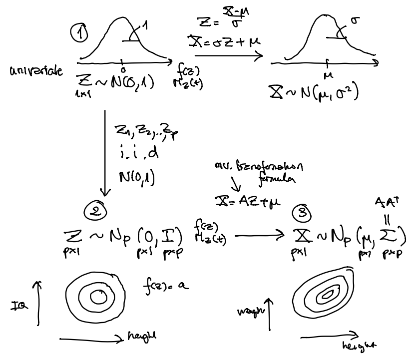
  $=\int\int\cdots\int_{-\infty}^{\infty} e^{t^Tx} f(x)\,dx_1\cdots dx_p$

MGF: easy for proofs.

Multivariate normal distribution

TMA4267 L4
20.01.2017

Plan: derive the pdf and moment generating function (MGF $M_Z(t)=E(e^{tZ})$) of the multivariate normal.

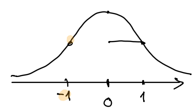


① $Z\sim N(0,1)$

$E(Z)=0$, $Var(Z)=1$

$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$ $\quad -\infty<z<\infty$

$f'(z)=0$ for $z=0$
$f''(z)=0$ for $z=1$ and $z=-1$
$z\to\pm\infty \Rightarrow f(z)\to 0$



mean, median, mode

MGF (univ) $M_Z(t)=E(e^{tZ})=\int_{-\infty}^{\infty}e^{tz}f(z)\,dz$

$=\dots=e^{\frac{1}{2}t^2}$

homework, pencast.

Cool: The law of the unconscious statistician (berustless?)

$E(g(X))=\int_{-\infty}^{\infty}g(x)f(x)\,dx$

Y: $E(Y)=\int_{-\infty}^{\infty}y f(y)\,dy$

① → ②

$Z_1, Z_2, \dots, Z_p$ independent $N(0,1)$ RVs

$$f(z) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} = \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{p} z_i^2\right\}$$

($f(z)$: $p \times 1$)

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} \qquad = \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} \exp\left\{-\frac{1}{2} z^T z\right\}$$

$$E(Z) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \quad (p \times 1) \quad \text{and} \quad Cov(Z) = I_{p \times p}$$

$Var(Z_i) = 1$

$Cov(Z_i, Z_j) = 0 \quad i \neq j$

MGF: multivariate version

$$M_Z(t) \underset{\text{DEF}}{=} E\left(\exp(t^T Z)\right)$$

($t^T Z = t_1 Z_1 + t_2 Z_2 + \dots + t_p Z_p$)

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(t_1 z_1 + t_2 z_2 + \dots + t_p z_p) \cdot f(z)\, dz_1 \dots dz_p$$

($f(z) = \prod_{i=1}^{p} f(z_i)$)

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t_1 z_1} \cdot e^{t_2 z_2} \cdots e^{t_p z_p} \cdot f(z_1) \cdot f(z_2) \cdots f(z_p)\, dz_1 \cdots dz_p$$

$$= \underbrace{\int_{-\infty}^{\infty} e^{t_1 z_1} f(z_1) dz_1}_{M_{Z_1}(t)} \cdots \underbrace{\int_{-\infty}^{\infty} e^{t_p z_p} f(z_p) dz_p}_{M_{Z_p}(t)}$$

$$M_Z(t) = \prod_{i=1}^{p} \underbrace{M_{Z_i}(t)}_{e^{\frac{1}{2}t_i^2}} = \prod_{i=1}^{p} e^{\frac{1}{2}t_i^2} = e^{\frac{1}{2}t^T t}$$

($e^{\frac{1}{2}t_i^2} \cdot e^{\frac{1}{2}t_2^2} \cdots e^{\frac{1}{2}t_p^2}$)

② → ③

$$\underset{p \times 1}{Z} \sim N_p(0, I) \qquad f(z) = \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} e^{-\frac{1}{2}z^T z}$$

$$M_Z(t) = e^{\frac{1}{2}t^T t}$$

Now: $\underset{p \times 1}{X} = \underset{p \times p}{A} \underset{p \times 1}{Z} + \underset{p \times 1}{\mu}$, where A has full rank ($Ax = 0$ has only solution $x = 0$)

$$E(X) = AE(Z) + E(\mu) = A \cdot 0 + \mu = \mu$$

$$Cov(X) = A \underbrace{Cov(Z)}_{I} A^T = AA^T \equiv \Sigma$$

since A has full rank AA^T is positive definite (can be proven using $(Ax)^T(Ax)$)

What can A be? $\Sigma^{\frac{1}{2}}$, but also other possibilities.

②→③ first MGF of X

$$M_X(t) = M_{AZ+\mu}(t) = E\left(e^{t^T(AZ+\mu)}\right)$$

(t^T)

$$= E\left(e^{t^T AZ} \cdot e^{t^T \mu}\right) = e^{t^T \mu} E\left(e^{t^T A Z}\right)$$

($t^* = A^T t$)

$$= e^{t^T \mu} M_Z(A^T t) \qquad M_Z(t) = E(e^{t^T Z})$$

$$= e^{t^T \mu} e^{\frac{1}{2} t^T A A^T t} = e^{t^T \mu + \frac{1}{2} t^T \Sigma t}$$

②→③ $f(x)$, and the mv. transformation formula

The mv. transf. formula

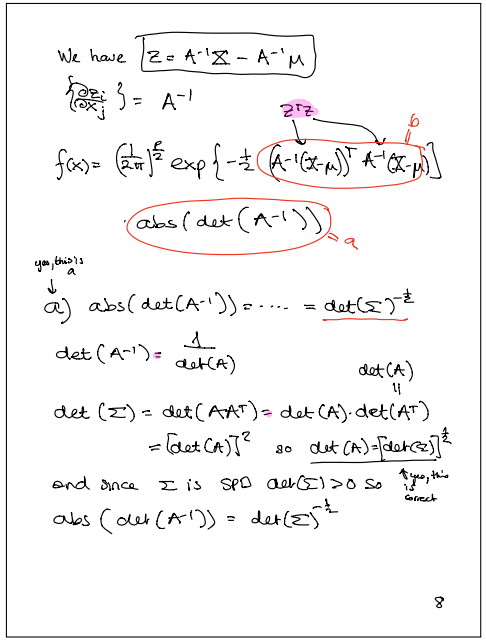
$$X = AZ + \mu \Leftrightarrow \boxed{Z = A^{-1}(X - \mu)}$$

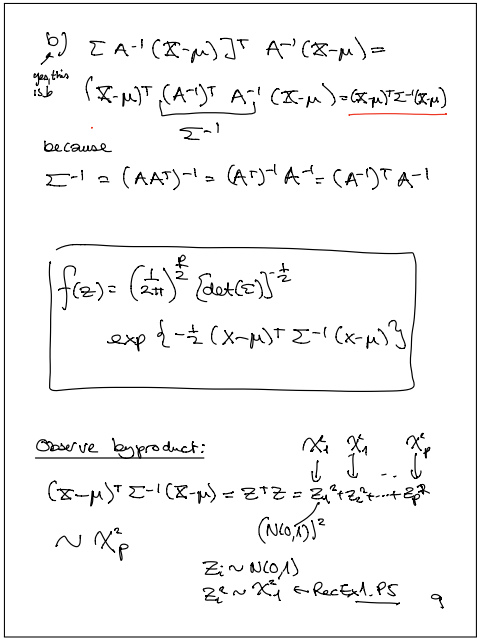
$$f_X(x) = f_Z(z(x)) \cdot abs(J)$$

$$J = \det\left(\left(\frac{\partial z_i}{\partial x_j}\right)_{(i,j)}\right)$$

element (i,j)

↑ a matrix with these as





Comment: When A does not have full rank
we can use a singular version of the pdf.

10

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. $\boldsymbol{X} = \boldsymbol{AZ} + \boldsymbol{\mu}$, $\boldsymbol{AA}^T = \boldsymbol{\Sigma}$ 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).

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).

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).

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).

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)

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 $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$.

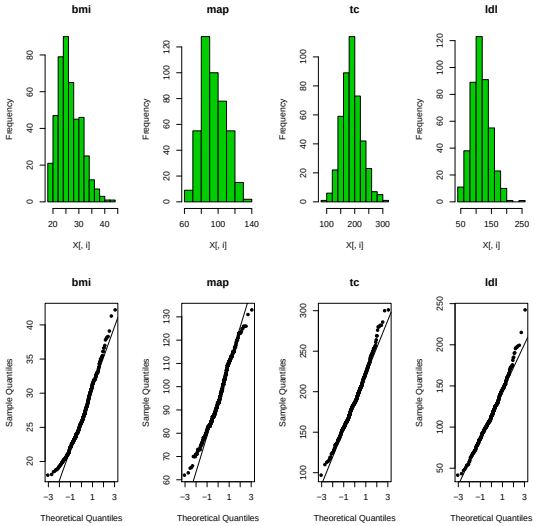
You find the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ on the next slide.

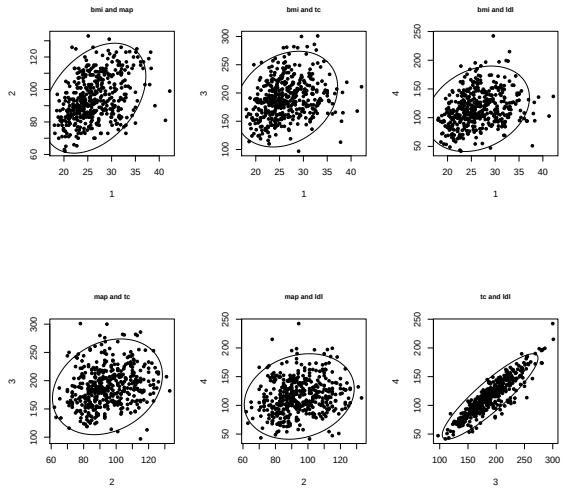
Describe the graph of the equation $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = b$ where $b > 0$ is a constant.
Make a drawing of the graph, for $b = 1$ found above.
What is the probability that a random sample from this distribution will be inside this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





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

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

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.

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. $\boldsymbol{A\Sigma B}^T = 0 \Leftrightarrow \boldsymbol{AX}$ 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 - 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?

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. $\boldsymbol{A\Sigma B}^T = 0 \Leftrightarrow \boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent (will be very important in Part 2). [CompEx1.2b]
6. The conditional distributions of the components are (multivariate) normal. $\boldsymbol{X}_2 \mid (\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim N_{p2}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$.

Properties of the multivariate normal distribution (H: 26, 44, 5.1)

TMA4267 LS 24.01.2017

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

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

$M_X(t) = E(e^{t^T X}) = e^{t^T\mu + \frac{1}{2}t^T\Sigma t}$

① Contours of the mvN

For some constant a ($a>0$) the solution to $f(x) = a$ are called the contours of $\boldsymbol{X}$.

$f(x) = \text{const} \cdot \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$

or work with

$(\boldsymbol{X}-\mu)^T\Sigma^{-1}(\boldsymbol{X}-\mu) = b \qquad (b>0)$

What is this graphical object?

Let $\Sigma = P\Lambda P^T$

$[e_1 \cdots e_p] \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \lambda_p \end{bmatrix} \begin{bmatrix} \frac{1}{\lambda_1} & & 0 \\ & \frac{1}{\lambda_2} & \\ 0 & & \ddots \frac{1}{\lambda_p} \end{bmatrix}$

Work with $\Sigma^{-1} = (P\Lambda P^T)^{-1} = P\Lambda^{-1}P^T$

$= \frac{1}{\lambda_1} e_1e_1^T + \frac{1}{\lambda_2} e_2e_2^T + \cdots + \frac{1}{\lambda_p} e_pe_p^T$

Study $p=2$:

$P = [e_1\ e_2]$, $\Lambda^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & 0 \\ 0 & \frac{1}{\lambda_2} \end{bmatrix}$

$\Sigma^{-1} = \frac{1}{\lambda_1}e_1e_1^T + \frac{1}{\lambda_2}e_2e_2^T$

$(\boldsymbol{X}-\mu)^T\Sigma^{-1}(\boldsymbol{X}-\mu) = b$ $\quad e_1 = 2\times 1$, $e_2 = 2\times 1$

$(\boldsymbol{X}-\mu)^T\left(\frac{1}{\lambda_1}e_1e_1^T + \frac{1}{\lambda_2}e_2e_2^T\right)(\boldsymbol{X}-\mu) = b$

$1\times 2 \quad 2\times 2 \quad 2\times 1$

$\frac{1}{\lambda_1}\underbrace{(\boldsymbol{X}-\mu)^T e_1}_{1\times 2\ \ 2\times 1}e_1^T(\boldsymbol{X}-\mu) + \frac{1}{\lambda_2}(\boldsymbol{X}-\mu)^T e_2 e_2^T(\boldsymbol{X}-\mu) = b$

$w_1^T \qquad\qquad w_2^T$

$\frac{1}{\lambda_1}\underbrace{w_1^Tw_1}_{w_1^2} + \frac{1}{\lambda_2}\underbrace{w_2^Tw_2}_{w_2^2} = b$

$\frac{1}{\lambda_1 b}\cdot w_1^2 + \frac{1}{\lambda_2 b}\cdot w_2^2 = 1$ What is this?

This is an ellipse ($p=2$)

Halflengths: $\sqrt{\lambda_1 b}$, $\sqrt{\lambda_2 b}$

Axis: are in the direction of the eigenvectors e_1, e_2

Centered: in μ

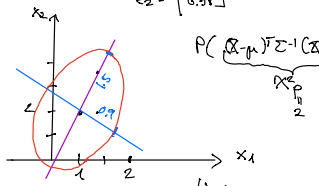
$P>2$: general: ellipsoide.

Exam K2014 Modified

$\underset{2\times 1}{\boldsymbol{X}}$ $\quad \mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix}$, $b=1$

$\lambda_1 = 2.2$ $\quad e_1 = \begin{bmatrix} 0.38 \\ 0.92 \end{bmatrix}$ $\quad \sqrt{\lambda_1 \cdot b} = 1.5$

$\lambda_2 = 0.8$ $\quad e_2 = \begin{bmatrix} -0.92 \\ 0.38 \end{bmatrix}$ $\quad \sqrt{\lambda_2 \cdot b} = 0.9$



$P((\boldsymbol{X}-\mu)^T\Sigma^{-1}(\boldsymbol{X}-\mu) \le 1) = \chi^2_{p=2} = 0.39$

pchisq(1,2) = 0.39

② Linear combinations

$\underset{p\times 1}{X} \sim N_p(\mu, \Sigma)$ and $\underset{q\times p}{B}$ and $\underset{\text{constants}}{c \in \mathbb{R}^q}$

$\underset{q\times 1}{Y} = BX + c \sim N_q(B\mu + c, B\Sigma B^T)$

$\mathrm{E}(e^{t^T X})$

PROOF: $M_X(t) = \exp\left(t^T\mu + \frac{1}{2}t^T\Sigma t\right)$

$M_Y(t) = M_{BX+c}(t) = \mathrm{E}\left(e^{t^T(BX+c)}\right)$

$= e^{t^Tc}\underbrace{\mathrm{E}\left(e^{t^TBX}\right)}_{M_X(B^Tt)} = e^{t^Tc}\, e^{t^TB\mu + \frac{1}{2}t^TB\Sigma B^Tt}$

$= \exp\left\{ t^T(B\mu + c) + \frac{1}{2}t^TB\Sigma B^Tt \right\}$

$\mathrm{E}(Y)$ $\quad$ $\mathrm{Cov}(Y)$

This is MVN with $\mathrm{E}(Y) = B\mu + c$ and $\mathrm{Cov}(Y) = B\Sigma B^T$

Exam K2014 1a

$Y = \underbrace{\begin{bmatrix} 3 & -2 \\ 1 & 1 \end{bmatrix}}_{C} X \sim N_2(C\mu, C\Sigma C^T)$

μ, Σ in text.

④ "Covariance = 0" implies independence for MVN

a) If $\underset{p\times 1}{X_1}$ and $\underset{p\times 1}{X_2}$ are independent.

$\Downarrow$

$\mathrm{Cov}(X_1, X_2) = \underset{p\times p}{0}$

for general X's.

But: $\mathrm{Cov}(X_1, X_2) = 0$ does not in general imply that X_1 and X_2 are independent, e.g.

$X_1 = Y$, $X_2 = Y^2$, $\mathrm{Cov}(Y, Y^2) = 0$

How does $M_X(t)$ look like when X_1 and X_2 are independent when $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_p(\mu, \Sigma)$.

p, p_1, p_2; $\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$

$M_X(t) = M_{X_1}(t_1) \cdot M_{X_2}(t_2)$

$= \exp\left(t_1^T\mu_1 + \frac{1}{2}t_1^T\Sigma_{11}t_1\right) \cdot \exp\left(t_2^T\mu_2 + \frac{1}{2}t_2^T\Sigma_{22}t_2\right)$

$t = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \overset{0}{\Sigma_{12}} \\ \underset{0}{\Sigma_{21}} & \Sigma_{22} \end{bmatrix}$

$M_X(t) = \exp\left(t^T\mu + \frac{1}{2}t^T\Sigma t\right)$

③ Subsets

$X \sim N_p(\mu, \Sigma)$, partition into

$\underset{p\times 1}{X} = \begin{bmatrix} X_1 \\ \underset{p_1\times 1}{} \\ X_2 \\ \underset{p_2\times 1}{} \end{bmatrix}$ $\quad p = p_1 + p_2$

$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

$\underset{p_1\times p_1}{\Sigma_{11}} = \mathrm{Cov}(X_1)$

$\underset{p_1\times p_2}{\Sigma_{12}} = \mathrm{Cov}(X_1, X_2)$

$\underset{p_2\times p_1}{\Sigma_{21}} = \Sigma_{12}^T$

$\underset{p_2\times p_2}{\Sigma_{22}} = \mathrm{Cov}(X_2)$

Let $\underset{p_1\times p}{B} = \begin{bmatrix} \underset{p_1\times p_1}{I} & \underset{p_1\times p_2}{0} \end{bmatrix}$

Then $X_1 = BX \sim N_{p_1}(B\mu = \mu_1, B\Sigma B^T = \Sigma_{11})$

$\begin{bmatrix} I & 0 \end{bmatrix}\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\begin{bmatrix} I \\ 0 \end{bmatrix}$

b) Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

X_1 and X_2 are independent iff $\underset{p_1\times p_2}{\Sigma_{12}} = \mathrm{Cov}(X_1, X_2) = \underset{p_1\times p_2}{0}$

Why: $\Sigma_{12} = 0 \Leftrightarrow \Sigma_{21} = 0$ and then

$M_X(t) = M_{X_1}(t_1) \cdot M_{X_2}(t_2)$.

NB only if N_p = multivariate normal.

Ex: Independent variables:
(1, 3), (1, 4), (2, 3)

$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ $\quad$ $\begin{bmatrix} X_3 \\ X_4 \end{bmatrix}$ not independent

Homework: K2014 1a on slide 17.

⑤ Independence of $A\mathbf{X}$ and $B\mathbf{X}$

$\mathbf{X} \sim N_p(\mu, \Sigma)$

$A\mathbf{X}$ and $B\mathbf{X}$ are independent iff $A\Sigma B^T = 0$.

Why? $Y = \begin{bmatrix} A \\ B \end{bmatrix} \mathbf{X}$ ($q \times p$, $p \times 1$; A: $q_1 \times p$, B: $q_2 \times p$, $q_1 + q_2 = q$)

$M_Y(t) = E\left[\exp\{t^T \begin{bmatrix} A \\ B \end{bmatrix} \mathbf{X}\}\right]$

know: $M_{\mathbf{X}}(t) = \exp(t^T\mu + \frac{1}{2} t^T \Sigma t)$

$M_Y(t) = \exp\{t^T \begin{bmatrix} A \\ B \end{bmatrix} \mu + \frac{1}{2} t^T \begin{bmatrix} A \\ B \end{bmatrix} \Sigma \begin{bmatrix} A^T & B^T \end{bmatrix} t\}$

which: $\begin{bmatrix} A \\ B \end{bmatrix} \mathbf{X} \sim N_q\left(\begin{bmatrix} A \\ B \end{bmatrix}\mu, \begin{bmatrix} A \\ B \end{bmatrix} \Sigma \begin{bmatrix} A^T & B^T \end{bmatrix}\right)$

$\begin{bmatrix} A\Sigma A^T & A\Sigma B^T \\ B\Sigma A^T & B\Sigma B^T \end{bmatrix}$

so if $A\Sigma B^T = 0 \Leftrightarrow A\mathbf{X}$ and $B\mathbf{X}$ are independent.

(This version did not rely on result ② and ④.)
→ as pointed out by student after class.

⑥ Conditional distribution of mvN

$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_p\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$

Q: What is the distribution of $\mathbf{X}_2 | \mathbf{X}_1 = x_1$.

A: $\mathbf{X}_2 | \mathbf{X}_1 = x_1 \sim N_{q_2}\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$

Observe $E(\mathbf{X}_2 | \mathbf{X}_1 = x_1)$ is linear in x_1
$\mathrm{Cov}(\mathbf{X}_2 | \mathbf{X}_1 = x_1)$ not dependent on x_1.

Added after the lecture: shorter version of "why" that builds on both ② and ④.

$\mathbf{X} \sim N_p(\mu, \Sigma)$ and A: $q_1 \times p$, B: $q_2 \times p$

$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} A\mathbf{X} \\ B\mathbf{X} \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} \mathbf{X} = C\mathbf{X}$ ($q \times p$).

From ② we know that $Y \sim N_q(C\mu, C\Sigma C^T)$.

From ④ we know that Y_1 and Y_2 are independent iff $\mathrm{Cov}(Y_1, Y_2) = 0$,

and $\mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(A\mathbf{X}, B\mathbf{X})$

$= A \underbrace{\mathrm{Cov}(\mathbf{X}, \mathbf{X})}_{\mathrm{Cov}(\mathbf{X}) = \Sigma} B^T = A\Sigma B^T$, so

$A\Sigma B^T = 0 \Leftrightarrow A\mathbf{X}$ and $B\mathbf{X}$ independent.

TMA4267 Linear Statistical Models V2017 [L6]

Part 1: Multivariate RVs, and the multivariate normal distribution
Estimators for mean and covariance
Quadratic forms [H:3.3,4.5,5.1,F:AppB3]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: January 27, 2017

Plan for today

- estimators for mean and covariance
- quadratic forms and rules for quadratic forms
- idempotent matrices
- more rules for quadratic forms - with idempotent matrices

Maximum likelihood estimators

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample of size n from the multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Maximum likelihood estimators

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample of size n from the multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The maximum likelihood estimators for are found by maximizing the likelihood:

$$
\begin{aligned}
L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{j=1}^{n} f(\boldsymbol{x}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \prod_{j=1}^{n} (\frac{1}{2\pi})^{\frac{p}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu})\}
\end{aligned}
$$

Maximum likelihood estimators

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample of size n from the multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The maximum likelihood estimators for are found by maximizing the likelihood:

$$
\begin{aligned}
L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{j=1}^{n} f(\boldsymbol{x}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \prod_{j=1}^{n} (\frac{1}{2\pi})^{\frac{p}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu})\}
\end{aligned}
$$

Could take ln and then partial derivatives, but easier to add and subtract the mean $\bar{\boldsymbol{x}}$ and rewrite (using trace-formulas)

Maximum likelihood estimators

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample of size n from the multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The maximum likelihood estimators for are found by maximizing the likelihood:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{j=1}^{n} f(\boldsymbol{x}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$= \prod_{j=1}^{n} (\frac{1}{2\pi})^{\frac{p}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu})\}$$

Could take ln and then partial derivatives, but easier to add and subtract the mean $\bar{\boldsymbol{x}}$ and rewrite (using trace-formulas)

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\frac{1}{2\pi})^{\frac{np}{2}} \det(\boldsymbol{\Sigma})^{-\frac{n}{2}}$$
$$\exp\{-\frac{1}{2}[\mathrm{tr}(\boldsymbol{\Sigma}^{-1} \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T) + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu})]\}$$

Maximum likelihood estimators: first for $\boldsymbol{\mu}$

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\frac{1}{2\pi})^{\frac{np}{2}} \det(\boldsymbol{\Sigma})^{-\frac{n}{2}}$$
$$\exp\{-\frac{1}{2}[\mathrm{tr}(\boldsymbol{\Sigma}^{-1} \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T) + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu})]\}$$

and see directly for SPD $\boldsymbol{\Sigma}$ that the maximum is achieved for $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$, so that the MLE for $\boldsymbol{\mu}$ is

$$\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{X}_j$$

Maximum likelihood estimators: then for $\boldsymbol{\Sigma}$

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\frac{1}{2\pi})^{\frac{np}{2}} \det(\boldsymbol{\Sigma})^{\frac{n}{2}}$$
$$\exp\{-\frac{1}{2}[\mathrm{tr}(\boldsymbol{\Sigma}^{-1} \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T) + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu})]\}$$

A maximization theorem for matrices it used to find that the MLE for $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^T$$

Properties of the ML estimators

- $\bar{\boldsymbol{X}}$ is distributed as $N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$
- $n\boldsymbol{S}$ is distributed as a Wishart random matrix with $n-1$ degrees of freedom.
- $\bar{\boldsymbol{X}}$ and $n\boldsymbol{S}$ are independent.

The Wishart distribution is not on the reading list for TMA4267. General properties of maximum likelihood estimation is covered in detail in TMA4295 Statistical Inference.

Quadratic forms - first results [F:B3.3, Theorem B.2]

We stay with our random vector $\boldsymbol{X}$ with $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and a symmetric constant matrix $\boldsymbol{A}$.

- What is a quadratic form? $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$
- The "trace-formula": $\mathrm{E}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X})$.

Useful facts about the trace [H:2.1] and [F:Theorem A.18]

Let A, B and C be conformable matrices

$$\begin{aligned}\mathrm{tr}(A+B) &= \mathrm{tr}(A)+\mathrm{tr}(B)\\ \mathrm{tr}(AB) &= \mathrm{tr}(BA)\\ \mathrm{tr}(ABC) &= \mathrm{tr}(CAB) = \mathrm{tr}(BCA)\end{aligned}$$

Exam V2014: Problem 1a

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ be a random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Further, let $\boldsymbol{A} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$ be a matrix of constants.

Define $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \boldsymbol{A}\boldsymbol{X}$.

Find $\mathrm{E}(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{Y})$. Are X_1 and X_2 independent? Are Y_1 and Y_2 independent? Justify your answers.

Find the mean of $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$.

Quadratic forms - last results [F:B3.3, Theorem B.2]

Now: $\boldsymbol{X}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{I}$, and we also have a symmetric and idempotent matrix $\boldsymbol{R}_{(p\times p)}$ with rank r.

- Properties of an idempotent matrix.
- Distribution of $\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X} \sim \chi^2_r$.
- Distribution of a ratio of two quadratic forms and the Fisher distribution.

Properties of symmetric idempotent matrices

A symmetric matrix $\boldsymbol{A}$ is idempotent, $\boldsymbol{A}^2 = \boldsymbol{A}$, and has the following properties (to be proven in RecEx1.P7).

Properties of symmetric idempotent matrices

A symmetric matrix $\boldsymbol{A}$ is idempotent, $\boldsymbol{A}^2 = \boldsymbol{A}$, and has the following properties (to be proven in RecEx1.P7).

1. The eigenvalues are 0 and 1.

Properties of symmetric idempotent matrices

A symmetric matrix $\boldsymbol{A}$ is idempotent, $\boldsymbol{A}^2 = \boldsymbol{A}$, and has the following properties (to be proven in RecEx1.P7).

1. The eigenvalues are 0 and 1.
2. The rank of a symmetric matrix (actually: a diagonalizable quadratic matrix) equals the number of nonero eigenvaluse of the matrix. Should be known from previous courses.

Properties of symmetric idempotent matrices

A symmetric matrix $\boldsymbol{A}$ is idempotent, $\boldsymbol{A}^2 = \boldsymbol{A}$, and has the following properties (to be proven in RecEx1.P7).

1. The eigenvalues are 0 and 1.
2. The rank of a symmetric matrix (actually: a diagonalizable quadratic matrix) equals the number of nonero eigenvaluse of the matrix. Should be known from previous courses.
3. (Combining 1+2). If a $(n \times n)$ symmetric idempotent matrix $\boldsymbol{A}$ has rank r then r eigenvalues are 1 and $n - r$ are 0.

Properties of symmetric idempotent matrices

A symmetric matrix $\boldsymbol{A}$ is idempotent, $\boldsymbol{A}^2 = \boldsymbol{A}$, and has the following properties (to be proven in RecEx1.P7).

1. The eigenvalues are 0 and 1.
2. The rank of a symmetric matrix (actually: a diagonalizable quadratic matrix) equals the number of nonero eigenvaluse of the matrix. Should be known from previous courses.
3. (Combining 1+2). If a $(n \times n)$ symmetric idempotent matrix $\boldsymbol{A}$ has rank r then r eigenvalues are 1 and $n - r$ are 0.
4. The trace and rank of a symmetric projection matrix are equal: $tr(\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$.

Properties of symmetric idempotent matrices

A symmetric matrix $\boldsymbol{A}$ is idempotent, $\boldsymbol{A}^2 = \boldsymbol{A}$, and has the following properties (to be proven in RecEx1.P7).

1. The eigenvalues are 0 and 1.
2. The rank of a symmetric matrix (actually: a diagonalizable quadratic matrix) equals the number of nonero eigenvaluse of the matrix. Should be known from previous courses.
3. (Combining 1+2). If a $(n \times n)$ symmetric idempotent matrix $\boldsymbol{A}$ has rank r then r eigenvalues are 1 and $n - r$ are 0.
4. The trace and rank of a symmetric projection matrix are equal: $tr(\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$.
5. The matrix $\boldsymbol{I} - \boldsymbol{A}$ is also idempotent, and $\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{A}) = 0$.

The Chi-square distribution

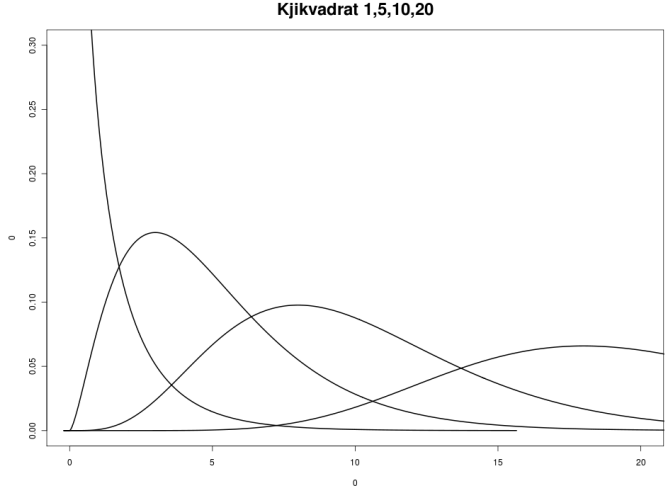
pdf χ^2_p:

$$f(y) = \frac{1}{2^{p/2}\Gamma(p/2)} y^{p/2-1} e^{(-y/2)} \text{ for } y > 0$$

MGF χ^2_p:

$$M_Y(t) = \frac{1}{(1-2t)^{p/2}}$$

Addition property:
Let $X_1 \sim \chi^2_p$ and $X_2 \sim \chi^2_q$, and let X_1 and X_2 be independent. Then $X_1 + X_2 \sim \chi^2_{p+q}$.
Subtraction property:
Let $X = X_1 + X_2$ with $X_1 \sim \chi^2_p$ and $X \sim \chi^2_{p+q}$. Assume that X_1 and X_2 are independent. Then $X_2 \sim \chi^2_q$.



The Fisher distribution [F: B.1 Def 8.14], RecEx2.P5+6

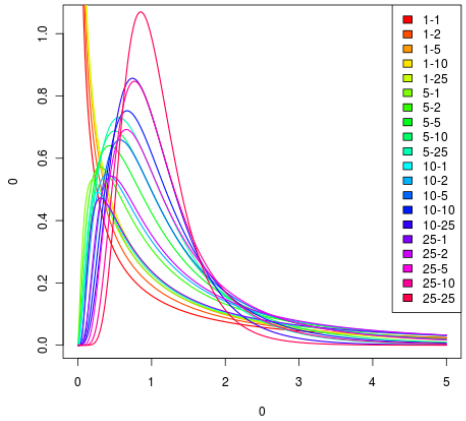
"Tabeller og formeler i statistikk":
If Z_1 and Z_2 are independent and χ^2-distributed with ν_1 and ν_2 degrees of freedom, then

$$F = \frac{Z_1/\nu_1}{Z_2/\nu_2}$$

is F(isher)-distributed with ν_1 and ν_2 degrees of freedom.

- The expected value of F is $\mathrm{E}(F) = \frac{\nu_2}{\nu_2 - 2}$.
- The mode is at $\frac{\nu_1 - 2}{\nu_1}\frac{\nu_2}{\nu_2 + 2}$.
- Identity:

$$f_{1-\alpha,\nu_1,\nu_2} = \frac{1}{f_{\alpha,\nu_2,\nu_1}}$$



The Fisher distribution with different degrees of freedom ν_1 and ν_2 (given in the legend).

Quadratic forms [F:B3.3, Theorem B.2]

Random vector $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, symmetric constant matrix $\boldsymbol{A}$.

- Quadratic form: $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$.
- The "trace-formula": $\mathrm{E}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) - \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$.

Then, let $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{R}$ is a symmetric and idempotent matrix with rank r.

$$\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X} \sim \chi^2_r$$

Now, also $\boldsymbol{S}$ is a symmetric and idempotent matrix with rank s, and $\boldsymbol{R}\boldsymbol{S} = \boldsymbol{0}$.

$$\frac{s\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}}{r\boldsymbol{X}^T\boldsymbol{S}\boldsymbol{X}} \sim F_{r,s}$$

Plan for the last week of Part 1

Supervision in lecture times.
See Blackboard: Part 1: dates and places for supervision.

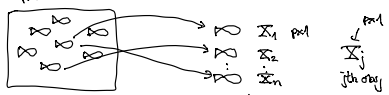
Kahoot!

Summing up the last three lectures with a few multiple choice questions.

Parameterestimation: μ and Σ — TMA4267 Lb — 27.01.2017

[H3.3, H4.5]

$\underset{p\times 1}{X}$ random vector ← quality of salmon



3 versions on notation:

a) X_j, $j=1,..,n$

b) $\underset{n\times p}{X} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}$ data matrix rows = obs cols = rv's

c) $\underset{np\times 1}{X^*} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ vector of rv's

How can we use X_j, $j=1,..,n$ to estimate $\mu = E(X)$ and $\Sigma = Cov(X)$?

① Estimating $\mu = E(X)$

$$\underset{p\times 1}{\bar{X}} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n}\sum_{j=1}^{n} \overbrace{E(X_j)}^{\mu} = \frac{1}{n}\cdot n\cdot\mu = \underline{\mu}$$

unbiased

$$Cov(\bar{X}) = Cov\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)$$

$Cov(X) = E((X-\mu)(X-\mu)^T)$

$$= \left(\frac{1}{n}\right)^2 \sum_{j=1}^{n} Cov(X_j)$$

$X_1,..,X_n$ independent

$$= \frac{1}{n^2}\sum_{j=1}^{n} \underbrace{\Sigma}_{Cov(X_j)} = \frac{1}{n^2}\cdot n\cdot\Sigma = \underline{\underline{\frac{1}{n}\Sigma}}$$

If $X_1, X_2, .., X_n$ i.i.d identically independent distributed $\sim N_p(\mu, \Sigma)$

$$\bar{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$$

② Estimating Σ

Motivation: $\Sigma = E\left((X-\mu)(X-\mu)^T\right)$

$$\underset{p\times p}{S} = \frac{1}{n-1}\sum_{j=1}^{n} \underbrace{(X_j-\bar{X})}_{p\times 1}(X_j-\bar{X})^T$$

$E(S) = \cdots = \Sigma$ unbiased

↑ see proof, separate file

Quadratic forms [F: B33, Theorem B.2]

$\underset{p\times 1}{X}$ random vector, $\underset{p\times p}{A}$ constant matrix

$X^TAX = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_i x_j$ is a quadratic form.

(X^T: $1\times p$, A: $p\times p$, X: $p\times 1$; result 1×1)

$\mu = E(X)$, $\Sigma = Cov(X)$

The "trace formula": $Cov(X)$

$$E(X^TAX) = tr(A\Sigma) + \mu^TA\mu$$

Ex: V2014 P1a

$p=3$, $\mu = \begin{bmatrix}1\\1\\1\end{bmatrix}$, $\Sigma = I$, $A = \begin{bmatrix}\frac{2}{3} & -\frac{1}{3} & -\frac{1}{3}\\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3}\\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3}\end{bmatrix}$

$$E(X^TAX) = tr(\underbrace{AI}_{A}) + \underbrace{(1\,1\,1)A\begin{bmatrix}1\\1\\1\end{bmatrix}}_{0}$$

$$= 0 + (1\,1\,1)\begin{bmatrix}0\\0\\0\end{bmatrix} = \underline{\underline{2}}$$

$\underset{n\times n}{I} - \frac{1}{n}11^T = \begin{bmatrix}1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n}\\ -\frac{1}{n} & 1-\frac{1}{n} & & \\ \vdots & & \ddots & \\ -\frac{1}{n} & & & 1-\frac{1}{n}\end{bmatrix}$

(11^T: $n\times 1$ · $1\times n$)

Distribution of quadratic form

$\underset{p\times 1}{X} \sim N_p(0, I)$ ← $X_1,\dots,X_p$ are independent

R is symmetric and idempotent with $rank(R) = r$.

Result: $X^TRX \sim \chi^2_r$

Proof via $R = P\Lambda P^T$

X^TRX = sum of r $N(0,1)^2 \sim \chi^2_r$

In Comp Ex 1.12c: $X \sim N_n(\mu 1, \sigma^2 I)$ ⇓ $N_n(0, I)$

Symmetric and idempotent matrices

A is symmetric; $A = A^T$

A is idempotent; $AA = A$

* The eigenvalues of A are 0 and 1

$$Ae = \lambda e$$
$$\underbrace{AA}_{A}e = A\lambda e$$
$$Ae = \lambda\underbrace{Ae}_{\lambda e} = \lambda^2 e$$

only 0 and 1 possible

* The rank of a matrix is the number of linearly independent rows, and also given as the number of nonzero eigenvalues.

* General rule: $tr(A) = \sum_{i=1}^{n}\lambda_i$, since $\lambda_i = 0$ or 1, $tr(A)$ must be the number of nonzero eigenvalues.

⇒ $tr(A) = rank(A)$ for symmetric and idempotent A.

Now: example = the centering matrix. (RecEx1. P4, CompEx1. 2a)

Finally: Ratio of quadratic forms (more in Part 2-3)

* $X \sim N_p(0, I)$
* R and S symmetric and idempotent with $rank(R) = r$, $rank(S) = s$
* $RS = 0$

Result: a) $X^TRX \sim \chi^2_r$, $X^TSX \sim \chi^2_s$ and independent.

Because: RX and SX are indep if $R\,Cov(X)\,S^T = 0$

We have $Cov(X) = I$ and $S = S^T$ so

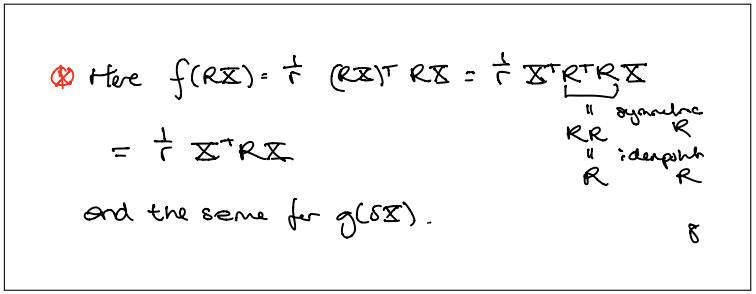
$RIS^T = RS = 0$ ⇒ yes!

Then a function of RX $f(RX)$ and function of SX $g(SX)$ are also independent ⇒ ⊛ see next page for a comment

X^TRX will be indep of X^TSX

b) $$\frac{\frac{X^TRX}{r}}{\frac{X^TSX}{s}} \sim \frac{\frac{\chi^2_r}{r}}{\frac{\chi^2_s}{s}} \sim F_{r,s}$$

This result will be used a lot in Part 2+3.



Multivariate normal pdf

The probability density function is $(\frac{1}{2\pi})^{\frac{p}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\{-\frac{1}{2}Q\}$ where Q is

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

B $(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu})^T$

C $\boldsymbol{\Sigma}-\boldsymbol{\mu}$

Trivariate normal pdf

What graphical form has the solution to $f(\boldsymbol{x}) = \text{constant}$?

A Circle

B Parabola

C Ellipsoid

D Bell shape

Multivariate normal distribution

$\boldsymbol{X}_p \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{C}$ is a $k \times p$ constant matrix. $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}$ is

A Chi-squared with k degrees of freedom

B Multivariate normal with mean $k\boldsymbol{\mu}$

C Chi-squared with p degrees of freedom

D Multivariate normal with mean $\boldsymbol{C}\boldsymbol{\mu}$

Independence

Let $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 3 & 1 \\ 0 & 2 & 5 \end{bmatrix}$. Which two variables are independent?

A X_1 and X_2

B X_1 and X_3

C X_2 and X_3

D None – but two are uncorrelated.

Constructing independent variables?

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. How can I construct a vector of independent standard normal variables from $\boldsymbol{X}$?

A $\boldsymbol{\Sigma}(\boldsymbol{X} - \boldsymbol{\mu})$

B $\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} + \boldsymbol{\mu})$

C $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})$

D $\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{X} + \boldsymbol{\mu})$

Conditional distribution: mean

$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is a bivariate normal random vector. What is true for the conditional mean of X_2 given $X_1 = x_1$?

A Not a function of x_1

B A linear function of x_1

C A quadratic function of x_1

Conditional distribution: variance

$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is a bivariate normal random vector. What is true for the conditional variance of X_2 given $X_1 = x_1$?

A Not a function of x_1

B A linear function of x_1

C A quadratic function of x_1

Estimator for mean

$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. What is the distribution of the estimator $\bar{\boldsymbol{X}}$ for the mean?

A $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

B $N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$

C χ^2_p

D χ^2_n

Unbiased estimators

$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample of size n of a p-dimensional random vector. An unbiased estimator for the covariance matrix $\boldsymbol{\Sigma}$ is.

A $\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^T$

B $\frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^T$

C $\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^T(\boldsymbol{X}_j - \bar{\boldsymbol{X}})$

D $\frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^T(\boldsymbol{X}_j - \bar{\boldsymbol{X}})$

Distribution of quadratic form

$\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{R}$ is a symmetric and idempotent matrix with rank r. What is the distribution of $\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}$?

A $N_p(\boldsymbol{\mu}, r\boldsymbol{I})$

B $N_r(\boldsymbol{0}, \boldsymbol{I})$

C χ^2_r

D χ^2_p

Correct?

Are you sure you want to read the correct answers? Maybe try first? The answers are explained on the next two slides.

Answers

1. A: exponent quadratic form is $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$.
2. C: contours are ellipsoids in general. In two dimensions we have ellipses. For two dimensions and equal variance and correlation 0 we have circles.
3. D: linear combinations of mvN are also mvN.
4. B: $\mathrm{Cov}(X_1, X_3) = 0$ and X_1 and X_3 are thus independent.
5. C: The Mahlanobis transform is $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{\mu})$.

Answers

6. B: Conditional mean is linear in x_1, which will be very useful when we start with multiple linear regression.
7. A: Conditional variance (covariance) is not a function of x_1.
8. B: The mean is also mvN with mean $\boldsymbol{\mu}$ and covariance $\frac{1}{n}\boldsymbol{\Sigma}$.
9. B: $\frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{X}_j-\bar{\boldsymbol{X}})(\boldsymbol{X}_j-\bar{\boldsymbol{X}})^T$ is the unbiased estimator for $\boldsymbol{\Sigma}$. Observe the $(n-1)$ and that the dimension is $p \times p$ (to place the transpose). Not a quadratic form.
10. C: Quadratic form is related to χ^2.