Statistical Data Analysis

Dr. Jana de Wiljes

8. Februar 2022

Universität Potsdam

Organisational stuff for exam

The exam will take place on the 15th of February 2021 from 12-14 in room 2.27.0.01 . As the exam will start at 12 exactly please ensure to be there at least 15-20 Minutes earlier. You are allowed to bring the following tools to the exam:

- a handwritten DinA4 Sheet (both sides) that contains any information you want on the lecture and other sources
- an old fashioned calculator (non graphing non programmable)
- Further it is required to wear a mask (medical or FFP2) at all times during the exam.

Topics NOT relevant for first exam

- Causal Graphs
- Householder transformation
- Iterative Methods
- Cross Validation
- Generalized Linear Models in particular logistic regression
- Support vector machine
- nonstationary clustering
- Kernighan-Lin algorithm
- Golub-Reinsch algorithm

Continuous Random Variables

Normal Distribution

A normal or Gaussian distributed random variable $X : \Omega \to \mathbb{R}$ with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ has the following density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}$$

and expected value and variance

$$\mathbb{E}[X] = \mu$$
$$Var(X) = \sigma^2$$

$X \sim \mathcal{N}(\mu, \sigma)$

Normal Distribution

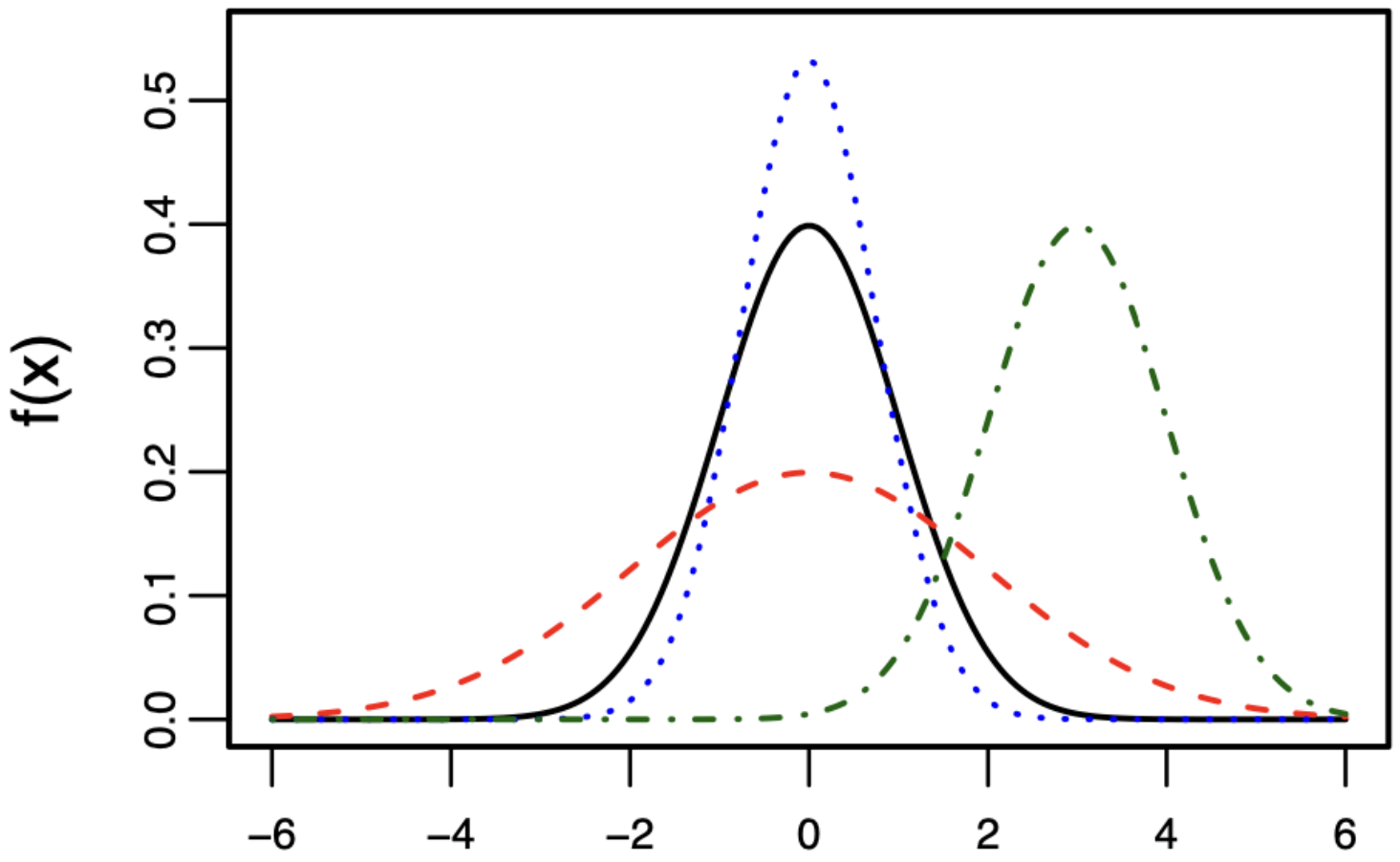


Abbildung 1: $\mu = 0$, $\sigma = 1$ (black), $\mu = 0$, $\sigma = 2$ (red), $\mu = 0$, $\sigma = 0.75$ (blue) and $\mu = 3$, $\sigma = 1$ (green)

Normal Distribution

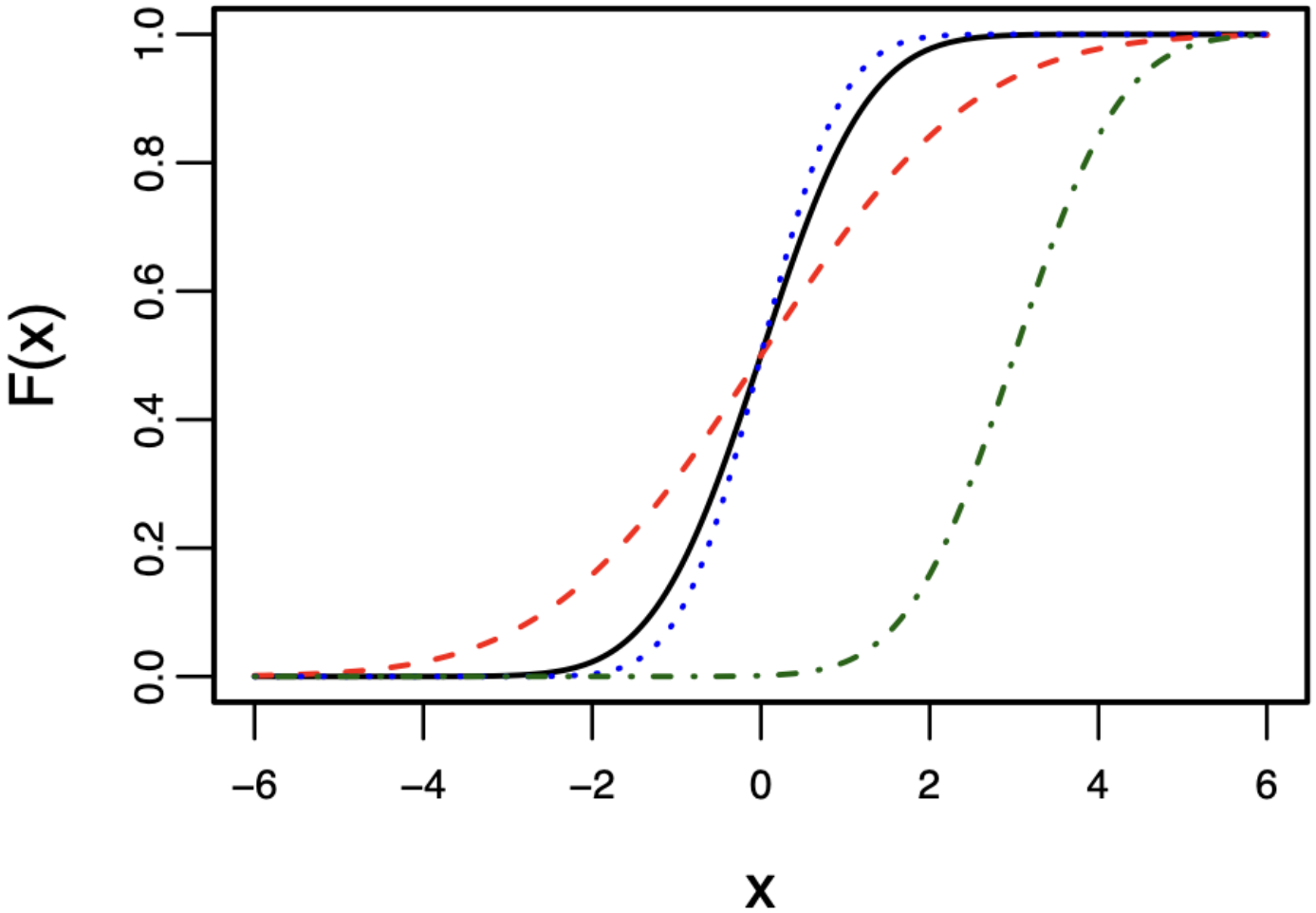


Abbildung 2: $\mu = 0$, $\sigma = 1$ (black), $\mu = 0$, $\sigma = 2$ (red), $\mu = 0$, $\sigma = 0.75$ (blue) and $\mu = 3$, $\sigma = 1$ (green)

Quantile

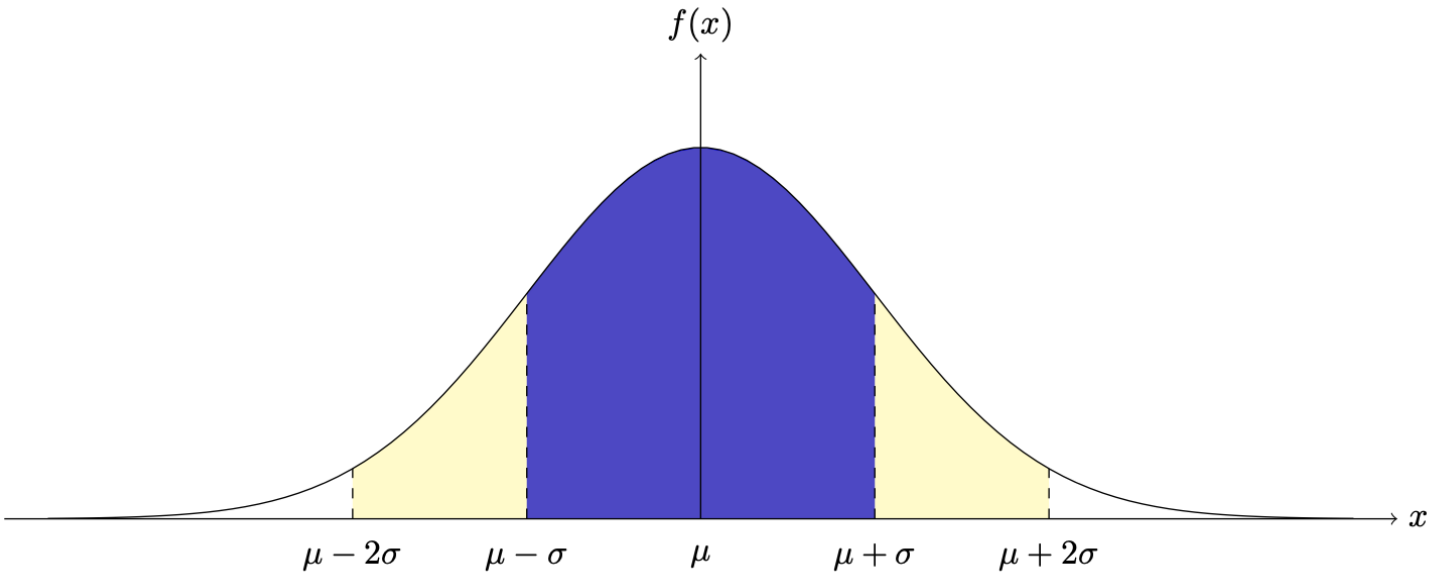


Abbildung 3: 60% of area under the curve (colored in blue) are in the $[\mu - \sigma, \mu + \sigma]$ interval and 95% of the area under the curve are in the interval $[\mu - \sigma, \mu + \sigma]$.

Standard normal distribution

A variable $X : \Omega \to \mathbb{R}$ follows a standard normal distribution, i.e., $X \sim \mathcal{N}(0, 1)$ if the associated density has the following form

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\left(\frac{x^2}{2}\right)\right\}$$

with the associate cumulative distribution

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)du \tag{1}$$

and quantile

$$z_\alpha = \Phi^{-1}(\alpha), \quad \alpha \in (0, 1) \tag{2}$$

Relationship between standard normal distribution and Normal distribution

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \tag{3}$$

Exponential Distribution

A random variable $X : \Omega \to \mathbb{R}$ follows the exponential distribution with parameters $\lambda > 0$ has the following density and cdf

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda \exp(-\lambda x) & x \geq 0 \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp(-\lambda x) & x \geq 0 \end{cases}$$

and expected value and variance

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

Notation: $X \sim \text{Exp}(\lambda)$ (often used for waiting times and lifetimes)

Exponential Distribution

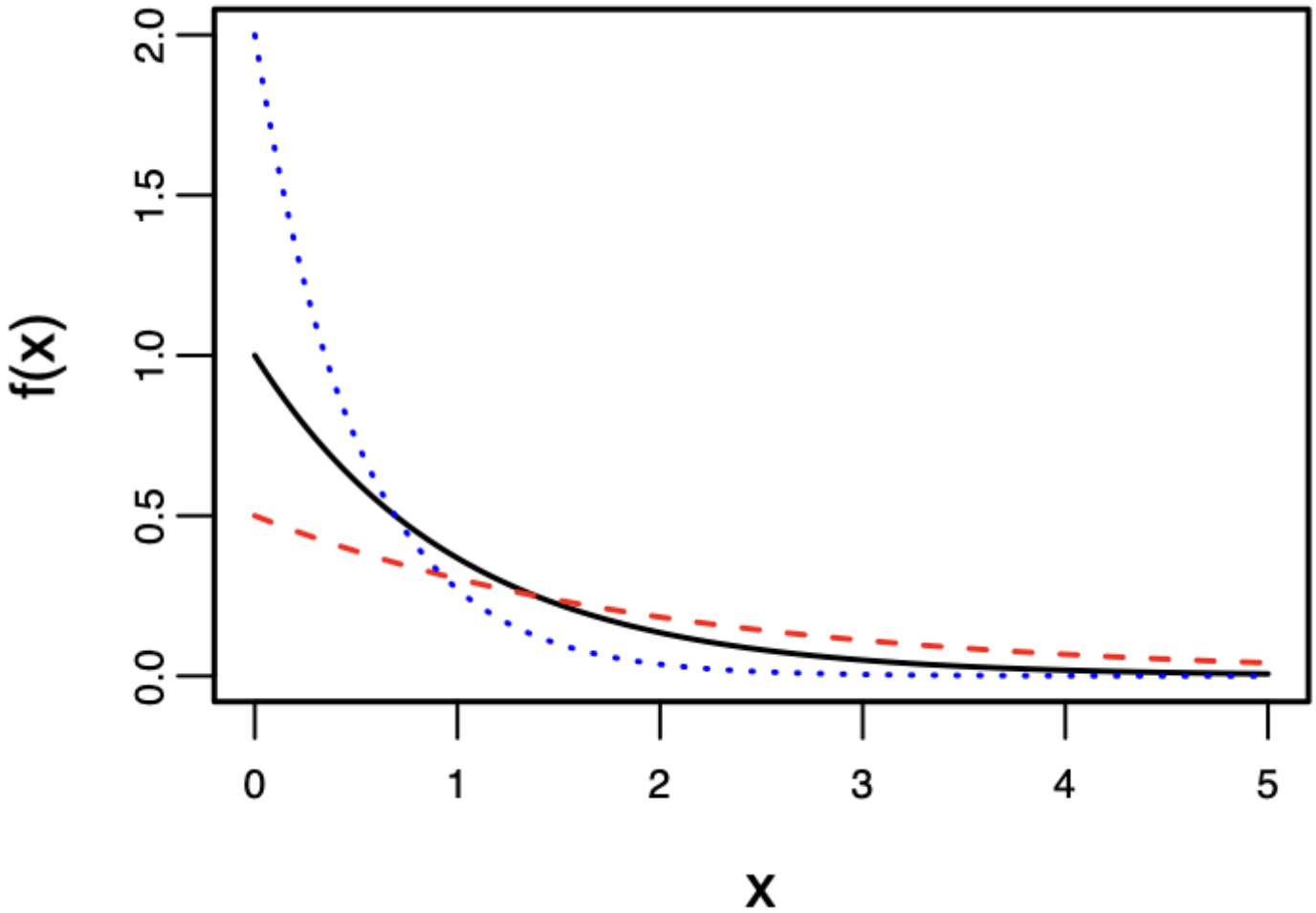


Abbildung 4: $\lambda = 1$ (black), $\lambda = 2$ (blue) and $\lambda = 1/2$ (red).

Exponential Distribution

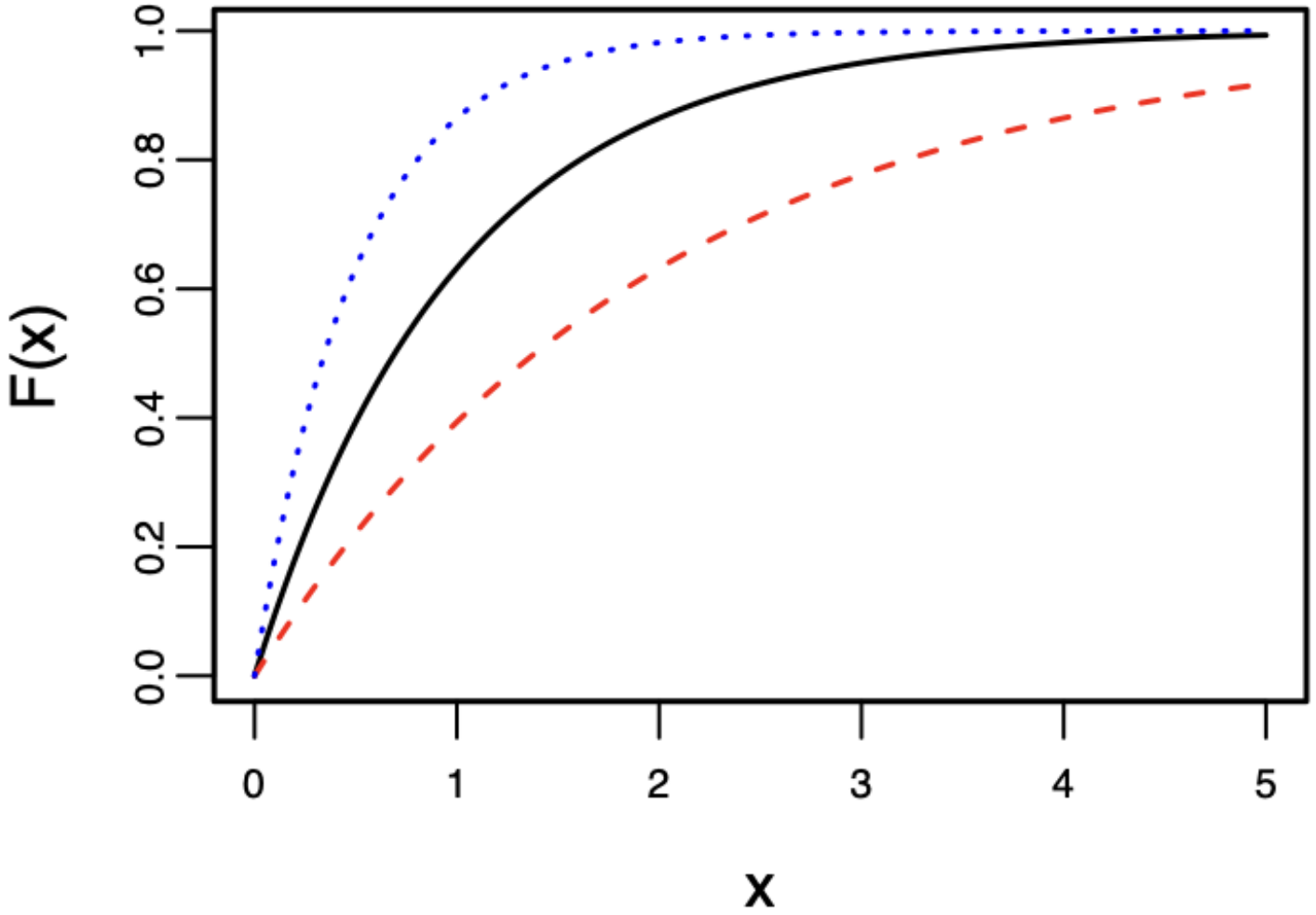


Abbildung 5: $\lambda = 1$ (black), $\lambda = 2$ (blue) and $\lambda = 1/2$ (red).

Example

Setting: The lifetime T of a computer chip is exponentially distributed, i.e., $T \sim \text{Exp}(\lambda)$ with expected lifetime of 15 weeks, i.e., parameter $\lambda = \frac{1}{15}$

Question:

- What is the probability that the computer chip is defect within the first 10 weeks?

- What is the probability that the computer chip will last at least 20 weeks?

Transformation

Reminder: for arbitrary g the following holds:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{4}$$

Proposition: Let g be a differentiable, strictly monoton function

and X and random variable. Then $Y = g(X)$ has the following density

$$f_Y(y) = \left| \frac{1}{g'(g^{-1})(y)} \right| f_X(g^{-1}(y)), y \in E_Y \tag{5}$$

E_Y is given by the value space of X via

$$E_Y = g(E_X) = \{g(x) : x \in E_X\} \tag{6}$$

Jensen's inequality

Proposition: Let g be a convex function and X random variable

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \tag{7}$$

Example:

Samples

Definition: Let $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and $X_1, \ldots, X_n$ be associated random variables. Realizations

$$x_1 := X_1(\omega), \ldots, x_n := X_n(\omega) \tag{8}$$

are referred to as *samples* and n the sample size.

Estimator

Definition: A measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is referred to as *sample function*, *estimator* or *statistic*.

Note: we will also consider the composition:

$$\varphi(X) : \Omega \to \mathbb{R}^m \tag{9}$$

$$\omega \mapsto \varphi(X_1(\omega), \ldots, X_n(\omega)) \tag{10}$$

Sample estimation

Given: $(x_1, \ldots, x_n) \in \mathbb{R}^n$ of independent and identical random variables $X_1, \ldots, X_n$ where

$$F(t) = \mathbb{P}[X_i \leq t], \quad t \in \mathbb{R} \tag{11}$$

but **unknown**

Goal: estimate $\mathbb{E}[X_i]$ or $Var[X_i]$

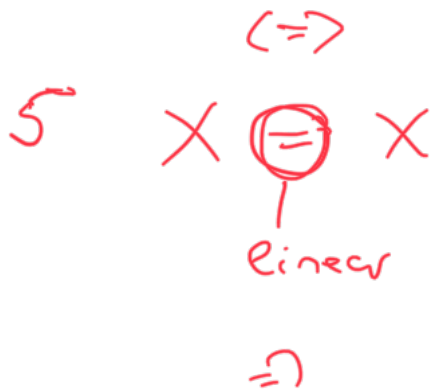
Empirical mean

Definition: The empirical mean is defined by

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{12}$$

Note: we will also use an analog notation for the random variables:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{13}$$



Random variables

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[\bar{X}_n] = \mu \text{ and } Var[\bar{X}_n] = \frac{\sigma^2}{n} \tag{14}$$

Law of large numbers

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$. Then

$$\bar{X}_n \to \mu \text{ for } n \to \infty \text{ (almost sure)} \tag{15}$$

Empirical variance

Definition: The empirical variance is defined by

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \tag{16}$$

Note: we will also use an analog notation for the random variables:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \tag{17}$$

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables. Then

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^2 - n\bar{X}_n^2) \tag{18}$$

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[S_n^2] = \sigma^2 \tag{19}$$

Random variables

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[\bar{X}_n] = \mu \text{ and } Var[\bar{X}_n] = \frac{\sigma^2}{n} \tag{20}$$

Law of large numbers

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$. Then

$$\bar{X}_n \to \mu \text{ for } n \to \infty \text{ (almost sure)} \tag{21}$$

Empirical variance

Definition: The empirical variance is defined by

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \tag{22}$$

Note: we will also use an analog notation for the random variables:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \tag{23}$$

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables. Then

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^2 - n\bar{X}_n^2) \tag{24}$$

Proof

Proof

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[S_n^2] = \sigma^2 \tag{25}$$

Empirical standard deviation

Def: The empirical standard deviation is defined by

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \tag{26}$$

Order statistic

Def: Let $(x_1, \ldots, x_n) \in \mathbb{R}^n$ be a sample set. One can order the elements in an increasing manner:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \tag{27}$$

Then $x_{(i)}$ is referred to as the i-th order statistic of the sample set.

Sample median

Def: The sample median of a set of samples if given by

$$\text{Med}_n = \text{Med}_n(x_1, \ldots, x_n) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \textit{in case } n \textit{ is uneven} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{in case } n \text{ is even} \end{cases}$$

Then $x_{(i)}$ is referred to as the i-th order statistic of the sample set.

Example

Truncated mean

Def: The truncated mean samples $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined by

$$\frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

Empirische α-Quantile

Def: Let $(x_1, \ldots, x_n) \in \mathbb{R}^n$ be a set of samples and $\alpha \in (0, 1)$. The empirical α Quantil is defined by

$$q_\alpha = \begin{cases} x_{\lfloor n\alpha \rfloor + 1} & \text{falls } n\alpha \notin \mathbb{N} \\ \frac{1}{2}\left(x_{\lfloor n\alpha \rfloor} + x_{\lfloor n\alpha \rfloor + 1}\right) & \text{falls } n\alpha \in \mathbb{N} \end{cases}$$

Distibution of the order statistic

Proposition: Let $X_1, X_2, \ldots, X_n$ be independent and identical distributed random variables, that are absolute continuous with a density f and cumulative distibution function F. Let

$$X_{(1)} \leq X_{(2)} \leq, \ldots, \leq X_{(n)} \tag{28}$$

be the order statistics. Then the density of the random variable $X_{(i)}$ is

$$f_{X_{(i)}}(t) = \frac{n!}{(i-1)!(n-1)!} f(t) F(t)^{(i-1)} (1 - F(t))^{n-i} \tag{29}$$

Beta distribution

Def: For a and b larger than zero and

$$f(x) = \frac{1}{\mathrm{B}(a,b)} x^{a-1}(1-x)^{b-1}.$$

where the normalization is given by

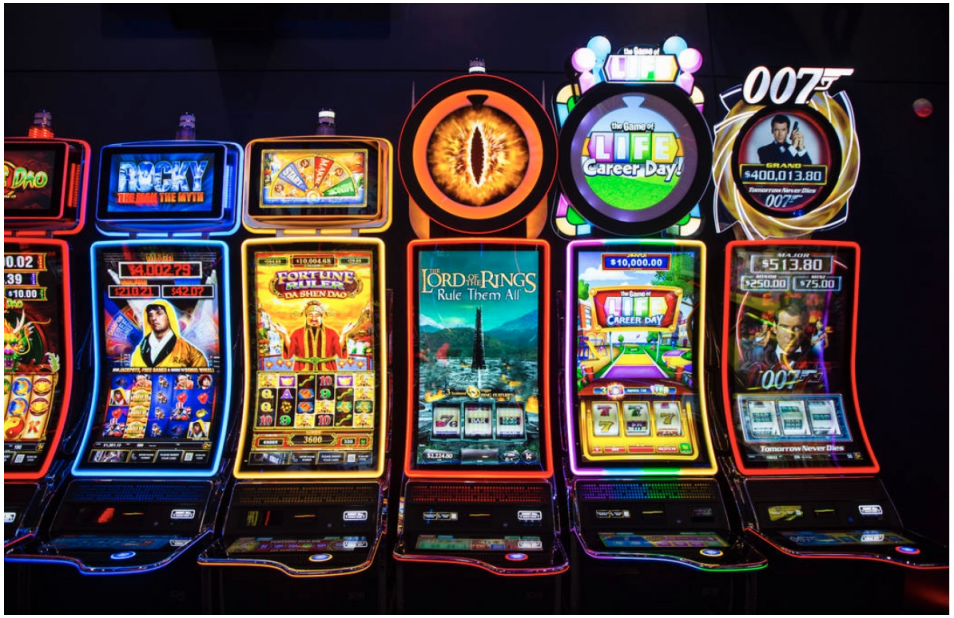
$$\mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$$

with $\Gamma(n) = (n-1)!$ being the gamma function.

Excurse to Bandits

Multi-armed bandits

Choose from K options
to receive a
high reward and
to educe loss after T rounds



Examples:

- Which advertising campaign generates the largest revenue
- Which restaurant to pick ?
- Which netflix series to streamen?
- Which vaccination should be further developed ?

Multi-armed bandits

A stochastic K-Armed Bandit is defined via the tuple $\langle \mathcal{A}, \mathcal{Y}, P, r \rangle$ where

- $\mathcal{A}$ is the set of actions (arms) and $|\mathcal{A}| = K$
- $\mathcal{Y}$ is the set of possible outcomes
- $P(\cdot|a) \in \mathcal{P}(\mathcal{Y})$ is the outcome probability, conditioned on action $a \in \mathcal{A}$ being taken,
- $r(\mathcal{Y}) \in \mathcal{R}$ represents the reward obtained when outcome $Y \in \mathcal{Y}$ is observed

Regret

Def: Let $a^* \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_{y \sim P(\cdot|a)}[r(y)]$ denote the optimal arm. The T-period regret of the sequence of actions $a_1, \ldots, a_T$ is the random variable

$$\textbf{Regret}(T) = \sum_{t=1}^{T} \Big[r(Y_t(a^*)) - r(Y_t(a_t)) \Big] \tag{30}$$

Thompson Sampling

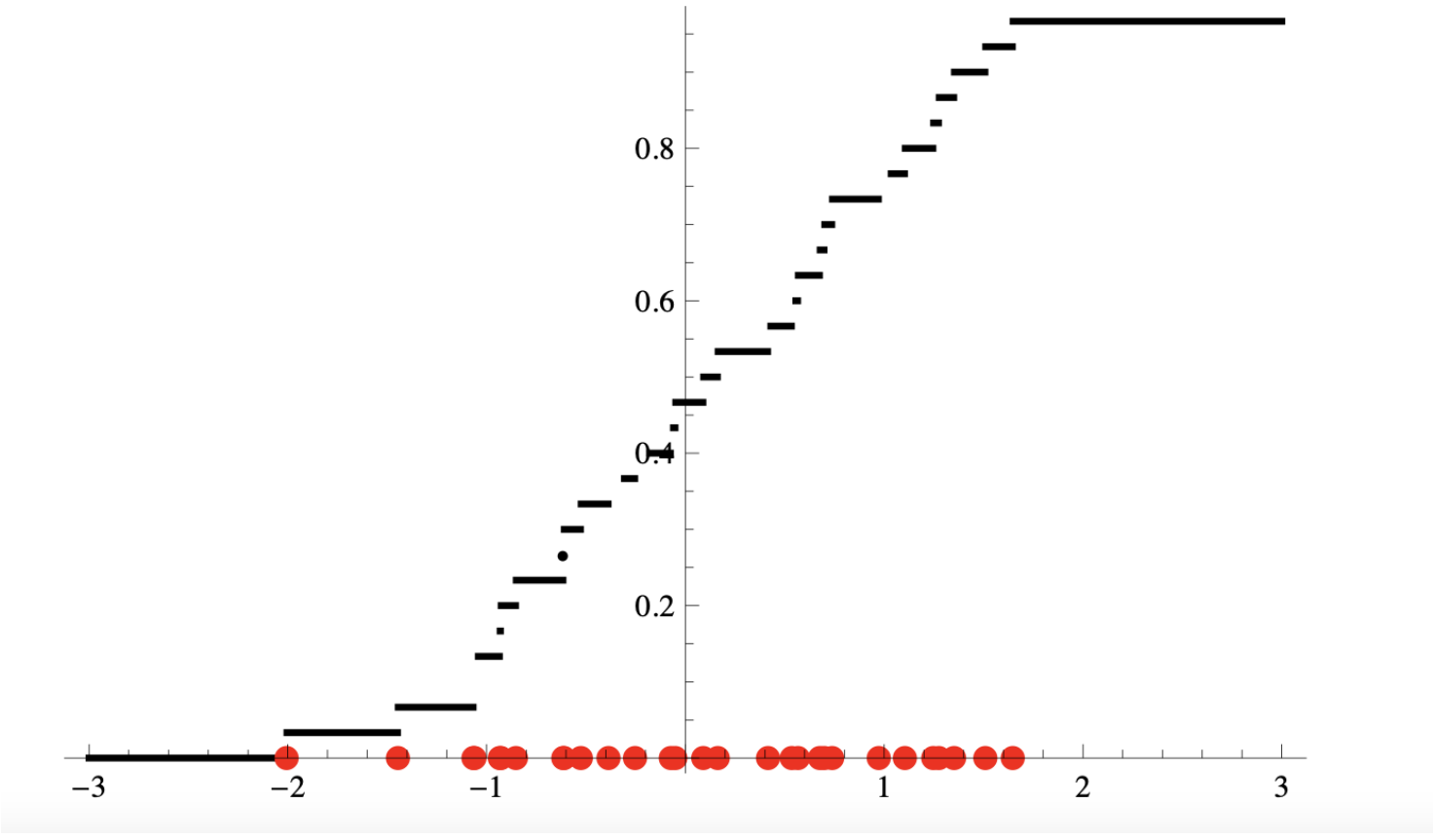
- **Problem setting:** Choose from K options to receive a high reward
- **Algorithm:** Iterated over the following steps:
  1. In each round save information on the choice of action and if a reward was received
  2. Draw from the beta distribution: defined via for each action by
     - a) how often performing action resulted in a reward
     - b) how often performing action did not resulted in a reward
  3. Choose the action that has the highest beta function value

Empirical cdf

The empirical cdf of a sample set $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined through

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i \leq 1} = \frac{1}{n} \#\{i \in \{1, \ldots, n\} : x_i \leq t\}, \quad t \in \mathbb{R} \tag{31}$$

Empirical cdf

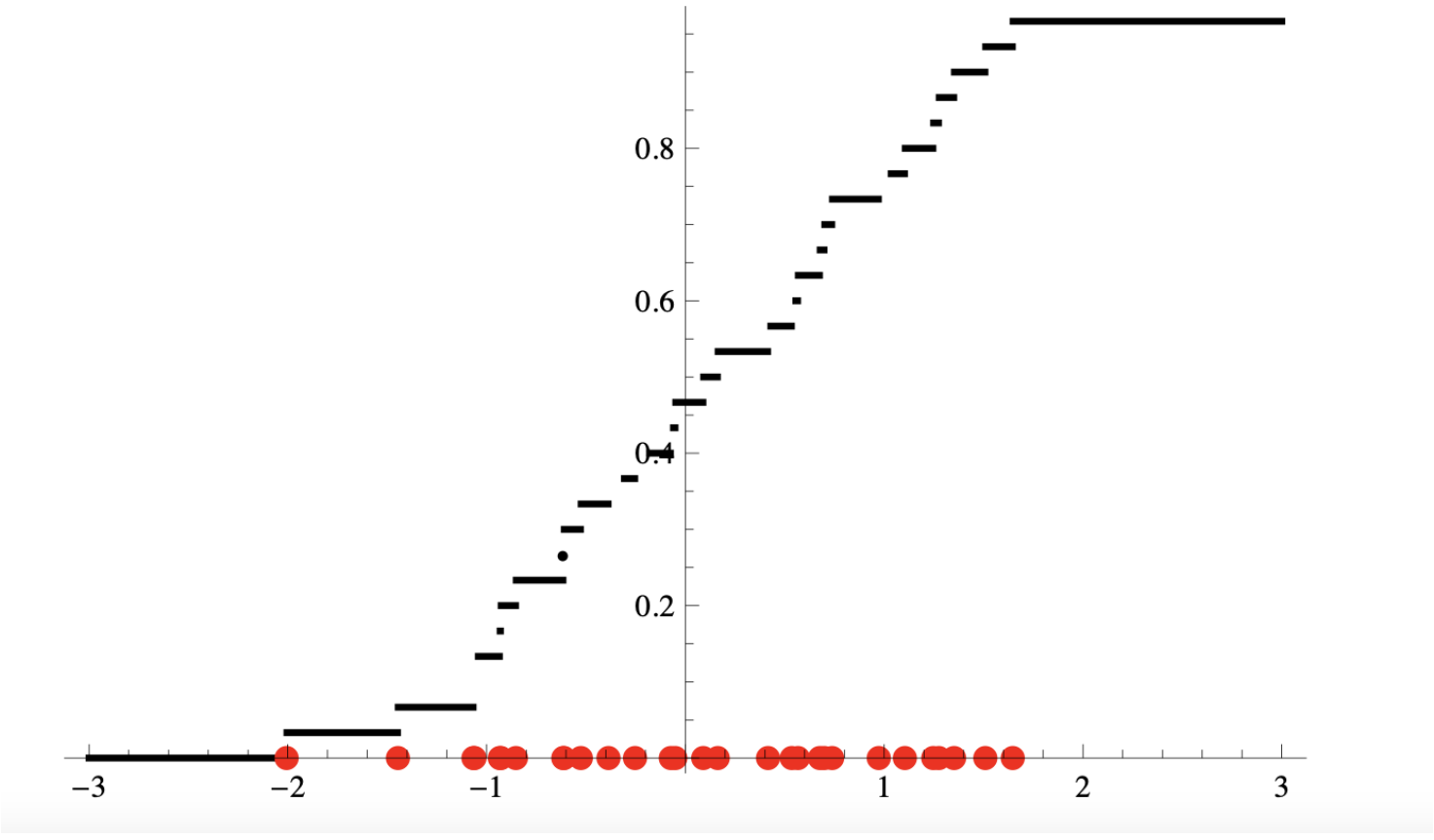


Empirical cdf of a sample set

The empirical cdf of a sample set $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined through

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i \leq 1} = \frac{1}{n} \#\{i \in \{1, \ldots, n\} : x_i \leq t\}, \quad t \in \mathbb{R} \tag{32}$$

Empirical cdf



Empirical cdf

The empirical cdf of a sample set $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined through

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq 1} \tag{33}$$

Proposition

Proposition: Let $(X_1, X_2 \ldots..)$ independent and identical distributed random variabel with cdf F. Then

- $$n\widehat{F}_n(t) \sim Bin(n, F(t)). \tag{34}$$

  This means
  $$\mathbb{P}\Big[\widehat{F}_n(t) = \frac{k}{n}\Big] = \binom{n}{k} F(t)^k (1 - F(t))^{n-k}, \quad k = 0, 1, \ldots, n.$$

- The expect value and variance of $\widehat{F}_n(t)$ are given by

  $$\mathbb{E}[\widehat{F}_n(t)] = F(t), \quad \text{Var}[\widehat{F}_n(t)] = \frac{F(t)(1 - F(t))}{n} \tag{36}$$

  i.e., $\widehat{F}_n(t)$ is an unbiased estimator of F(t).

- For all $t \in \mathbb{R}$ it holds that
  $$\widehat{F}_n(t) \to F(t) \quad n \to \infty \text{ almost everywhere} \tag{37}$$

- For all $t \in \mathbb{R}$ with $F(t) \neq 0$ or 1 the following holds:

  $$\sqrt{n}\frac{\widehat{F}_n(t) - F(t)}{\sqrt{F(t)(1 - F(t))}} \to \mathcal{N}(0, 1) \text{ for } n \to \infty \text{ (in distribution)} \tag{38}$$

Theoretical distribution

Def: Let X be a random. The theoretical distribution of X is a probability measure μ on $(\mathbb{R}, \mathcal{B})$ with

$$\mu(A) = \mathbb{P}[X \in A] \text{ for every Borel set A} \subset \mathbb{R} \tag{39}$$

Note: the relationship between the theoretical distribution μ and the theoretical cdf F is;

$$F(t) = \mu((-\infty, t]), \quad t \in \mathbb{R} \tag{40}$$

Empirical distribution

Def: The empirical distribution of a sample set $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined through

$$\widehat{\mu}_n(A) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i \in A} = \frac{1}{n} \#\{i \in \{1, \ldots, n\} : x_i \in A\}, \qquad (41)$$

for every Borel set $A \subset \mathbb{R}$

Dirac δ measure

Def: Let $x \in \mathbb{R}$ be a real number. The dirac-δ measure δ_x is a probability measure on $(\mathbb{R}, \mathcal{B})$ with with

$$\delta_x(A) \begin{cases} 1, & \text{for } x \in A \\ 0, & \text{for } x \notin A \end{cases} \tag{42}$$

for all Borel set $A \subset \mathbb{R}$

Remark: Then the empirical measure $\widehat{\mu}_n$ can be written as

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \tag{43}$$

and further note that

$$\widehat{F}_n(t) = \widehat{\mu}_n((-\infty, t]) \tag{44}$$

Proposition

Proposition: Let $(X_1, X_2 \ldots..)$ independent and identical distributed random variabel with distribution μ and let $A \subset \mathbb{R}$ a Borel set. Then

- $$n\widehat{\mu}_n(A) \sim Bin(n, \mu(A)). \tag{45}$$
- The expect value and variance of $\widehat{\mu}_n(A)$ are given by

$$\mathbb{E}[\widehat{\mu}_n(A)] = \mu(A), \quad \text{Var}[\widehat{\mu}_n(A)] = \frac{\mu(A)(1-\mu(A))}{n} \tag{46}$$

  i.e., $\widehat{\mu}_n(A)$ is an unbiased estimator of $\mu(A)$.
- Further it follows that $\widehat{\mu}_n$ is a consistent estimator, i.e.,

$$\widehat{\mu}_n(A) \to \mu(A) \quad n \to \infty \text{ almost everywhere} \tag{47}$$

- For $\mu(A) \neq 0$ or 1 the following holds:

$$\sqrt{n}\frac{\widehat{\mu}_n(A) - \mu(A)}{\sqrt{\mu(A)(1-\mu(A))}} \to \mathcal{N}(0,1) \text{ for } n \to \infty \text{ (in distribution)} \tag{48}$$

Plugin Estimator

Setting: Let $(X_1, \ldots, X_n)$ be independent and identical distributed random variables with the distribution μ. Further we assume that a reliasation $(x_1, \ldots, x_n) \in \mathbb{R}^n$ of the respective random variables

Goal: approximate $\Psi(\mu)$ where $\Psi : \mathcal{M} \to \mathbb{R}$

Def: $\Psi(\widehat{\mu}_n)$ is called the plugin estimator of $\Psi(\mu)$.

Kolmogorov-distance

Def: The Kolmogorov-distance between the empirical cdf $\widehat{F}_n(t)$ and the theoretical cdf F is defined as follows

$$D_n := \sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F(t)| \tag{49}$$

Theorem of Gliwenko-Cantelli

Theorem: For the Kolmogorov-distance D_n the following holds

$$D_n \to 0 \text{ for } n \to \infty \text{ almost everywhere} \tag{50}$$

i.e.,

$$\mathbb{P}\left[\lim_{n\to} D_n = 0\right] = 1 \tag{51}$$

Kolmogorov-distance

Def: The Kolmogorov-distance between the empirical cdf $\widehat{F}_n(t)$ and the theoretical cdf F is defined as follows

$$D_n := \sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F(t)| \tag{52}$$

Theorem of Gliwenko-Cantelli

Theorem: For the Kolmogorov-distance D_n the following holds

$$D_n \to 0 \text{ for } n \to \infty \text{ almost everywhere} \tag{53}$$

i.e.,

$$\mathbb{P}\left[\lim_{n\to} D_n = 0\right] = 1 \tag{54}$$

A statistical model

Def: A statistical model is a triple $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ where

- $\mathcal{X}$ is the sample space
- $\mathcal{A} \subset 2^{\mathcal{X}}$ is a σ-algebra on $\mathcal{X}$
- Θ is the parameter space
- for every $\theta \in \Theta$ $\mathbb{P}_\theta$ is a probability measure on $(\mathcal{X}, \mathcal{A})$

Estimator

Def: Let $\Theta \subset \mathbb{R}^r$. A estimator is a measurable map

$$\widehat{\theta} : \mathcal{X} \to \Theta, \quad x \mapsto \widehat{\theta}(x) \tag{55}$$

Maximum-Likelihood estimator

Def: The Maximum-Likelihood estimator is defined via

$$\widehat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} L(\theta) \tag{56}$$

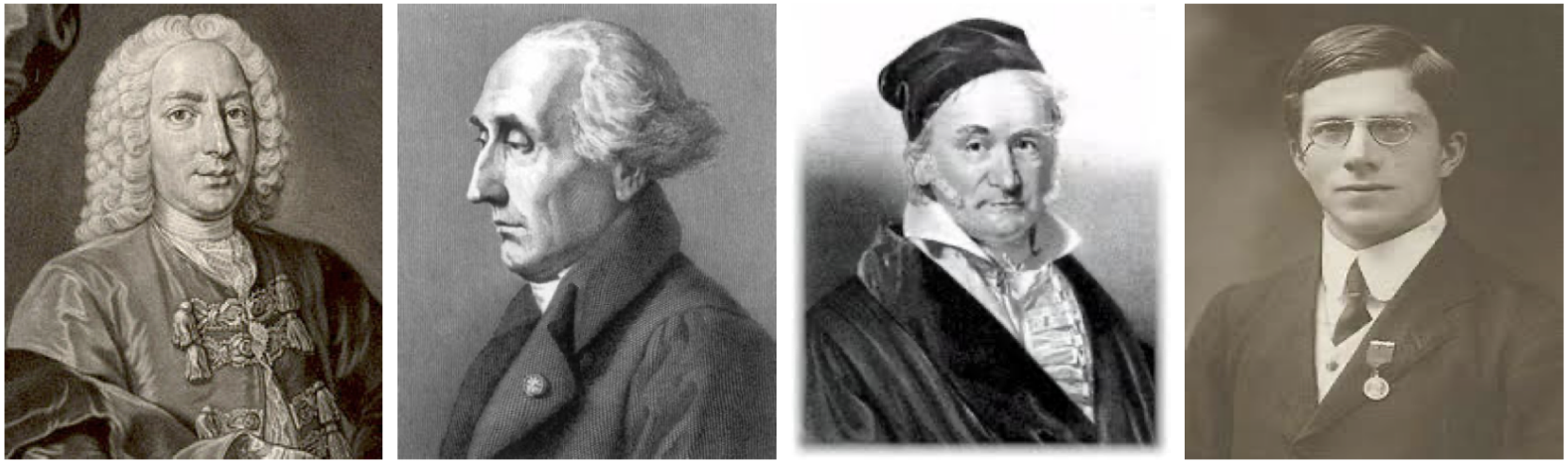


Abbildung 6: Daniel Bernoulli, Joseph-Louis Lagrange, Carl-Friedrich Gau and Ronald Fisher

A-posteriori-distribution

Def: The a-posteriori-distribution of θ is the conditional distribution given the information $X_1 = x_1, \ldots, X_n = x_n$, i.e.,

$$q(\theta_i | x_1, \ldots, x_n) := \mathbb{P}[\theta = \theta_i | X_1 = x_1, \ldots, X_n = x_n], \quad i = 1, 2, \ldots$$

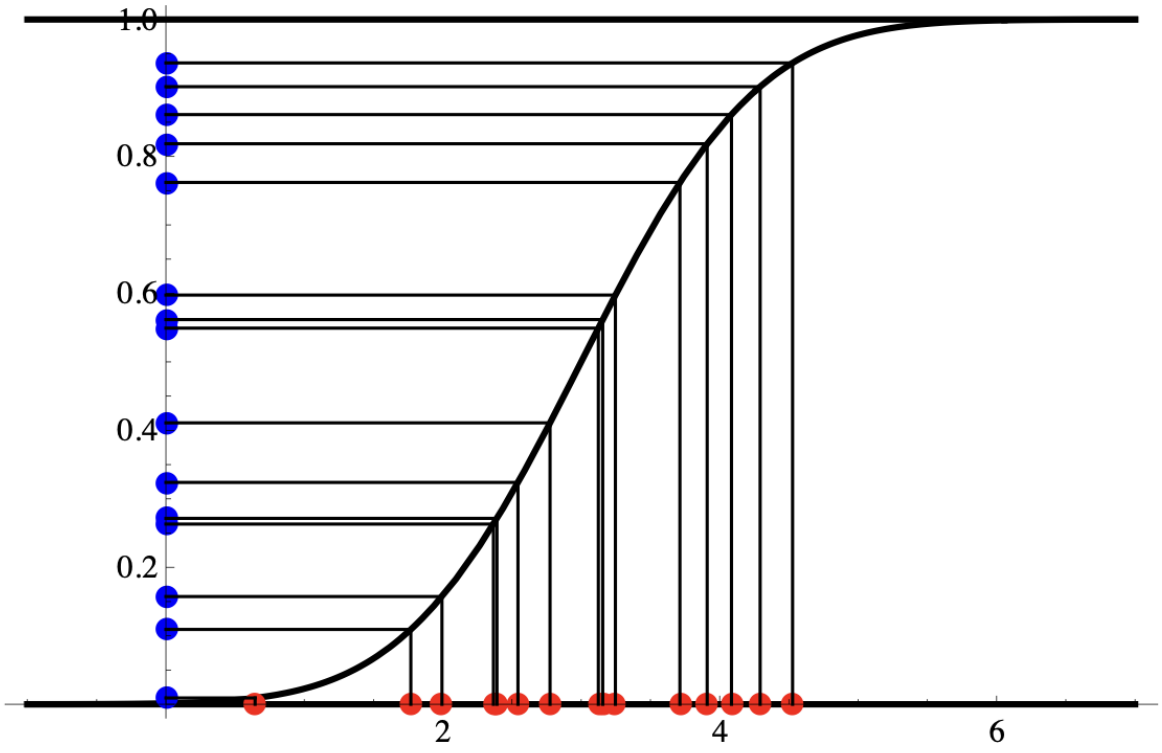
Bayes method

Def: The Bayes estimator is defined as the expectation of the a-posteriori-distribution

$$\widehat{\theta}_{\text{Bayes}} = \sum_i \theta_i q(\theta_i | x_1, \ldots, x_n)$$

Maximum-spacing method

Lemma: Let the cdf F_θ be continuous and strictly monoton increasing. Under $\mathbb{P}_\theta$ the random variables $F_\theta(X_1), \ldots, F_\theta(X_n)$ are independent and uniformly distributed on the $(0, 1)$ interval.



Maximum-spacing method

Lemma: Let $z_1, \ldots, z_k \in [0, 1]$ be number that are subject to the condition $z_1 + \cdots + z_k = 1$. Then

$$z_1 \cdot \cdots \cdot z_k \leq \frac{1}{k^k}. \tag{57}$$

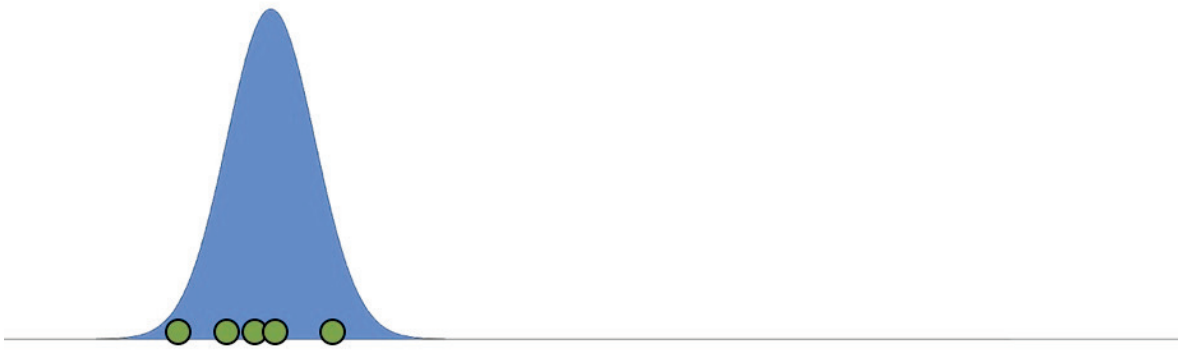
Equality is attained only if all the numbers all the numbers are equal to $\frac{1}{k}$.

Maximum-spacing method

Lemma: The maximum-spacing method is defined via

$$\widehat{\theta}_{MS} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n+1} (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})) \tag{58}$$

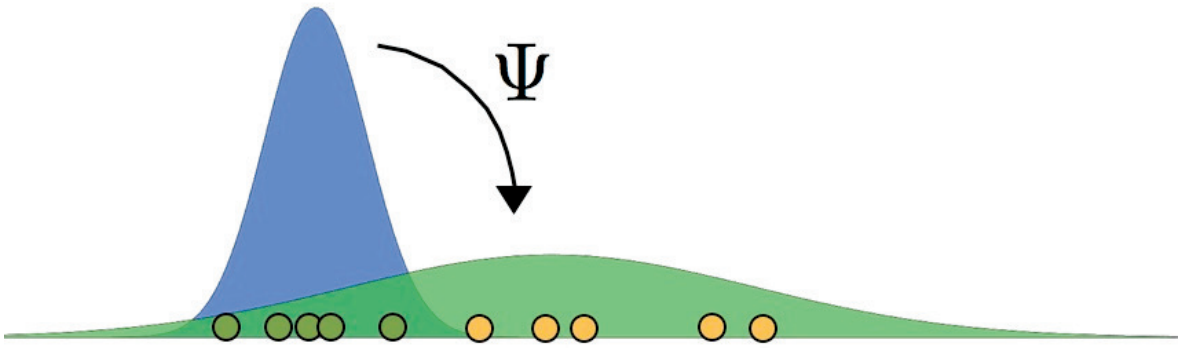
Ensemble Kalman filter



$$\mathcal{N}(m_0, C_0) \text{ with } m_0 \approx \frac{1}{M}\sum_{i=1}^{M} z_0^i$$

$$C_0 \approx \frac{1}{M}\sum_{i=1}^{M}(z_0^i - m_0)(z_0^i - m_0)^\top$$

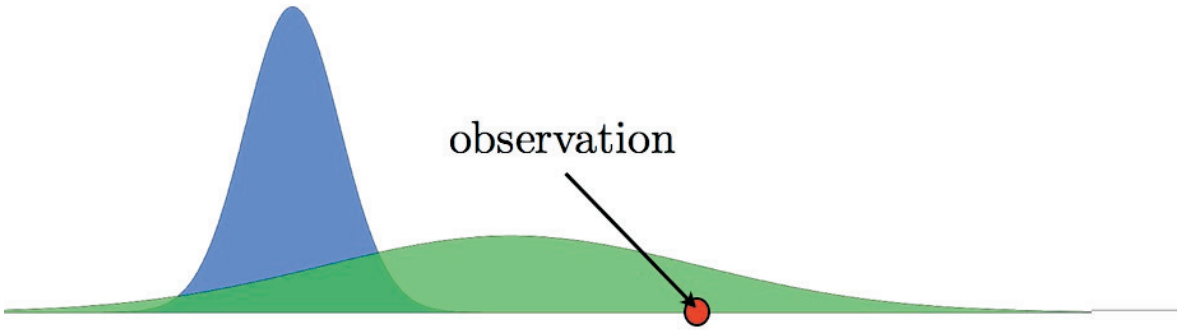
Ensemble Kalman filter



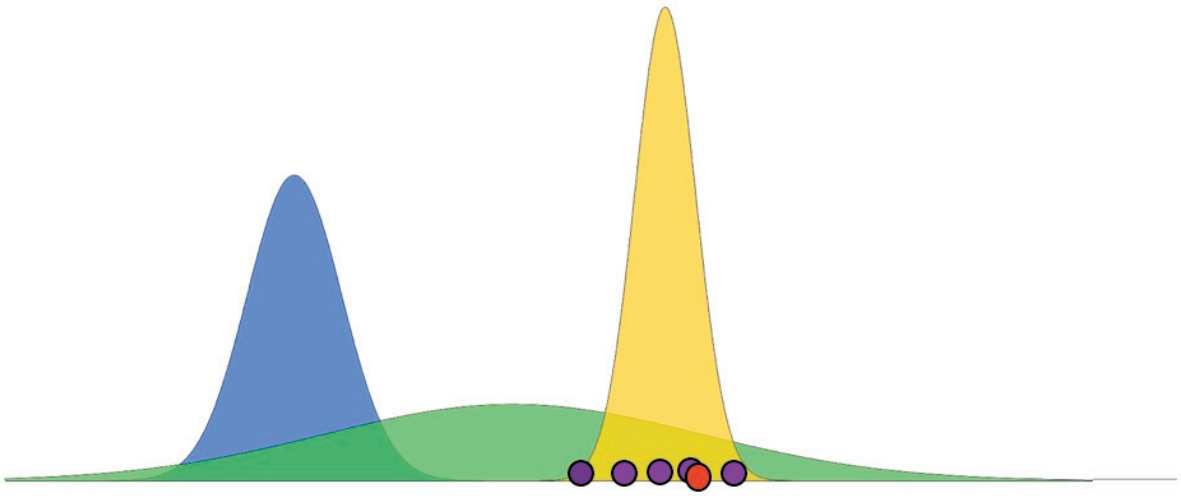
$$\mathcal{N}(\hat{m}_1, \hat{C}_1) \text{ with } \hat{m}_1 \approx \frac{1}{M}\sum_{i=1}^{M} \hat{z}_1^i = \frac{1}{M}\sum_{i=1}^{M} \Psi(z_0^i)$$

$$\hat{C}_1 \approx \frac{1}{M}\sum_{i=1}^{M} (\hat{z}_1^i - \hat{m}_1)(\hat{z}_1^i - \hat{m}_1)^\top$$

Ensemble Kalman filter



Ensemble Kalman filter



$$\mathcal{N}(m_1, C_1) \text{ with } m_1 \approx \frac{1}{M}\sum_{i=1}^{M} z_1^i$$

$$C_1 \approx \frac{1}{M}\sum_{i=1}^{M}(z_1^i - m_1)(z_1^i - m_1)^\top$$

Ensemble Kalman filter

Goal: approximate $\pi(z_n|y_{1:n})$

Ensemble Kalman filter

Goal: approximate $\pi(z_n|y_{1:n})$

Approach: propagate samples $\hat{z}^i_{n+1}$ with Kalman formula

$$z^i_{n+1} = \hat{z}^i_{n+1} - K_{n+1}(H\hat{z}^i_{n+1} - \tilde{y}^i_{n+1})$$

Ensemble Kalman filter

Goal: approximate $\pi(z_n|y_{1:n})$

Approach: propagate samples $\hat{z}^i_{n+1}$ with Kalman formula

$$z^i_{n+1} = \hat{z}^i_{n+1} - K_{n+1}(H\hat{z}^i_{n+1} - \tilde{y}^i_{n+1})$$

Need: perturbed observations

$$\tilde{y}^i_{n+1} = y_{n+1} + \epsilon^i_{n+1}$$

with $\epsilon^i_{n+1} \sim \mathcal{N}(0, R)$ i.i.d. to get the correct mean and covariance in the linear case for $M \to \infty$

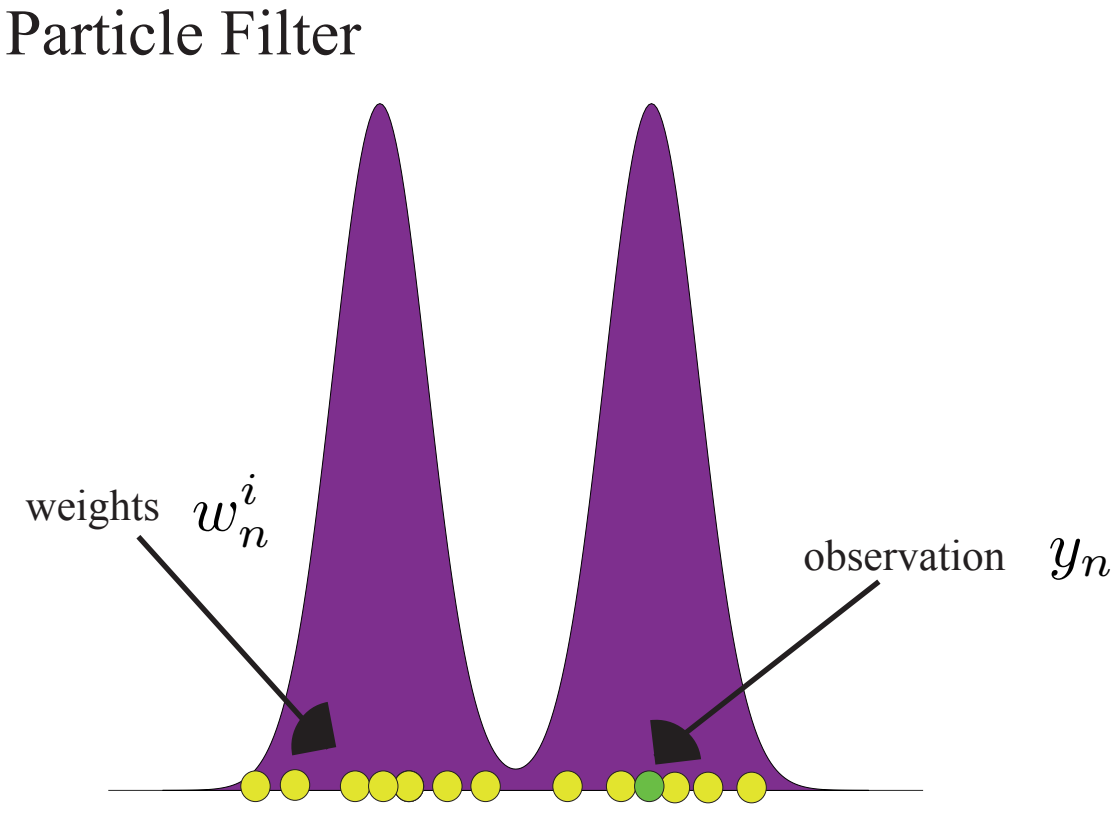
Ensemble Kalman filter

Works well in practice: e.g., EnKF is used for operational NWP for z_n^i with dimension 10^8 only using $M = 100$

Yet: mathematical foundation largely missing

Recent study: accuracy results for EnKF for idealized setting: $H = Id$ and observational error small

Particle filter



Particle filter

Problem: sampling from $\pi(z_n|y_{1:n})$ to approximate posterior via

$$\pi(z_n|y_{1:n}) = \frac{1}{M}\sum_{i=1}^{M}\delta(z - z_n^i)$$

Idea: sampling from $\pi(z_n|y_{1:n-1})$ instead i.e.,

$$\pi(z_n|y_{1:n}) = \sum_{i=1}^{M} w_n^i \delta(z - \hat{z}_n^i)$$

Bayes:

$$\pi(z_{n+1}|y_{1:n}) \propto \pi(y_n|z_n)\pi(z_n|y_{1:n-1}) \tag{59}$$

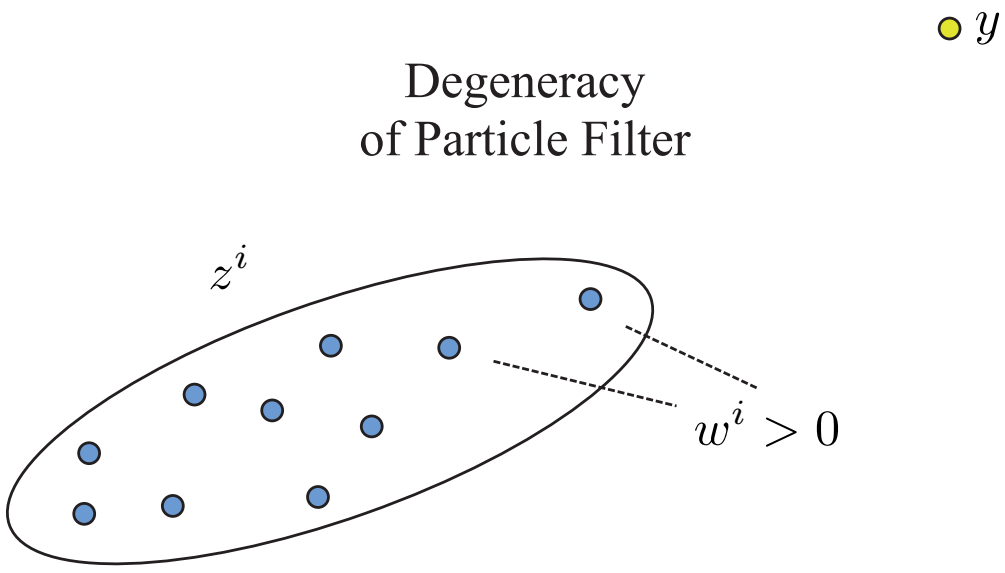
Weighting: unnormalized weights

$$\tilde{w}_n^i = \pi(y_n|z_n^i)w_{n-1}^i \text{ with } w_0^i = \frac{1}{M}$$

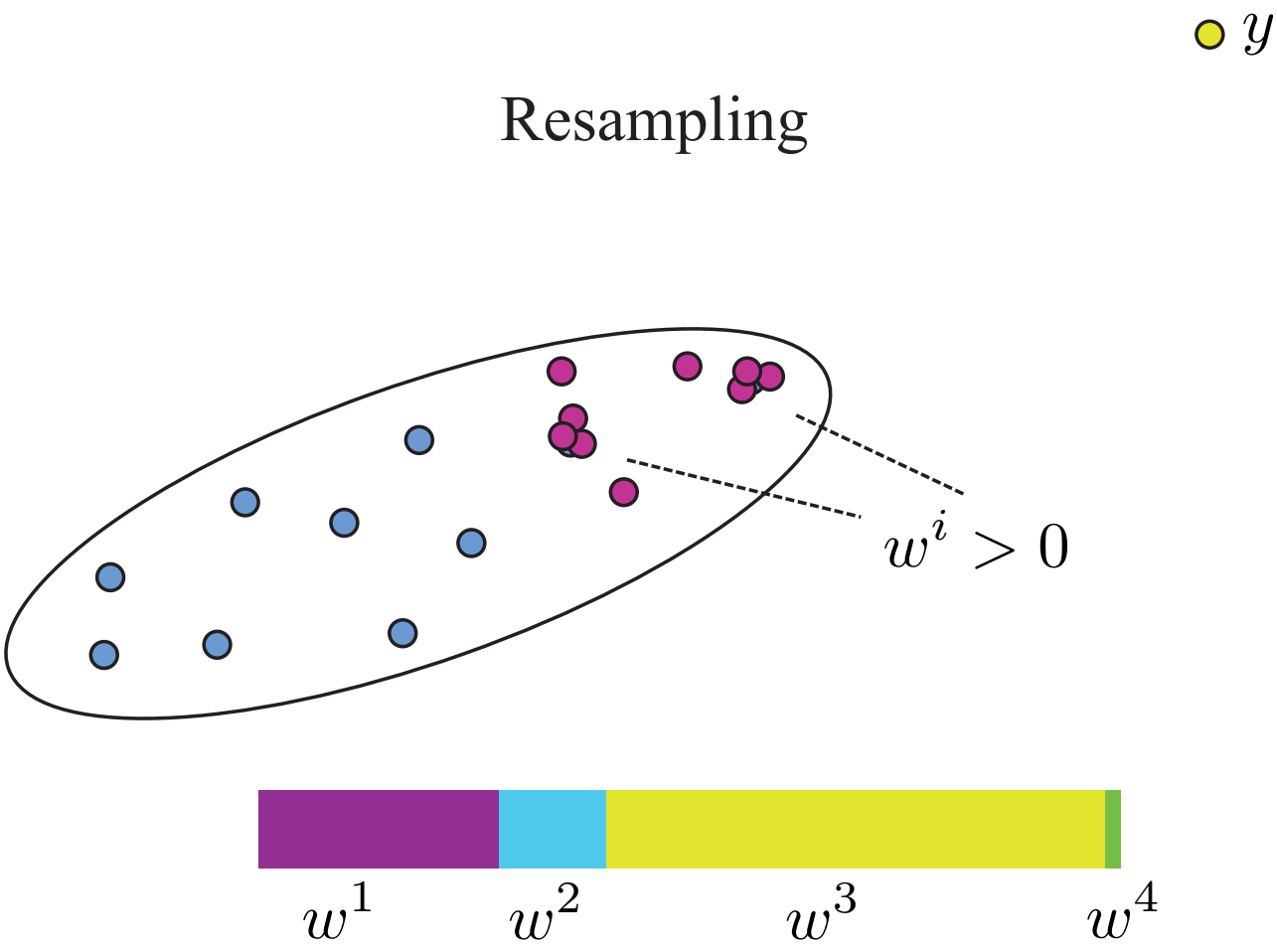
and normalized weights

$$w_n^i = \frac{\tilde{w}_n^i}{\sum_{j=1}^{M}\tilde{w}_n^j}$$

Particle collapse



Resampling



Resampling

Problem: weights w_n^i become very small

Ansatz: resampling

Input: w_n^i

For$(k = 1 : M)$

1. Draw a number $u \in [0, 1]$ from the uniform distribution $U[0, 1]$
2. Compute $i^* \in \{1, \ldots, M\}$ which satisfies

$$i^* = \arg\min_{i \geq 1} \sum_{j=1}^{i} w_j \geq u \tag{60}$$

3. Set $\xi_{i^*} = \xi_{i^*} + 1$

Return ξ_i

Still a lot of challenges....

Ansatz: approximative via empirical measure

$$\pi(z_n|y_{1:n}) = \frac{1}{M}\sum_{i=1}^{M}\delta(z - z_n^i)$$

where

$$z_n^i \sim \pi(z_n|y_n)$$

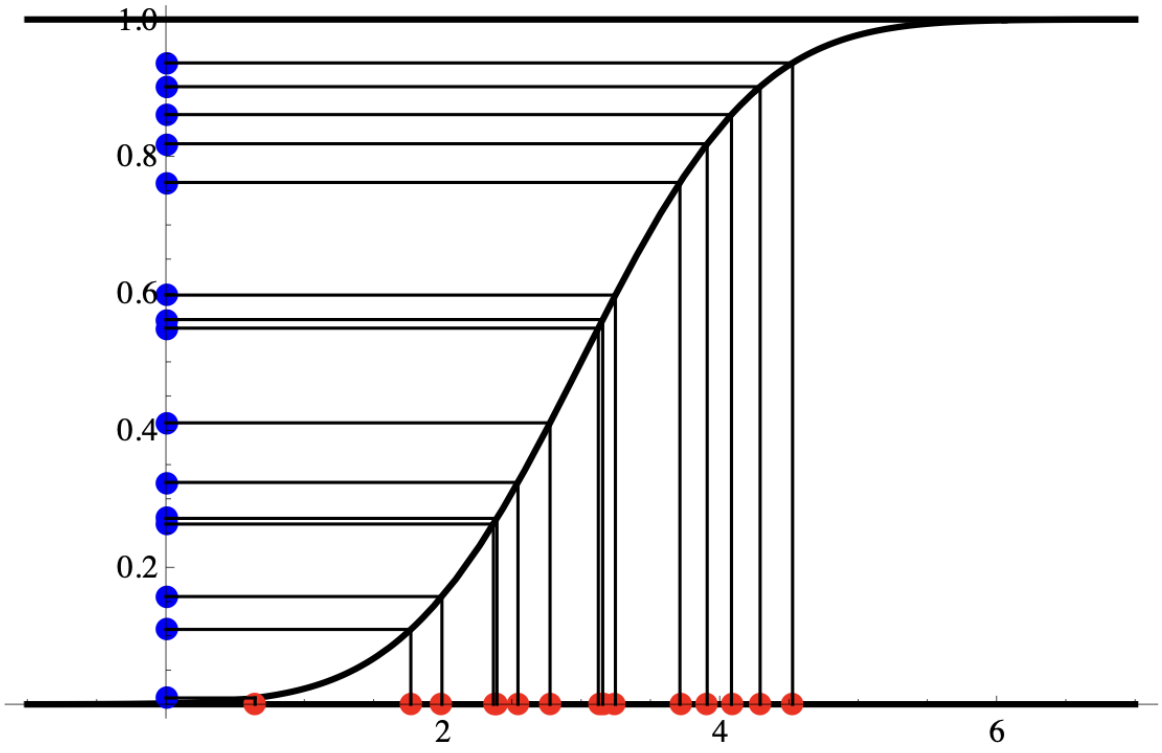
Monte Carlo approximation leads to a variety of filters e.g.,

- Particle filters (curse of dimensionality)
- Ensemble Kalman filter (underlying Gaussian assumption)

Maximum-spacing method

Maximum-spacing method

Lemma: Let the cdf F_θ be continuous and strictly monoton increasing. Under $\mathbb{P}_\theta$ the random variables $F_\theta(X_1), \ldots, F_\theta(X_n)$ are independent and uniformly distributed on the $(0, 1)$ interval.



Proof

Proof

Maximum-spacing method

Lemma: Let $z_1, \ldots, z_k \in [0, 1]$ be number that are subject to the condition $z_1 + \cdots + z_k = 1$. Then

$$z_1 \cdot \cdots \cdot z_k \leq \frac{1}{k^k}. \tag{61}$$

Equality is attained only if all the numbers all the numbers are equal to $\frac{1}{k}$.

Example

Maximum-spacing method

Lemma: The maximum-spacing method is defined via

$$\widehat{\theta}_{MS} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n+1} (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})) \tag{62}$$

Unbiased estimators

Estimator

Def:

- An estimator is an arbitrary (Borel-measurable) function

$$\widehat{\theta} : \mathcal{X} \to \Theta, \quad x \mapsto \widehat{\theta}(x) \tag{63}$$

- An estimator $\widehat{\theta}$ is called unbiased, if

$$\mathbb{E}_\theta[\widehat{\theta}(X)] = \theta \tag{64}$$

  for all $\theta \in \Theta$.

- The bias of an estimator $\widehat{\theta}$ is

$$\text{Bias}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[\widehat{\theta}(X)] - \theta \tag{65}$$

Note: $\text{Bias}_\theta(\widehat{\theta})$ is a function in $\widehat{\theta}$

Mean square error

Def: Let $\Theta = (a, b) \subset \mathbb{R}$ be an interval. The mean square error (MSE) of an estimator $\widehat{\theta} : \mathcal{X} \to \Theta$

$$\mathrm{MSE}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta}(X) - \theta)^2] \tag{66}$$

Mean square error

Lemma: The relationship between the mean square error (MSE) of an estimator $\widehat{\theta} : \mathcal{X} \to \Theta$ and the BIAS is given by

$$\mathsf{MSE}_\theta(\widehat{\theta}) = \mathsf{Var}_\theta \widehat{\theta} + (\mathsf{Bias}_\theta(\widehat{\theta}))^2 \tag{67}$$

Consistently better

Let $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$ be two estimators. The estimator θ_1 is consistently better than θ_2 if,

$$MSE_\theta(\widehat{\theta}_1) \leq MSE_\theta(\widehat{\theta}_2) \quad \forall \theta \in \Theta \tag{68}$$

Minimum-variance unbiased estimator

Def: An unbiased estimator $\widehat{\theta}$ is called minimum-variance unbiased estimator if all unbiased estimators $\tilde{\theta}$ the following inequality holds

$$\mathrm{Var}_\theta \widehat{\theta} \leq \mathrm{Var}_\theta \tilde{\theta} \tag{69}$$

for all $\theta \in \Theta$.

Minimum-variance unbiased estimator

Lemma: Let $\widehat{\theta}_1, \widehat{\theta}_2 : \mathcal{X} \to \Theta$ are two minimum-variance unbiased estimator the

$$\widehat{\theta}_1 = \widehat{\theta}_2 \quad \text{almost surely under } \mathbb{P} \text{ for all } \theta \in \Theta \tag{70}$$

for all $\theta \in \Theta$.

Sufficient statistic

Def: A function $T : \mathcal{X} \to \mathbb{R}^r$ is called a sufficient statistic if the function

$$\theta \mapsto \mathbb{P}_\theta[X = x | T(X) = t] \tag{71}$$

is constant for all $x \in \mathcal{X}$ and for all $t \in \mathbb{R}^r$,i.e.,

$$\mathbb{P}_{\theta_1}[X = x | T(X) = t]\mathbb{P}_{\theta_2}[X = x | T(X) = t] \tag{72}$$

for all $t \in \mathbb{R}^r$ and all $\theta_1, \theta_2 \in \Theta$ with $\mathbb{P}_{\theta_1}[T(X) = t] \neq 0$ and $\mathrm{P}_{\theta_2}[T(X) = t] \neq 0$

Unbiased estimators

Estimator

Def:

- An estimator is an arbitrary (Borel-measurable) function

$$\widehat{\theta} : \mathcal{X} \to \Theta, \quad x \mapsto \widehat{\theta}(x) \tag{73}$$

- An estimator $\widehat{\theta}$ is called unbiased, if

$$\mathbb{E}_\theta[\widehat{\theta}(X)] = \theta \tag{74}$$

  for all $\theta \in \Theta$.

- The bias of an estimator $\widehat{\theta}$ is

$$\text{Bias}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[\widehat{\theta}(X)] - \theta \tag{75}$$

Note: $\text{Bias}_\theta(\widehat{\theta})$ is a function in $\widehat{\theta}$

Example

Example

Example

Mean square error

Def: Let $\Theta = (a, b) \subset \mathbb{R}$ be an interval. The mean square error (MSE) of an estimator $\widehat{\theta} : \mathcal{X} \to \Theta$

$$\mathrm{MSE}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta}(X) - \theta)^2] \tag{76}$$

Mean square error

Lemma: The relationship between the mean square error (MSE) of an estimator $\widehat{\theta} : \mathcal{X} \to \Theta$ and the BIAS is given by

$$\mathsf{MSE}_\theta(\widehat{\theta}) = \mathsf{Var}_\theta \widehat{\theta} + (\mathsf{Bias}_\theta(\widehat{\theta}))^2 \tag{77}$$

Proof

Proof

Consistently better

Def: Let $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$ be two estimators. The estimator θ_1 is called consistently better than θ_2 if,

$$MSE_\theta(\widehat{\theta}_1) \leq MSE_\theta(\widehat{\theta}_2) \quad \forall \theta \in \Theta \tag{78}$$

Minimum-variance unbiased estimator

Def: An unbiased estimator $\widehat{\theta}$ is called minimum-variance unbiased estimator if all unbiased estimators $\tilde{\theta}$ the following inequality holds

$$\mathrm{Var}_\theta \widehat{\theta} \leq \mathrm{Var}_\theta \tilde{\theta} \tag{79}$$

for all $\theta \in \Theta$.

Minimum-variance unbiased estimator

Lemma: Let $\widehat{\theta}_1, \widehat{\theta}_2 : \mathcal{X} \to \Theta$ are two minimum-variance unbiased estimator the

$$\widehat{\theta}_1 = \widehat{\theta}_2 \quad \text{almost surely under } \mathbb{P} \text{ for all } \theta \in \Theta \tag{80}$$

for all $\theta \in \Theta$.

Bernoulli Experiment MVUE

Lemma: The estimator $\widehat{\theta}(x_1, \ldots, x_n) = \bar{x}_n$ is the minimum-variance unbiased estimator of θ in n Bernoulli experiments.

Sufficient statistic

Def: A function $T : \mathcal{X} \to \mathbb{R}^r$ is called a sufficient statistic if the function

$$\theta \mapsto \mathbb{P}_\theta[X = x | T(X) = t] \tag{81}$$

is constant for all $x \in \mathcal{X}$ and for all $t \in \mathbb{R}^r$,i.e.,

$$\mathbb{P}_{\theta_1}[X = x | T(X) = t] = \mathbb{P}_{\theta_2}[X = x | T(X) = t] \tag{82}$$

for all $t \in \mathbb{R}^r$ and all $\theta_1, \theta_2 \in \Theta$ with $\mathbb{P}_{\theta_1}[T(X) = t] \neq 0$ and $\mathrm{P}_{\theta_2}[T(X) = t] \neq 0$

Sufficient statistic

Lemma: Let $T_{\mathcal{X}} \to \mathbb{R}^r$ be a sufficient statistic and let $g : Im(T) \to \mathbb{R}^k$ an injective function. Then the concatenation

$$g \circ T : \mathcal{X} \to \mathbb{R}^k, \quad x \mapsto g(T(x)) \tag{83}$$

a sufficient statistic.

Rao-Blackwell

Proposition: Let $(\mathbb{P}_\theta)_{\theta\in\Theta}$ a family of probability measures on the sample space $(\mathcal{X}, \mathcal{A})$, where $\Theta \subset \mathbb{R}$ is an interval. Furthermore let

- $T : \mathcal{X} \to \mathbb{R}^m$ a sufficient statistic and
- $\hat{\theta} : \mathcal{X} \to \mathcal{R}$ an unbiased estimator of θ with $\mathbb{E}_\theta[\theta^2] \leq \infty$ for all $\theta \in \Theta$.

Define $\tilde{\theta} := \mathbb{E}_\theta[\hat{\theta}|\, T]$. Then $\tilde{\theta}$ is an unbiased estimator of θ and the following holds

$$Var_\theta \tilde{\theta} \leq Var_\theta \hat{\theta} \tag{84}$$

for all $\theta \in \Theta$.

Causal relations

Model for simple linear regression

Model:

$$Y_i = f(X_i, \beta) + \epsilon_i, \quad i = 1, \ldots, n \tag{85}$$

where ϵ_i are iid with $\mathbb{E}[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$

Data: it is possible to observe realisations

$$(y_i, x_i) \quad i = 1, \ldots, n \tag{86}$$

Goal: estimate parameters β of the function to obtain approximative $f(x, \hat{\beta})$

Note: note that f approximates $\mathbb{E}[Y_i|X_i]$

Linear regression

Model for simple linear regression

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n \tag{87}$$

where ϵ_i are iid with $\mathbb{E}[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$

Data:

$$(y_i, x_i) \quad i = 1, \ldots, n \tag{88}$$

Goal: estimate $f(x, \hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 x$

The Ordinary Multiple Linear Regression Model

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \cdots + \beta_p X_{i,p} + \epsilon_i, \quad i = 1, \ldots, n \tag{89}$$

where ϵ_i are iid with $\mathbb{E}[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$

Data:

$$(y_i, x_i) \quad i = 1, \ldots, n \tag{90}$$

Goal: estimate $\hat{f}(x_1, \ldots, x_p, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots, \hat{\beta}_p x_p$

Multivariate Random Variables

Def: Let $\mathbf{X}$ be a vector of (univariate) random variables, i.e., $\mathbf{X} = (X_1, \ldots, X_p)^\top$ with $\mathbb{E}[X_i] = \mu_i$. $\mathbf{X}$ is called a multivariate random variable and we denote $\mathbb{E}[\mathbf{X}] = \mu$

Note:

- Variance
  $\text{Var}(X_i) = \mathbb{E}[(X_i - \mathbb{E}(X_i))^2] = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_i - \mathbb{E}(X_i))]$
- Covariance $\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))]$

Covariance

Def: The covariance of the multivariate random variable $\mathbf{X}$ is defined by

$$\Sigma := \text{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] \tag{91}$$

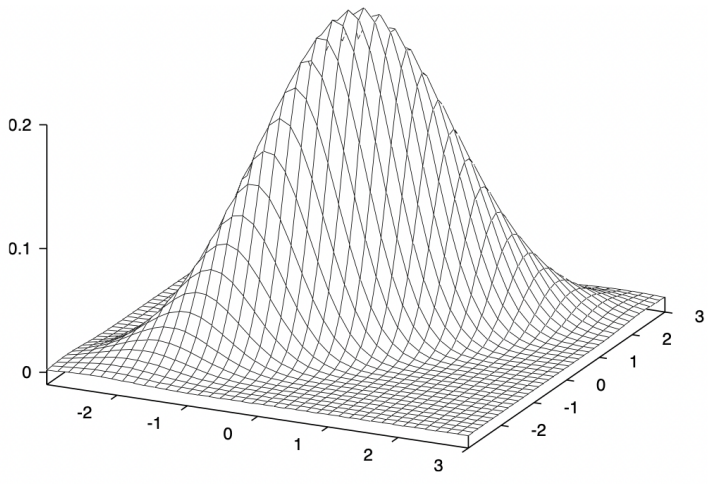
Example:

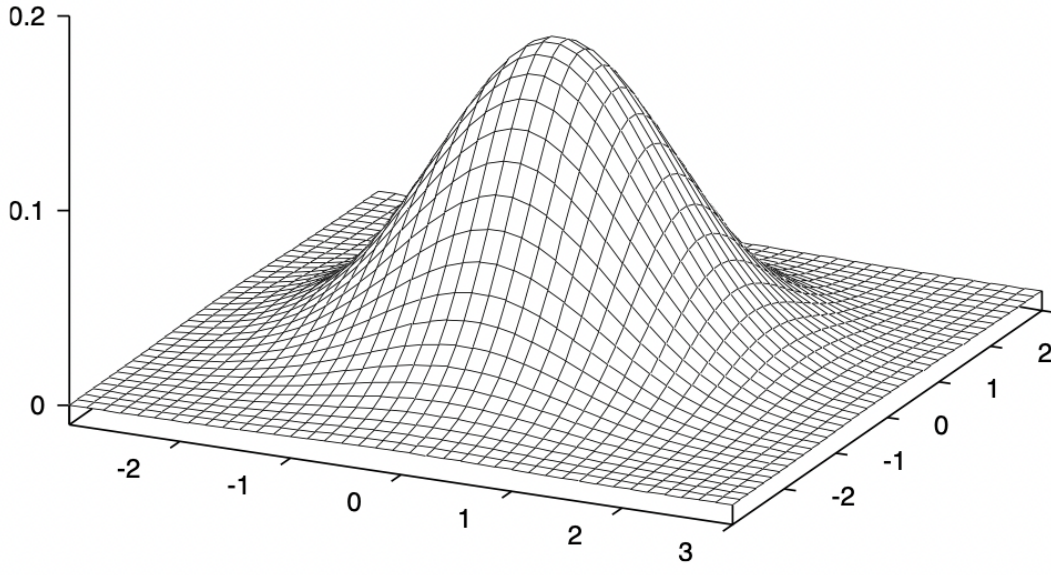
$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) \end{pmatrix} \tag{92}$$

Properties of Σ:

- quadractic
- symmetric
- positive-semidefinite

Multivariate Normal Distribution





$$\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma) \tag{93}$$

Positive semi-definite

Lemma: Let $\mathbf{B}$ be an $n \times (p+1)$ matrix. Then the matrix $\mathbf{B}^\top\mathbf{B}$ is symmetric and postive semi-definite. It is positive definite, if $\mathbf{B}$ has full column rank. Then, besides $\mathbf{B}^\top\mathbf{B}$ also $\mathbf{B}\mathbf{B}^\top$ is postive semi-definite.

LS-estimator

Theorem: The LS-estimator of the unknown parameters β is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{94}$$

if $\mathbf{X}$ has full column rank p+1.

Positive semi-definite

Proposition: The hat-matrix $\mathbf{H} = (h_{ij})_{1 \leq i,j, \leq b}$ has the following properties:

1. $\mathbf{H}$ is symmetric
2. $\mathbf{H}$ is idempotent, i.e., $\mathbf{HH} = \mathbf{H}$
3. $rk(\mathbf{H}) = tr(\mathbf{H}) = p + 1$
4. $0 \leq h_{ii} \leq 1, \quad \forall i = 1, \ldots, n$
5. the matrix $\mathbf{I}_n - \mathbf{H}$ is also symmetric and idempotent with $rk(\mathbf{I}_n - \mathbf{H}) = n - p - 1$

ML-estimator

Theorem: The ML-estimator of the unknown parameters σ^2 is $\hat{\sigma}^2_{ML} = \frac{\hat{\epsilon}\hat{\epsilon}}{n}$ with $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$.

ML-estimator

Proposition: For the ML-estimator $\hat{\sigma}^2_{ML}$ of σ^2 the following property holds:

$$\mathbb{E}[\sigma^2_{ML}] = \frac{n-p-1}{n}\sigma^2 \tag{95}$$

Adjusted estimator

Proposition: The adjusted estimator

$$\hat{\sigma}^2_{ad} = \frac{\hat{\epsilon}\hat{\epsilon}}{n-p-1} \tag{96}$$

of the unknown parameter σ^2 can be written as

$$\hat{\sigma}^2_{ad} = \frac{\mathbf{y}^\top \mathbf{y} - \hat{\beta}^\top \mathbf{X}^\top \mathbf{y}}{n-p-1} \tag{97}$$

ML estimator

Proposition: The LS-estimator $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ is equivalent to the ML-estimator based on maximization of the log-likelihood

$$l(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) \quad (98)$$

LS estimator

Proposition: The LS-estimator $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ and the REML-estimator $\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\epsilon}^\top \hat{\epsilon}$ the following properties hold:

1. $\mathbb{E}[\hat{\beta}] = \beta$, $\mathrm{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$
2. $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$

Best linear unbiased estimator (BLUE)

Linear estimator

Def: A linear estimator has the form

$$\hat{\beta}^L = \mathbf{b} + \mathbf{A}\mathbf{y} \tag{99}$$

where $\mathbf{b} \in \mathbb{R}^{(p+1)\times 1}$ and $\mathbf{A} \in \mathbb{R}^{(p+1)\times n}$.

Example: The LS-estimator:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \tag{100}$$

is a linear estimator with $\mathbf{b} = \mathbf{0}$ and $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$

Gauss-Markov

Theorem: The LS-estimator is BLUE. This means that the LS-estimator has minimal variance among all linear and unbiased estimators $\hat{\beta}^L$

$$\text{Var}(\hat{\beta}_j) \leq \text{Var}(\hat{\beta}_j^L), \quad j = 0, \ldots, p. \tag{101}$$

Furthermore, for an arbitrary linear combination $\mathbf{c}^\top \hat{\beta}$ it holds that

$$\text{Var}(\mathbf{c}^\top \hat{\beta}) \leq \text{Var}(\mathbf{c}^\top \hat{\beta}^L) \tag{102}$$

Coefficient of determination

Def: The coefficient of determination is defined by

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{103}$$

and measures the proportion of variability in y that is accounted for by the statistical model from the overall variation in y.

Coefficient of determination

Lemma: The method of least squares yields the following geometrical results:

- The fitted values $\hat{\mathbf{y}}$ are orthogonal to the residuals $\hat{\epsilon}$, i.e., $\hat{\mathbf{y}}^\top \hat{\epsilon} = 0$.
- The columns of $\mathbf{X}$ are orthogonal to the residuals $\hat{\epsilon}$, i.e., $\mathbf{X}^\top \hat{\epsilon} = 0$
- The residuals are zero on average, i.e.,

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0 \quad \text{and} \quad \bar{\hat{\epsilon}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i = 0 \tag{104}$$

- The mean of the estimated values

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \bar{y} \tag{105}$$

Proof

Proof

Coefficient of determination

Lemma: The following decomposition holds:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2 \tag{106}$$

Coefficient of determination

Lemma: The coefficient of determination R^2 can be transformed into

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\hat{\beta}^\top \mathbf{X}^\top \mathbf{y} - n\bar{y}^2}{\mathbf{y}^\top \mathbf{y} - n\bar{y}^2} \tag{107}$$

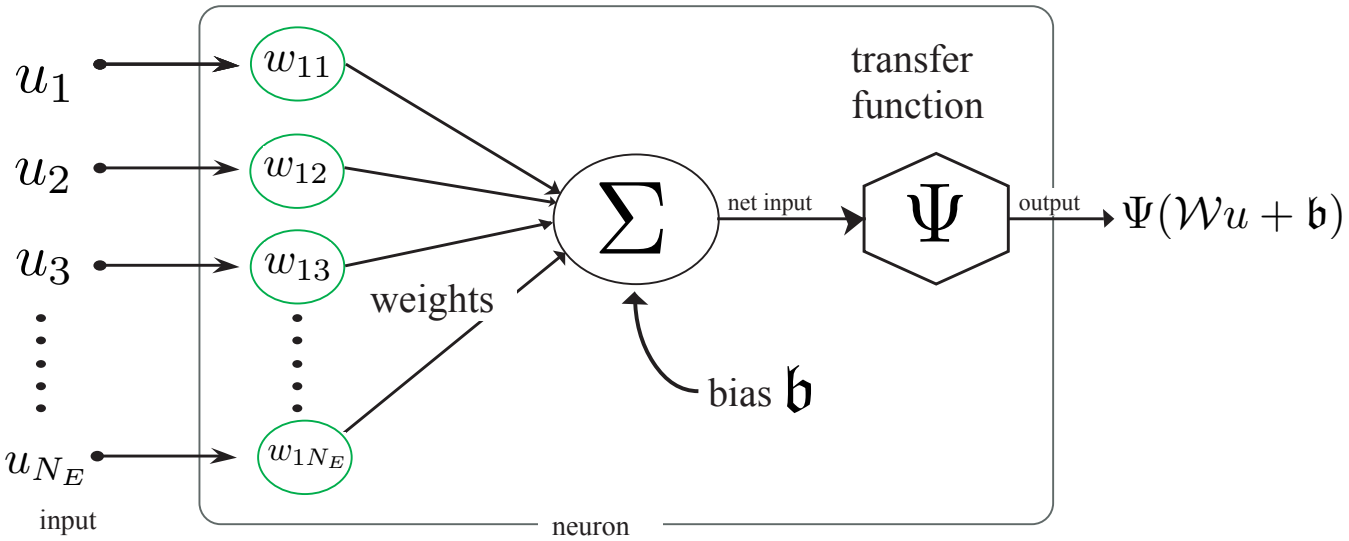
Coefficient of determination

Def: The corrected coefficient of determination $\bar{R}^2$ is defined by

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2) \tag{108}$$

Neural Networks

Neuron


u_1
u_2
u_3
u_{N_E}
input
w_{11}
w_{12}
w_{13}
w_{1N_E}
weights
Σ
net input
transfer function
Ψ
output
$\Psi(\mathcal{W}u + \mathfrak{b})$
bias $\mathfrak{b}$
neuron

Ill-posedness and Regularization

If the least squares problem is ill-posed, i.e., solution does not exist or is unstable.

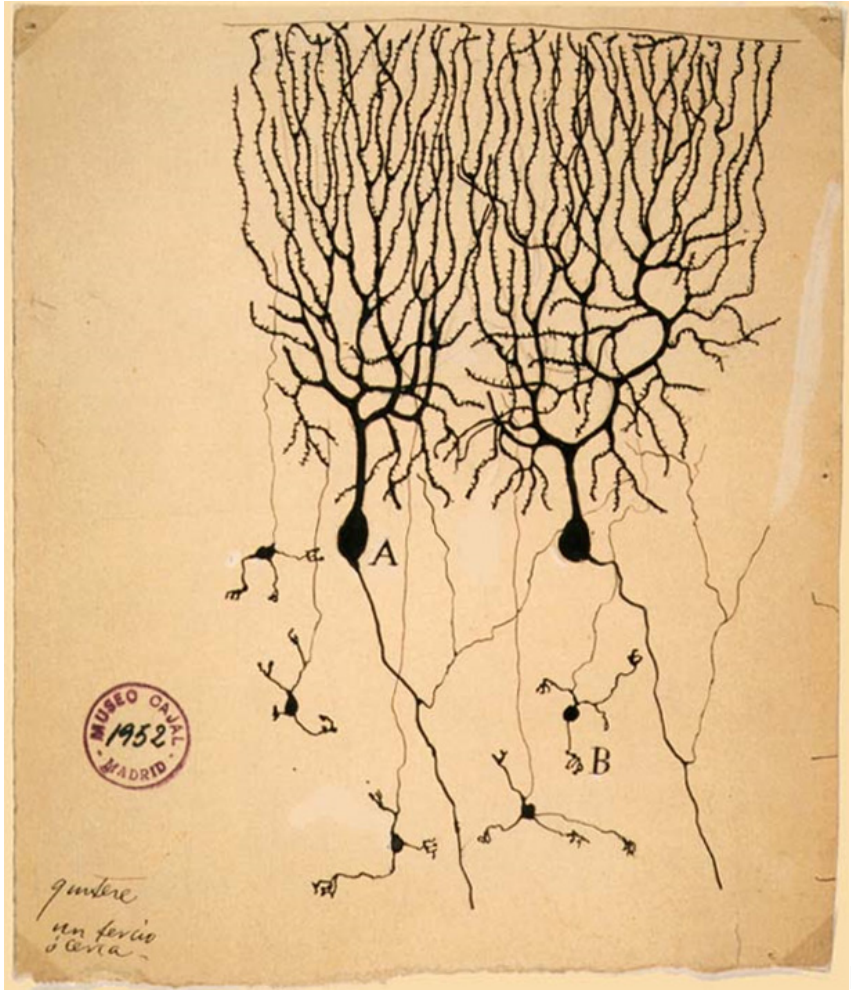
Small perturbations in $\mathbf{y}$ or $\mathbf{X}$ yield large perturbations in β

Solve regularized problem: For some $\lambda > 0$ and matrix $\mathbf{G}$

$$\min_{\beta} \frac{1}{2}\|\mathbf{X}\beta - \mathbf{y}^\top\|^2 + \frac{\lambda}{2}\|\mathbf{G}\beta\|^2$$

Neural Networks

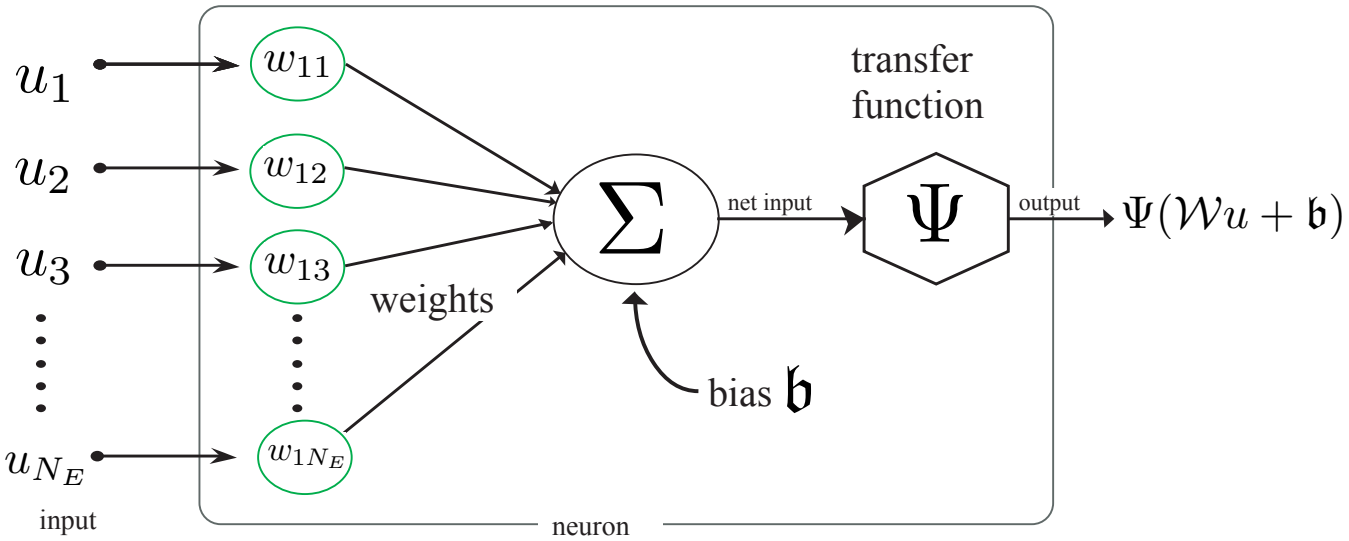
Motivation from biology



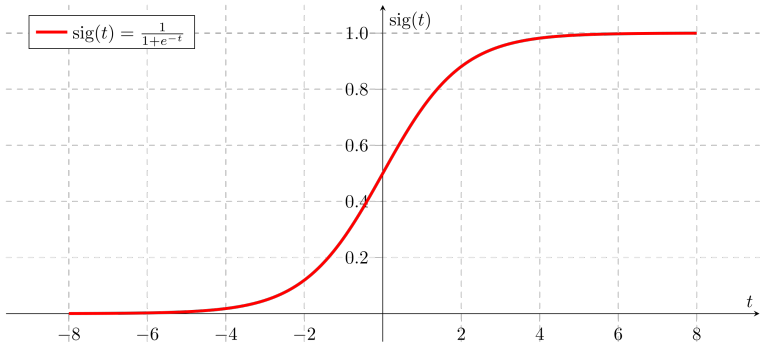
By Santiago Ramn y Cajal in 1899 see

https://de.wikipedia.org/wiki/Santiago_Ramn_y_Cajalfordetails

Neuron


u_1
u_2
u_3
u_{N_E}
input
w_{11}
w_{12}
w_{13}
w_{1N_E}
weights
Σ
net input
transfer function
Ψ
output
$\Psi(\mathcal{W}u + \mathfrak{b})$
bias $\mathfrak{b}$
neuron

Activation function example: sigmoid



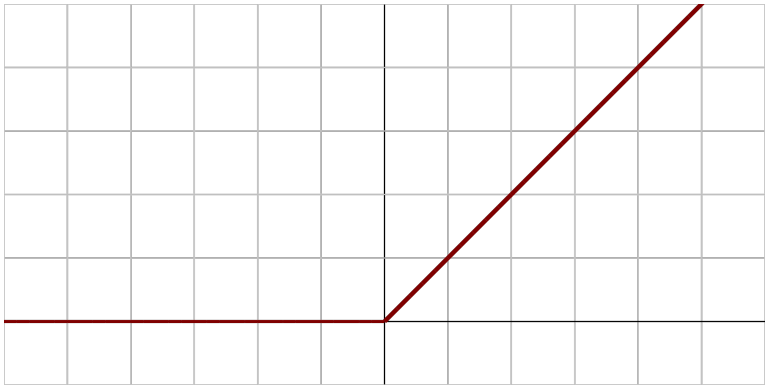
Sigmoid function:

$$\mathrm{sig}(t) = \frac{1}{1+e^{-t}}$$

Properties:

- Derivative: $\dfrac{1 + e^{-x} + xe^{-x}}{(1 + e^{-x})^2}$
- $\mathrm{sig}'(t) = \mathrm{sig}(t)\,(1 - \mathrm{sig}(t))$

Activation function example: ReLu



Rectified linear unit:

$$\text{ReLu}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

$$= \max\{0, x\}$$

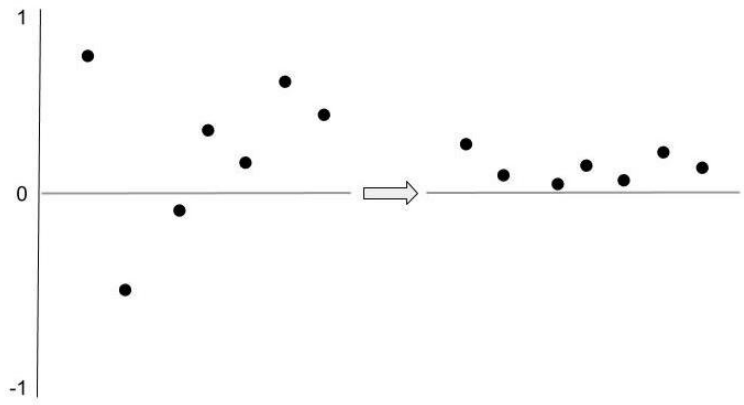
Properties:

- Derivative:

$$\begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ \text{undefined} & \text{if } x = 0 \end{cases} \tag{109}$$

- very popular for Deep RL
- Dying ReLU problem - vanishing gradient problem.

Activation function example: Softmax



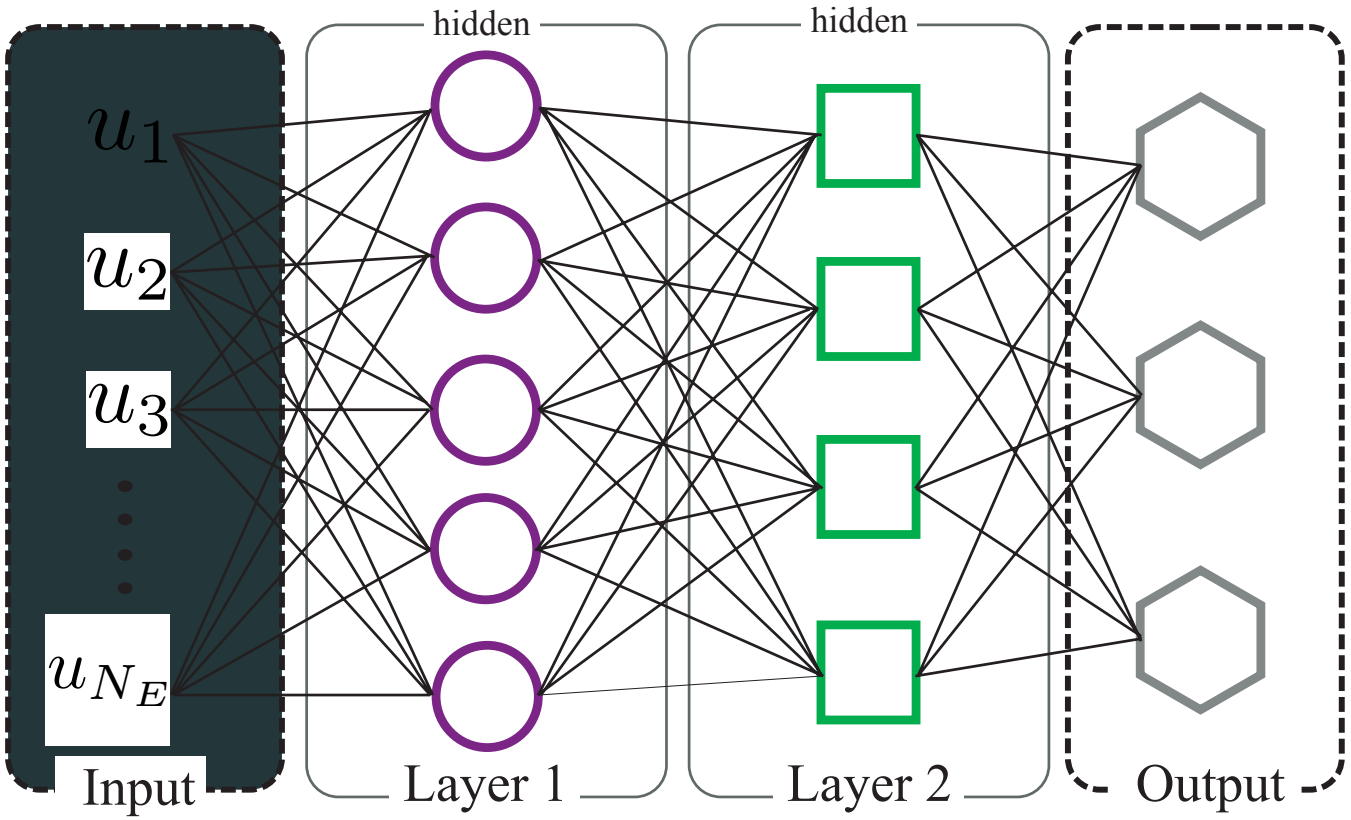
Softmax:

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad \text{for } i = 1, \ldots, K$$
$$\text{and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K.$$

Properties:

- Derivative:
$$\frac{\partial}{\partial q_k}\sigma(\mathbf{q}, i) = \sigma(\mathbf{q}, i)(\delta_{ik} - \sigma(\mathbf{q}, k)). \tag{110}$$
- used in to normalize the output (map to a probability distribution)
- also used in RL to convert action values into action

Multilayer perceptron



Training Neural Network

1. Choose network architecture:
   - activation functions
   - hidden layers (shallow or deep)
   - number of neurons
   - etc.
2. Choose appropriate loss function E, e.g., least squares
3. Find minima via:
   - stochastic gradient descent
   - Backpropagation

Stochastic Gradient Descent

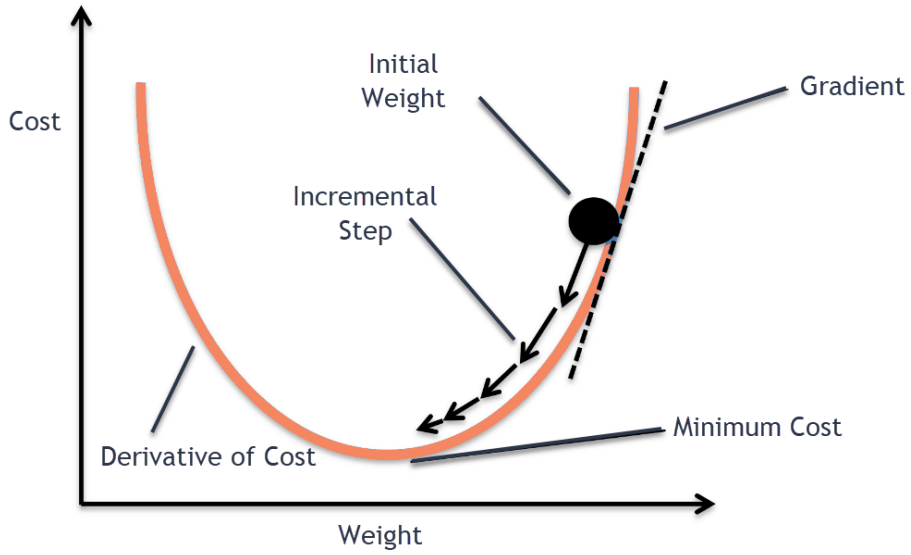


Image ref: https://morioh.com/p/bc6bc20e9739 and

https://medium.com/38th-street-studios/exploring-stochastic-gradient-descent-with-restarts-sgdr-fa206c38a74e

Iterative weight improvement:

$$w := w - \eta \nabla E_i(w). \tag{111}$$

Backpropagation

Backpropagation

Backpropagation

Backpropagation

Backpropagation

Asymptotic Properties of the LS-Estimator

Proposition: Consider the setting

$$\mathbf{y}_n = \mathbf{X}_n\beta + \epsilon_n \quad \text{with } \mathbb{E}[\epsilon_n] = \mathbf{0} \quad \text{and } Cov(\epsilon_n) = \sigma^2\mathbf{I}_n \tag{112}$$

with the following assumption being fulfilled:

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{X}_n^\top\mathbf{X}_n = \mathbf{V} \tag{113}$$

where $\mathbf{V}$ is positive definite. Then

- The LS-estimator $\hat{\beta}_n$ for β as well as the ML- and REML-estimators $\hat{\sigma}_n^2$ for σ^2 are consistent. ($\text{MSE}_\theta(\hat{\theta}) \to 0$ $n \to \infty$)
- The LS-estimator $\hat{\beta}_n$ for β is asymptotically normally distributed:

$$\sqrt{n}(\hat{\beta}_n - \beta) \to \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{V}^{-1}) \text{ (in distribution)} \tag{114}$$

Asymptotic Properties of the LS-Estimator

Proposition: Hence, for sufficiently large n it follows that $\hat{\beta}_n$ is approximately normally distributed with

$$\hat{\beta}_n \to \mathcal{N}(\beta, \sigma^2 \mathbf{V}^{-1}/n) \text{ (almost surely)} \tag{115}$$

Asymptotic Properties of the LS-Estimator

Proposition:

- Similar to the error terms, also the residuals have expectation zero.
- In contrast to the error terms, the residuals are not uncorrelated.

Asymptotic Properties of the LS-Estimator

Proposition: Beside the usual assumptions, additionally assume that the error terms are normally distributed. Then the following properties hold:

- The distribution of the squared sum of residuals is given by:

$$\frac{\hat{\epsilon}^\top \hat{\epsilon}}{\sigma^2} = (n - p - 1)\frac{\hat{\sigma}^2}{\sigma^2} \tag{116}$$

- The squared sum of residuals $\hat{\epsilon}^\top \hat{\epsilon}$ and the LS-estimator $\hat{\beta}$ are independent.

Prediction

Proposition:

1. The expected prediction error is zero i.e., $\mathbb{E}[\hat{\mathbf{y}}_0 - \mathbf{y}_0] = 0$, i.e., $\mathbb{E}[\hat{\mathbf{y}}_0 - \mathbf{y}_0] = 0$
2. Prediction error covariance matrix is given by:

$$\mathbb{E}[(\hat{\mathbf{y}}_0 - \mathbf{y}_0)(\hat{\mathbf{y}}_0 - \mathbf{y}_0)^\top] = \sigma^2(\mathbf{X}_0(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_0^\top + \mathbf{I}_{T_0}) \quad (117)$$

Proof

Proof

Hypotheses Testing and Confidence Intervals

Primaries

Primaries

Primaries

Primaries

Primaries

Gamma-distribution

Def: A continuous, non-negative random variable X is called gamma-distributed with parameters $a > 0$ and $b > 0$, abbreviated by the notation $X \sim \mathcal{G}(a, b)$, if it has a density function of the following form

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \quad x > 0 \tag{118}$$

with $\Gamma(n) = (n-1)!$.

Gamma-distribution

Lemma: Let $X \sim \mathcal{G}(a, b)$ be a continuous, non-negative random variable. Then its expectation and variance are given by:

- $\mathbb{E}[X] = \frac{a}{b}$
- $Var(X) = \frac{a}{b^2}$

χ^2-distribution

Def: A continuous, non-negative random variable X with density

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp\Big(-\frac{1}{2}x\Big), \quad x > 0 \qquad (119)$$

is called χ^2-distributed with n degrees of freedom, abbreviated by the notation $X \sim \chi^2_n$.

χ^2-distribution

Lemma: Let $X \sim \chi^2_n$ be a continuous, non-negative random variable. Then its expectation and variance are given by:

- $\mathbb{E}[X] = n$
- $Var(X) = 2n$

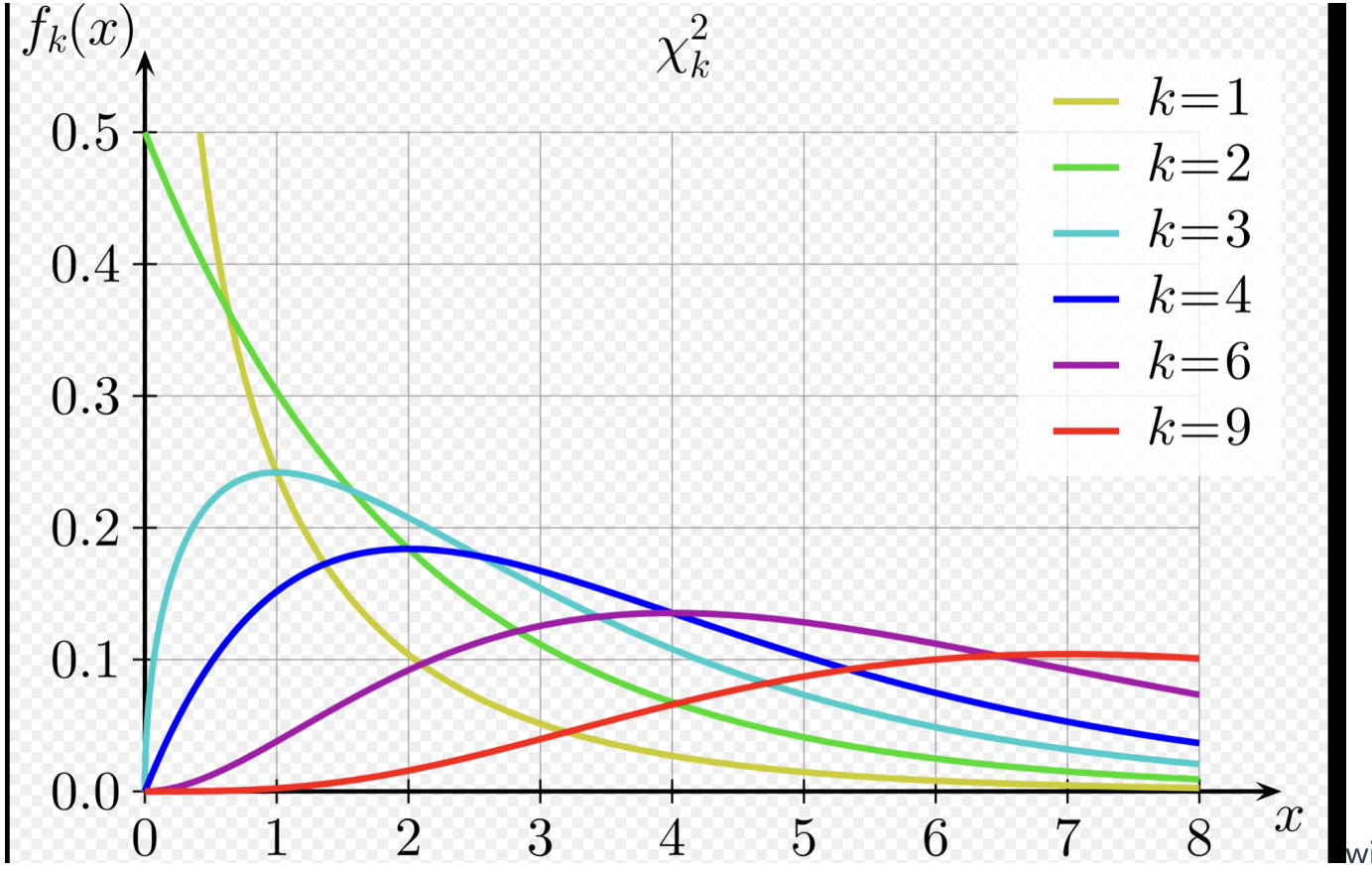
χ^2-distribution

Lemma: Let $X_1, ..., X_n$ be independent and identically standard normally distributed, then

$$Y_n = \sum_{i=1}^{n} X_i^2 \tag{120}$$

is χ^2- distributed with n degrees of freedom.

χ^2-distribution



wikipedia

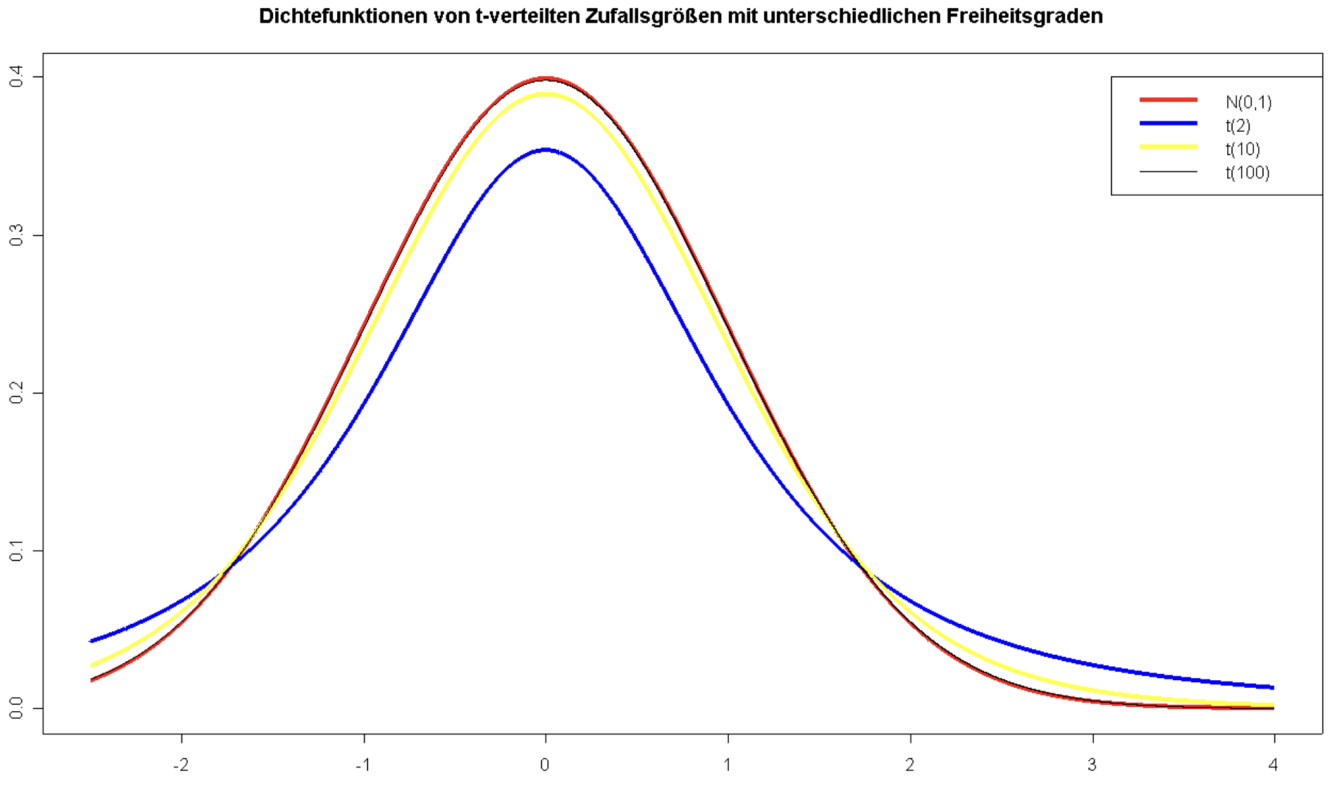
t-distribution

Def: A continuous random variable X with density

$$f(x) = \frac{\Gamma(n+1)/2}{\sqrt{n\pi}\Gamma(n/2)(1+x^2/n)^{(n+1)/2}} \tag{121}$$

is called t-distributed with n degrees of freedom, abbreviated by the notation $t \sim t_n$

t-distribution



wikipedia

t-distribution

Lemma: Let $X \sim t_n$ be a continuous, non-negative random variable. Then its expectation and variance are given by:

- $\mathbb{E}[X] = n \quad n > 1$
- $Var(X) = n/(n-2), \quad n > 2$

The t_1-distribution is also called Cauchy-distribution. If $X_1, ..., X_n$ are iid with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, it follows that

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \sim t_{n-1} \tag{122}$$

with

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ and } \bar{X} = \sum_{i=1}^{n} X_i \tag{123}$$

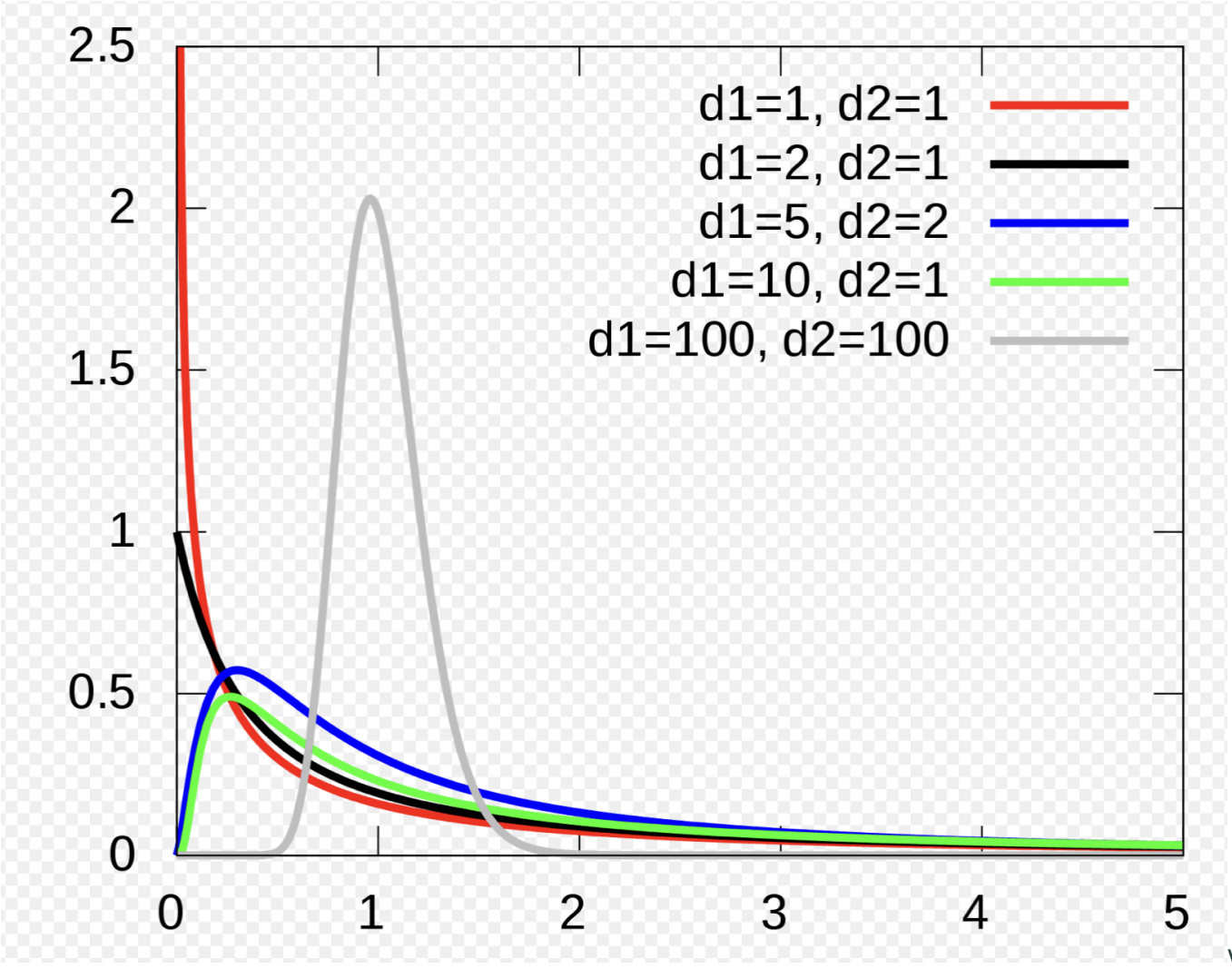
F-distribution

Def: Let X_1 and X_2 be independent random variables χ^2_n- and χ^2_m distributions respectively. Then the random variable

$$F = \frac{X_1/n}{X_2/m} \tag{124}$$

is called F-distributed with n and m degrees of freedom, abbreviated with the notation $F \sim F_{n,m}$.

F-distribution



wikipedia

Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_k$ be independent random variables. Then therandom variable

$$T := \frac{Z}{\sqrt{\frac{X}{k}}} \tag{125}$$

is t-distributed with k degrees of freedom.

Hypothesis testing for Regression parameters

In general, the following hypotheses can be tested:

1. $H_0 : \mathbf{R}_1\beta = r$ vs. $H_1 : \mathbf{R}_1\beta \neq r$
2. $H_0 : \mathbf{R}_1\beta \geq r$ vs. $H_1 : \mathbf{R}_1\beta < r$
3. $H_0 : \mathbf{R}_1\beta \leq r$ vs. $H_1 : \mathbf{R}_1\beta > r$

Under H_0:

$$\mathbf{R}_1\hat{\beta} \sim \mathcal{N}(r, \sigma^2\mathbf{R}_1(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}_1^\top) \tag{126}$$

holds. For unknown σ^2, a reasonable test-statistic is

$$T = \frac{\mathbf{R}_1\hat{\beta} - r}{\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}_1^\top}} \sim t_{n-p-1} \tag{127}$$

The corresponding rejection areas are:

1. $|T| > t_{1-\alpha/2,n-p-1}$
2. $T < t_{1-\alpha,n-p-1}$
3. $T > t_{1-\alpha,n-p-1}$

$(1-\alpha)$-confidence intervals for $\mathbf{R}_1\hat{\beta}$ are:

$$\mathbf{R}_1\hat{\beta} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}_1^\top} \tag{128}$$

Hypothesis for general parameter identification problems

Setting:

1. Let $(\mathbb{P})_{\theta \in \Theta}$ a family of probability measures on the sample space $(\mathcal{X}, \mathcal{A})$.
2. Find disjunct subsets Θ_1 and Θ_2 of parameter space $\Theta = \Theta_1 \cup \Theta_2$ and $\Theta_1 \cap \Theta_2 = \emptyset$
3.

Hypothesis:

1. Null hypothesis H_0: $\theta \in \Theta_0$
2. alternative hypothesis H_1: $\theta \in \Theta_1$

Setting: $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$, $\Theta = \{\Theta_0, \Theta_1\}$

Assumption: The associate probability measures $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_{\theta_1}$ have densities h_0 and h_1 for a measure λ on $(\mathcal{X}, \mathcal{A})$

Def: Let $k \in [0, \infty]$ and $\gamma \in [0, 1]$. A likelihood-quotient-test (LQ-test) is of the form

$$\phi(x) = \begin{cases} 1, & \text{if } \frac{h_1(x)}{h_0(x)} > k \\ 1, & \text{if } \frac{h_1(x)}{h_0(x)} < k \\ \gamma, & \text{if } \frac{h_1(x)}{h_0(x)} = k. \end{cases} \tag{129}$$

Neyman-Pearson Lemma

Lemma: Let ϕ be a LQ-test with $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$. Then

$$\mathbb{E}_{\theta_1}[\phi(X)] = \sup_{\psi:\mathbb{E}_{\theta_0}[\psi(X)]\leq\alpha} \mathbb{E}_{\theta_1}[\psi(X)] \tag{130}$$

Further for every $\alpha \in (0, \infty)$ it is possible to find $k \in [0, \infty]$ and $\gamma \in [0, 1]$ so that for a predefined Test ϕ

$$\mathbb{E}_{\theta_0}[\phi(X)] = \alpha \tag{131}$$

Regularization

Ridge Regularization (L_2)

$$\hat{\beta}^{Ridge} = \arg\min_{\beta\mathbb{R}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{132}$$

- decreases variance but increases bias (for increasing λ)
- Can improve predictive performance
-

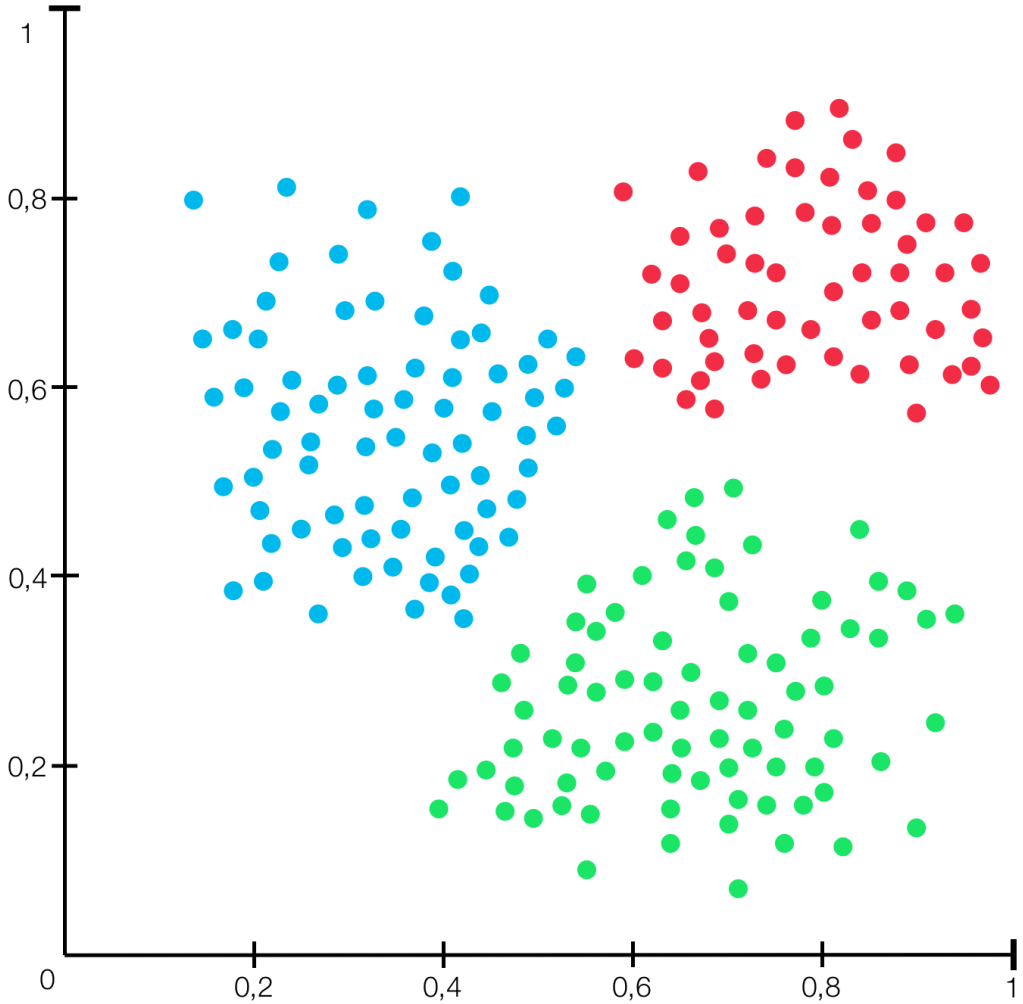
$$\hat{\beta}^{Ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^\top\mathbf{y} \tag{133}$$

Lasso Regularization (L_1)

$$\hat{\beta}^{Lasso} = \arg\min_{\beta \mathbb{R}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \tag{134}$$

- LASSO=Least Absolute Shrinkage and Selection Operator
- This penalty allows coefficients to shrink towards exactly zero
- LASSO usually leads to sparse models, that are easier to interpret

Clustering



K-means clustering

Input:

- Number of Clusters K
- Set of points $\{x_1, \ldots, x_M\}$ in vector space that need to be classified

Output:

- Sets $\mathcal{M}_k$ of the clusters

1. Initialize the centre of the cluster $\theta_1, \ldots, \theta_K \in \mathbb{R}^n$ randomly
2. Repeat till a stopping criterion is fulfilled $\{$
   for all $k = 1 : K$
   $\mathcal{M}_k := \{\ \}$
   for all $m = 1 : M$
   $j = \arg\min_h ||\theta_h - x_m||_2^2$
   $\mathcal{M}_j = \mathcal{M}_j \cup \{x_m\}$
   for all $k = 1 : K$
   $\theta_k = \frac{1}{|\mathcal{M}_k|} \sum_{x_m \in \mathcal{M}_k} x_m$
3. **return** $\theta_1, \ldots, \theta_K$

Initialisation

- Random Partition Method
- Forgy Initialization
- kmeans++
  1. choose θ_1 uniformly at random from set of points
  2. Choose new center θ_i with probability

$$\frac{D(x_m)^2}{\sum_{x_l} D(x_l)^2} \tag{135}$$

     where $D(x_m)$ denotes the shortest distance from data point x_m to the closest center we have already chosen
  3. Repeat Step 2 until we have all K centers

Disadvantages

- true number of clusters K unknow (requires tuning)
- K-means algorithm dependents on the chosen initial values
- Clustering data of varying sizes and density
- Centroids can be dragged by outliers

Using the Triangle Inequality to Accelerate k-Means

Using the Triangle Inequality to Accelerate k-Means

Using the Triangle Inequality to Accelerate k-Means

Using the Triangle Inequality to Accelerate k-Means

Algorithm:

1. Initialize the centre of the cluster $\theta_1, \ldots, \theta_K \in \mathbb{R}^n$ randomly
2. Set lower bounds to $l(x_m, \theta_i) = 0$ for all θ_i and x_m
3. Assign each x_m to its closest initial center $\theta(x_m) = \arg\min_h ||\theta_h - x_m||_2^2$ (avoid redundant calculations using Lemma 1)
4. Each time $||\theta_h - x_m||_2^2$ is computed, set $l(x_m, \theta_h) = ||\theta_h - x_m||_2^2$
5. Assign upper bounds $u(x_m) = min_i||\theta_i - x_m||_2^2$
6. Repeat till a stopping criterion is fulfilled {
   - 6.1 **for all** θ_i and θ_j, compute $||\theta_i - \theta_j||_2^2$. **For all** centers θ_i, compute $s(\theta_i) = \frac{1}{2}\min_j ||\theta_i - \theta_j||_2^2$
   - 6.2 Identify all points x_m such that $u(x_m) \leq s(\theta(x_m))$.
   - 6.3 **for all** centers θ_i **for all** remaining points x_m check
     - $\theta_i \neq \theta(x_m)$ and
     - $u(x_m) > l(x_m, \theta_i)$ and
     - $u(x_m) > \frac{1}{2}||\theta(x_m) - \theta_i||_2^2$

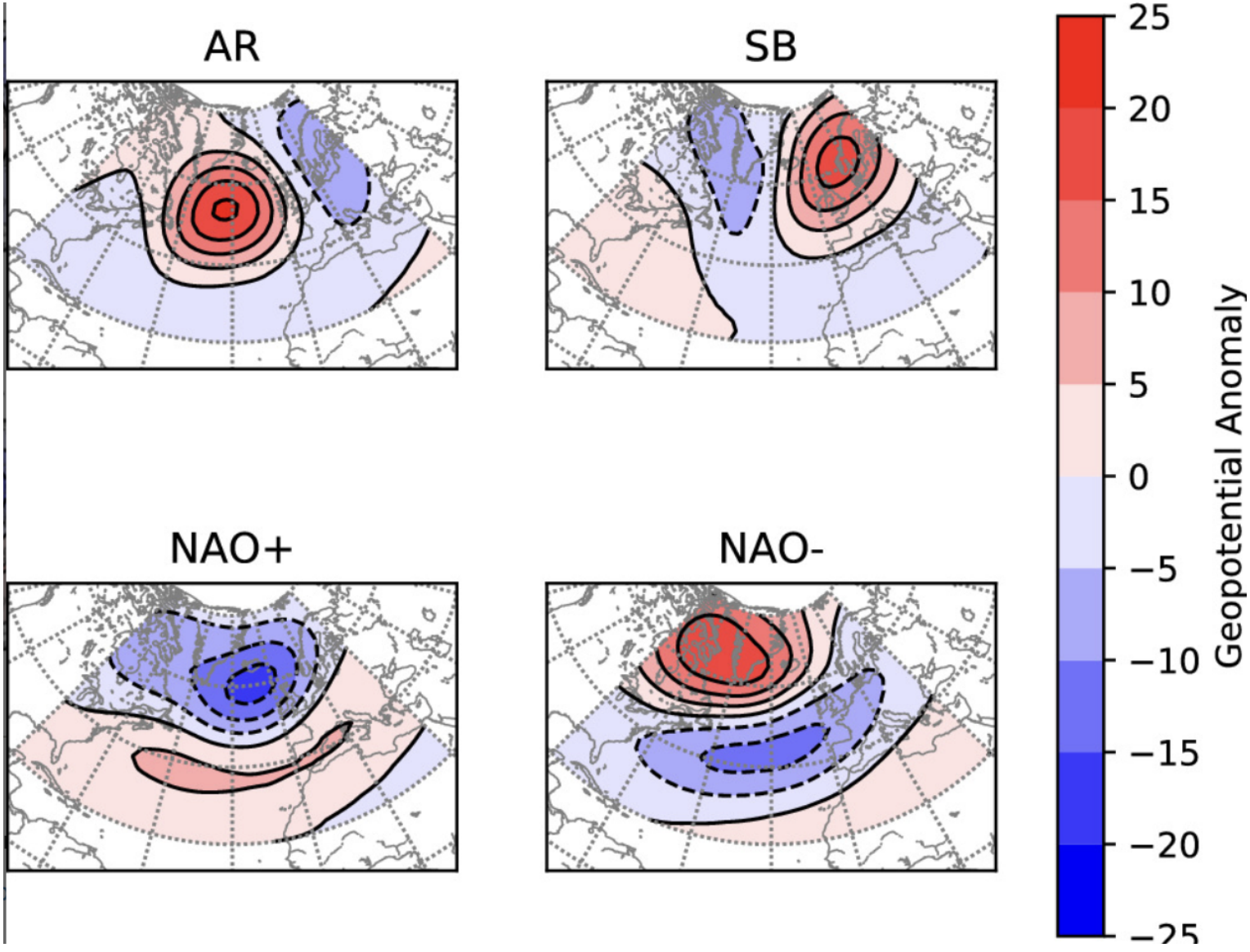
     If conditions $r(x_{m)=\text{true}}$ are true compute $\|x_m - \theta(x_m)\|$ and assign $r(x_m) =$ false. Otherwise $\|x_m - \theta(x_m)\|_2^2 = u(x_m)$.

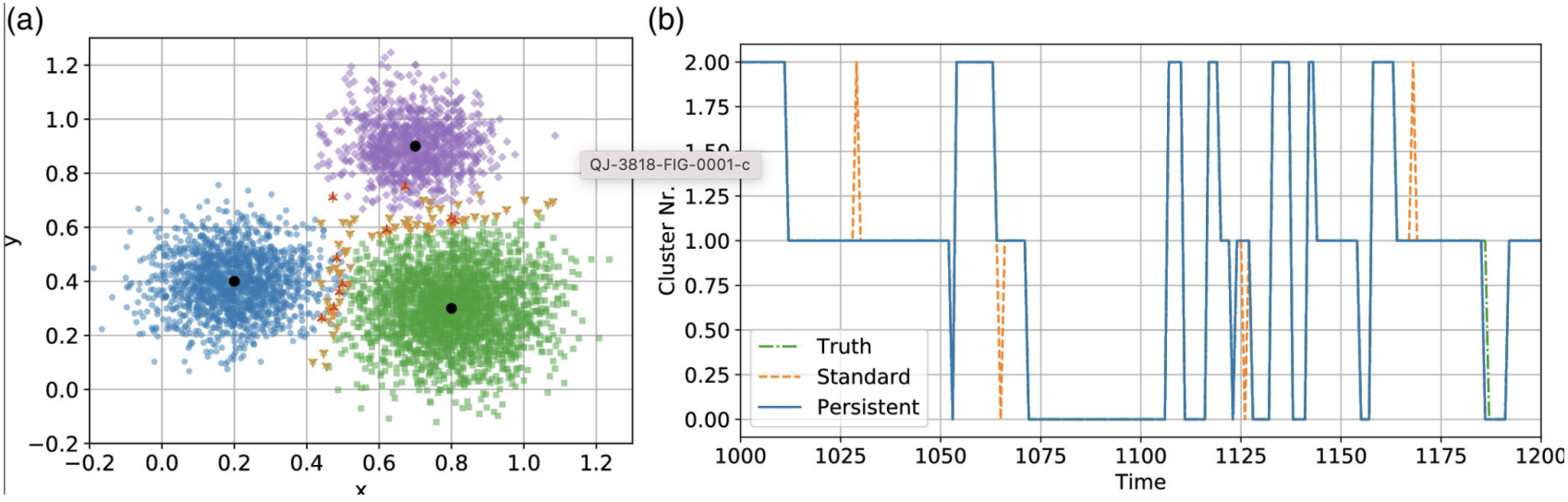
   - 6.4 if $\|x_m - \theta(x_m)\|_2^2 > l(x_m, \theta_i)$ or $\|x_m - \theta(x_m)\} > \frac{1}{2}||\theta(x_m) - \theta_i||_2^2$ then
     - compute $\|(x_m - \theta_i)\|_2^2$
     - if $\|(x_m - \theta_i)\|_2^2 < \|(x_m - \theta(x_m))\|_2^2$ then assign $\theta(x_m) = \theta_i$
7. **for all** centers θ_i, let $m(\theta_i)$ be the mean of the points assigned to θ_i
8. **for all** points x_m and **for all** centers θ_i assign $l(x_m, \theta_i) = \max\{l(x_m, \theta_i) - \|\theta_i - m(\theta_i)\|_2^2, 0\}$
9. **for all** points x_m, assign $u(x_m) = u(x_m) + \|m(\theta(x_m)) - \theta(x_m)\|$ and $r(x_m) =$ true
10. replace each center θ_i with $m(\theta_i)$
11. **return** $\theta_1, \ldots, \theta_K$

Example: pattern recognition for atmospheric circulation regimes

Regime

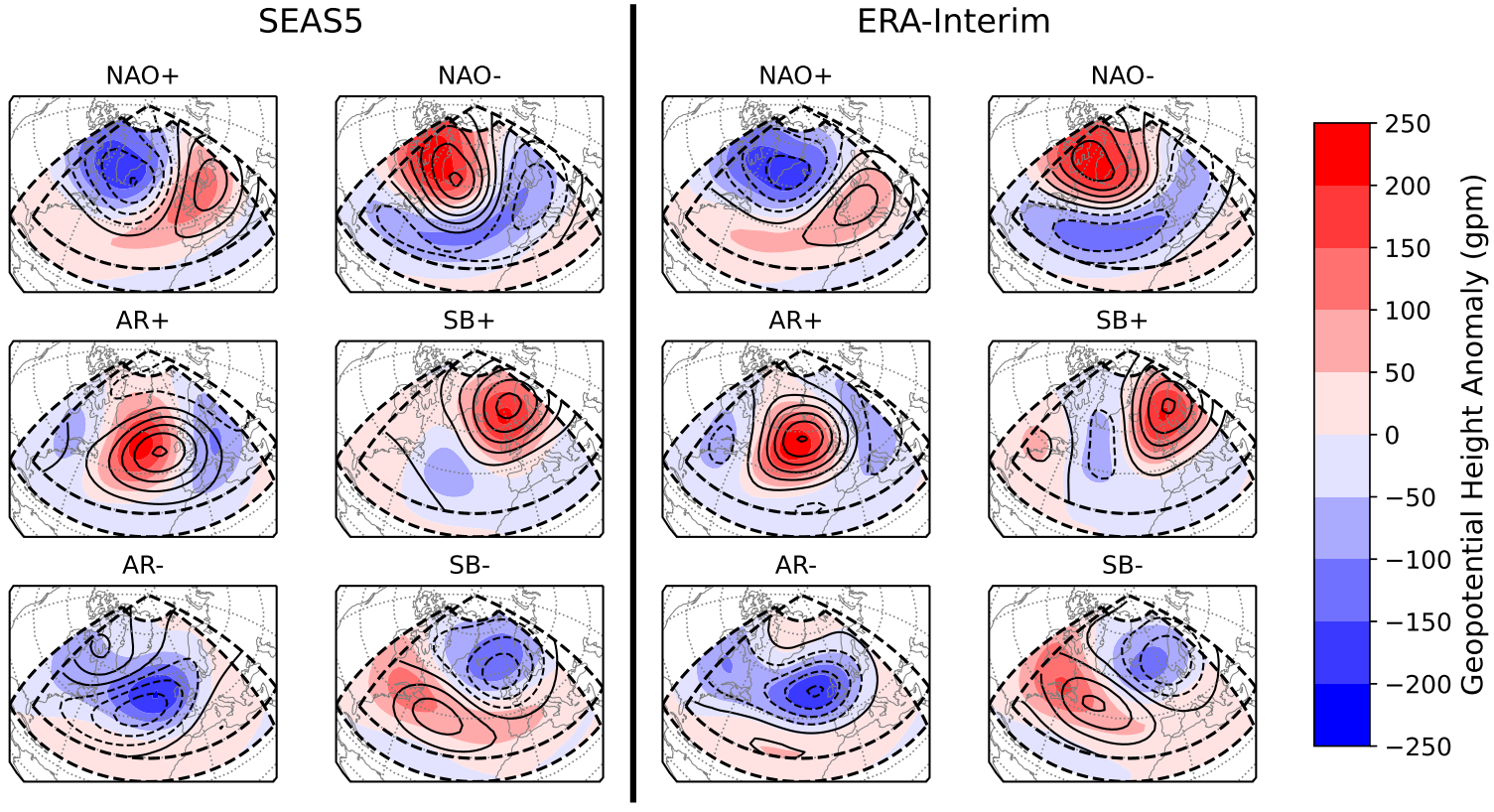


Time persistency constraint

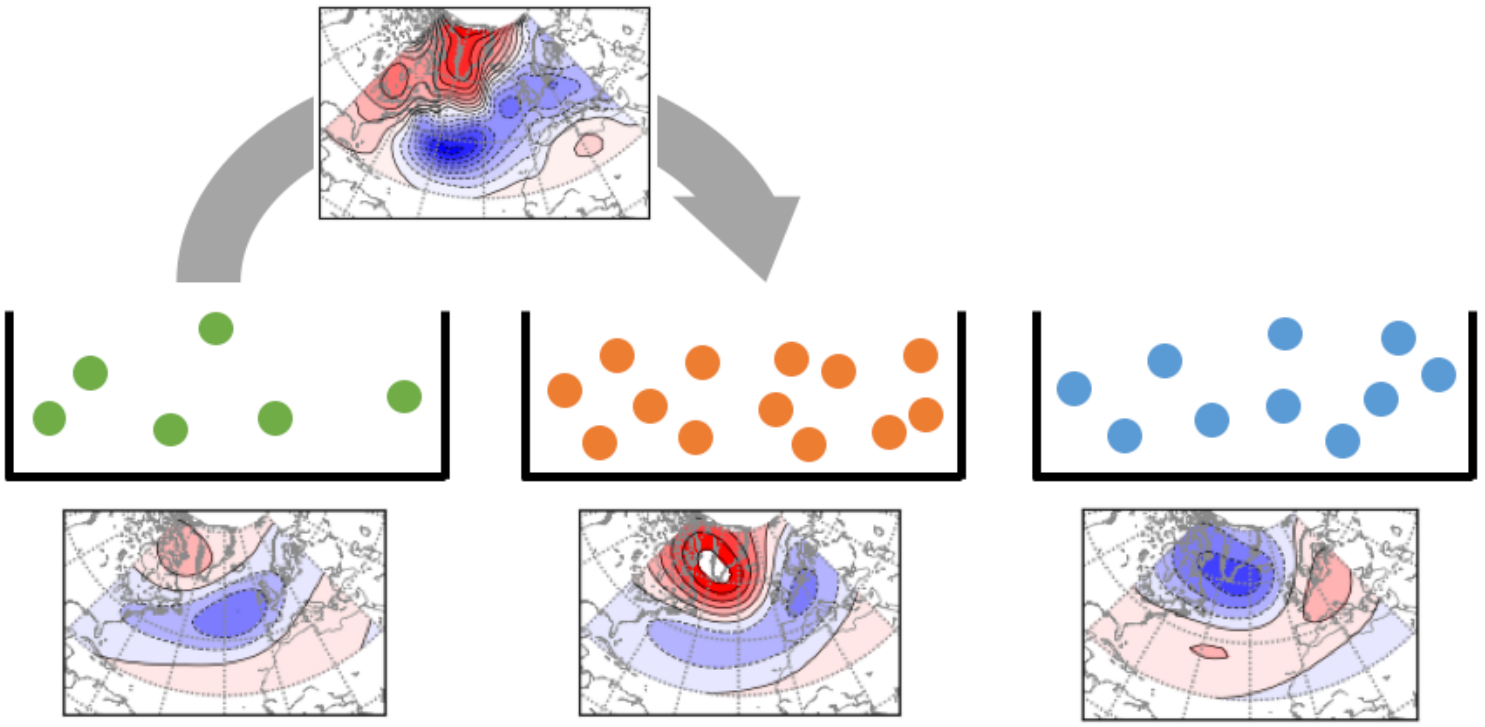


$$\sum_{t=1}^{T-1} |\gamma_k(t+1) - \gamma_k(t)| \leq N_C \quad \forall k$$

k-means clustering for different domains



k-means clustering for different domains



Optimisation problem

$$\mathbf{L}(\Theta, \Gamma) = \sum_{t=0}^{T} \sum_{n=1}^{N} \sum_{i=1}^{k} \gamma_i(t, n) \|x_{t,n} - \theta_i\|^2$$

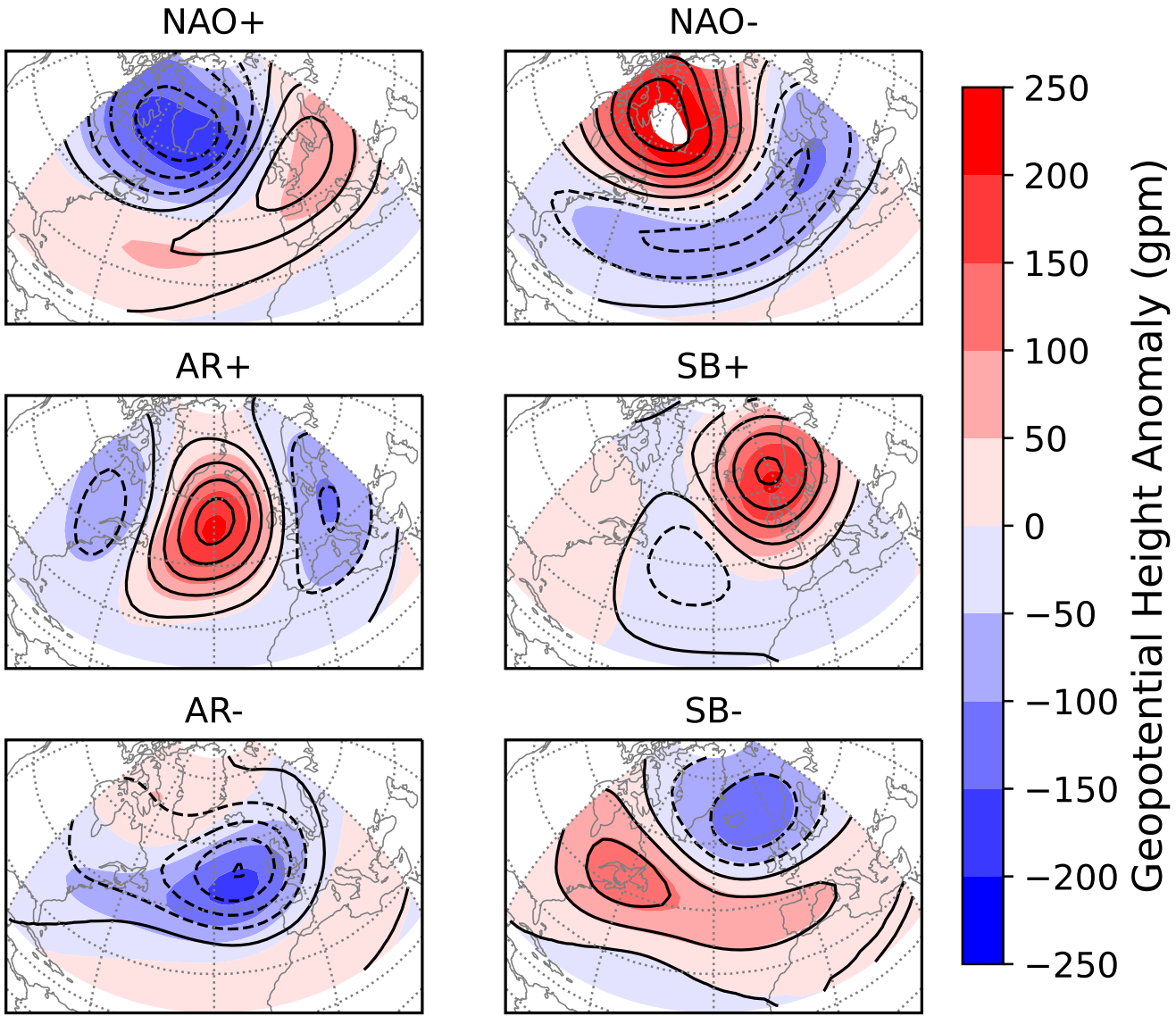
with

$$\sum_{i=1}^{k} \gamma_i(t, n) = 1, \qquad \forall t \in [0, T], \quad \forall n \in [1, N].$$

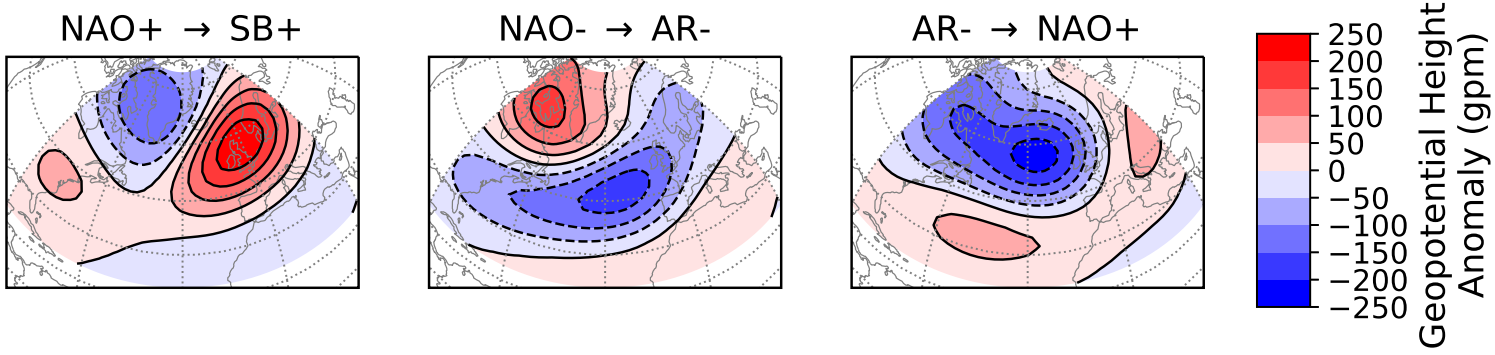
and

$$\sum_{i=1}^{k} \sum_{n_1, n_2} |\gamma_i(t, n_1) - \gamma_i(t, n_2)| \leq \phi \cdot C_{\text{eq}}, \qquad \forall t \in [0, T],$$

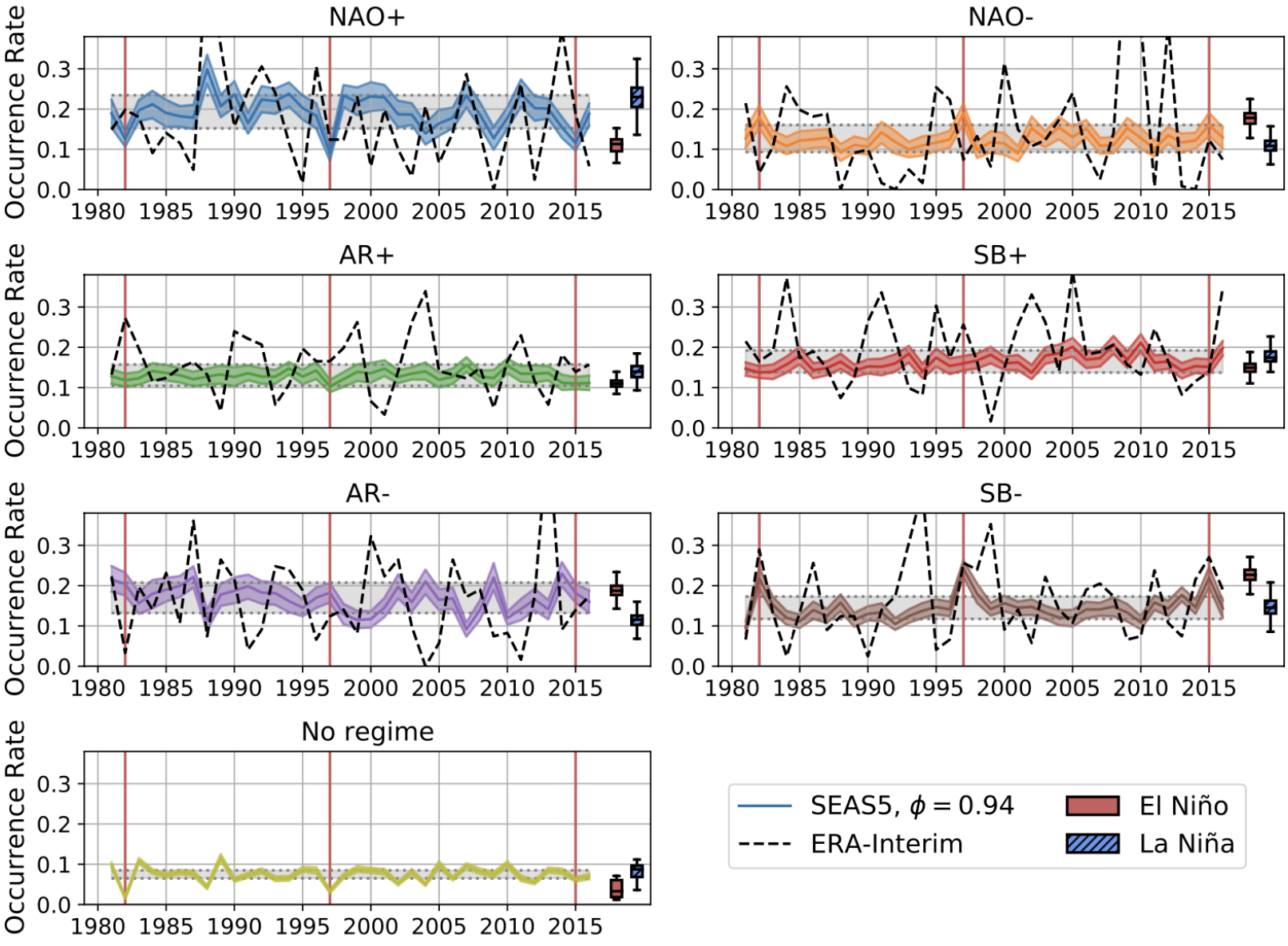
Ensemble persistency constraint



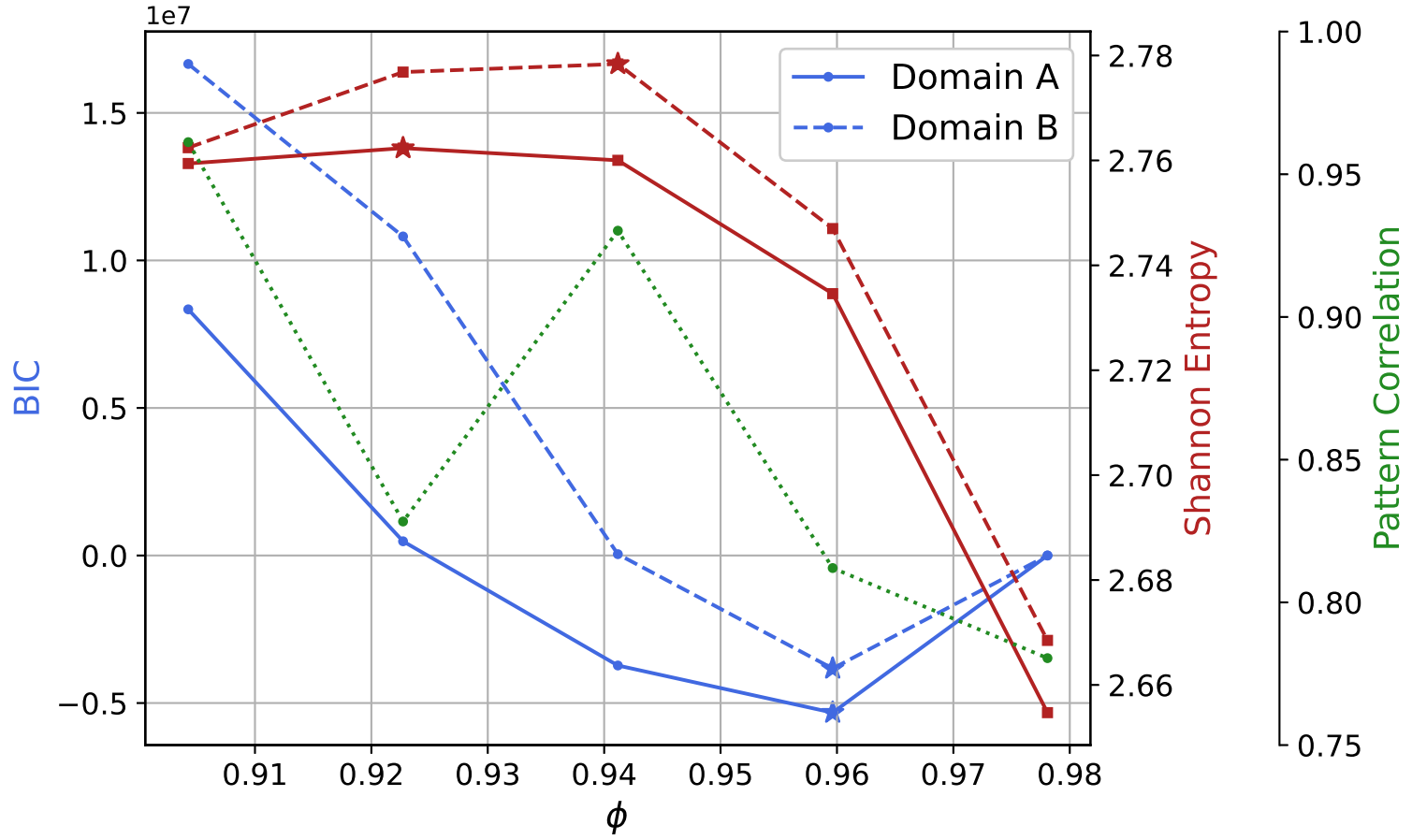
Ensemble persistency constraint



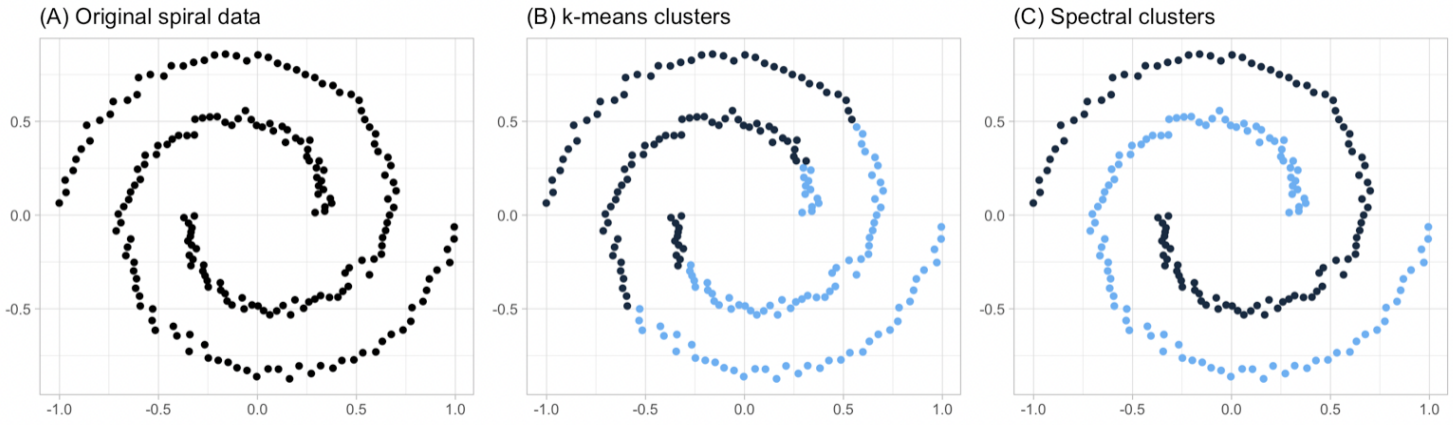
Occurrence rates



Optimal ϕ


1e7
Domain A
Domain B
BIC
Shannon Entropy
Pattern Correlation
ϕ
1.5
1.0
0.5
0.0
−0.5
2.78
2.76
2.74
2.72
2.70
2.68
2.66
1.00
0.95
0.90
0.85
0.80
0.75
0.91
0.92
0.93
0.94
0.95
0.96
0.97
0.98

K-Means vs Spectral Clustering



Eigenvalues and Eigenvectors

Definition

Let V be a K-Vector space, $f : V \to V$ an Endomorphismus, $\lambda \in K$. The scalar λ is called **Eigenvalue** of f, if there is a vector $v \in V, v \neq 0$, so that

$$f(v) = \lambda \cdot v.$$

The vector v is called **Eigenvector** of f an Eigenvalue λ.

Note: An Eigenvalue λ can be $0 \in K$, but an Eigenvector is always $\neq 0$.

Theorem

Let V be a K-vector space, $n = \dim V < \infty$ and $f : V \to V$ an Endomorphismus. The following two are equivalent:

1. *V has a basis of Eigenvectors of f.*
2. *There is a Basis $\mathcal{B}$ of V, so that*

$$M_{\mathcal{B}}^{\mathcal{B}}(f) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \text{ with } \lambda_i \in K.$$

$D = S A S^{-1}$

$(1\,1\,1)$

Characteristic Polynom

Definition

Let $A \in K^{n \times n}$ and $\lambda \in K$ abitrary. Then

$$\text{Eig}(A, \lambda) := \{v \in K^n \mid Av = \lambda v\}$$

is called the **Eigenspace** of A with respect to λ.

$$\chi_A(t) := \det(A - tE) \in K[t]$$

is called the **charakteristisches Polynom** of A.

Remark: For a matrix $A \in K^{n \times n}$ the following holds:

$$\lambda \in K \text{ is an Eigenvalue of } A \Leftrightarrow \text{Eig}(A, \lambda) \neq 0.$$

Theorem

Let $A \in K^{n \times n}$ and $\lambda \in K$. Then:

$$\lambda \text{ is an Eigenvalue of } A \Leftrightarrow \lambda \text{ is a root of } \chi_A(t).$$

Multiplicity

Definition

Let $P(t) \in K[t]$ be a Polynom. $P(t)$ can be decomposed over K in **Linear factors** if and only if there are $\lambda_1, \ldots, \lambda_n \in K, c \in K$, so that

$$P(t) = c \cdot (t - \lambda_1) \cdots (t - \lambda_n) = c \cdot \prod_{j=1}^{r} (t - \lambda'_j)^{m_j},$$

where $m_j \in \mathbb{N}$ and $\lambda'_1, \ldots, \lambda'_r \in \{\lambda_1, \ldots, \lambda_n\}$ are pairwise different. m_j is called the **Multiplicity** of the root λ'_j. It holds that

$$\sum_{j=1}^{r} m_j = n.$$

Using the Triangle Inequality to Accelerate k-Means

Algorithm:

1. Initialize the centre of the cluster $\theta_1, \ldots, \theta_K \in \mathbb{R}^n$ randomly
2. Set lower bounds to $l(x_m, \theta_i) = 0$ for all θ_i and x_m
3. Assign each x_m to its closest initial center $\theta(x_m) = \arg\min_h ||\theta_h - x_m||_2^2$ (avoid redundant calculations using Lemma 1)
4. Each time $||\theta_h - x_m||_2^2$ is computed, set $l(x_m, \theta_h) = ||\theta_h - x_m||_2^2$
5. Assign upper bounds $u(x_m) = min_i||\theta_i - x_m||_2^2$
6. Repeat till a stopping criterion is fulfilled {
   - 6.1 **for all** θ_i and θ_j, compute $||\theta_i - \theta_j||_2^2$. **For all** centers θ_i, compute $s(\theta_i) = \frac{1}{2}\min_j ||\theta_i - \theta_j||_2^2$
   - 6.2 Identify all points x_m such that $u(x_m) \leq s(\theta(x_m))$.
   - 6.3 **for all** centers θ_i **for all** remaining points x_m check
     - $\theta_i \neq \theta(x_m)$ and
     - $u(x_m) > l(x_m, \theta_i)$ and
     - $u(x_m) > \frac{1}{2}||\theta(x_m) - \theta_i||_2^2$

     If conditions $r(x_m) =$ true are true compute $\|x_m - \theta(x_m)\|$ and assign $r(x_m) =$ false. Otherwise $\|x_m - \theta(x_m)\|_2^2 = u(x_m)$.
   - 6.4 if $\|x_m - \theta(x_m)\|_2^2 > l(x_m, \theta_i)$ or $\|x_m - \theta(x_m)\} > \frac{1}{2}||\theta(x_m) - \theta_i||_2^2$ then
     - compute $\|(x_m - \theta_i)\|_2^2$
     - if $\|(x_m - \theta_i)\|_2^2 < \|(x_m - \theta(x_m))\|_2^2$ then assign $\theta(x_m) = \theta_i$
7. **for all** centers θ_i, let $m(\theta_i)$ be the mean of the points assigned to θ_i
8. **for all** points x_m and **for all** centers θ_i assign $l(x_m, \theta_i) = \max\{l(x_m, \theta_i) - \|\theta_i - m(\theta_i)\|_2^2, 0\}$
9. **for all** points x_m, assign $u(x_m) = u(x_m) + \|m(\theta(x_m)) - \theta(x_m)\|$ and $r(x_m) =$ true
10. replace each center θ_i with $m(\theta_i)$
11. **return** $\theta_1, \ldots, \theta_K$

Eigenvalues and Eigenvectors

Definition

Let V be a K-Vector space, $f : V \to V$ an Endomorphismus, $\lambda \in K$. The scalar λ is called **Eigenvalue** of f, if there is a vector $v \in V, v \neq 0$, so that

$$f(v) = \lambda \cdot v.$$

The vector v is called **Eigenvector** of f an Eigenvalue λ.

Note: An Eigenvalue λ can be $0 \in K$, but an Eigenvector is always $\neq 0$.

Theorem

Let V be a K-vector space, $n = \dim V < \infty$ and $f : V \to V$ an Endomorphismus. The following two are equivalent:

1. *V has a basis of Eigenvectors of f.*
2. *There is a Basis $\mathcal{B}$ of V, so that*

$$M_{\mathcal{B}}^{\mathcal{B}}(f) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \text{ with } \lambda_i \in K.$$

Characteristic Polynom

Definition

Let $A \in K^{n \times n}$ and $\lambda \in K$ abitrary. Then

$$\text{Eig}(A, \lambda) := \{v \in K^n \mid Av = \lambda v\}$$

is called the **Eigenspace** of A with respect to λ.

$$\chi_A(t) := \det(A - tE) \in K[t]$$

is called the **charakteristisches Polynom** of A.

Remark: For a matrix $A \in K^{n \times n}$ the following holds:

$$\lambda \in K \text{ is an Eigenvalue of } A \Leftrightarrow \text{Eig}(A, \lambda) \neq 0.$$

Theorem

Let $A \in K^{n \times n}$ and $\lambda \in K$. Then:

$$\lambda \text{ is an Eigenvalue of } A \Leftrightarrow \lambda \text{ is a root of } \chi_A(t).$$

Multiplicity

Definition

Let $P(t) \in K[t]$ be a Polynom. $P(t)$ can be decomposed over K in **Linear factors** if and only if there are $\lambda_1, \ldots, \lambda_n \in K, c \in K$, so that

$$P(t) = c \cdot (t - \lambda_1) \cdots (t - \lambda_n) = c \cdot \prod_{j=1}^{r} (t - \lambda'_j)^{m_j},$$

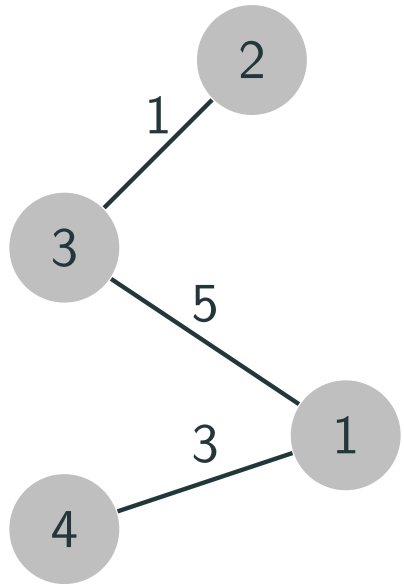
where $m_j \in \mathbb{N}$ and $\lambda'_1, \ldots, \lambda'_r \in \{\lambda_1, \ldots, \lambda_n\}$ are pairwise different. m_j is called the **Multiplicity** of the root λ'_j. It holds that

$$\sum_{j=1}^{r} m_j = n.$$

What is a graph (formally)?

The objects on the following slides will play a major role in this course.

- $G = (V, E, \omega)$, where $V \neq \emptyset$ is a set (called the **vertex set**), $E \subset \binom{V}{2} = \{\{u, v\} : u, v \in V\}$ (called the **edge set**) and $\omega : E \to \mathbb{R}^+$, is called a **(weighted) graph**
- usually we choose (or rename) $V = \{1, 2, \ldots, n\}$ and use the notations $ij = \{i, j\}$ for $\{i, j\} \in E$ and $\omega_{ij} = \omega(ij)$
- for every $i \in V$ define $N(i) := \{j \in V : ij \in E\}$, called the **neighbourhood** of i (in G); elements of $N(i)$ are called **neighbours** of i (those elements are **adjacent** to i)



$w(23) = 1$,
$N(4) = \{1\}$,
$d(1) = |\{3, 4\}| = 2$

Graph classes

Well known graph classes are:

- the **path graph** P_n has vertex set $\{1, 2, \ldots, n\}$ and edge set $\{\{1, 2\}, \{2, 3\}, \ldots, \{n-1, n\}\}$
- the **cycle graph** C_n has vertex set $\{1, 2, \ldots, n\}$ and edge set $\{\{1, 2\}, \{2, 3\}, \ldots, \{n-1, n\}, \{n, 1\}\}$
- the **complete graph** K_n consists of n vertices which are all adjacent to each other
- the **complete bipartite graph** $K_{m,n}$ has two sets V_1 and V_2 of vertices of sizes m and n, such that the edge set consists of all possible edges between V_1 and V_2

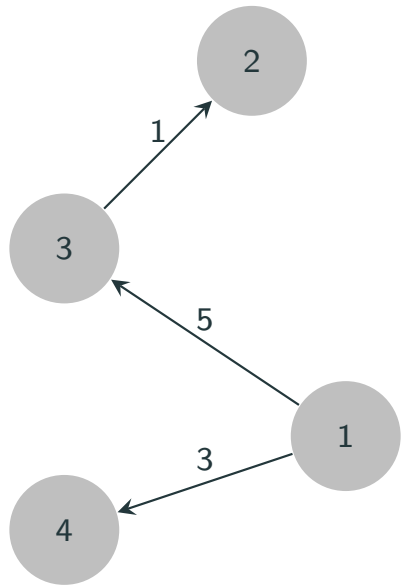
A set of vertices in a graph which are all adjacent to each other (they **induce** a complete (sub)graph), is called **clique**.

The graph $K_{1,n}$ is called a **star**.

What is a digraph (formally)?

Edges can have a direction.

- $G = (V, E, \omega)$, where $V \neq \emptyset$ is a set, $E \subset V \times V$ (this is sometimes also called the **set of arcs**) and $\omega : E \to \mathbb{R}^+$, is called a **(weighted) digraph**
- for $(i, j) \in E$ the vertex i is called **predecessor** of j and j is called **successor** of i
- similar notation simplifications as before
- $N^+(i) := \{j \in V : (i, j) \in E\}$ is the **out-neighbourhood** of i, $N^-(i) := \{j \in V : (j, i) \in E\}$ is the **in-neighbourhood** of i
- $d^+(i) := |N^+(i)|$ is the **out-degree** of i and $d^-(i) := |N^-(i)|$ is the **in-degree** of i



$N^-(3) = \{1\}$,
$N^+(4) = \emptyset$,
$d^+(1) = 2$,
$d^-(2) = 1$

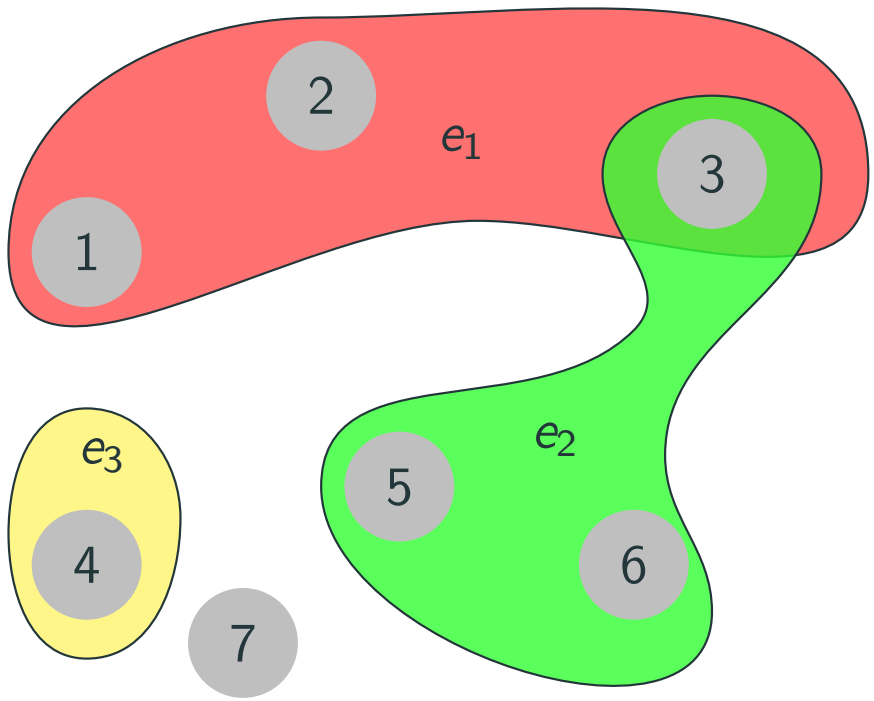
Example of a multigraph

It is sometimes necessary to allow multiple edges between two vertices or a **loop** (a self-edge). In that case we use the term **multigraph**.

What is a hypergraph (formally)?

Sometimes more than two vertices need to form an edge (certain real life situations' have this property).

- natural generalisation is a **hypergraph** $H = (V, E)$, where
  - $V \neq \emptyset$ is (also) a set, but
  - E can be an arbitrary subset (the elements are called **hyperedges**) of the power set $\mathcal{P}(V)$
- if all hyperedges are of the same size r, then H is called r-**uniform**



Storing graphs

Certain matrices and lists can be associated with a graph (we will see more examples later).

- **affinity matrix** $W(G)$:

$$w_{ij} = \begin{cases} \omega_{ij} & \text{if } \{i,j\} \in E, \\ 0 & \text{else.} \end{cases}$$

- **adjacency matrix** $A(G)$: special case of $W(G)$, where $w_{ij} = 1$ for all $ij \in E$.
- **adjacency list**:
  - associate list to every vertex containing its neighbours
  - call list of these lists adjacency list of the graph (treated differently in the literature)
  - not very useful for mathematical arguments
  - especially useful (for storing) when $A(G)$ is sparse

All the above constructions are valid for directed graphs.

How to transform a hypergraph into a graph?

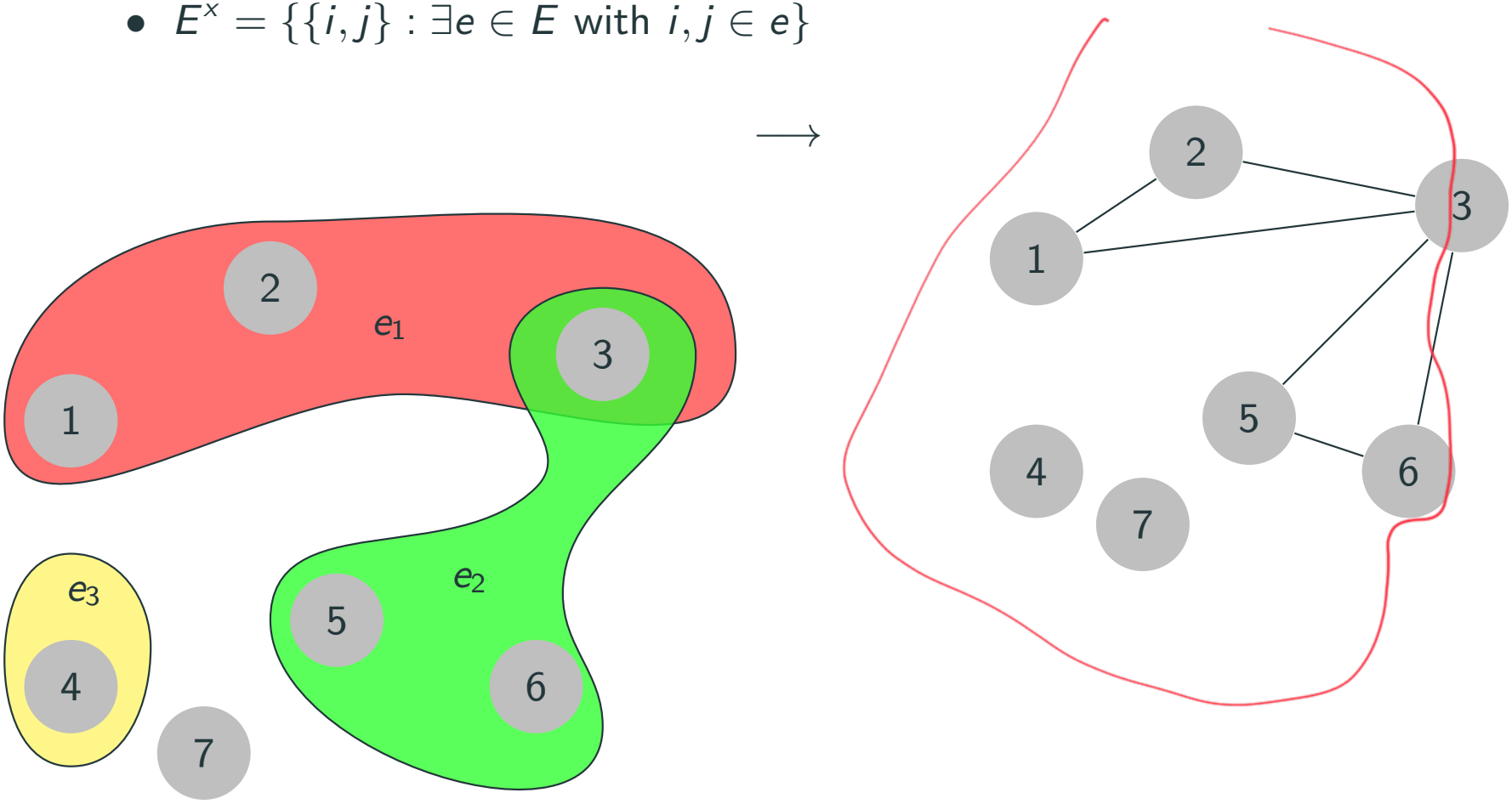
The following constructions are standard.

- clique expansion
  - the vertex set is V
  - each hyperedge e is replaced by an edge for every pair of vertices in e
  - this construction yields cliques for every hyperedge
- star expansion
  - vertex set is $V \cup E$
  - edge between u and e iff $u \in e$
  - every hyperedge corresponds to a star
- there are more...

Clique expansion

The clique expansion $G^x = (V^x, E^x)$ is constructed from $H = (V, E)$ via:

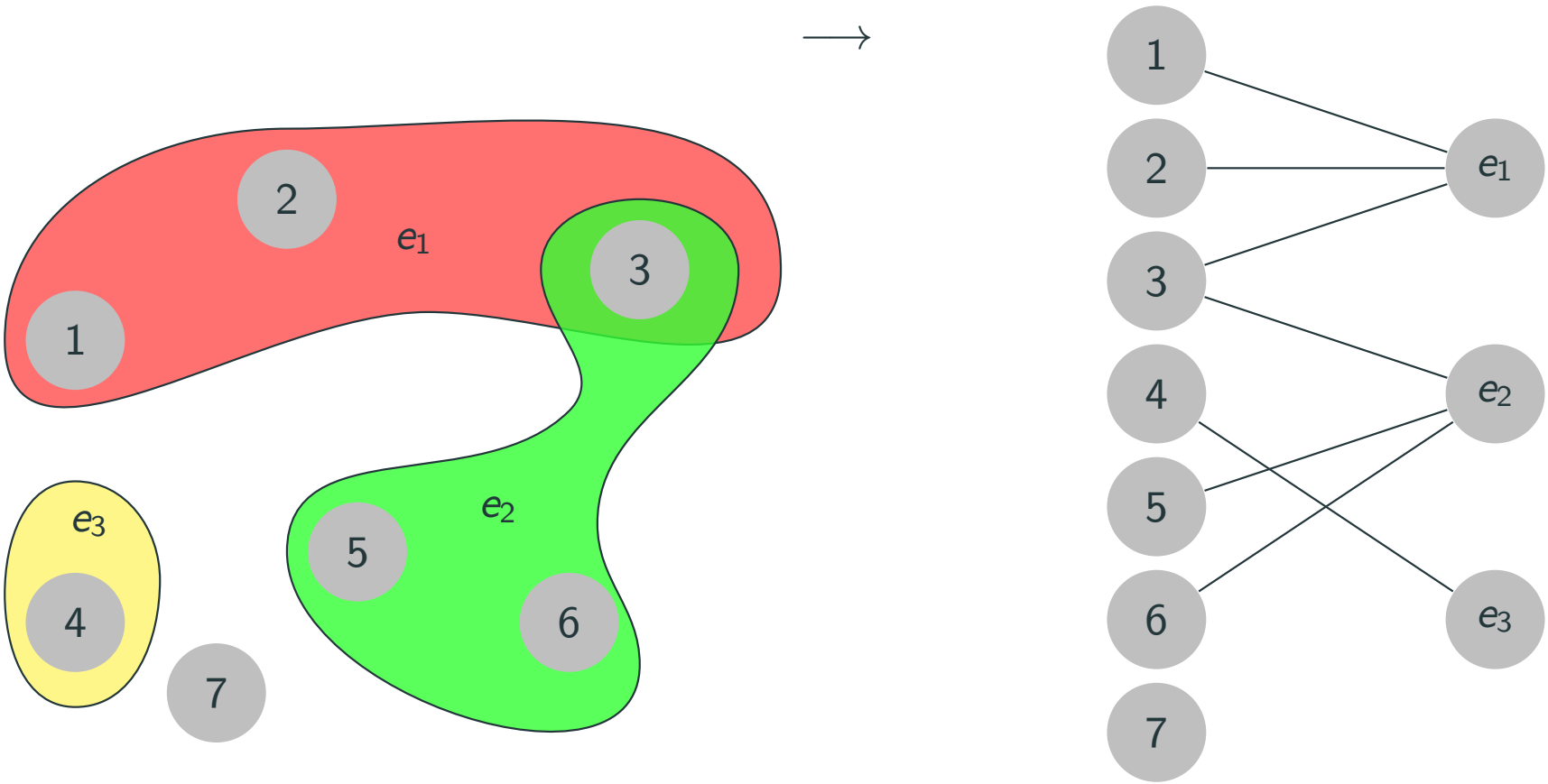
- $V^x = V$
- $E^x = \{\{i, j\} : \exists e \in E \text{ with } i, j \in e\}$



Star expansion

The star expansion $G^* = (V^*, E^*)$ is constructed from $H = (V, E)$ via:

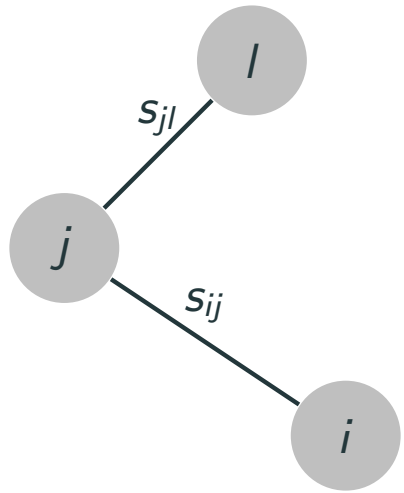
- $V^* = V \cup E$
- $E^* = \{\{i, e\} : i \in e, e \in E\}$



What if data without network structure is given?

Solution: Build your own graph!

- given a set of data points $x_1, x_2 \ldots, x_n$ and some notion of similarity[1] $s_{ij} \geq 0$ between all pairs of data points x_i and x_j
- build graph $G = (V, E)$, where the vertex i represents the data point x_i, so $V = \{1, 2, \ldots, n\}$
- $\{i, j\} \in E$ if $s_{ij} > 0$
- edge weight $\omega_{ij} = s_{ij}$ (edge weights represent similarities)
- G is called **similarity graph** (although with this particular choice of edges it is often referred to as the **fully connected graph**)



graph for $\{x_i, x_j, x_l\}$ with $s_{ij}, s_{jl} > 0$ and $s_{il} = 0$

The ε-neighbourhood graph

The ε-**neighbourhood graph** is constructed as follows:

- vertices are data points
- fix some $\varepsilon > 0$
- connect all vertices whose similarities are smaller than ε
- since ε is usually small, values of existing edges are roughly of the same scale
- hence usually unweighted

The (mutual) k-nearest neighbour graph

The k-**nearest neighbour graph** is constructed as follows:

- vertices are data points
- fix some $k > 0$
- connect i to the k nearest (w.r.t. s_{ij}) k vertices via an edge starting at i
- obtain an undirected graph by ignoring the directions

The **mutual** k-**nearest neighbour graph** is constructed as follows:

- vertices are data points
- fix some k
- connect i to the k nearest (w.r.t. s_{ij}) k vertices via an edge starting at i
- obtain an undirected graph by deleting all non symmetric edges

Laplacian matrix (and another graph definition)

The degree matrix $D(G)$ is given by

$$d_{ij} = \begin{cases} \sum_{l \in N(i)} w_{il} & \text{if } i = j, \\ 0 & \text{else.} \end{cases}$$

Laplacian matrix:

$$L(G) = D(G) - W(G)$$

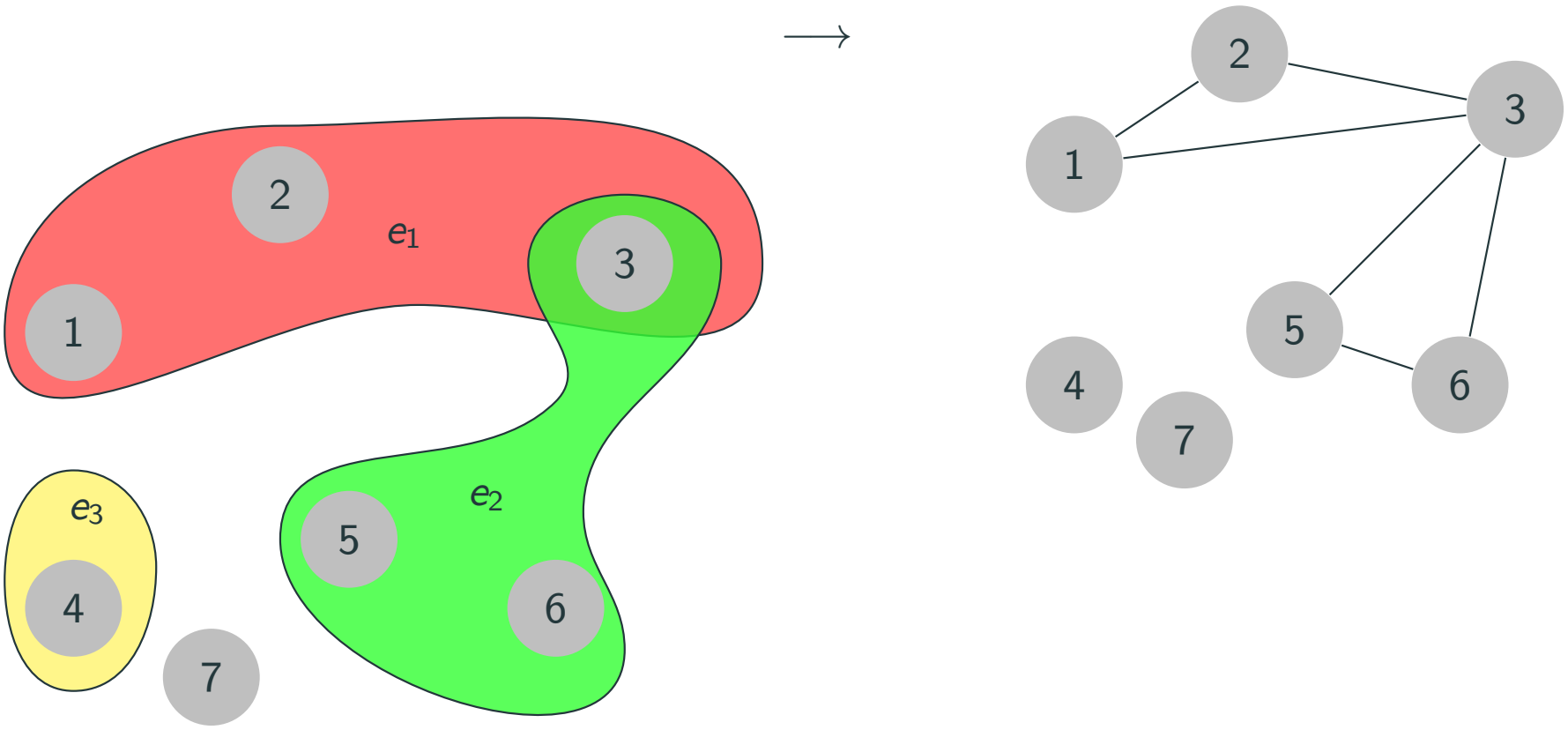
We also need:

$$(A) = \sum_{ij \in E, i,j \in A} \omega_{ij} \text{ for } A \subset V \text{ (no double counting!)}$$

Clique expansion

The clique expansion $G^x = (V^x, E^x)$ is constructed from $H = (V, E)$ via:

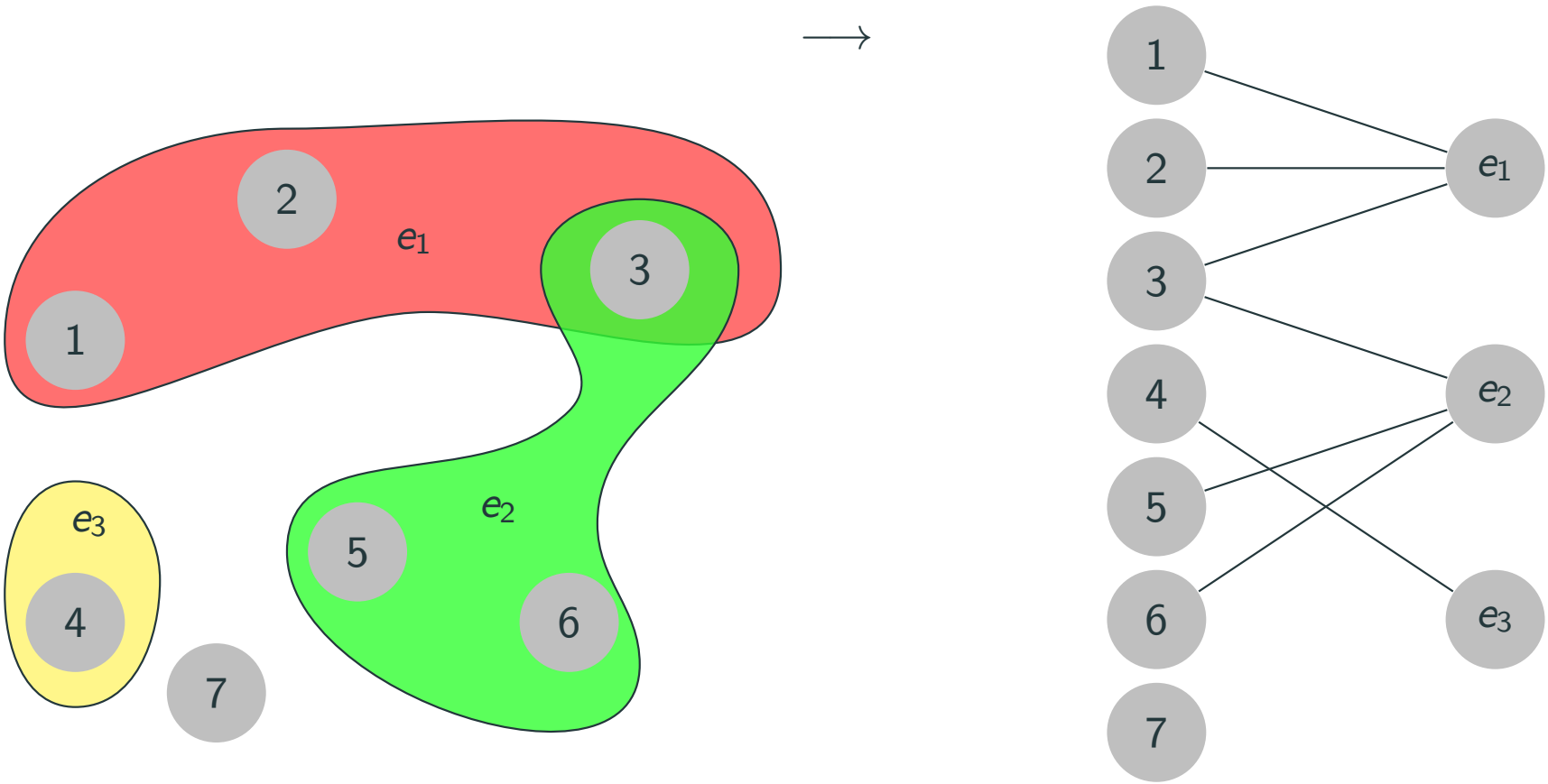
- $V^x = V$
- $E^x = \{\{i, j\} : \exists e \in E \text{ with } i, j \in e\}$



Star expansion

The star expansion $G^* = (V^*, E^*)$ is constructed from $H = (V, E)$ via:

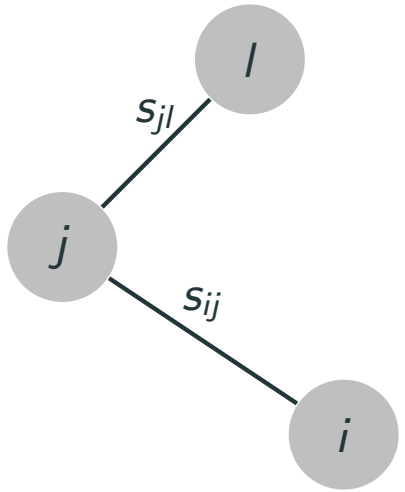
- $V^* = V \cup E$
- $E^* = \{\{i, e\} : i \in e, e \in E\}$



What if data without network structure is given?

Solution: Build your own graph!

- given a set of data points $x_1, x_2 \ldots, x_n$ and some notion of similarity[2] $s_{ij} \geq 0$ between all pairs of data points x_i and x_j
- build graph $G = (V, E)$, where the vertex i represents the data point x_i, so $V = \{1, 2, \ldots, n\}$
- $\{i, j\} \in E$ if $s_{ij} > 0$
- edge weight $\omega_{ij} = s_{ij}$ (edge weights represent similarities)
- G is called **similarity graph** (although with this particular choice of edges it is often referred to as the **fully connected graph**)



graph for $\{x_i, x_j, x_l\}$ with $s_{ij}, s_{jl} > 0$ and $s_{il} = 0$

The ε-neighbourhood graph

The ε-**neighbourhood graph** is constructed as follows:

- vertices are data points
- fix some $\varepsilon > 0$
- connect all vertices whose similarities are smaller than ε
- since ε is usually small, values of existing edges are roughly of the same scale
- hence usually unweighted

The (mutual) k-nearest neighbour graph

The k-**nearest neighbour graph** is constructed as follows:

- vertices are data points
- fix some $k > 0$
- connect i to the k nearest (w.r.t. s_{ij}) k vertices via an edge starting at i
- obtain an undirected graph by ignoring the directions

The **mutual** k-**nearest neighbour graph** is constructed as follows:

- vertices are data points
- fix some k
- connect i to the k nearest (w.r.t. s_{ij}) k vertices via an edge starting at i
- obtain an undirected graph by deleting all non symmetric edges

Graph Partitioning and Community Detection

Difference

Graph Partitioning (GP)

- partition vertices into given number of groups
- sizes of groups are (roughly) fixed
- many edges inside groups, few edges between groups
- goal: dividing network into smaller more manageable pieces
- example:
    - numerical solution of network processes on a parallel computer

Community Detection (CD)

- partition vertices into groups
- sizes of groups are not fixed
- many edges inside groups, few edges between groups
- goal: understanding structure of a network
- examples:
    - collaboration
    - related web pages

Why is partitioning hard?

Problem

Partition vertex set into two parts (*graph bisection*).

n vertices into parts of sizes n_1 and n_2 ($n_1 + n_2 = n$):

- $\frac{n!}{n_1!n_2!}$ possibilities (half of it if order is ignored and $n_1 = n_2$)
- using Stirling's formula $n! \approx \sqrt{2\pi n}(n/e)^n$ we get

$$\frac{n!}{n_1!n_2!} \approx \frac{n^{n+1/2}}{n_1^{n_1+1/2} n_2^{n_2+1/2}}$$

- for a balanced partition ($n_1 \approx n_2$):

$$\text{roughly } \frac{2^{n+1}}{\sqrt{n}} \text{ possibilities}$$

Therefore, exhausitive search is usually unfeasible.

Arbitrary number of classes

Methods for graph bisection can be generalised (other ways are possible):

- number of vertices: n
- number of classes: k
- define $l = k - 2^{\lfloor \log_2 k \rfloor}$
- for $r = 1, \ldots, l$ (ascending order) apply graph bisection method with

$$n_1^{(r)} = n - \frac{r \cdot n}{k} \quad \text{and} \quad n_2^{(r)} = \frac{n}{k}$$

- apply (equally sized) graph bisection $\lfloor \log_2 k \rfloor$ times to the $n - \frac{l \cdot n}{k}$ vertices in the large class
- results in k classes of (almost) same size $\frac{n}{k}$

Graph cuts

Given a graph $G = (V, E, \omega)$.

- for disjoint $A, B \subset V$ define the *cut size*

$$(A, B) = \sum_{i \in A, j \in B} \omega_{ij}$$

- for a partition $\mathcal{A} = A_1, A_2, \ldots, A_k$ define

$$(A_1, A_2, \ldots, A_k) = \sum_{i=1}^{k} (A_i, \overline{A}_i)$$

- basic problem: find $\mathcal{A}$ minimizing $(A_1, A_2, \ldots, A_k)$

Spectral clustering

- mathematical foundation by Donath & Hoffman and Fiedler in 1973
- applications in various fields/for various problems
  - image segmentation
  - educational data mining
  - entity resolution
  - speech separation
  - ...

Laplacian matrix (and another graph definition)

The degree matrix $D(G)$ is given by

$$d_{ij} = \begin{cases} \sum_{l \in N(i)} w_{il} & \text{if } i = j, \\ 0 & \text{else.} \end{cases}$$

Laplacian matrix:

$$L(G) = D(G) - W(G)$$

We also need:

$$(A) = \sum_{ij \in E, i,j \in A} \omega_{ij} \text{ for } A \subset V \text{ (no double counting!)}$$

Spectral clustering and graph cuts

Clustering corresponds to finding cuts in graphs (see next slides; other interpretations possible), there are (usually) two types:

- RatioCut (balanced by number of vertices in each cluster)
- NCut (balanced by sum of edge weights in each cluster)

Formalization of RatioCut and NCut

- $\text{RatioCut}(A_1, A_2, \ldots, A_k) = \sum_{i=1}^{k} \frac{(A_i, \overline{A}_i)}{|A_i|}$
  - $\sum_{i=1}^{k}(1/|A_i|)$ is minimized if all A_i have the same size
  - hence minimizing RadioCut balances clusters by their number of vertices (as desired)
- $\text{NCut}(A_1, A_2, \ldots, A_k) = \sum_{i=1}^{k} \frac{(A_i, \overline{A}_i)}{(A_i)}$
  - $\sum_{i=1}^{k}(1/(A_i))$ is minimized if all (A_i) coincide
  - hence minimizing NCut balances clusters by their edge weights (as desired)
- corresponding optimization problems are NP hard
- spectral clustering is a way to solve relaxed versions of those problems

Useful properties of the Laplacian

Proposition: The Laplacian matrix L $(= D - W)$ satisfies the following properties:

i For every vector $f \in \mathbb{R}^n$ we have

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2.$$

ii L is symmetric and positive semi-definite.

iii The smallest eigenvalue of L is 0, the corresponding eigenvector is the constant one vector.

iv L has n non-negative, real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

Useful properties of the Laplacian

Proof:

i By the definition of d_i,

$$f^T L f = f^T D f - f^T W f = \sum_{i=1}^{n} d_i f_i^2 - \sum_{i,j=1}^{n} f_i f_j w_{ij}$$

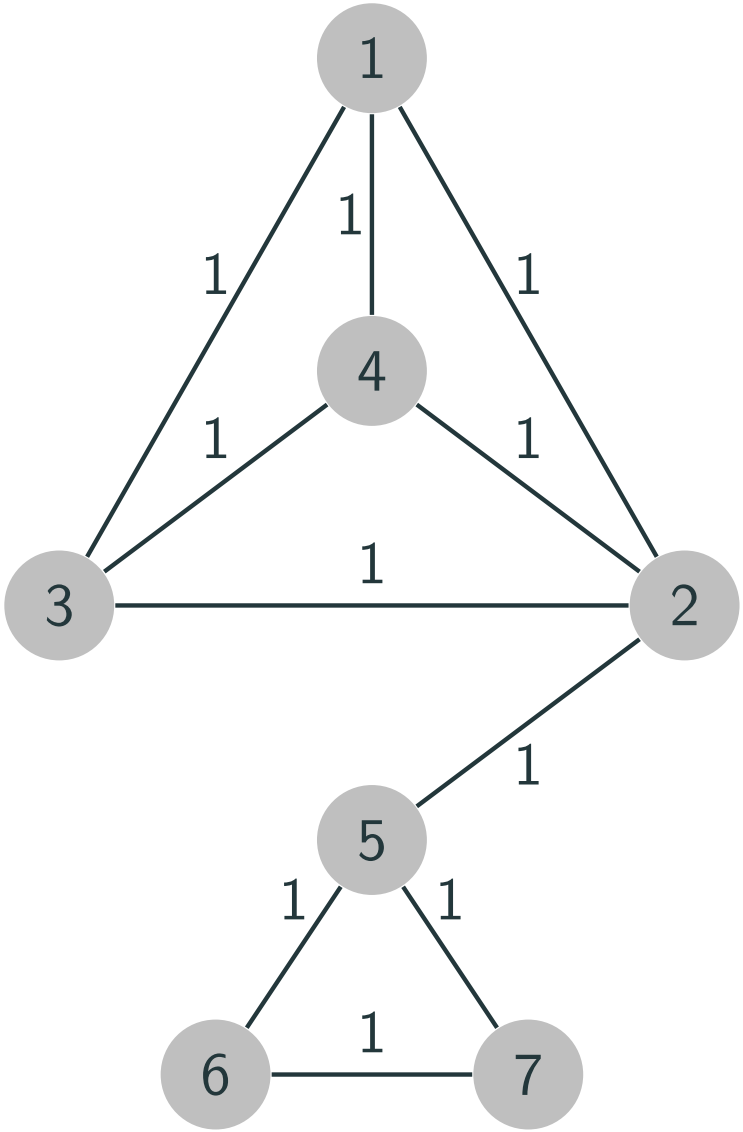
$$= \frac{1}{2}\left(\sum_{i=1}^{n} d_i f_i^2 - 2\sum_{i,j=1}^{n} f_i f_j w_{ij} + \sum_{j=1}^{n} d_j f_j^2\right) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2.$$

ii The symmetry of L follows directly from the symmetry of W and D. The positive semi-definiteness is a direct consequence of Part (i), which shows that $f^T L f \geq 0$ for all $f \in \mathbb{R}^n$.

iii All eigenvalues are real (symmetric matrix). The rest follows easily from Part (ii) and the defining equation of eigenvalues.

iv This is a direct consequence of (i)-(iii).

Graph cuts and spectral clustering



$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2$$

Connectivity

Proposition: The number of connected components of a graph is equal to the multiplicity of the first eigenvalue (which is 0) of the graph Laplacian matrix.

Relaxation of RatioCut for $k = 2$

Goal is to solve

$$\min_{A \subset V} \text{RatioCut}(A, \overline{A})$$

Reformulation: for $A \subset V$ define $f = (f_1, \ldots, f_n) \in \mathbb{R}^n$ (some kind of indicator vector) via

$$f_i = \begin{cases} \sqrt{|\overline{A}|/|A|} & \text{if } v_i \in A, \\ -\sqrt{|A|/|\overline{A}|} & \text{if } v_i \in \overline{A}. \end{cases}$$

It can be shown that

$$f^T L f = |V| \cdot \text{RatioCut}(A, \overline{A})$$

Relaxation of RatioCut for $k = 2$

Proof:

$$
\begin{aligned}
2f^T Lf &= \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \\
&= \sum_{i\in A, j\in \overline{A}} w_{i,j} \left(\sqrt{\frac{|\overline{A}|}{|A|}} \right. \\
&\quad \left. + \sqrt{\frac{|A|}{|\overline{A}|}} \right)^2 + \sum_{i\in \overline{A}, j\in A} w_{i,j} \left(-\sqrt{\frac{|\overline{A}|}{|A|}} - \sqrt{\frac{|A|}{|\overline{A}|}} \right)^2 \\
&= 2(A, \overline{A}) \left(\frac{|\overline{A}|}{|A|} + \frac{|A|}{|\overline{A}|} + 2 \right) \\
&= 2(A, \overline{A}) \left(\frac{|A| + |\overline{A}|}{|A|} + \frac{|A| + |\overline{A}|}{|\overline{A}|} \right) \\
&= 2|V| \cdot \mathrm{RatioCut}(A, \overline{A})
\end{aligned}
$$

Relaxation of RatioCut for $k = 2$

Further, f is orthogonal to 1:

$$\sum_{i=1}^{n} f_i = \sum_{i \in A} \sqrt{\frac{|\overline{A}|}{|A|}} - \sum_{i \in \overline{A}} \sqrt{\frac{|A|}{|\overline{A}|}} = |A| \sqrt{\frac{|\overline{A}|}{|A|}} - |\overline{A}| \sqrt{\frac{|A|}{|\overline{A}|}} = 0.$$

Finally

$$\|f\|^2 = \sum_{i=1}^{n} f_i^2 = |A| \frac{|\overline{A}|}{|A|} + |\overline{A}| \frac{|A|}{|\overline{A}|} = |\overline{A}| + |A| = n.$$

Hence the minimization problem is equivalent to

$$\min_{A \subset V} f^T L f$$

subject to the given f.

Relaxation of RatioCut for $k = 2$

This problem is NP-hard since the solution vector only takes two particular values.

Relaxing this problem is possible:

$$\min_{f \in \mathbb{R}^n} f^T L f \text{ subject to } f \perp 1, \|f\| = \sqrt{n}$$

The Rayleigh-Ritz theorem (see e.g. Strang - Linear algebra and its applications) gives the solution of this problem via an eigenvector corresponding to the second smallest eigenvalue of L.

Remark: Solution has to be transformed!

Special case ($k = 2$) of unnormalized spectral clustering

Algorithm: Unnormalized spectral clustering ($k = 2$)

Input: Weight matrix (or any similarity matrix) $S \in \mathbb{R}^{n \times n}$

1 Construct similarity graph G;
2 Compute $L(G)$;
3 Compute Eigenvectors X_1 and X_2 of $L(G)$;
4 Build $U \in \mathbb{R}^{n \times 2}$ with X_1 and X_2 as columns;
5 Rows of U are $y_1, y_2, \ldots, y_n \in \mathbb{R}^2$;
6 Cluster $y_1, y_2, \ldots, y_n$ into Clusters C_1 and C_2 (k-means);

Output: Clusters $A_1 = \{j : y_j \in C_1\}$ and $A_2 = \{j : y_j \in C_2\}$

Relaxation of RatioCut for $k > 2$ (sketch)

- Define k indicator vectors $h_j = (h_{1,j}, h_{2,j}, \ldots, h_{n,j})^T$ via

$$h_{i,j} = \begin{cases} 1/\sqrt{|A_j|} & \text{if } v_i \in A_j, \\ 0 & \text{else} \end{cases}$$

- Matrix $H \in \mathbb{R}^{n \times k}$ with columns $h_1, h_2, \ldots, h_k$ is orthogonal
- We have $h_i^T L h_i = \frac{(A_i, \overline{A}_i)}{|A_i|}$ and $h_i^T L h_i = (H^T L H)_{ii}$
- Together this yields $\text{RatioCut}(A_1, A_2, \ldots, A_k) = \text{tr}(H^T L H)$
- Hence the minimum of $\text{RatioCut}(A_1, A_2, \ldots, A_k)$ can be approximated by solving (the relaxed version)

$$\min_{H \in \mathbb{R}^{n \times k}} \text{tr}(H^T L H) \text{ subject to } H^T H = I$$

General case of unnormalized spectral clustering

Algorithm: Unnormalized spectral clustering

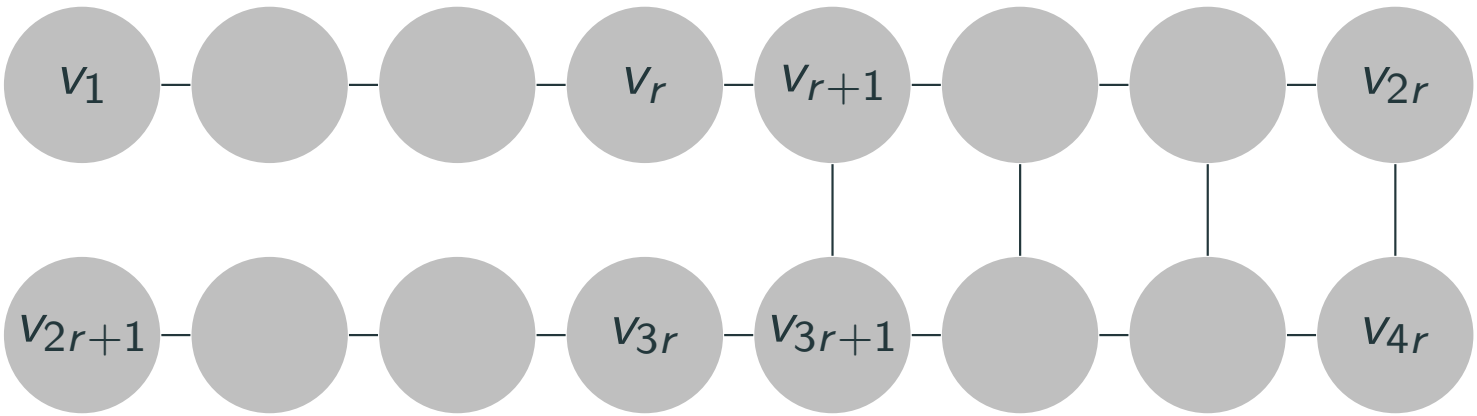
Input: Weight matrix (or any similarity matrix) $S \in \mathbb{R}^{n \times n}$, number k of clusters

1 Construct similarity graph G;
2 Compute $L(G)$;
3 Compute Eigenvectors $X_1, X_2, \ldots, X_k$ of $L(G)$;
4 Build $U \in \mathbb{R}^{n \times k}$ with $X_1, X_2, \ldots, X_k$ as columns;
5 Rows of U are $y_1, y_2, \ldots, y_n \in \mathbb{R}^k$;
6 Cluster $y_1, y_2, \ldots, y_n$ into Clusters $C_1, C_2, \ldots, C_k$ (k-means);

Output: Clusters $A_1, A_2, \ldots, A_k$ with $A_i = \{j : y_j \in C_i\}$

Notes on spectral clustering

- faster ($\mathcal{O}(mn)$ or $\mathcal{O}(n^2)$ for the eigenvector calculation) than Kernighan-Lin
- quality of approximated solution can be arbitrarily far from exact solution (cockroach graphs for $k = 2$)



- other relaxations possible (and probably useful)
- any other clustering algorithm may be used instead of k-means

Computing eigenvalues

- find roots of characteristic polynomial (computationally expensive)
- power method: iterate (with any starting vector $X_{(0)}$)

$$X_{(l)} = A^l X_{(0)}$$

- power method converges to eigenvector X corresponding to (w.r.t. absolute value) largest eigenvalue
- power method is fast but
  - method does not work if $X_{(0)}$ is orthogonal to X (can be avoided by choosing all entries of $X_{(0)}$ to be positive, since X has only entries of same sign)
  - entries of $X_{(l)}$ become large during the iterative process (renormalization helps)
  - When are we done? (option is to start with two different vectors)

Computational complexity of the power method

Two aspects:

- complexity of one multiplication
- required number of multiplications

First aspect:

- n^2 multiplications if stored in adjacency matrix
- less if matrix is stored in adjacency lists and matrix is sparse
  - compute
    $$\sum_{j \in N(i)} X_{(l)_j}$$
    which gives the i-th entry of $X_{(l+1)}$
  - therefore a total of
    $$\sum_i d(i) = 2m$$
    operations (m being the number of edges in the graph)

Computational complexity of the power method

Second aspect:

- it can be shown that this is $\mathcal{O}(n)$
- we will come back to this

Total computational complexity is $\mathcal{O}(mn)$, which means

- $\mathcal{O}(n^2)$ if graph is sparse
- $\mathcal{O}(n^3)$ if graph is dense

$\rightsquigarrow$ use adjacency list

Computing other eigenvalues

We have

$$(\lambda_n I - L)X_i = (\lambda_n - \lambda_i)X_i$$

hence eigenvalues are reversed for matrix $\lambda_n I - L$

$\rightsquigarrow$ smallest eigenvalue can be calculated with power method

Computing other eigenvalues

Trick to compute second largest eigenvalue:

- X_n normalised eigenvector corresponding to largest eigenvalue λ_n
- choose starting vector X and define

$$Y = X - (X_n^T X) X_n$$

- we have

$$X_i^T Y = \begin{cases} 0 & \text{if } i = n, \\ X_i^T X & \text{otherwise} \end{cases}$$

- therefore

$$Y = \sum_{i=1}^{n-1} c_i X_i$$

  where $c_i = X_i^T Y$

$\rightsquigarrow$ use Y as starting vector for power method

Compute all eigenvalues and eigenvectors

Combining methods for given (symmetric) matrix A with eigenvectors X_i and eigenvalues λ_i:

- Find orthogonal matrix Q with $B = Q^T A Q$ being a tridiagonal matrix
  - $Q^T X_i$ is eigenvector of B
  - B can be found efficiently, e.g. by Householder algorithm or Lancosz algorithm
- compute eigenvalues and eigenvectors of B
  - these give eigenvalues and eigenvectors of A
  - can be done using for example the QL algorithm

Singular value decomposition

Singular value decomposition

Before: for quadratic matrices we had Eigenvalues and Eigenvectors that can be used to diagonalise a matrix

Now:

- similar concept for non quadratic matrices
- the corresponding scalars are called Singular values which opposed to the Eigenvalues are **always** real
- Although similarity exist singular value decomposition is not an generalization of Eigenvalues/Eigenvector approach
- the rank of a matrix can be determined in a numerical stable way

Group of orthogonal matrices

Definition

- $\mathrm{GL}(n, \mathbb{R})$ general linear group of degree n is the set of $n \times n$ invertible matrices
- $\mathrm{O}(n) = \left\{ Q \in \mathrm{GL}(n, \mathbb{R}) \mid Q^{\mathsf{T}} Q = Q Q^{\mathsf{T}} = I \right\}.$

Singular value decomposition

Theorem

Let $A \in \mathbb{R}^{m \times n}$ be a matrix. Then $\sigma_1, \ldots, \sigma_p \in \mathbb{R}$ with $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ as well as $U \in O(m)$ and $V \in O(n)$ exist, so that

$$U^t A V = \Sigma := \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \\ \vdots & & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix},$$

wobei $p = \min(m, n)$. The values σ_i are called **singular values** *of A. A representation of the form $A = U \Sigma V^t$ is called* **singular value decomposition** *(**SVD**).*

Example

- For a quadratic matrix:

$$A_1 = \begin{pmatrix} 4 & 12 \\ 12 & 11 \end{pmatrix} = \begin{pmatrix} 3/5 & 4/5 \\ 4/5 & -3/5 \end{pmatrix} \cdot \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix} \cdot \begin{pmatrix} 3/5 & 4/5 \\ -4/5 & 3/5 \end{pmatrix}$$

- SDV of orthogonal matrizes:

$$A_2 = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

-

$$A_3 = \begin{pmatrix} 0.36 & 1.60 & 0.48 \\ 0.48 & -1.20 & 0.64 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.6 \\ -0.6 & 0.8 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 0 \\ 0.6 & 0 & 0.8 \\ -0.8 & 0 & 0.6 \end{pmatrix}$$

Remark

- Note that
$$\operatorname{rang}\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0, \quad \operatorname{rang}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = 1.$$
- yet the two eigenvalues in both cases are 0 and 0. The singular values on the other hand are 0, 0 and 0, 1 respectively, i.e., in this case the eigenvalues do not tell you anything about the rank of the matrix but the number of singular values of the matrix correspond to its rank
- Consider for $\varepsilon > 0$:
$$A = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}.$$
Since $\chi_A(t) = t^2 - \varepsilon$ the corresponding eigenvalues are $\pm\sqrt{\varepsilon}$. The singular values are $\sigma_1 = 1, \sigma_2 = \varepsilon$ and for ε converging towards 0, the rank of matrix is converging towards 1

Singular value decomposition

Theorem

Let $A = U\Sigma V$ be the singular value decomposition of $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_p$ für $p = \min(m, n)$. Let $u_1, \ldots, u_m$ and $v_1, \ldots, v_n$ denote the columns of U and V respectively. Then the following holds:

- *$Av_i = \sigma_i u_i$ and $A^t u_i = \sigma_i v_i$ für $i = 1, 2, \ldots, p$.*
- *For $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$ follows that* $\operatorname{rang} A = r$*. Furthermore,*

$$\mathbf{Ker}(A) = \langle v_{r+1}, \ldots, v_n \rangle \text{ und } \mathbf{Im}(A) = \langle u_1, \ldots, u_r \rangle.$$

- *the squares $\sigma_1^2, \ldots, \sigma_p^2$ of the singular values are the eigenvalues of A^tA and of AA^t to the corresponding eigen vectors $v_1, \ldots, v_p$ and $u_1, \ldots, u_p$ respectively.*

Singular value decomposition

Remark: For symmetric matrices A the singular values are the absolut values of the eigenvalues of A. In case all eigenvalues are non-negativ, $A = S^t \Lambda S$ is the SVD.

Pseudoinverse

Definition

Let $A \in \mathbb{R}^{m \times n}$. A matrix $A^+ \in \mathbb{R}^{n \times m}$ is called the **pseudoinverse** of A, if $\forall b \in \mathbb{R}^m$ the vector $x = A^+ b$ is the solution of the minimalisation problem

$$\text{Find } x, \text{ so that } \|b - Ax\|_2 \text{ is minimal}$$

i.e., $\|b - AA^+ b\| = \min_{x \in \mathbb{R}^n} \|b - Ax\|$.

Motivation

Note: for a quadratic invertible matrix A the pseudoinverse is: $A^+ = A^{-1}$

Application: in case the system $Ax = b$ does not have a solution, it is possible to obtain the best approximation $\tilde{x} = A^+ b$ via the pseudoinverse A^+ i.e., the one that minimizes the error $\|Ax - b\|$ (note that is the solution of the least squares problem.

Note that A^+ can be consider as a lineare mapping. Then the following holds

- $$AA^+ \colon \mathbb{R}^m \to \mathbf{Im}(A)$$
  is the orthogonal projection to image of A and
- $$A^+A \colon \mathbb{R}^n \to (\mathbf{Ker}A)^\perp$$
  ist the orthogonal projection to the orthogonal complement von the kernel of A.

SVD and Pseudoinverse

Theorem

Let $A \in \mathbb{R}^{m \times n}$ and let $A = U\Sigma V^t$ be the corresponding singular value decomposition with singular value $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$. Then we define

$$\Sigma^+ = \begin{pmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ & & \frac{1}{\sigma_r} \\ & & \\ 0 & & 0 \end{pmatrix}$$

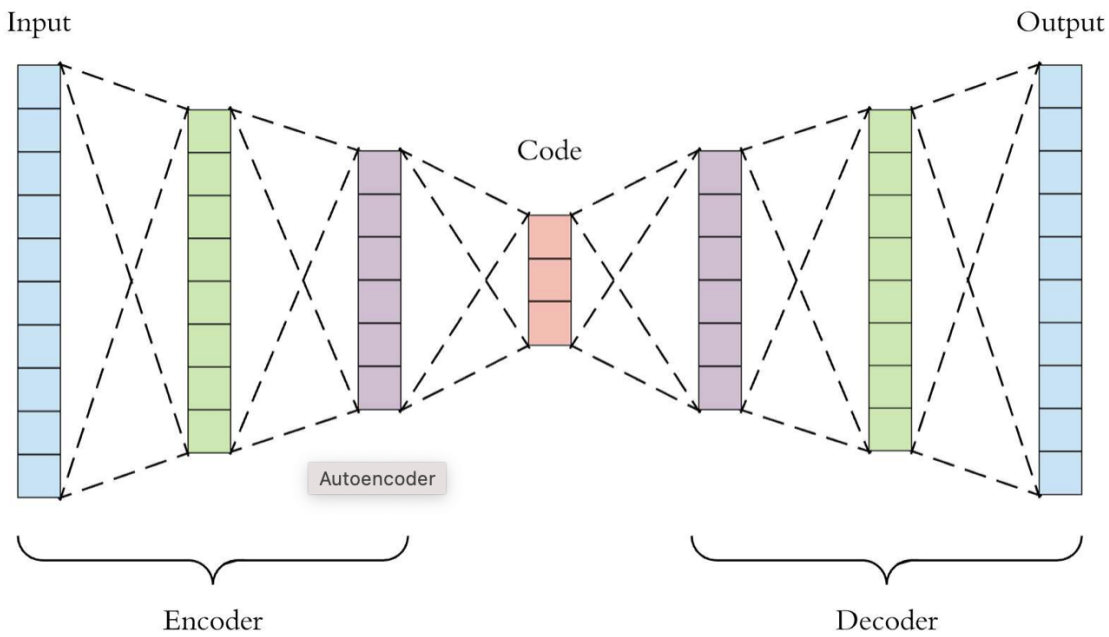
and the matrix $A^+ = V\Sigma^+ U^t \in \mathbb{R}^{n \times m}$ is the pseudo inverse of A.

Autoencoders

Autoencoders

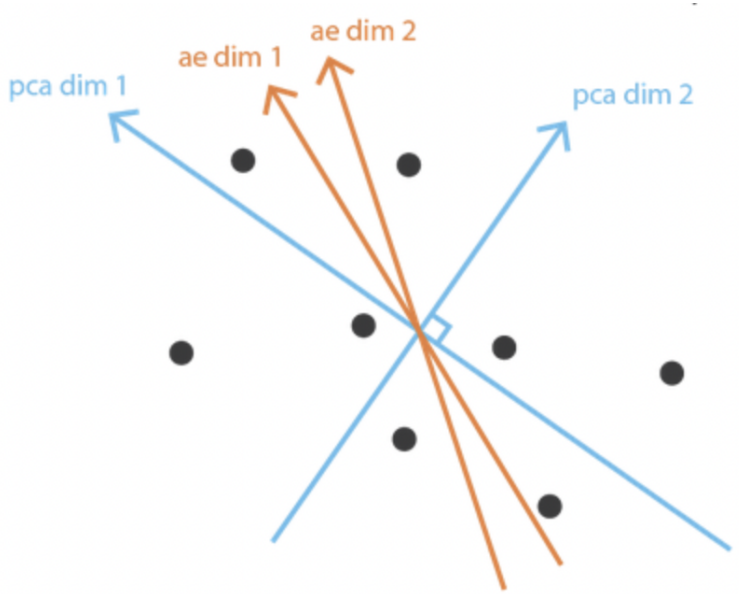
Snapshot information

- unsupervised artificial neural network (feed forward)
- Two steps:
  - **Encoder:** learns how to efficiently compress and encode data
  - **Decoder:** learns how to reconstruct the data back from the reduced encoded representation to a representation that is as close to the original input as possible.

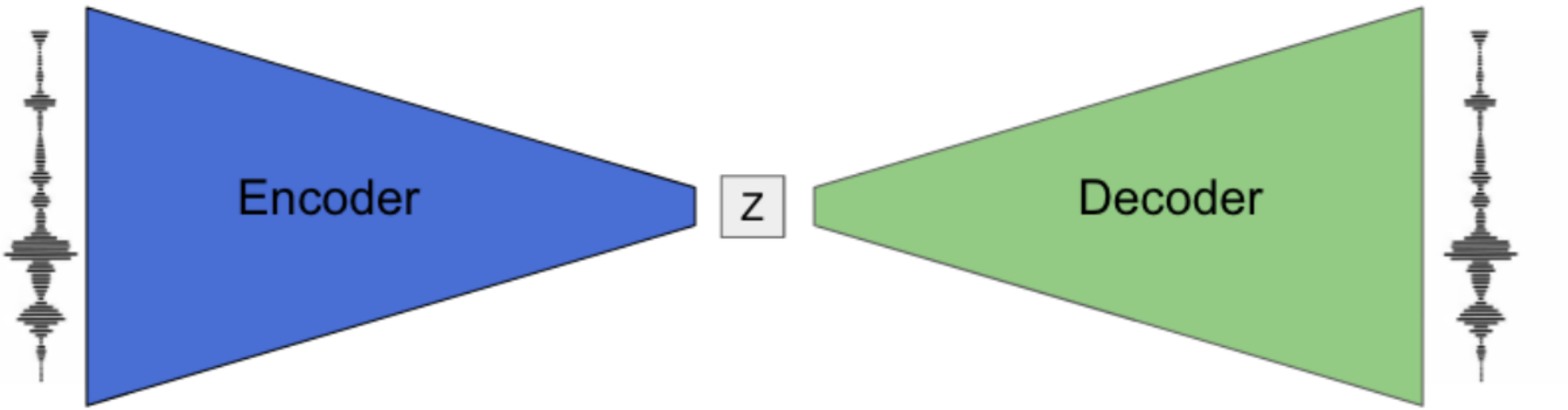


PCA vs Autoencoders

- PCA is essentially a linear transformation but Auto-encoders are capable of modelling complex non linear functions.
- PCA features are totally linearly uncorrelated while autoencoded features might have correlations
- PCA is faster and computationally cheaper than autoencoders.
- Autoencoder is prone to overfitting due to high number of parameters. (though regularization and careful design can avoid this)



General Autoencoders



Regularised Autoencoders

$$L(x, g(f(x))) + \Omega(h, x) \tag{136}$$

where $h = f(x)$ is the encoder and $g(h)$ the decoder, L is a choice of loss function. The regularisation term can have the form

$$\Omega(h, x) = \lambda \sum_i \|\nabla_x h_i\| \tag{137}$$

An autoencoder with this regularisation is known as contractive autoencoder.

Denoising Autoencoders

Traditionally, autoencoders minimize some function

$$L(x, g(f(x))) \tag{138}$$

while a so called denoising autoencoder (DAE) instead minimizes

$$L(x, g(f(\tilde{x}))) \tag{139}$$