Machine Learning for Insurers and Actuaries

Arthur Charpentier

2025

Learning, with a mathematical perspective

(lecture 1)

Probabilities and Statistics

- Probability and statistics are foundational for understanding and modeling uncertainty in data.
- In machine learning, we often work with data that is uncertain or incomplete.
  - **Probability**: Quantifies uncertainty and models random events (e.g., likelihood of events, predictions).
  - **Statistics**: Analyzes data to infer properties of a population from a sample (e.g., parameter estimation, hypothesis testing).
- Key tools in ML:
  - Probability distributions.
  - Expectation and variance.
  - Maximum likelihood estimation (MLE).
  - Bayesian inference.
- A **probability distribution** describes how the values of a random variable are distributed.

Probabilities and Statistics

- Common probability distributions used in ML:
  - Gaussian Distribution
  - Bernoulli Distribution: Used for binary outcomes.
  - Poisson Distribution: Models the number of events in a fixed interval.
  - Multinomial Distribution: Generalization of the Bernoulli distribution to multiple outcomes.
- Application in Machine Learning:
  - Gaussian Naive Bayes: Assumes features follow a Gaussian distribution.
  - Generative Models: Many generative models use probability distributions to model data (e.g., Variational Autoencoders).
- **Expectation** (Mean): The expected value of a random variable, representing the central tendency.

$$\mathbb{E}[X] = \sum_x x \cdot \mathbb{P}[X = x] = \boldsymbol{p}^\top \boldsymbol{x} = \langle \vec{\boldsymbol{p}}, \vec{\boldsymbol{x}} \rangle,$$

Probabilities and Statistics

- **Variance**: Measures the spread of the distribution, i.e., how far values are from the mean.

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

- **Covariance**: Measures the joint variability of two random variables.

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- **Application in Machine Learning**:
  - Gaussian Distribution / Linear Regression: Uses covariance to understand the relationship between variables and compute coefficients.
  - Gaussian Distribution / Principal Component Analysis (PCA): Uses the covariance matrix to identify directions of maximum variance.
- **Maximum Likelihood Estimation** is a method for estimating the parameters of a statistical model.

Probabilities and Statistics

- The likelihood function represents the probability of the observed data given the model parameters θ:

$$\mathcal{L}(\theta) = \mathbb{P}(\mathcal{D} \mid \theta)$$

- **MLE** finds the parameter values $\hat{\theta}$ that maximize the likelihood (or the log likelihood):

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\{\mathcal{L}(\theta)\} \text{ or } \hat{\theta} = \underset{\theta}{\text{argmax}}\{\log \mathcal{L}(\theta)\}$$

- **Application in Machine Learning**:
  - Gaussian Distribution
  - Logistic Regression: Uses MLE to estimate model parameters by maximizing the likelihood of the observed data.
  - Gaussian Distribution
  - Gaussian Mixture Models (GMM): Uses MLE to estimate the parameters of Gaussian distributions.

Probabilities and Statistics

- **Bayesian Inference** updates the probability estimate for a hypothesis as more evidence becomes available.
- The Bayes' Theorem is central to Bayesian inference:

$$\mathbb{P}(H \mid \mathcal{D}) = \frac{\mathbb{P}(\mathcal{D} \mid H) \cdot \mathbb{P}(H)}{\mathbb{P}(\mathcal{D})}$$

where:

  - Gaussian Distribution
  - $\mathbb{P}(H \mid \mathcal{D})$ is the posterior (the updated belief about H after seeing data $\mathcal{D}$).
  - Gaussian Distribution
  - $\mathbb{P}(\mathcal{D} \mid H)$ is the likelihood (how likely the data is given the hypothesis).
  - Gaussian Distribution
  - $\mathbb{P}(H)$ is the prior (the belief about H before seeing the data).

- **Application in Machine Learning**:

- Naive Bayes Classifier: Applies Bayes' Theorem to classify data.
- Bayesian Neural Networks: Use Bayesian methods to update the weights of the neural network.

- **Statistical Inference** is the process of drawing conclusions from data, typically based on probability models.
- **Hypothesis Testing** involves testing a null hypothesis (H_0) against an alternative hypothesis (H_1):

$$\text{p-value} = \mathbb{P}(\mathcal{D} \mid H_0)$$

If the p-value is small (e.g., < 0.05), we reject H_0.

- **Application in Machine Learning**:
  - **Model Evaluation**: Statistical tests are used to evaluate and compare the performance of machine learning models (e.g., A/B testing).
  - **Overfitting Check**: Hypothesis testing can be used to assess whether a model is too complex and overfitting the data.

Probabilities and Statistics

- Machine Learning without Probabilistic Assumptions,
  - Many ML algorithms, such as decision trees, support vector machines (SVMs), or neural networks, can be applied without explicit probabilistic assumptions.
  - These algorithms focus on finding patterns in data through optimization, without assuming any underlying distribution.
- **Why is probability important**?
  - Probabilities help quantify uncertainty in predictions.
  - Provide mathematical guarantees about the performance of algorithms.
  - Essential for understanding generalization and overfitting
- Without probabilistic assumptions, we cannot guarantee the generalization of our models to unseen data.

Probabilities and Statistics

- **Generalization** refers to how well a model trained on a sample of data performs on new, unseen examples.
- **Probability** allows us to express and prove mathematical guarantees, such as:
  - **PAC (Probably Approximately Correct) Learning**: Ensures that a model will be approximately correct with high probability, provided certain conditions are met.
- PAC Learning formalizes the relationship between:
  - The error of the hypothesis (how far off the model's predictions are from the true distribution).
  - The confidence in the model (the probability that the model's error is small).
  - The number of training examples required to achieve these guarantees.
- PAC Learning is a formal framework that allows us to assess how well a hypothesis learned from a finite sample will generalize to unseen data.
- The key components of PAC learning:

- A learning algorithm that outputs a hypothesis H.
- A target distribution $\mathcal{D}$ over examples.
- A hypothesis H is said to be probably approximately correct if:

$$\mathbb{P}(\text{error}(H) \leq \varepsilon) \geq 1 - \delta$$

where:

- ε is the allowable error.
- δ is the probability of failure (the confidence level).

- This implies that with high probability (at least $1 - \delta$), the hypothesis h will have an error less than ε.
- Sample Complexity refers to the number of training examples needed to achieve a PAC guarantee.
- The number of examples required depends on:
  - The complexity of the hypothesis class (the set of models we consider).

  - The desired error ε and confidence δ.
- For a given hypothesis class H, the sample complexity can be bounded as:

$$\text{complexity}(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$$

- This means that to guarantee with high probability that the error is less than ε, the required sample size increases as the allowed error ε decreases.
- Without probability, we cannot assess the risk of overfitting (i.e., fitting the model too closely to training data).
- Overfitting occurs when a model captures noise or irrelevant patterns in the training data.
- Probabilistic models help in controlling overfitting by:
  - Estimating uncertainty.
  - Regularizing models to avoid fitting to noise.

  - Using Bayesian methods to incorporate prior knowledge and regularize the model.

- **Example:** In Bayesian Neural Networks, the model's weights are treated as distributions, which allows us to quantify uncertainty in predictions and improve generalization.
- Machine Learning without Probabilistic Assumptions:
  - Many models, such as decision trees and SVMs, can be learned without making probabilistic assumptions.
  - These models often rely on optimization techniques to fit data and are effective for many practical applications.
- However, Probability Provides Mathematical Guarantees:
  - Probability gives us tools to assess generalization and make predictions about model performance on unseen data.

  - PAC learning formalizes the trade-off between model complexity, sample size, and generalization error.
  - Probabilistic assumptions allow us to quantify uncertainty, handle noise, and ensure the reliability of the model.
- Ultimately, while deterministic models can work in practice, probabilistic reasoning is essential for understanding the theoretical guarantees of machine learning algorithms.

Probabilities and random variables

"Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means," Russell (1929), quoted in Bell (1945)

Probability and statistics rely on the concept of probability spaces, $(\Omega, \mathcal{F}, \mathbb{P})$,

- Ω (or S in some textbooks) is the sample space, the set of all possible outcomes
- $\mathcal{F}$ a set of events on Ω, $A \in \mathcal{F}$ is an "event"
- $\mathbb{P}$ is a function $\mathcal{F} \to [0, 1]$, called probability, satisfying some properties

e.g. $\mathbb{P}(\Omega) = 1$; for disjoint events, an additiviy property: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$; a subset property, if $A \subset B$, $\mathbb{P}(A) \leq \mathbb{P}(B)$, as inCardano (1564) or Bernoulli (1713), or for multiple disjoint events as in Kolmogorov (1933), $A_1, \cdots, A_n, \cdots$,

$$\mathbb{P}(A_1 \cup \cdots \cup A_n \cup \cdots) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) + \cdots$$

inspired by Lebesgue (1918), etc. In this (mathematical) framework, we can finally define random variables

- X is a function $\Omega \to \mathbb{R}$ or more generally $\Omega \to \mathcal{X}$.

Probabilities and random variables

Definition 1.1: Cumulative Distribution Function

Let Y denote a (real-valued) random variable, then its cumulative distribution function is $F(y) = \mathbb{P}[Y \leq y]$, for $y \in \mathbb{R}$. We write $Y \sim F$.

Definition 1.2: Density

Absolutely continuous distributions have a density (with respect to Lebesgue measure),

$$F(y) = \int_{-\infty}^{y} f(z)\mathrm{d}z \text{ or } f(y) = \left.\frac{\mathrm{d}f(z)}{\mathrm{d}z}\right|_{z=y}$$

Thus, $\mathbb{P}[Y \in [a, b]] = F(b) - F(a) = \int_a^b f(z)\mathrm{d}z$.

Probabilities and random variables

We have formal objects, mathematically well defined, but in a context of modeling does one have a univocal sense of interpretation of the result of the calculation?
See "Is the probability inherent to the event, or to our judgment?," in Martin (2009)

There are many philosophical paradoxes when we talk about probability (and chance), e.g. I throw a coin, which falls back, out of my sight

- $\mathbb{P}(X = \text{heads}) = \mathbb{P}(X = \text{tails}) = 1/2$?
- $\mathbb{P}(X = \text{heads}) = 1$ or $\mathbb{P}(X = \text{tails}) = 1$?

Or in a legal context, "Look, the guy either did it or he didn't do it. If he did then he is 100% guilty and if he didn't then he is 0% guilty; so giving the chances of guilt as a probability somewhere in between makes no sense and has no place in the law," quoted in Fenton and Neil (2018)

See also Hájek (2002) on the philosophical approach of "probability".

Probabilities and random variables

As said by Martin (2009), "To attribute an objective meaning to the probability that an event will occur is to admit that this event is not necessary, in other words, that it is not completely determined," [...] "If we suppose an integral and universal determinism, the probability can only receive a subjective meaning, and the probability depends on our knowledge and our ignorance"

A lot of importance is attributed to this supposedly objective probability $\mathbb{P}$.

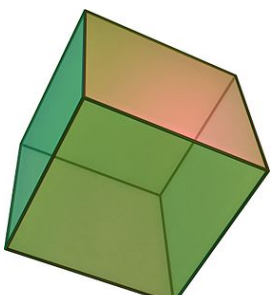
The (mathematical) probability was not born as a well defined concept within the framework of a mathematical formalism mathematical formalism, but as a tool to quantify and control situations of uncertainty, applied to the measurement of the probability of life mortality tables (for the calculation of life annuities), the calculation of the risks of error (in of error (in measurement operations), the study of the probability of testimonies and judgments, etc.

Probabilities and random variables

Cournot (1843) thus distinguished a objective meaning of the probability (as measure of the physical possibility of realization of a random event) and a subjective meaning (the probability being a judgement made on an event, this judgement being linked to the ignorance of judgment being linked to the ignorance of the conditions of the realization of the event).

Note: a probability not defined in terms of frequency can still have an objective meaning: :

There is no need to repeat throws of dice to affirm that (with a perfectly balanced die) the probability of obtaining 6 at the time of a throw is equal to 1/6 (by symmetry of the cube)



Probabilities and random variables

But very often, the "physical" probabilities receive an objective value only posterior on the basis of the law of large numbers, the empirical frequency converge towards the probability (frequentist theory of probabilities)

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \in A)}_{\text{(empirical) frequency}} \overset{\text{a.s.}}{\to} \underbrace{\mathbb{P}(X \in A)}_{\text{probability}} \text{ as } n \to \infty$$

(in some textbooks, there is a confusion between "probability" and "frequency")

$$\text{Law of large numbers : } \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{a.s.}}{\to} \mathbb{E}(X) \text{ as } n \to \infty \text{ or } \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}(X)$$

Probabilities and random variables

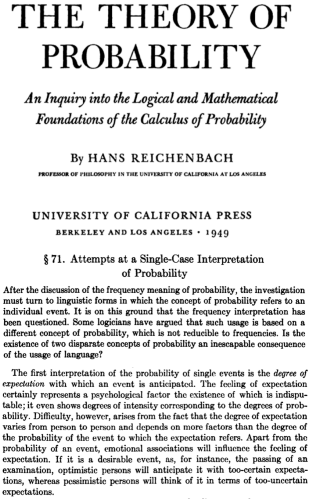
But this approach is unable to make sense of the probability of a "**single singular event**", as noted by von Mises (1928, 1939).

"When we speak of the 'probability of death', the exact meaning of this expression can be defined in the following way only. We must not think of an individual, but of a certain class as a whole, e.g., 'all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations'. A probability of death is attached to the class of men or to another class that can be defined in a similar way. We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning for us at all."

See also "**Single-Case Interpretation of Probability**," in Reichenbach (1971)

Probabilities and random variables

"What must be asked is whether an interpretation is adequate to account for the use of probabilities as degrees of reliability or as instruments suitable for an evaluation of predictions" [...] "The degree of probability, therefore, will be of no use after the truth about the occurrence of the event is known: probability is used as a substitute for truth so long as the truth is unknown. If the event is to happen in the future, the degree of its probability qualifies the reliability of a prediction," Reichenbach (1971).



THE THEORY OF PROBABILITY

An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability

By HANS REICHENBACH

PROFESSOR OF PHILOSOPHY IN THE UNIVERSITY OF CALIFORNIA AT LOS ANGELES

UNIVERSITY OF CALIFORNIA PRESS

BERKELEY AND LOS ANGELES · 1949

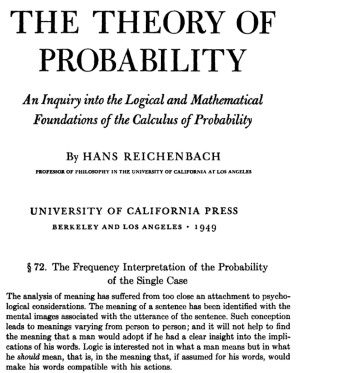
§ 71. Attempts at a Single-Case Interpretation of Probability

After the discussion of the frequency meaning of probability, the investigation must turn to linguistic forms in which the concept of probability refers to an individual event. It is on this ground that the frequency interpretation has been questioned. Some logicians have argued that such usage is based on a different concept of probability, which is not reducible to frequencies. Is the existence of two disparate concepts of probability an inescapable consequence of the usage of language?

The first interpretation of the probability of single events is the *degree of expectation* with which an event is anticipated. The feeling of expectation certainly represents a psychological factor the existence of which is indisputable; it even shows degrees of intensity corresponding to the degrees of probability. Difficulty, however, arises from the fact that the degree of expectation varies from person to person and depends on more factors than the degree of the probability of the event to which the expectation refers. Apart from the probability of an event, emotional associations will influence the feeling of expectation. If it is a desirable event, as, for instance, the passing of an examination, optimistic persons will anticipate it with too-certain expectations, whereas pessimistic persons will think of it in terms of too-uncertain expectations.

Probabilities and random variables

"If we are asked to find the probability holding for an individual future event, we piust first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes, from which different probabilities will result. This ambiguity has been called the problem of the reference class'' [...] "I regard the statement about the probability of the single case, not as having a meaning of its own, but as representing an elliptic mode of speech. In order to acquire meaning, the statement must be translated into a statement about a frequency in a sequence of repeated occurrences. The statement concerning the probability of the single case thus is given a fictitious meaning, constructed by a transfer of meaning from the general to the particular case," Reichenbach (1971).



THE THEORY OF PROBABILITY

An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability

By HANS REICHENBACH

PROFESSOR OF PHILOSOPHY IN THE UNIVERSITY OF CALIFORNIA AT LOS ANGELES

UNIVERSITY OF CALIFORNIA PRESS

BERKELEY AND LOS ANGELES · 1949

§ 72. The Frequency Interpretation of the Probability of the Single Case

The analysis of meaning has suffered from too close an attachment to psychological considerations. The meaning of a sentence has been identified with the mental images associated with the utterance of the sentence. Such conception leads to meanings varying from person to person; and it will not help to find the meaning that a man would adopt if he had a clear insight into the implications of his words. Logic is interested not in what a man means but in what he *should* mean, that is, in the meaning that, if assumed for his words, would make his words compatible with his actions.

If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes, from which different probabilities will result. This ambiguity has been called the *problem of the reference class*. Assume that a case of illness can be characterized by its inclusion in the class of cases of tuberculosis. If additional information is obtained from an X-ray, the same case may be incorporated in the class of serious cases of tuberculosis. Depending on the classification, different probabilities will result for the prospective issue of the illness.

We then proceed by considering *the narrowest class for which reliable statistics can be compiled*. If we are confronted by two overlapping classes, we shall choose their common class. Thus, if a man is 21 years old and has tuberculosis, we shall regard the class of persons of 21 who have tuberculosis. Classes that are known to be irrelevant for the statistical result may be disregarded. A class C is irrelevant with respect to the reference class A and the attribute class B if the transition to the common class $A \cdot C$ does not change the probability, that is, if $P(A \cdot C, B) = P(A, B)$. For instance, the class of persons having the same initials is irrelevant for the life expectation of a person.

Probabilities and random variables

For Popper (1959), probabilities correspond to physical dispositions ("propensions") inherent to the system. This propensity has a physical existence, but it is not directly observable.

The frequencies of occurrence are manifestations of these propensities. In the contrary case, it is nevertheless possible to estimate the probability of realization of the singular event, by considering this one as measured not by an "actual" frequency, but by a "potential" (or "virtual") frequency.

Finally, when an individual makes a judgment, the degree of credibility or belief that he or she gives it depends on the knowledge that the individual has (Pettigrew (2016)). depends on the knowledge that this individual has (Pettigrew (2016)). This degree of belief will be associated with a probability, which will then only have a subjective meaning.

Probabilities and random variables

"The probability of a diagnosis, a testimony, etc., does not measure the conformity of this judgment to reality, but the degree to which one can hypothesize this conformity. This conformity can be hypothesized", Martin (2009).
This subjectivity raises concerns about their use, especially in criminal matters,

"Sometimes the 'balance of probability' standard is expressed mathematically as '50+% probability', but this can carry with it a danger of pseudo-mathematics, as the argument in this case demonstrated. When judging whether a case for believing that an event was caused in a particular way is stronger than the case for not so believing, the process is not scientific (although it may obviously include evaluation of scientific evidence) and to express the probability of some event having happened in percentage terms is illusory, *Nulty & Ors v Milton Keynes Borough Council* cited in Hunt and Mostyn (2020).

See also Jonakait (1983), Saini (2011) or Fenton et al. (2016).

Statistics, Frequentist Approach

Strong law of large numbers (also called Kolmogorov's law), see Loève (1977)

Proposition 1.1: Law of Large Numbers

Consider an infinite collection of i.i.d. random variables $X, X_1, X_2, \cdots, X_n, \cdots$ in a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$, then

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \in \mathcal{A})}_{\text{(empirical) frequency}} \overset{\text{a.s.}}{\to} \underbrace{\mathbb{P}(\{X \in \mathcal{A}\})}_{\text{probability}} = \mathbb{P}[\mathcal{A}], \text{ as } n \to \infty.$$

Likelihood

- Fisher (1921) introduced what is today called a "likelihood interval"
- Fisher (1922) introduced the term "method of maximum likelihood"

"In 1922, I proposed the term "likelihood," in view of the fact that, with respect to [the parameter], it is not a probability, and does not obey the laws of probability, while at the same time it bears to the problem of rational choice among the possible values of [the parameter] a relation similar to that which probability bears to the problem of predicting events in games of chance... Whereas, however, in relation to psychological judgment, likelihood has some resemblance to probability, the two concepts are wholly distinct (...) I stress this because in spite of the emphasis that I have always laid upon the difference between probability and likelihood there is still a tendency to treat likelihood as though it were a sort of probability."

Likelihood

Likelihood function

A likelihood function (often simply called the likelihood) measures how well a statistical model explains observed data by calculating the probability of seeing that data under different parameter values of the model. The likelihood function, parameterized by a (possibly multivariate) parameter $\boldsymbol{\theta}$, is usually defined differently for discrete and continuous probability distributions. Given a probability density or mass function $x \mapsto f(x \mid \boldsymbol{\theta})$, where x is a realization of the random variable X, the likelihood function is $\boldsymbol{\theta} \mapsto f(x \mid \boldsymbol{\theta})$, often written $\mathcal{L}(\boldsymbol{\theta} \mid x)$. W

Likelihood

Definition 1.3: Maximum Likelihood (1)

Let $\boldsymbol{y} = (y_1, \cdots, y_n)$ be an sample from i.i.d. variables with distribution f_θ. The likelihood is

$$\mathcal{L}(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f_\theta(y_i)$$

And the maximum likelihood is

$$\widehat{\theta}(\boldsymbol{y}) = \underset{\theta\in\Theta}{\text{argmax}}\{\mathcal{L}(\theta|\boldsymbol{y})\} \text{ and } \widehat{\theta}(\boldsymbol{Y}) = \underset{\theta\in\Theta}{\text{argmax}}\{\mathcal{L}(\theta|\boldsymbol{Y})\}$$

Likelihood

Definition 1.4: Maximum Likelihood (2)

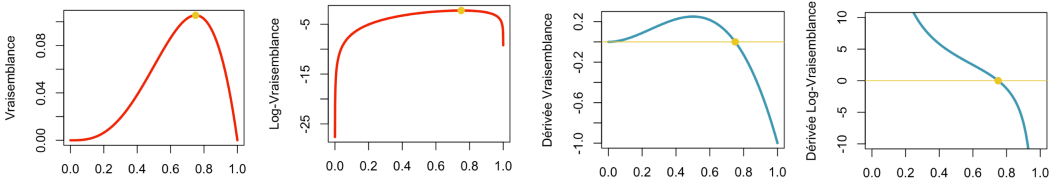
Let $\boldsymbol{y} = (y_1, \cdots, y_n)$ be an sample from i.i.d. variables with distribution f_θ. The log-likelihood is

$$\log \mathcal{L}(\theta|\boldsymbol{y}) = \sum_{i=1}^{n} \log[f_\theta(y_i)];$$

And the maximum likelihood is

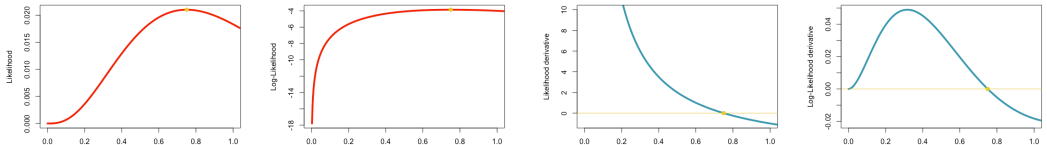
$$\widehat{\theta}(\boldsymbol{y}) = \underset{\theta\in\Theta}{\text{argmax}}\{ \log \mathcal{L}(\theta|\boldsymbol{y})\} \text{ and } \widehat{\theta}(\boldsymbol{Y}) = \underset{\theta\in\Theta}{\text{argmax}}\{ \log \mathcal{L}(\theta|\boldsymbol{Y})\}$$

Likelihood



- $\theta \mapsto \prod_{i=1}^{n} f(y_i;\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^3(1-\theta)$, $\boldsymbol{y} = \{0,1,1,1\}$, $Y_i \sim \mathcal{B}(\theta)$.
- $\theta \mapsto \sum_{i=1}^{n} \log f(y_i;\theta) = \sum_{i=1}^{n} y_i \log(\theta) + (1-y_i)\log(1-\theta) = 3\log\theta + \log(1-\theta)$
- $\theta \mapsto \dfrac{\partial}{\partial\theta} \prod_{i=1}^{n} f(y_i;\theta) = \cdots = 3\theta^2 - 4\theta^3$
- $\theta \mapsto \dfrac{\partial}{\partial\theta} \sum_{i=1}^{n} \log f(y_i;\theta) = \cdots = \dfrac{3}{\theta} - \dfrac{1}{1-\theta}$ (score function)

Likelihood



- $\theta \mapsto \prod_{i=1}^{n} f(y_i;\theta) = \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{y_i}}{y_i!} = e^{-4\theta}\theta^3$, where $\boldsymbol{y} = \{0,1,1,1\}$, $Y_i \sim \mathcal{P}oisson(\theta)$.
- $\theta \mapsto \sum_{i=1}^{n} \log f(y_i;\theta) = \sum_{i=1}^{n} -\theta + y_i \log\theta - \log(y_i!) = -4\theta + 3\log\theta$
- $\theta \mapsto \frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(y_i;\theta) = \cdots = -4 + \frac{3}{\theta}$
- $\theta \mapsto \frac{\partial}{\partial\theta}\sum_{i=1}^{n} \log f(y_i;\theta) = \cdots = e^{-4\theta}(3\theta^2 - 4\theta^3)$ (score function)

Likelihood, $\mathcal{B}(p)$ Binary Sample

```
> y = c(1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0)
```

- what theory (or mathematics) tells us

$$\mathcal{L}(p;\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i;p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{s_n}(1-p)^{n-s_n},\ s_n = \sum_{i=1}^{n} y_i$$

$$\log \mathcal{L}(p;\boldsymbol{x}) = s_n \log(p) + (n-s_n)\log(1-p)$$

$$\frac{\partial}{\partial p}\log \mathcal{L}(p;\boldsymbol{y}) = \frac{\partial}{\partial p} s_n \log(p) + (n-s_n)\log(1-p) = \frac{s_n}{p} - \frac{n-s_n}{1-p}$$

$$\left.\frac{\partial}{\partial p}\log \mathcal{L}(p;\boldsymbol{y})\right|_{p=\widehat{p}} = 0 \text{ if and only if } \frac{s_n}{\widehat{p}} = \frac{n-s_n}{1-\widehat{p}},\ \text{i.e. } \widehat{p} = \frac{s_n}{n} = \overline{y}$$

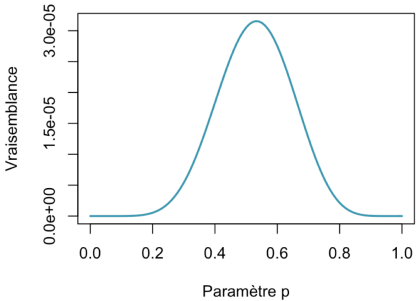
Likelihood, $\mathcal{B}(p)$ Binary Sample

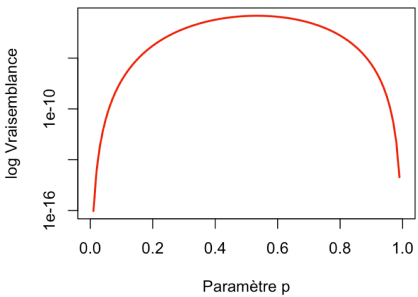
- what a computer tells us ?

One can plot the likelihood function $p \mapsto \mathcal{L}(p; \boldsymbol{x})$, or log-likelihood

```
> n = 15
> y = c(1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0)
> likelihood = function(p) prod(dbinom(y, size = 1, prob = p))
> vect_p = seq(0,1, by=0.01)
> plot(vect_p,Vectorize(likelihood)(vect_p))
```

Likelihood, $\mathcal{B}(p)$ Binary Sample





Likelihood, $\mathcal{B}(p)$ Binary Sample

- what a computer tells us ?

We can compute (numerically) the maximum of $p \mapsto \mathcal{L}(p; \boldsymbol{x})$

```
> optim(par = .5,fn = function(z) -likelihood(z))
$par
[1] 0.5333252

$value
[1] -3.155276e-05
```

(from the maths, we know that actually $\widehat{p} = \overline{x}$)

```
> mean(y)
[1] 0.5333333
```

Likelihood, $\mathcal{B}(p)$ Binary Sample

- what theory (or mathematics) tells us

since $\widehat{p}(\boldsymbol{x}) = \bar{y}$, we can use the law of large numbers,

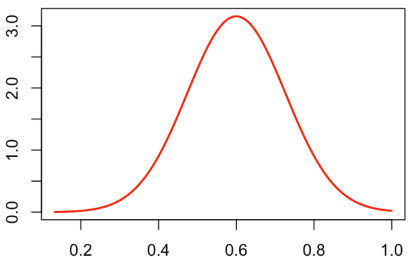
$$Z_n = \sqrt{n}\frac{\widehat{p}(\boldsymbol{Y}) - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

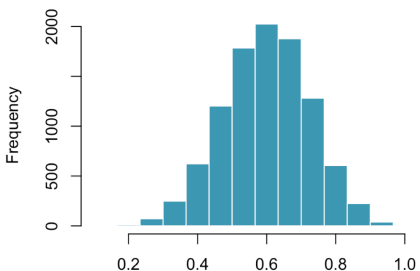
but since $n = 15$ this Gaussian assumption is maybe not valid...

If $p = 60\%$ the approximated distribution of $\widehat{p}(\boldsymbol{Y})$ would be

```
> u = seq(2/15,1,by=.001)
> plot(u, dnorm(u,.6,sqrt(.4*.6/15))
```

Likelihood, $\mathcal{B}(p)$ Binary Sample



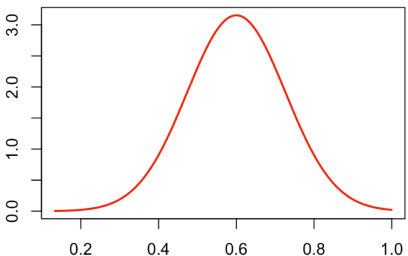


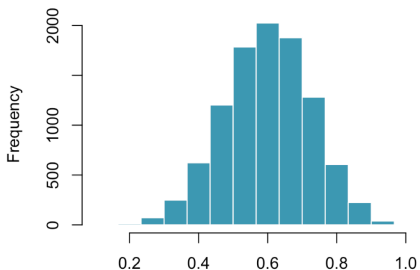
Likelihood, case $\mathcal{B}(p)$

- what a computer tells us ? (suppose that $\theta = 60\%$)

```
> theta=rep(NA,1e4)
> for(s in 1:1e4){
+      x=sample(0:1, size = n, prob = c(.4,.6),r eplace=TRUE)
+      neglogL = function(p) -sum(log(dbinom(y, size = 1, prob = p)))
+      theta[s] = optim(par = .5,fn = neglogL)$par
+ }
> hist(theta)
```

Likelihood, case $\mathcal{B}(p)$





Confidence Interval

This can be used to get some confidence...

Pointwise estimation: compute (simply) $\widehat{\theta}(\boldsymbol{y})$ as a single numerical value

Definition 1.5: Confidence Interval

Let $\boldsymbol{Y}$ be a random sample of i.i.d. variables with distribution f_θ. A confidence interval of level $1-\alpha$ for the parameter θ is a (random) interval $\left[\widehat{a}(\boldsymbol{Y}), \widehat{b}(\boldsymbol{Y})\right]$ such that

$$\mathbb{P}\Big[\theta \in \left[\widehat{a}(\boldsymbol{Y}), \widehat{b}(\boldsymbol{Y})\right]\Big] = 1-\alpha$$

Classically, α is 10%, 5% or 1% (or less).

The smaller α, the wider the confidence interval.

Confidence Interval, Gaussian Sample

Let $\{y_1, \cdots, y_n\}$ be an i.i.d. sample with a $\mathcal{N}(\mu, \sigma_0^2)$ distribution, where σ_0^2 is assumed to be known.

$\widehat{\mu}(\boldsymbol{Y}) = \overline{Y}$, then $\widehat{\mu}(\boldsymbol{Y}) = \mathcal{N}\left(\mu, \dfrac{\sigma_0^2}{n}\right)$.

Set $Z = \dfrac{\widehat{\mu}(\boldsymbol{Y}) - \mu}{\sigma_0/\sqrt{n}}$, $Z \sim \mathcal{N}(0,1)$.

The bilateral confidence interval for μ, with level $1-\alpha$ is

$$\left[\widehat{\mu}(\boldsymbol{Y}) - u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \widehat{\mu}(\boldsymbol{Y}) + u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]$$

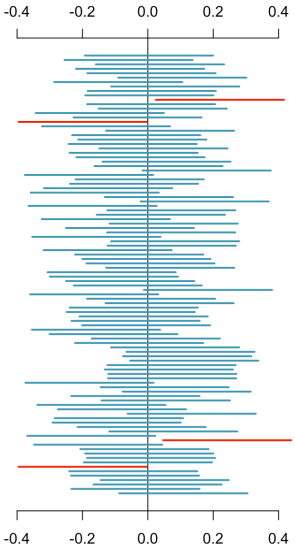
where $u_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, where $\Phi(z) = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{z} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x$.

Confidence Interval, level α

Sample $\mathcal{N}(0,1)$ with size n,

$$IC = \left[\widehat{\mu}(\boldsymbol{Y}) \pm u_{\alpha/2}\frac{1}{\sqrt{n}}\right]$$

```
alpha = .05
set.seed(1)
n=100
IC = matrix(NA,100,2)
for(s in 1:100){
  x = rnorm(100,0,1)
  m = mean(x)
  IC[s,1] = m-qnorm(1-alpha/2)*1/sqrt(n)
  IC[s,2] = m+qnorm(1-alpha/2)*1/sqrt(n)
}
idx=which((IC[,1]<0)&(IC[,2]>0))
```

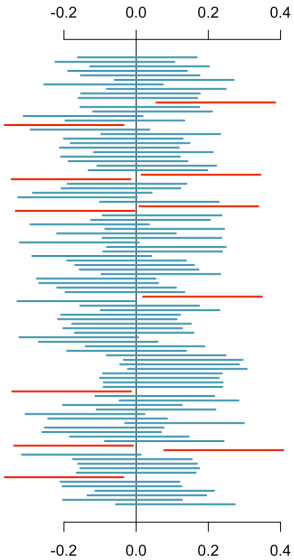


Confidence Interval, level α

Sample $\mathcal{N}(0,1)$ with size n,

$$IC = \left[\widehat{\mu}(\boldsymbol{Y}) \pm u_{\alpha/2}\frac{1}{\sqrt{n}}\right]$$

```
alpha = .1
set.seed(1)
n=100
IC = matrix(NA,100,2)
for(s in 1:100){
  x = rnorm(100,0,1)
  m = mean(x)
  IC[s,1] = m-qnorm(1-alpha/2)*1/sqrt(n)
  IC[s,2] = m+qnorm(1-alpha/2)*1/sqrt(n)
}
idx=which((IC[,1]<0)&(IC[,2]>0))
```

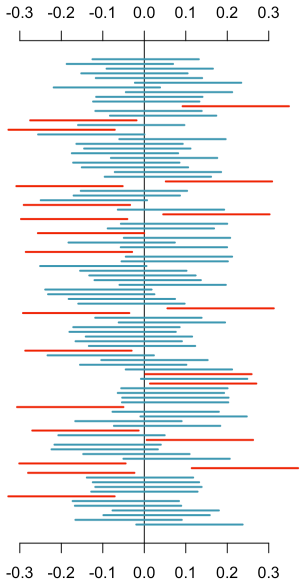


Confidence Interval, level α

Sample $\mathcal{N}(0,1)$ with size n,

$$IC = \left[\widehat{\mu}(\boldsymbol{Y}) \pm u_{\alpha/2}\frac{1}{\sqrt{n}}\right]$$

```
alpha = .2
set.seed(1)
n=100
IC = matrix(NA,100,2)
for(s in 1:100){
  x = rnorm(100,0,1)
  m = mean(x)
  IC[s,1] = m-qnorm(1-alpha/2)*1/sqrt(n)
  IC[s,2] = m+qnorm(1-alpha/2)*1/sqrt(n)
}
idx=which((IC[,1]<0)&(IC[,2]>0))
```



Binomial distribution and Confidence

If the variables $Y_1, \cdots, Y_n$ are i.i.d. of law $\mathcal{B}(p)$, and if $\widehat{p} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$,

$$\mathbb{E}[\widehat{p}] = p \text{ and } \text{Var}[\widehat{p}] = \frac{p(1-p)}{n}$$

More precisely, since $n\widehat{p} \sim \mathcal{B}(n, p)$,

$$\mathbb{P}\left(F = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}, \ \forall k = 0, 1, \cdots, n.$$

If n is large enough, according to the central limit theorem

$$Z_n = \sqrt{n}\frac{\widehat{p} - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Binomial distribution and Confidence

In practice, we assume the normal approximation is valid if $n \geq 30$, $np \geq 15$ and $n(1-p) \geq 15$. Binomial model, with n large enough

Definition 1.6: Confidence Interval $\{y_1, \cdots, y_n\}$, $\mathcal{B}(p)$, n large

Let $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ denote an i.i.d. sample, from $\mathcal{B}(p)$.

$$\left[\widehat{p} \pm u_{1-\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \right], \text{ where } \widehat{p} = \overline{y}$$

Note sometimes the confidence interval exceeds 0 or 1

```
> sum_y = 12
> n = 14
> sum_y/n + qnorm(c(.025,.975))*sqrt(sum_y*(n-sum_y)/n^3)
[1] 0.6738432 1.0404425
```

Binomial distribution and Confidence

dans un modèle binomial, avec n assez grand

Definition 1.7: Confidence Interval $\{y_1, \cdots, y_n\}$, $\mathcal{B}(p)$, Wilson

Let $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ denote an i.i.d. sample, from $\mathcal{B}(p)$. A confidence interval with level α for p is

$$\left[\frac{1}{1+\frac{u_{1-\alpha/2}^2}{n}} \left(\hat{p} + \frac{u_{1-\alpha/2}^2}{2n} \right) \pm \frac{u_{1-\alpha/2}}{1+\frac{u_{1-\alpha/2}^2}{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{u_{1-\alpha/2}^2}{4n^2}} \right]$$

On obtient ces bornes en notant qu'elle correspondent aux p tels que $(\hat{p}-p)^2 = u_{1-\alpha/2}^2 \cdot \frac{p(1-p)}{n}$ qui est l'équation de degré 2

$$\left(1+\frac{u_{1-\alpha/2}^2}{n}\right) p^2 + \left(-2\hat{p} - \frac{u_{1-\alpha/2}^2}{n}\right) p + \left(\hat{p}^2\right) = 0 \ .$$

Binomial distribution and Confidence

Modèle binomial, avec n assez grand, comme

$$\frac{\left(\overline{X}-\overline{Y}\right)-(p_x-p_y)}{\sqrt{\dfrac{\overline{X}(1-\overline{X})}{m}+\dfrac{\overline{Y}(1-\overline{Y})}{n}}}\approx\mathcal{N}(0,1)$$

Definition 1.8: Intervalle de confiance pour p_x-p_y $\mathcal{B}(p_x)$ et $\mathcal{B}(p_y)$, Wald

Soient $\boldsymbol{x}=\{x_1,\cdots,x_m\}$ de loi $\mathcal{B}(p_x)$ et $\boldsymbol{y}=\{y_1,\cdots,y_n\}$ de loi $\mathcal{B}(p_y)$. Un intervalle de confiance de niveau α pour p_x-p_y est

$$\left[\overline{x}-\overline{y}\pm u_{1-\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{m}+\frac{\overline{y}(1-\overline{y})}{n}}\right]$$

Binomial distribution and Confidence

Definition 1.9: Margin of Error ($N = \infty$)

The margin of error is the precision of the result obtained given the confidence level we're willing to accept (α). to accept (α). The (absolute) margin of error is then equal to

$$z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

```
> qnorm(.975)*sqrt(.5*(1-.5)/1000)
[1] 0.03098975
```

i.e. 3.1% (95% level) when $n = 1,000$ and $p \sim 50\%$

Binomial distribution and Confidence

Definition 1.10: Margin of Error ($N < \infty$)

The margin of error for a sample n drawn from a population of finite size N is written as

$$z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

```
> qnorm(.975)*sqrt(.5*(1-.5)/1000)*sqrt(5000/5999)
[1] 0.028292
```

i.e. 2.8% (95% level) when $n = 1,000$, $p \sim 50\%$, out of $N = 6,000$.

Binomial distribution and Confidence

We can invert these formulas to determine the sampling size, based on the (absolute) margin of error

$$n = \frac{z^2_{1-\alpha/2} \times p(1-p)}{\text{margin of error}^2}$$

when $N = \infty$, or if $N < \infty$

$$n = \frac{p(1-p) + \dfrac{\text{margin of error}^2}{z^2_{1-\alpha/2}}}{\dfrac{\text{margin of error}^2}{z^2_{1-\alpha/2}} + \dfrac{p(1-p)}{N}}$$

Maximum Likelihood and Confidence Intervals

- Consider a parametric model with parameter $\theta \in \Theta \subset \mathbb{R}$.
- Let $\log \mathcal{L}(\theta)$ be the log-likelihood function.
- We test the null hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
- The likelihood ratio test statistic is:

$$\Lambda = -2 \left[\log \mathcal{L}(\theta_0) - \log \mathcal{L}(\hat{\theta}) \right]$$

  where $\hat{\theta}$ is the MLE of θ.
- Under regularity conditions and H_0, we have:

$$\Lambda \xrightarrow{\mathcal{L}} \chi^2_1$$

  as $n \to \infty$ (Wilks' theorem, Wilks (1938)).

Maximum Likelihood and Confidence Intervals

- Define the $(1-\alpha)$ confidence region as the set of values of θ for which:

$$\Lambda(\theta) = -2\left[\ell(\theta) - \ell(\hat{\theta})\right] \leq \chi^2_{1,1-\alpha}$$

- This gives the confidence interval:

$$\left\{\theta \in \Theta : -2\left[\log \mathcal{L}(\theta) - \log \mathcal{L}(\hat{\theta})\right] \leq \chi^2_{1,1-\alpha}\right\}$$

- Interpretation:
  - It includes all values of θ not significantly worse than $\hat{\theta}$.
  - Useful even when the distribution of $\hat{\theta}$ is skewed or unknown.
- This interval is invariant under reparameterization.

For example

- Let $y_1, \ldots, y_n$ i.i.d. sample from $\mathcal{P}\text{oisson}(\lambda)$

Maximum Likelihood and Confidence Intervals

- Log-likelihood function:

$$\log \mathcal{L}(\lambda) = \sum_{i=1}^{n} [-\lambda + X_i \log \lambda - \log(X_i!)] = -n\lambda + (\sum X_i) \log \lambda + \text{const}$$

- MLE: $\hat{\lambda} = \bar{X}$
- LRT statistic for testing $H_0 : \lambda = \lambda_0$:

$$\Lambda = -2 \left[\log \mathcal{L}(\lambda_0) - \log \mathcal{L}(\hat{\lambda}) \right] = 2n \left(\lambda_0 - \bar{y} + \bar{X} \log \frac{\bar{y}}{\lambda_0} \right)$$

- Under H_0: $\Lambda \xrightarrow{d} \chi^2_1$
- For a confidence level $1 - \alpha$, find all λ such that:

$$\Lambda(\lambda) = 2n \left(\lambda - \bar{y} - \bar{y} \log \frac{\lambda}{\bar{y}} \right) \leq \chi^2_{1,1-\alpha}$$

Maximum Likelihood and Confidence Intervals

- This defines the confidence interval:

$$\left\{\lambda > 0 \ : \ \Lambda(\lambda) \leq \chi^2_{1,1-\alpha}\right\}$$

(and solve this numerically).

Gamma Distribution: Maximum Likelihood

Consider some i.i.d. sample $\{y_i\}$ from $Y_i \sim \mathcal{G}amma(\alpha, \beta)$,

$$f(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$$

$$\log \mathcal{L}(\alpha,\beta|\boldsymbol{y}) = n\alpha \log(\beta) - n\log[\Gamma(\alpha)] + (\alpha-1)\sum_{i=1}^n \log(y_i) - \beta \sum_{i=1}^n y_i$$

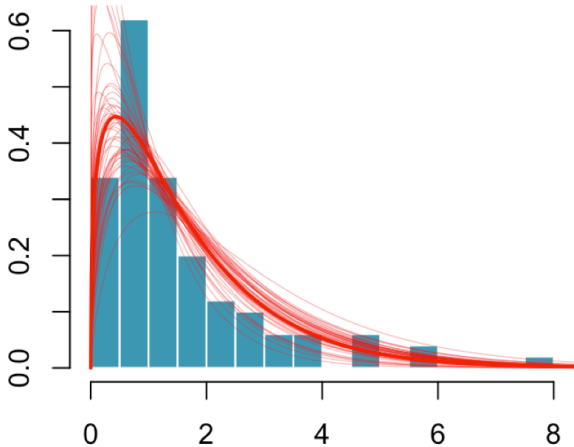
$$\nabla \log \mathcal{L}(\alpha,\beta) = \begin{pmatrix} n\log(\beta) - n[\log\Gamma(\alpha)]' + \sum_{i=1}^n \log(y_i) \\ \frac{n\alpha}{\beta} - \sum_{i=1}^n y_i \end{pmatrix}$$

$$H\log \mathcal{L}(\alpha,\beta) = \frac{1}{n(1-\alpha[\log\Gamma(\alpha)]'')} \begin{pmatrix} \alpha & \beta \\ \beta & \beta^2[\log(\Gamma(\alpha))]'' \end{pmatrix}$$

Gamma Distribution: Maximum Likelihood

Gamma distribution with parameter, $\boldsymbol{\theta} = (\alpha, \beta)$, $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha}, \widehat{\beta}) = \underset{\boldsymbol{\theta} \in \mathbb{R}_+^2}{\text{argmax}} \Big\{ \log \mathcal{L}(\boldsymbol{\theta}) \Big\}$

```
> y = exp(rnorm(100))
> library(MASS)
> (F = fitdistr(x,"gamma"))
     shape        rate
  1.3562877   0.8207035
 (0.1732663) (0.1263275)
> log_lik = function(theta){
+   a = theta[1]
+   b = theta[2]
+   logL = sum(log(dgamma(y,a,b)))
+   return(-logL)
+ }
> optim(c(1,1),log_lik)
$par
[1] 1.3558113 0.8206505
```



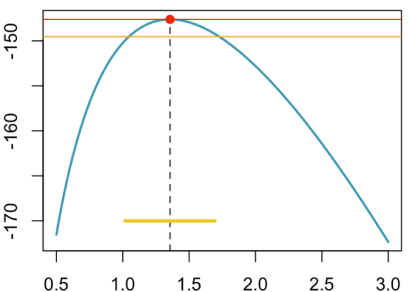
Gamma Distribution: Profile Likelihood

We can also consider if, somehow, α is the parameter of interest

$$\widehat{\alpha}^\star = \underset{\alpha\in\mathbb{R}_+}{\text{argmax}}\Big\{ \underbrace{\max_{\beta\in\mathbb{R}_+}\Big\{\log\mathcal{L}(\alpha,\beta)\Big\}}_{\text{function of }\alpha}\Big\}$$

Function $\alpha \mapsto \max_{\beta\in\mathbb{R}_+}\Big\{\log\mathcal{L}(\alpha,\beta)\Big\}$ called profile likelihood

```
> prof_log_lik = function(a){
+    b = (optim(1,function(z) -sum(log
     (dgamma(x,a,z)))))$par
+    return(-sum(log(dgamma(x,a,b))))
+ }
> optim(1,prof_log_lik)
$par
[1] 1.356445
```



Gamma Distribution: Profile Likelihood

This can be used to get a confidence interval for the parameter of interest.
In a statistical context, suppose that unknown parameter can be partitioned $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$ where α is the parameter of interest, and $\boldsymbol{\beta}$ is a nuisance parameter.
Consider $\{y_1, \cdots, y_n\}$, a sample from distribution $F_{\boldsymbol{\theta}}$, so that the log-likelihood is

$$\log \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(y_i)$$

$\widehat{\boldsymbol{\theta}}^{MLE}$ is defined as $\widehat{\boldsymbol{\theta}}^{MLE} = \text{argmax}\{\log \mathcal{L}(\boldsymbol{\theta})\}$..
Rewrite the log-likelihood as $\log \mathcal{L}(\boldsymbol{\theta}) = \log \mathcal{L}_\alpha(\boldsymbol{\beta})$. Define

$$\widehat{\boldsymbol{\beta}}_\alpha^{pMLE} = \underset{\boldsymbol{\beta}}{\text{argmax}}\{\log \mathcal{L}_\alpha(\boldsymbol{\beta})\}$$

Gamma Distribution: Profile Likelihood

and then $\widehat{\alpha}^{pMLE} = \underset{\alpha}{\text{argmax}} \left\{ \log \mathcal{L}_\alpha(\widehat{\beta}_\alpha^{pMLE}) \right\}$. Observe that

$$\sqrt{n}(\widehat{\alpha}^{pMLE} - \alpha) \xrightarrow{\mathcal{L}} \mathcal{N}(0, [\mathbb{I}_{\alpha,\alpha} - \mathbb{I}_{\alpha,\beta}\mathbb{I}_{\beta,\beta}^{-1}\mathbb{I}_{\beta,\alpha}]^{-1})$$

The (profile) likelihood ratio test is based on

$$2\left(\max\left\{\mathcal{L}(\alpha,\beta)\right\} - \max\left\{\mathcal{L}(\alpha_0,\beta)\right\}\right)$$

If (α_0, β_0) are the true value, this difference can be written

$$2\left(\max\left\{\mathcal{L}(\alpha,\beta)\right\} - \max\left\{\mathcal{L}(\alpha_0,\beta_0)\right\}\right) - 2\left(\max\left\{\mathcal{L}(\alpha_0,\beta)\right\} - \max\left\{\mathcal{L}(\alpha_0,\beta_0)\right\}\right)$$

Using Taylor's expension

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \alpha}\bigg|_{(\alpha_0,\widehat{\beta}_{\alpha_0})} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{I}_{\alpha_0\alpha_0}) - \mathbb{I}_{\alpha_0\beta_0}\mathbb{I}_{\beta_0\beta_0}^{-1}\mathbb{I}_{\beta_0\alpha_0}$$

and $2\left(\mathcal{L}(\widehat{\alpha},\widehat{\beta}) - \mathcal{L}(\alpha_0,\widehat{\beta}_{\alpha_0})\right) \xrightarrow{\mathcal{L}} \chi^2(\dim(\alpha))$.

Gamma Distribution: Profile Likelihood

Consider some lognormal sample, and fit a Gamma distribution,

$$f(x;\alpha,\beta)=\frac{x^{\alpha-1}\beta^{\alpha}\,e^{-\beta x}}{\Gamma(\alpha)} \text{ with } x>0 \text{ and } \boldsymbol{\theta}=(\alpha,\beta).$$

```
> y = exp(rnorm(100))
```

Maximum-likelihood, $\widehat{\boldsymbol{\theta}}=\text{argmax}\{\log\mathcal{L}(\boldsymbol{\theta})\}$.

```
> library(MASS)
> (F = fitdistr(x,"gamma"))
     shape       rate
  1.4214497   0.8619969
 (0.1822570) (0.1320717)
> F$estimate[1]+c(-1,1)*1.96*F$sd[1]
[1] 1.064226 1.778673
```

Gamma Distribution: Profile Likelihood

See also

```
> log_lik=function(theta){
+   a = theta[1]
+   b = theta[2]
+   logL = sum(log(dgamma(y,a,b)))
+   return(-logL)
+ }
> optim(c(1,1),log_lik)
$par
[1] 1.4214116 0.8620311
```

We can also use profile likelihood,

$$\widehat{\alpha} = \text{argmax}\left\{ \max_{\beta} \{ \log \mathcal{L}(\alpha, \beta) \} \right\} = \text{argmax}\left\{ \log \mathcal{L}(\alpha, \widehat{\beta}_\alpha) \right\}$$

Gamma Distribution: Profile Likelihood

```
> prof_log_lik = function(a){
+   b = (optim(1,function(z) -sum(log(dgamma(x,a,z)))))$par
+   return(-sum(log(dgamma(x,a,b))))
+ }


> vx = seq(.5,3,length=101)
> vl = -Vectorize(prof_log_lik)(vx)
> plot(vx,vl,type="l")
> optim(1,prof_log_lik)
$par
[1] 1.421094
```

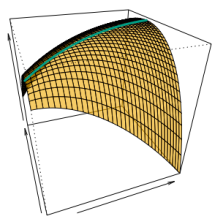
We can use the likelihood ratio test

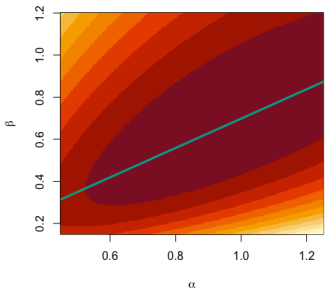
$$2(\log \mathcal{L}_p(\widehat{\alpha}) - \log \mathcal{L}_p(\alpha)) \sim \chi^2(1)$$

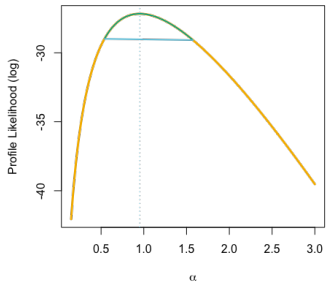
Gamma Distribution: Profile Likelihood

The implied 95% confidence interval is

```
> (b1 = uniroot(function(z) Vectorize(prof_log_lik)(z)+borne,c(.5,1.5))
    $root)
[1] 1.095726
> (b2 = uniroot(function(z) Vectorize(prof_log_lik)(z)+borne,c(1.25,2.5))
    $root)
[1] 1.811809
```







Random Vectors

Let $\boldsymbol{X} = (X_1, \cdots, X_d)$ denote a random vector in dimension d,

- The expectation of $\boldsymbol{X}$, denoted $\mathbb{E}(\boldsymbol{X})$ is defined (if it exists) by the vector of dimension d, $\mathbb{E}(\boldsymbol{X}) = (\mathbb{E}(\boldsymbol{X}_1), \ldots, \mathbb{E}(\boldsymbol{X}_d))^\top$.
- The covariance matrix (also called variance-covariance matrix of $\boldsymbol{X}$) is defined (if it exists) by the matrix of size (d, d).

$$\text{Var}(\boldsymbol{X}) = \mathbb{E}\left((\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))^\top \right).$$

So the i, j term of this matrix represents the covariance between X_i and X_j,

$$\text{Cov}(X_i, X_j) = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j)) \right].$$

Random Vectors

Let $\boldsymbol{X}$ be a random vector of dimension d, mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two reel matrices of size (d, p) and (d, q) and let $\mathbf{a} \in \mathbb{R}^p$ then

- $\text{Var}(\boldsymbol{X}) = \mathbb{E}\left((\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top\right) = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top$.
- $\mathbb{E}\left(\boldsymbol{A}^\top\boldsymbol{X} + \mathbf{a}\right) = \boldsymbol{A}^\top\boldsymbol{\mu} + \mathbf{a}$.
- $\text{Var}\left(\boldsymbol{A}^\top\boldsymbol{X} + \mathbf{a}\right) = \boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{A}$.
- $\text{Cov}\left(\boldsymbol{A}^\top\boldsymbol{X}, \boldsymbol{B}^\top\boldsymbol{X}\right) = \boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{B}$.

The Gaussian Distribution

Definition 1.11: Gaussian Univariate Distributoin

A Gaussian variable, with distribution $\mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$, has density

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

- Then $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$.
- Observe that if $Z \sim \mathcal{N}(0, 1)$, $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.

The Gaussian Distribution

Definition 1.12: Gaussian Multivariate Distributoin

The Gaussian **vector** $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: $\boldsymbol{X} = (X_1, ..., X_d)$ is a Gaussian vector with mean $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $d \times d$ matrix $\text{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \mathbb{E}\left((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\right)$ non-degenerated ($\boldsymbol{\Sigma}$ is invertible) if its density is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \ \boldsymbol{x} \in \mathbb{R}^n,$$

see multivariate Gaussian distribution

Gaussian (multivariate) distribution

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with density

$$f_{\mathbf{X}}(x_1, \ldots, x_d) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

where $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

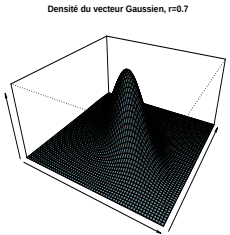
Estimates are $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$

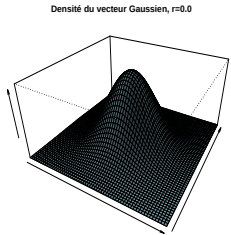
In dimension 2, $f(x, y)$ is proportional to

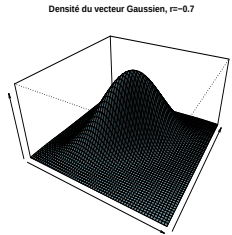
$$\exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} \right] \right)$$

levels curves (isodensities) are ellipses.

Gaussian (multivariate) distribution







Quadratic Forms

Consider $\boldsymbol{M} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$,

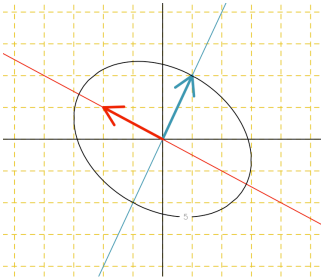
and function $\boldsymbol{z} \mapsto \boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z}$, i.e.

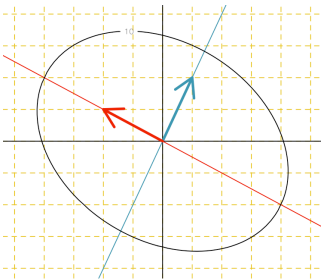
$$f : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

or $ax^2 + 2bxy + cy^2$ is a quadratic form.

If $\boldsymbol{M} > 0$, points $\boldsymbol{z} = (x, y)$ such that $\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z} = \gamma$, for some $\gamma > 0$, are on an ellipse (centered on $\mathbf{0}$)

Let $\lambda_1 \geq \lambda_2 > 0$ denote the eigenvalues of $\boldsymbol{M}$ and $\vec{\boldsymbol{v}}_1$ and $\vec{\boldsymbol{v}}_2$ denote the eigenvectors.





Quadratic Forms

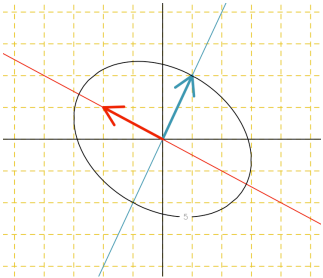
On the picture, $\boldsymbol{M} = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.9 \end{pmatrix}$

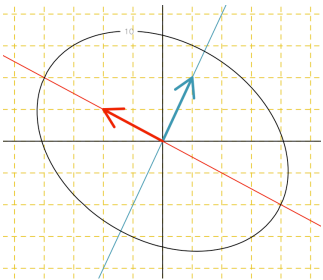
```
> M=matrix(c(.6,.2,.2,.9),2,2)
> eigen(M)
eigen() decomposition
$values
[1] 1.0 0.5
$vectors
           [,1]       [,2]
[1,] 0.4472136 -0.8944272
[2,] 0.8944272  0.4472136
```

i.e. $\lambda_1 = 1$ and $\lambda_2 = 1/2$, and

$$\vec{v}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \ \vec{v}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} -2 \\ 1 \end{pmatrix}$$

Note that $\|\vec{v}_1\| = \|\vec{v}_2\| = 1$ and $\vec{v}_1 \perp \vec{v}_2$





The Gaussian Distribution

If $\boldsymbol{X}$ is a Gaussian vector, then for any i, X_i has a (univariate) Gaussian distribution, but its converse it not necessarily true.

Let $\boldsymbol{X} = (X_1, ..., X_d)$ be a random vector with mean $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and with covariance matrix $\boldsymbol{\Sigma}$, if $\boldsymbol{A}$ is a $k \times d$ matrix, and $\boldsymbol{b} \in \mathbb{R}^k$, then $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ is a Gaussian vector $\mathbb{R}^k$, with distribution $\mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\right)$.

Observe that if (X_1, X_2) is a Gaussian vector, X_1 and X_2 are independent if and only if

$$\text{Cov}(X_1, X_2) = \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))) = 0.$$

Beta and Dirichlet distribution

Definition 1.13: Beta distribution

X, random variable in $[0,1]$, has distribution $\mathcal{B}eta(\alpha,\beta)$ if its density is

$$f(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \text{ where } \mathrm{B}(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},\ x\in[0,1].$$

Proposition 1.2: Beta distribution

X_1, X_2 be two independent Gamma variables, $X_1 \sim \mathcal{G}amma(\alpha,\gamma)$ and $X_2 \sim \mathcal{G}amma(\beta,\gamma)$, then

$$Y=\frac{X_1}{X_1+X_2}\sim \mathcal{B}eta(\alpha,\beta).$$

Beta and Dirichlet distribution

Definition 1.14: Dirichlet distribution

$\boldsymbol{X}$, random vector in $\mathcal{S}_d \subset [0,1]^d$, has distribution $\mathcal{D}irichlet(\boldsymbol{\alpha})$ if its density is

$$f(\boldsymbol{x}) = \frac{x_1^{\alpha_1-1} x_2^{\alpha_2-1} \cdots x_d^{\alpha_d-1}}{\mathrm{B}(\boldsymbol{\alpha})} \text{ where } \mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{d} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{d} \alpha_i\right)}, \ \boldsymbol{x} \in \mathcal{S}_d.$$

Proposition 1.3: Dirichlet distribution

$X_1, X_2, \cdots, X_d$ be d independent Gamma variables, $X_i \sim \mathcal{G}amma(\alpha_i, \gamma)$,

$$\boldsymbol{Y} = (Y_1, \cdots, Y_d) = \left(\frac{X_1}{X_1 + \cdots + X_d}, \cdots, \frac{X_d}{X_1 + \cdots + X_d} \right) \sim \mathcal{D}irichlet(\boldsymbol{\alpha}).$$

Statistics, Bayesian Approach

Proposition 1.4: Bayes formula

Given two events $\mathcal{A}$ and $\mathcal{B}$ such that $\mathbb{P}[\mathcal{B}] \neq 0$,

$$\mathbb{P}[\mathcal{A}|\mathcal{B}] = \frac{\mathbb{P}[\mathcal{B}|\mathcal{A}] \cdot \mathbb{P}[\mathcal{A}]}{\mathbb{P}[\mathcal{B}]} \propto \mathbb{P}[\mathcal{B}|\mathcal{A}] \cdot \mathbb{P}[\mathcal{A}].$$

Bayes (1763) and Laplace (1774).

Bayesian statistics ?

- Bayes formula (the "inverse problem"),
  Bayes (1763), Laplace (1774)

Given two events $\mathcal{A}$ and $\mathcal{B}$ such that $\mathbb{P}(\mathcal{B}) \neq 0$,

$$\mathbb{P}[\mathcal{A}|\mathcal{B}] = \frac{\mathbb{P}[\mathcal{B}|\mathcal{A}] \cdot \mathbb{P}[\mathcal{A}]}{\mathbb{P}[\mathcal{B}]}.$$

"If a person has an expectation depending on the happening of an event, the probability of the event is [in the ratio] to the probability of its failure as his loss if it fails [is in the ratio] to his gain if it happens ", Proposition 2, Bayes (1763)

"The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the chance of the thing expected upon its happening ", Bayes (1763)

Bayesian statistics ?

- Bayes formula (the "inverse problem"),
  Bayes (1763), Laplace (1774)

Given two events $\mathcal{A}$ and $\mathcal{B}$ such that $\mathbb{P}(\mathcal{B}) \neq 0$,

$$\mathbb{P}[\mathcal{A}|\mathcal{B}] = \frac{\mathbb{P}[\mathcal{B}|\mathcal{A}] \cdot \mathbb{P}[\mathcal{A}]}{\mathbb{P}[\mathcal{B}]} \propto \mathbb{P}[\mathcal{B}|\mathcal{A}] \cdot \mathbb{P}[\mathcal{A}].$$

- subjective probabilities,
  De Finetti (1937), Anscombe et al. (1963), Kahneman and Tversky (1972) Savage (1972), Jeffrey (2004)
- Non-frequentist approach of probablities,
  Neyman (1977), Bayarri and Berger (2004)
- Credibility and "*experience rating*"
  Whitney (1918), Longley-Cook (1962), Bühlmann (1967), Klugman (1991)

Bayesian statistics ?

- Bayes formula (the "inverse problem"),
  Bayes (1763), Laplace (1774)

Given two events $\mathcal{A}$ and $\mathcal{B}$ such that $\mathbb{P}(\mathcal{B}) \neq 0$,

$$\mathbb{P}[\mathcal{A}|\mathcal{B}] = \frac{\mathbb{P}[\mathcal{B}|\mathcal{A}] \cdot \mathbb{P}[\mathcal{A}]}{\mathbb{P}[\mathcal{B}]}.$$

- An inverse problem (we try to determine the causes of a phenomenon of a phenomenon from the experimental observation of its effects)
- An update of beliefs (from a *prior* distribution $\mathbb{P}(\mathcal{A})$ to a *posterior* distribution $\mathbb{P}(\mathcal{A}|\mathcal{B})$)

Bayesian statistics ?

A person coughs (event B). Which hypothesis is the most credible?

$$\begin{cases} \mathcal{A}_1 : \text{she has lung cancer} \\ \mathcal{A}_2 : \text{she has gastroenteritis} \\ \mathcal{A}_3 : \text{she has the flu} \end{cases}$$

With Bayes' rule $\mathbb{P}[\text{disease}|\text{symptom}] \propto \mathbb{P}[\text{symptom}|\text{disease}] \cdot \mathbb{P}[\text{disease}]$

$$\begin{cases} \mathcal{A}_1 : \mathbb{P}[\text{disease}] \approx 0 \text{ (even if } \mathbb{P}[\text{symptom}|\text{disease}] \approx 1) \\ \mathcal{A}_2 : \mathbb{P}[\text{symptom}|\text{disease}] \approx 0 \text{ (even if } \mathbb{P}[\text{symptom}|\text{disease}] \text{ high)} \\ \mathcal{A}_3 : \text{two reasonable probabilities} \end{cases}$$

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

$$\mathbb{P}(\text{treasure behind the door}) = \frac{1}{3}$$

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

$$\mathbb{P}(\text{treasure behind the door}) = \frac{1}{3}$$

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

- strategy 1 : always switch the door
- strategy 2 : never switch the door

$$\begin{aligned}
&\mathbb{P}(\text{strategy 2 winning})\\
&= \mathbb{P}(\text{treasure behind the door choisie initialement})\\
&= \frac{1}{3}
\end{aligned}$$

(making the goat appear behind the third door does not bring no information on what's behind the first door)

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

- strategy 1 : always switch the door
- strategy 2 : never switch the door

$$
\begin{aligned}
&\mathbb{P}(\text{strategy 1 winning})\\
&= \mathbb{P}(\text{treasure behind the other door})\\
&= \mathbb{P}(\text{treasure behind the other door}|\times \text{ correct }) \cdot \mathbb{P}(\times \text{ correct })\\
&+ \mathbb{P}(\text{treasure behind the other door}|\times \text{ false }) \cdot \mathbb{P}(\times \text{ false })\\
&= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}
\end{aligned}
$$

Practice of Bayesian Statistics

"Do doctors understand test results? ", from Kremer (2014):

1 percent of adults have cancer. The vast majority of these cancers (90 percent) can be detected by a test. There is a 9 percent chance that the test will be positive in a person who does not have cancer. If the test is positive, what is the likelihood that the person actually has cancer?

A) 9 out of 10
B) 8 out of 10
C) 1 out of 2
D) 1 out of 10
E) 1 out of 100

Practice of Bayesian Statistics

"Do doctors understand test results? ", from Kremer (2014):

1 percent of adults have cancer. The vast majority of these cancers (90 percent) can be detected by a test. There is a 9 percent chance that the test will be positive in a person who does not have cancer. If the test is positive, what is the likelihood that the person actually has cancer?

A) 9 out of 10 (chosen by 50% gynecologists)

B) out of 10

C) 1 out of 2

D) 1 out of 10

E) 1 out of 100

Practice of Bayesian Statistics

1 percent of adults have cancer. The vast majority of these cancers (90 percent) can be detected by a test. There is a 9 percent chance that the test will be positive in a person who does not have cancer. If the test is positive, what is the likelihood that the person actually has cancer?

Answer: when formalizing

$$\begin{cases} \mathbb{P}[\text{cancer}] = 1\% \\ \mathbb{P}[\text{test positive}|\text{cancer}] = 90\% \\ \mathbb{P}[\text{test positive}|\text{no cancer}] = 9\% \end{cases}$$

then, using Bayes' rule

$$\mathbb{P}[\text{cancer}|\text{test positive}] = \frac{\mathbb{P}[\text{test positive}|\text{cancer}] \cdot \mathbb{P}[\text{cancer}]}{\mathbb{P}[\text{test positive}]} = \frac{90\% \times 1\%}{9\% \times 99\% +} = \frac{9}{9+89} \simeq \frac{1}{10}$$

valid answer is D, "1 out of 10".

Practice of Bayesian Statistics

For Gigerenzer and Hoffrage (1995), the Bayesian formulation is (too) complex.

Another presentation of the problem:

Out of 10,000 people, 100 have cancer. Of these 100, 90%, or 90, will test positive. Of the remaining 9,900, 9 percent, or 899, will test positive. Of a sample of people who test positive, what fraction actually have cancer?

Answer: 90 among (90+899), i.e. about "1 out of 10".

Bayesianism, statistics and calculus

$$\text{posterior } = \pi(\theta|\boldsymbol{y}) = \frac{\pi(\theta)\cdot\mathbb{P}(\boldsymbol{y}|\theta)}{\mathbb{P}(\boldsymbol{y})} = \frac{\text{prior } \cdot \text{likelihood}}{\text{evidence}}$$

$$\text{posterior } = \pi(\theta|\boldsymbol{y}) \propto \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \cdot \binom{s}{n}\theta^s(1-\theta)^{n-s}$$

- Conjugate distributions: **Binomial - Beta**

The likelihood for binomial (Bernoulli) variables

$$\begin{cases} \boldsymbol{x} \mapsto f(\boldsymbol{x}; p) = p^s(1-p)^{n-s} \text{ where } s = \boldsymbol{x}^\top \mathbf{1} = x_1 + \cdots + x_n \\ p \mapsto \pi(p; \boldsymbol{x}) = p^s(1-p)^{n-s} \text{ on } [0,1] \text{ is a Beta distribution} \end{cases}$$

$$\text{If } \begin{cases} x_i|\theta \sim \mathcal{B}(\theta) \\ \theta \sim \mathcal{B}eta(a,b) \text{ prior} \end{cases} \quad \text{then } \theta|\boldsymbol{x} \sim \mathcal{B}eta(a+s, b+n-s) \text{ posterior}$$

(that can be extended to **Multinomial - Dirichlet**)

Bayesianism, statistics and calculus

- Conjugate distributions : **Poisson - Gamma**

The likelihood for Poisson variables is

$$\begin{cases} \boldsymbol{x} \mapsto f(\boldsymbol{x};\lambda) = \dfrac{e^{n\lambda}\lambda^s}{x_1!\cdots x_n!} \text{ where } s = \boldsymbol{x}^\top \mathbf{1} = x_1 + \cdots + x_n \\ \lambda \mapsto \pi(\lambda;\boldsymbol{x}) \propto e^{n\lambda}\lambda^s \text{ on } \mathbb{R}_+ \text{ is a Gamma distribution} \end{cases}$$

If

$$\begin{cases} x_i|\lambda \sim \mathcal{P}(\lambda) \\ \theta \sim \mathcal{G}amma(a,b) \text{ a priori} \end{cases} \quad \text{then } \lambda|\boldsymbol{x} \sim \mathcal{G}amma(a+s, b+n) \text{ a posteriori}$$

Hence

$$\text{a priori } \mathbb{E}(\lambda) = \frac{a}{b} \text{ and a posteriori } \mathbb{E}(\lambda|\boldsymbol{x}) = \frac{a+s}{b+n}$$

intensively used in credibility theory Bühlmann (1967).

Bayesianism, statistics and calculus

- Conjugate distributions : **Normal - Normal**

If variance $\boldsymbol{\Sigma}$ is known

$$\begin{cases} \boldsymbol{x}_i|\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{cases} \quad \text{then } \boldsymbol{\mu}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$$

$$\text{where } \begin{cases} \boldsymbol{\mu}_x = \left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}\right) \\ \boldsymbol{\Sigma}_x = \left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \end{cases}$$

used classically in Bayesian econometrics.

Bayesianism, statistics and calculus

- Conjugate distributions : **Normal - Inverse Wishart**

If mean $\boldsymbol{\mu}$ is known

$$\begin{cases} \boldsymbol{x}_i|\boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma} \sim IW(\nu_0, \boldsymbol{\Psi}_0) \end{cases} \quad \text{then } \boldsymbol{\Sigma}|\boldsymbol{x} \sim IW(\nu_x, \boldsymbol{\Psi}_x)$$

$$\text{where } \begin{cases} \nu_x = n + \nu \\ \boldsymbol{\Psi}_x = \boldsymbol{\Psi} + \sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^\top \end{cases}$$

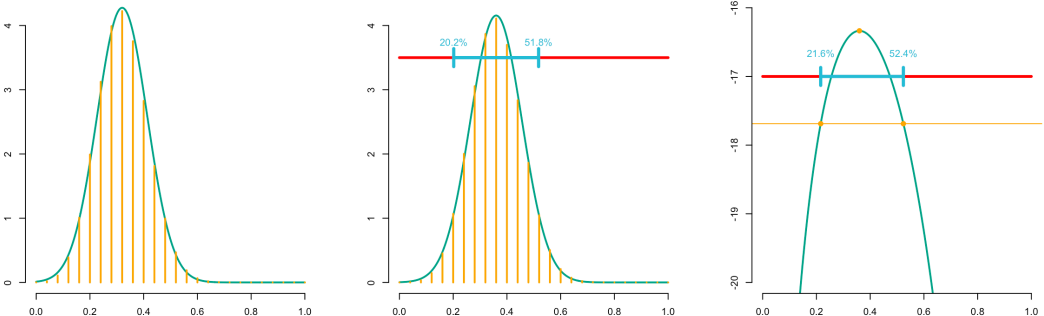
Classically used in Bayesian econometrics, for VAR models, Adjemian and Pelgrin (2008), or in portfolio management, Black and Litterman (1990, 1992) (see also Satchell and Scowcroft (2000) for a perspective).

Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0,0,0,1,0,1,1,0,0,0,0,0,1,0,1,0,1,1,0,1,1,0,0,0,0\}$, $\mathcal{B}(\theta)$

Frequentist approach, $\widehat{\theta} \approx \mathcal{N}\left(\theta, \dfrac{\theta(1-\theta)}{n}\right)$, $\mathbb{P}\left(\theta \in \left[\overline{x} \pm 1.64\sqrt{\dfrac{\overline{x}(1-\overline{x})}{n}}\right]\right) \approx 90\%$

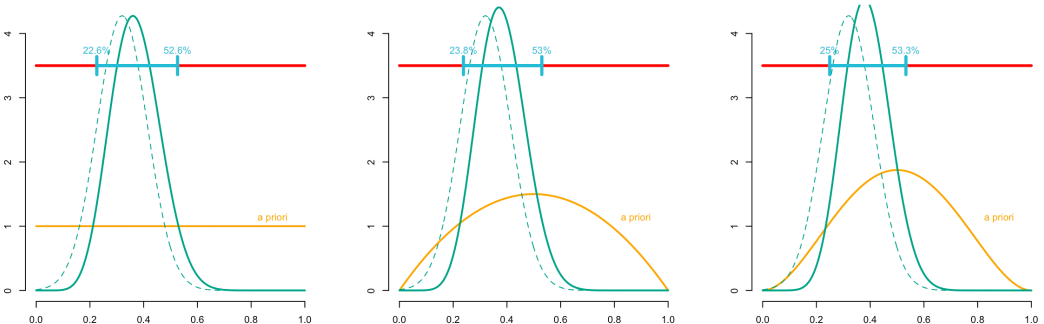


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

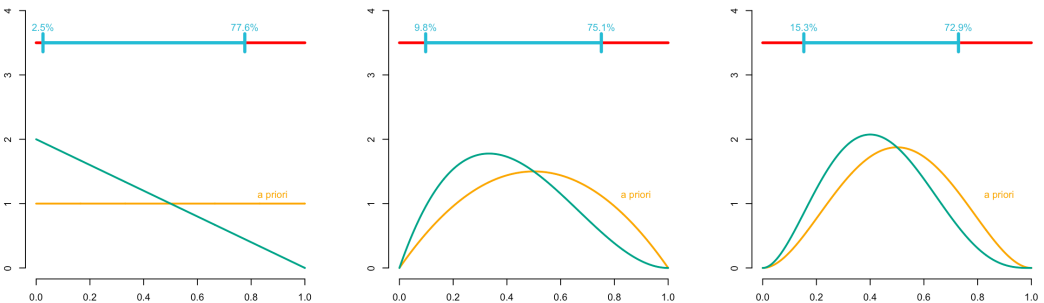


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

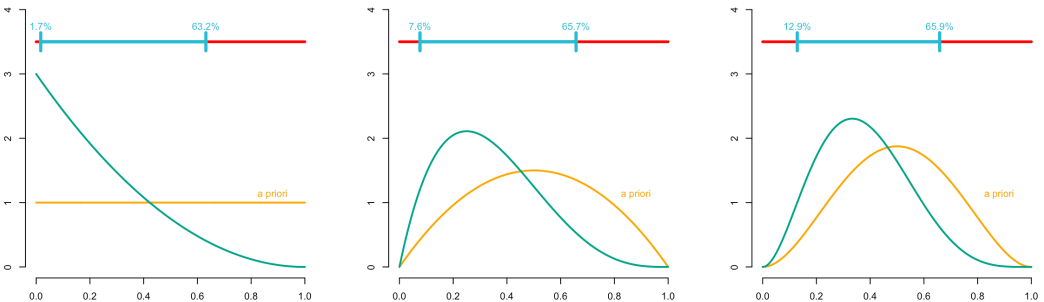


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach , $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

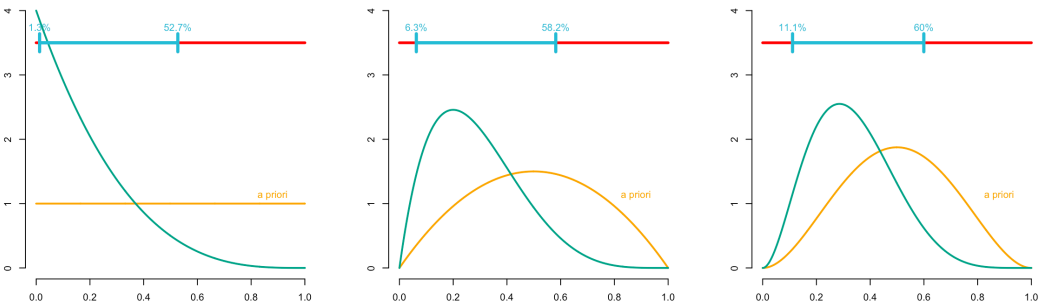


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach , $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

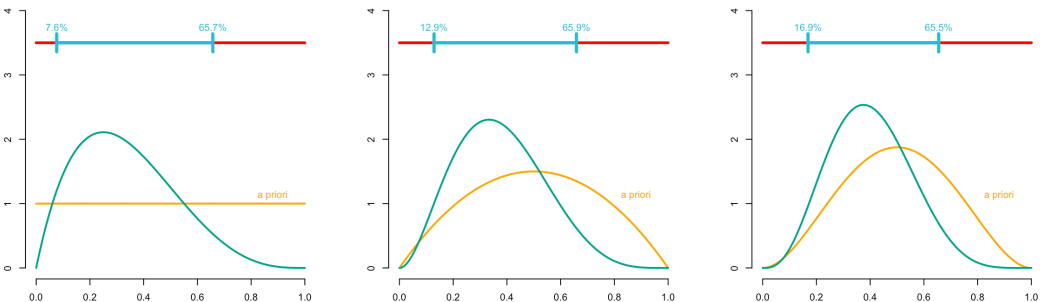


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1\}$, $\mathcal{B}(\theta)$

Bayesian approach , $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

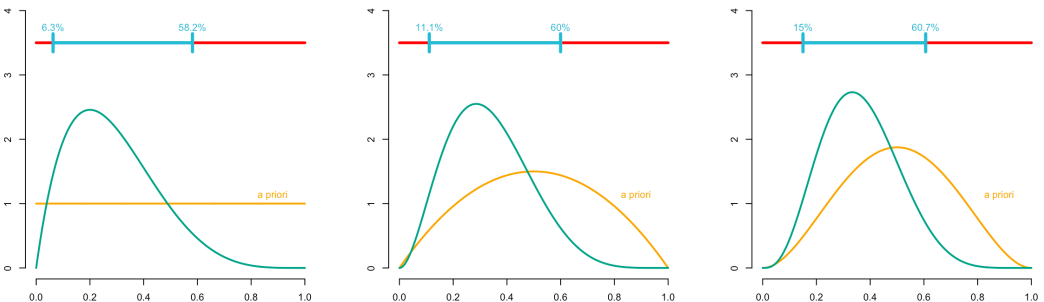


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach , $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

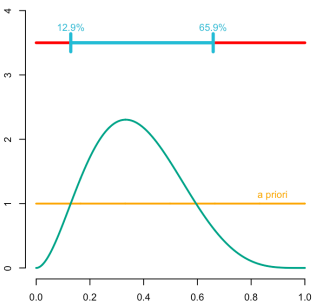


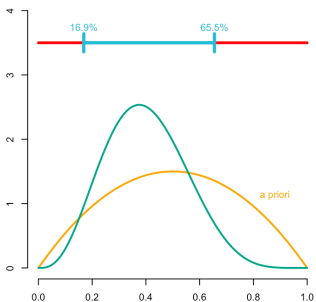
Bayesianism, statistics and calculus

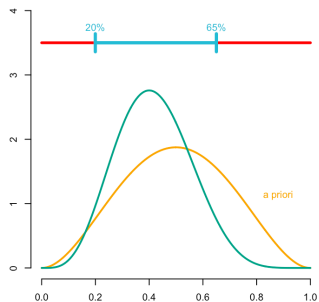
- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$





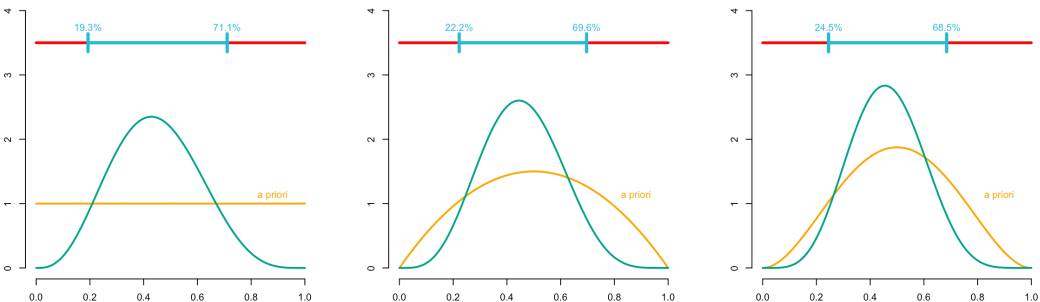


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

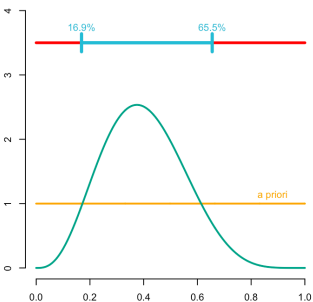


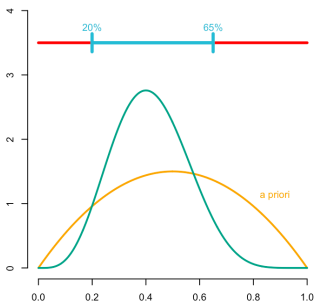
Bayesianism, statistics and calculus

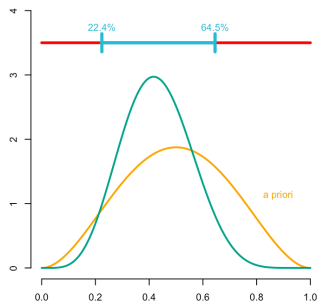
- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$





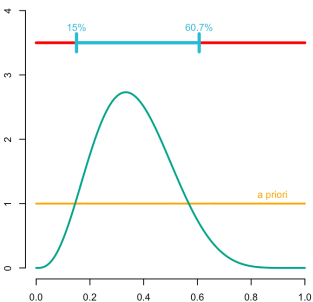


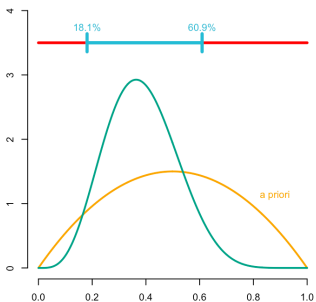
Bayesianism, statistics and calculus

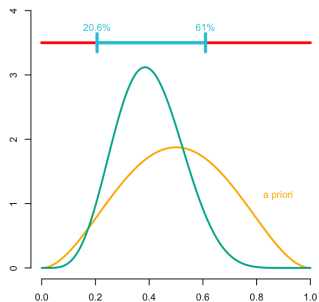
- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$





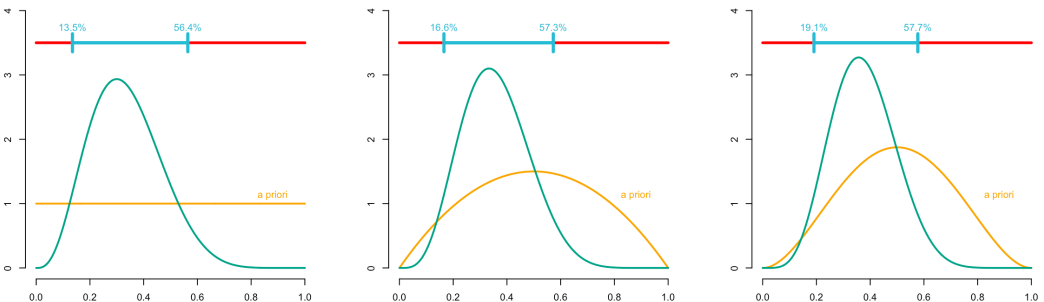


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

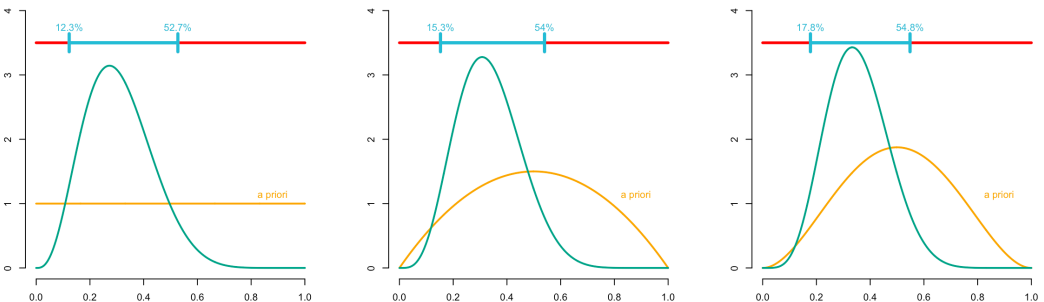


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

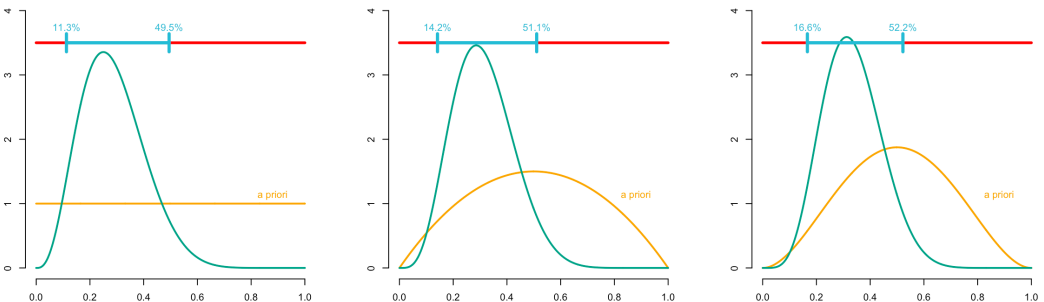


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

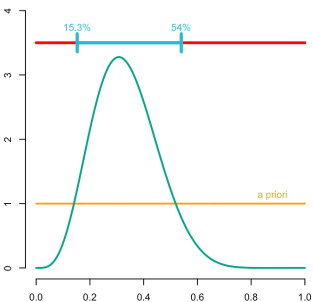


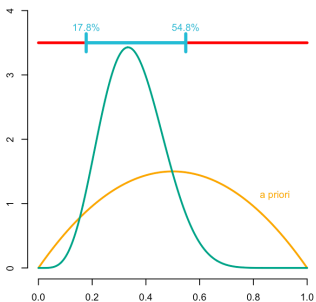
Bayesianism, statistics and calculus

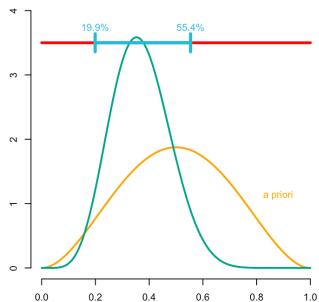
- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$





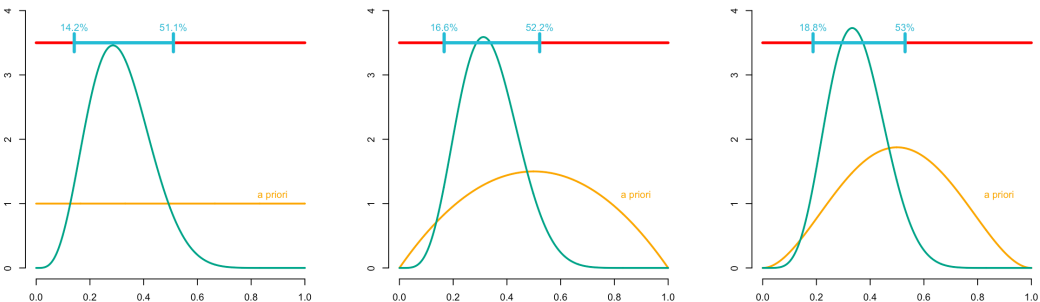


Bayesianism, statistics and calculus

- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

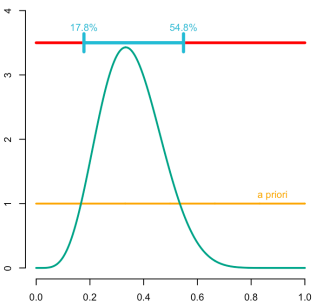


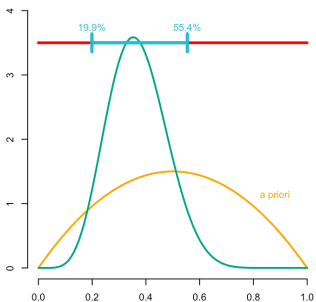
Bayesianism, statistics and calculus

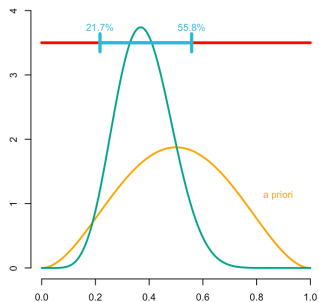
- Posterior distribution

Suppose $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$





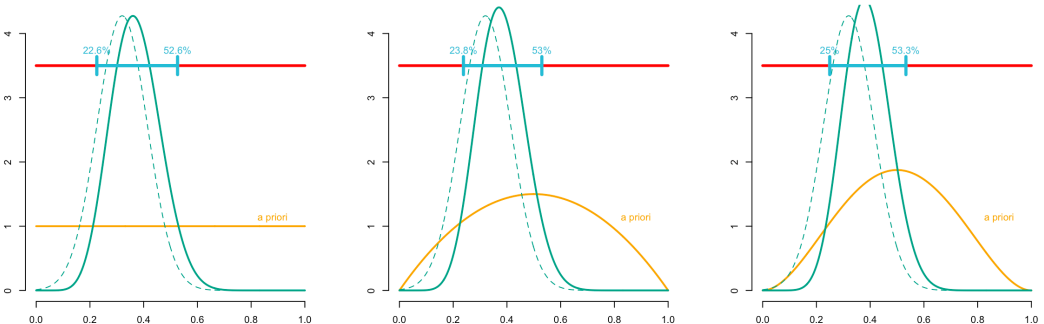


Bayesianism, statistics and calculus

- Posterior distribution

and finally $\boldsymbol{x} = \{0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \sum_{i=1}^{n} x_i$

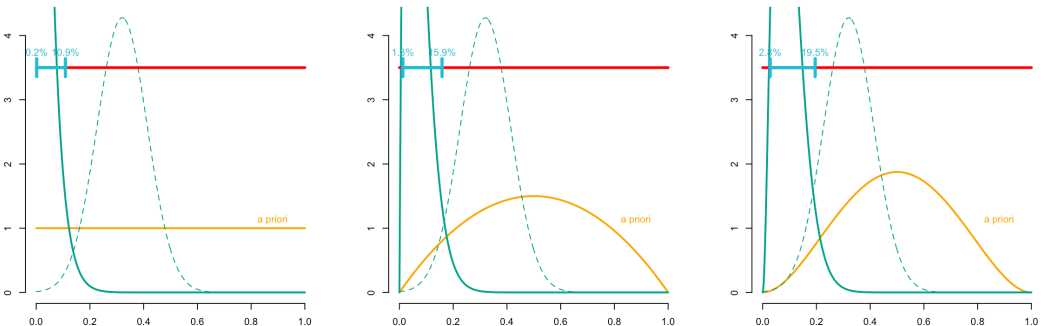


Bayesianism, statistics and calculus

- Posterior distribution

What if $\boldsymbol{x} = \{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$, $\mathcal{B}(\theta)$?

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0, \beta_0 + n)$, since $\sum_{i=1}^{n} x_i = 0$



Bayesianism, statistics and calculus

- From the distribution to the estimator

$$\begin{cases} \text{posterior average} & \widehat{\theta} = \mathbb{E}[\theta|\mathcal{D}] \\ \text{maximum a posteriori (MAP)} & \widehat{\theta} = \max\{\pi(\theta|\mathcal{D})\} \text{ i.e. the mode} \end{cases}$$

The average posterior is also the solution of the problem

$$\widehat{\theta} = \underset{\tau}{\text{argmin}} \left\{ \mathbb{E}[(\theta - \tau)^2|\mathcal{D}] \right\} = \underset{\tau}{\text{argmin}} \left\{ \int (\theta - \tau)^2 \pi(\theta|\mathcal{D}) d\theta \right\}$$

- "confidence interval" or "credibility interval"

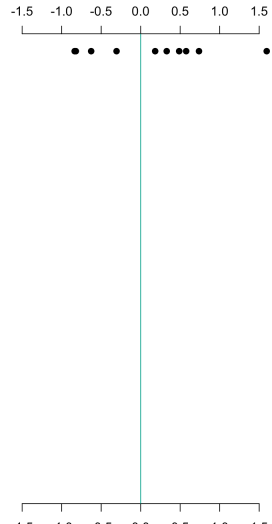
For the confidence interval, we look for $[\widehat{a}_{\mathcal{D}}, \widehat{b}_{\mathcal{D}}]$ such that $\mathbb{P}[\theta \in [\widehat{a}_{\mathcal{D}}, \widehat{b}_{\mathcal{D}}]] \geq 95\%$.

For the credibility interval, we look for $[a, b]$ such that $\mathbb{P}[\theta \in [a, b]|\mathcal{D}] \geq 95\%$.

Bayesianism, statistics and calculus

- "confidence interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
(here $\theta = 0$)

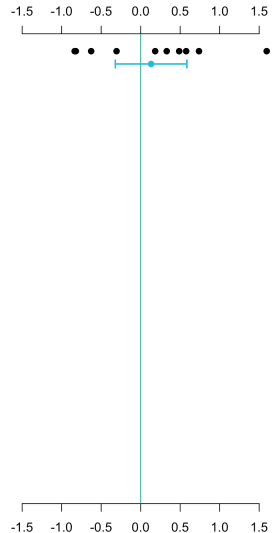


Bayesianism, statistics and calculus

- "confidence interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
(here $\theta = 0$)

Consider $[a, b] = \left[\overline{x} \pm q_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}\right]$



Bayesianism, statistics and calculus

- "confidence interval"

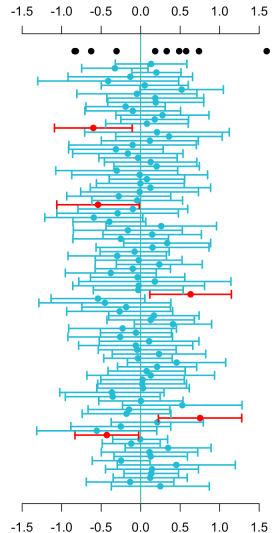
Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
(here $\theta = 0$)

Consider $[a, b] = \left[\overline{x} \pm q_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}\right]$

Generate $\mathcal{D}' = \{x'_1, \cdots, x'_n\}$ from $\mathcal{N}(\theta, \sigma^2)$, we want

$$\mathbb{P}\left[\theta \notin \left[\overline{x}' \pm q_\alpha \frac{\widehat{\sigma}}{\prime} \sqrt{n}\right]\right] \approx \alpha$$

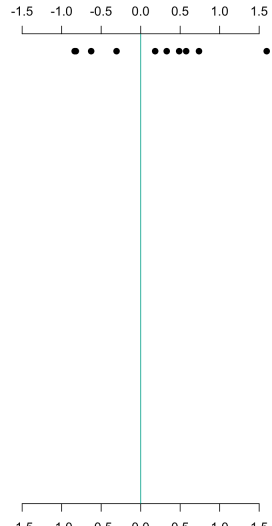
interpreted as a frequency, and repeating the experience.
Here, $\alpha = 5\%$: in 5% of the simulations, 0 is not in $[a, b]$.



Bayesianism, statistics and calculus

- "credibility interval"

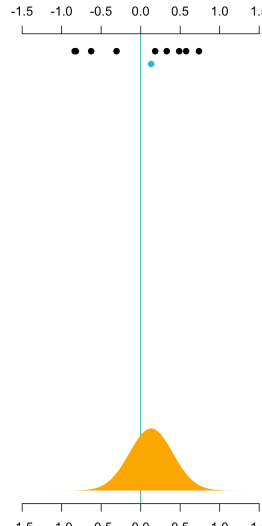
Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ



Bayesianism, statistics and calculus

- "credibility interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ
and $\pi(\cdot|\mathcal{D})$ is the posterior distribution
(potentially complicated)

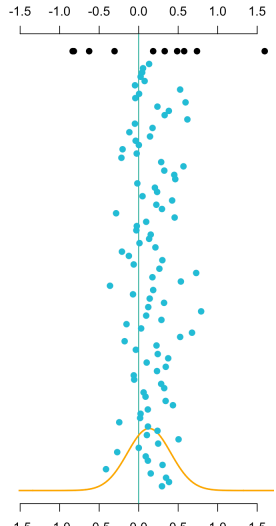


Bayesianism, statistics and calculus

- "credibility interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ
and $\pi(\cdot|\mathcal{D})$ is the posterior distribution
(potentially complicated)

Suppose we generate $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ given $\pi(\cdot|\mathcal{D})$.



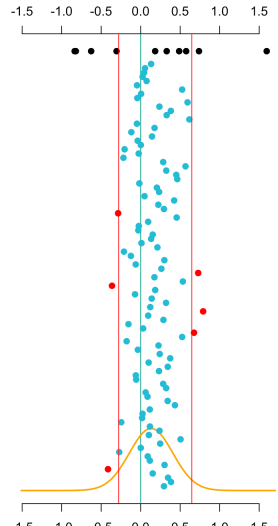
Bayesianism, statistics and calculus

- "credibility interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ
and $\pi(\cdot|\mathcal{D})$ is the posterior distribution
(potentially complicated)

Suppose we generate $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ given $\pi(\cdot|\mathcal{D})$.
Consider

$$\begin{cases} a = \widehat{\Pi}^{-1}(\alpha/2|\mathcal{D}) \text{ quantile with level } \alpha/2 \\ b = \widehat{\Pi}^{-1}(1-\alpha/2|\mathcal{D}) \text{ quantile with level } 1-\alpha/2 \end{cases}$$



Bayesianism, statistics and calculus

- "credibility interval"

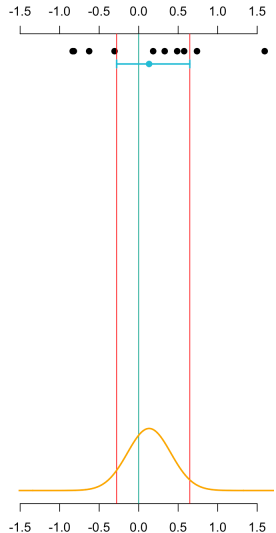
Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ
and $\pi(\cdot \mid \mathcal{D})$ is the posterior distribution
(potentially complicated)

Suppose we generate $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ given $\pi(\cdot|\mathcal{D})$.
Consider

$$\begin{cases} a = \widehat{\Pi}^{-1}(\alpha/2 \mid \mathcal{D}) \text{ quantile with level } \alpha/2 \\ b = \widehat{\Pi}^{-1}(1-\alpha/2 \mid \mathcal{D}) \text{ quantile with level } 1-\alpha/2 \end{cases}$$

then

$$\mathbb{P}\left[\theta \notin \left[\widehat{\Pi}^{-1}(\alpha/2|\mathcal{D}); \widehat{\Pi}^{-1}(1-\alpha/2|\mathcal{D})\right]\right] \approx \alpha$$



Bayesianism, statistics and calculus

We can also evoke the nonparametric Bayesian modeling, Ferguson (1973). Instead of assuming $X_i \sim f \in \mathcal{F}_\Theta$ where $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$, we consider a more general family,

$$X_i \sim f \in \mathcal{F} = \left\{ f : \int_{\mathbb{R}} [f''(y)]^2 dy < \infty \right\}$$

We can always compute a posterior law,

$$\pi(f \in A \mid \mathcal{D}) = \mathbb{P}(X \in A \mid \mathcal{D}) = \frac{\int_A \mathcal{L}_n(f) \mathrm{d}\pi(f)}{\int_{\mathcal{F}} \mathcal{L}_n(f) \mathrm{d}\pi(f)}, \text{ where } \mathcal{L}_n(f) = \prod_{i=1}^{n} f(x_i)$$

where π is an a prior distribution on $\mathcal{F}$. Very close to the Pólya urn problems (infinite), to the Chinese restaurant process and to the Dirichlet processes, Blackwell and MacQueen (1973), Ghosh and Ramamoorthi (2003), Orbanz and Teh (2010).

Bayesianism, statistics and calculus

For example, if $X_1, \cdots, X_n$ i.i.d. of distribution F. The a priori law π is a Dirichlet process, $D(\alpha, F_0)$, where $F_0 \in \mathcal{F}$ is a prior distribution for X, while α indicates the dispersion around F_0.

To draw according to $D(\alpha, F_0)$,

- we draw $z_1, z_2, \cdots$ according to F_0,
- we draw $v_1, v_2, \cdots$ according to a Beta law $\mathcal{B}(1, \alpha)$,
- we define iteratively weights, $\omega_1 = v_1$ and $\omega_j = v_j(1 - v_{j-1}) \cdots (1 - v_1)$
- $F(x) = \sum_{j \geq 1} \omega_j \mathbf{1}(x \leq z_j)$

If prior $\pi \sim D(\alpha, F_0)$, then the posterior is, $\pi | \mathcal{D} \sim D(\alpha + n, F_n)$ where

$$F_n = \frac{n}{n+\alpha} \widehat{F}_n + \frac{\alpha}{n+\alpha} F_0, \text{ where } \widehat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}(x \leq x_j)$$

Statistics with Weights

Suppose that instead of $\{x_1, x_2, \cdots, x_n\}$ we have $\{(x_1, \omega_1), (x_2, \omega_2), \cdots, (x_n, \omega_n)\}$

Suppose that weights are normalized, i.e., $\omega_1 + \omega_2 + \cdots + \omega_n = 1$.

- Weighted mean (or barycenter)

$$\overline{x} = \text{argmin}\left\{\sum_{i=1}^{n} (x_i - x)^2\right\} \text{ while } \overline{x}_{\boldsymbol{\omega}} = \text{argmin}\left\{\sum_{i=1}^{n} \omega_i (x_i - x)^2\right\}$$

- Weighted quantile

Given a dataset $\{x_1, x_2, \cdots, x_n\}$ with corresponding weights $\{\omega_1, \omega_2, \cdots, \omega_n\}$, the weighted quantile $Q_{\boldsymbol{\omega},p}$ for a probability p is determined as follows:

- Sort the observations in ascending order, $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ with corresponding reordered weights.
- Compute the cumulative sum of weights: $W_j = \sum_{i=1}^{j} \omega_{(i)}$

Statistics with Weights

- **Find the quantile position**: The weighted quantile Q_p corresponds to the value where the cumulative weight first reaches or exceeds pW_n, i.e., $W_j \geq pW_n$. The value at this index is the weighted quantile: $Q_p = x_{(j)}$.
- Linear Interpolation (if needed): If pW_n falls between two cumulative weights, we interpolate:

$$Q_{\omega,p} = x_{(j-1)} + \frac{pW_n - W_{j-1}}{W_j - W_{j-1}}(x_{(j)} - x_{(j-1)})$$

```
> library(cNORM)
> set.seed(1234)
x = runif(10)
> x
[1] 0.114 0.622 0.609 0.623 0.861 0.640 0.009 0.233 0.666 0.514
> w = dbeta(x,2,2)
> w
[1] 0.605 1.410 1.428 1.409 0.718 1.382 0.056 1.071 1.334 1.499
weighted.quantile(x, probs = .9, weights = w)
```

Statistics with Weights

```
[1] 0.7555
```

Independence

Independence

Independence is a fundamental notion in probability theory, as in statistics and the theory of stochastic processes. Two events are independent if, informally speaking, the occurrence of one does not affect the probability of occurrence of the other or, equivalently, does not affect the odds. $\mathbb{W}$

Definition 1.15: Independence (dimension 2)

X and Y are independent, denoted $X \perp\!\!\!\perp Y$, if for any sets $\mathcal{A}, \mathcal{B} \subset \mathbb{R}$,

$$\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B}] = \mathbb{P}[X \in \mathcal{A}] \cdot \mathbb{P}[Y \in \mathcal{B}].$$

Definition 1.16: Linear Independence (dimension 2)

Consider two random variables X and Y. $X \perp Y$ if and only if $\text{Cov}[X, Y] = 0$.

Independence

> **Correlation**
> in the broadest sense, "correlation" may indicate any type of association, in statistics it usually refers to the degree to which a pair of variables are linearly related. $\mathbb{W}$

Definition 1.17: Correlation (dimension 2), Pearson (1895)

X and Y are two random variables

$$\text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}}.$$

where $\text{Cov}[X, Y] = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Independence

From Cauchy-Schwarz theorem, $-1 \leq \text{Corr}[X, Y] \leq +1$ but those bounds are rarely sharp,

Proposition 1.5: Correlation bounds (dimension 2)

For any random variables X and Y (with finite variances), $r_{\min} \leq \text{Corr}[X, Y] \leq r_{\max}$, where

$$r_{\min} = \frac{\text{Cov}[F_x^{-1}(U), F_y^{-1}(1-U)]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}} \text{ and } r_{\max} = \frac{\text{Cov}[F_x^{-1}(U), F_y^{-1}(U)]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}}$$

Maximal correlation is obtained when X and Y are comonotonic (minimal correlation when X and $-Y$ are comonotonic).

Related to optimal transport, see also Knott and Smith (1984).

Independence

Proposition 1.6

Consider two random variables X and Y. $X \perp\!\!\!\perp Y$ if and only if for any functions $\varphi : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ (such that the expected values below exist and are well-defined) $\text{Cov}[\varphi(X), \psi(Y)] = 0$, i.e.,

$$\mathbb{E}[\varphi(X) \cdot \psi(Y)] = \mathbb{E}[\varphi(X)] \cdot \mathbb{E}[\psi(Y)].$$

Definition 1.18: Maximal Correlation, HGR

Consider two random variables X and Y,

$$r^\star(X, Y) = \max_{\varphi,\psi} \{\text{Corr}[\varphi(X), \psi(Y)]\}.$$

Independence

HGR because of Hirschfeld (1935), Gebelein (1941) and Rényi (1959) (also Sarmanov (1958a,b)).

$$r^\star(X,Y) = \max_{\varphi\in\mathcal{F}_x,\ \psi\in\mathcal{G}_y} \mathbb{E}[\varphi(X)\psi(Y)],$$

where

$$\begin{cases} \mathcal{F}_x = \{\varphi : \mathcal{X} \to \mathbb{R} : \mathbb{E}[\varphi(X)] = 0 \text{ and } \mathbb{E}[\varphi^2(X)] = 1\} \\ \mathcal{G}_y = \{\psi : \mathcal{Y} \to \mathbb{R} : \mathbb{E}[\psi(Y)] = 0 \text{ and } \mathbb{E}[\psi^2(Y)] = 1\} \end{cases}$$

See either `ccaPP` or `acepack` package,

```
> ccaPP::maxCorProj(x = x, y = y, method = "pearson")
> corstar = acepack::ace(x = x, y = y)
> cor(corstar$tx, corstar$ty)
```

Independence

Proposition 1.7

Consider two random variables X and Y. $X \perp\!\!\!\perp Y$ if and only if $r^\star(X, Y) = 0$.

Proof: Given a random variable X, its characteristic function is $\phi_X(t) = \mathbb{E}[e^{itX}]$. Recall that

$$\begin{cases} \phi_X(t) = \phi_Y(t), \ \forall t \in \mathbb{R} \text{ if and only if } X \overset{\mathcal{L}}{=} Y \\ \phi_{X,Y}(s,t) = \mathbb{E}[e^{i(sX+tY)}] = \phi_X(s) \cdot \phi_Y(t), \ \forall s,t \in \mathbb{R} \text{ if and only if } X \perp\!\!\!\perp Y \end{cases}$$

If $r^\star(X, Y) = 0$, let $s, t \in \mathbb{R}$ and consider $\varphi(x) = \phi_X(x) = \mathbb{E}[e^{ixX}]$ and $\psi(y) = \phi_Y(y) = \mathbb{E}[e^{isY}]$, then $\text{Cov}[e^{isX}, e^{itY}] = \text{Cov}[X'_s, Y'_t] = 0$, i.e. $\mathbb{E}[X'_s Y'_t] = \mathbb{E}[X'_s]\mathbb{E}[Y'_t]$,

$$\underbrace{\mathbb{E}[e^{i(sX+tY)}]}_{\phi_{XY}(s,t)} = \underbrace{\mathbb{E}[e^{isX}] \cdot \mathbb{E}[e^{itY}]}_{\phi_X(s)\cdot\phi_Y(t)}, \ \forall s,t \in \mathbb{R} \text{ i.e. } X \perp\!\!\!\perp Y.$$

Independence

Proposition 1.8

Consider two random variables X and Y such that (X, Y) is a Gaussian vector. Then $r^\star(X, Y) = |\text{Corr}[X,Y]|$.

See Lancaster (1957, 1958), and Gauss-Hermite decomposition

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2\left[1-\rho^2\right]}\right) = \phi(x)\phi(y) \cdot \sum_{i=0}^{\infty} r^i H_i(x) H_i(y)$$

where H_i's are Hermite polynomial.

Independence

Instead of

$$r^\star(X,Y) = \max_{\varphi\in\mathcal{F}_x,\ \psi\in\mathcal{G}_y} \mathbb{E}[\varphi(X)\psi(Y)],$$

where

$$\begin{cases} \mathcal{F}_x = \{\varphi : \mathcal{X} \to \mathbb{R} : \mathbb{E}[\varphi(X)] = 0 \text{ and } \mathbb{E}[\varphi^2(X)] = 1\} \\ \mathcal{G}_y = \{\psi : \mathcal{Y} \to \mathbb{R} : \mathbb{E}[\psi(Y)] = 0 \text{ and } \mathbb{E}[\psi^2(Y)] = 1\} \end{cases}$$

Definition 1.19: Constrained Maximal Correlation, Bach and Jordan (2002), Gretton et al. (2005)

Consider two random variables X and Y, as well as some Hilbert spaces $\bar{\mathcal{F}}_x \subset \mathcal{F}_x$ and $\bar{\mathcal{G}}_y \subset \mathcal{G}_y$,

$$\bar{r}^\star(X,Y) = \max_{\varphi\in\bar{\mathcal{F}}_x,\psi\in\bar{\mathcal{G}}_y} \{\text{Corr}[\varphi(X),\psi(Y)]\}.$$

Independence

Kimeldorf and Sampson (1978) and Kimeldorf et al. (1982) suggested to consider for $\bar{\mathcal{F}}_x$ and $\bar{\mathcal{G}}_y$ as subsets of monotone functions.

$$\begin{cases} \bar{\mathcal{F}}x = \{\varphi \in \mathcal{F}_x : \varphi \text{ monotone}\} \\ \bar{\mathcal{G}}_y = \{\psi \in \mathcal{G}_y : \psi \text{ monotone}\} \end{cases}$$

See Mourier (1953), Hannan (1961), Jensen and Mayer (1977) and Lin (1987).

Independence

Definition 1.20: Linear Independence

In a general context, consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively. $\boldsymbol{X} \perp \boldsymbol{Y}$ if and only if for any $\boldsymbol{a} \in \mathbb{R}^{d_x}$ and $\boldsymbol{b} \in \mathbb{R}^{d_y}$

$$\text{Cov}[\boldsymbol{a}^\top \boldsymbol{X}, \boldsymbol{b}^\top \boldsymbol{Y}] = 0.$$

Definition 1.21: Independence

In a general context, consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ if and only if for any $\mathcal{A} \subset \mathbb{R}^{d_x}$ and $\mathcal{B} \subset \mathbb{R}^{d_y}$,

$$\mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\} \cap \{\boldsymbol{Y} \in \mathcal{B}\}] = \mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\}] \cdot \mathbb{P}[\{\boldsymbol{Y} \in \mathcal{B}\}].$$

Independence

Proposition 1.9: Independence

Consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ if and only if for any functions $\varphi : \mathbb{R}^{d_x} \to \mathbb{R}$ and $\psi : \mathbb{R}^{d_y} \to \mathbb{R}$ (such that the expected values below exist and are well-defined)

$$\mathbb{E}[\varphi(\boldsymbol{X})\psi(\boldsymbol{Y})] = \mathbb{E}[\varphi(\boldsymbol{X})] \cdot \mathbb{E}[\psi(\boldsymbol{Y})],$$

or equivalently

$$\text{Cov}[\varphi(\boldsymbol{X}), \psi(\boldsymbol{Y})] = 0.$$

Independence

Definition 1.22: Mutual Independence

Let $\boldsymbol{Y} = (Y_1, \cdots, Y_k)$ denote some random vector. All components of $\boldsymbol{Y}$ are (mutually) independent if for any $\mathcal{A}_1, \cdots, \mathcal{A}_k \subset \mathbb{R}$

$$\mathbb{P}\left[\{(Y_1, \cdots, Y_k) \in \bigcap_{i=1}^{k} \mathcal{A}_i\}\right] = \prod_{i=1}^{k} \mathbb{P}[\{Y_i \in \mathcal{A}_i\}].$$

Definition 1.23: Conditional Independence (dimension 2)

X and Y are independent conditionally on Z, denoted $X \perp\!\!\!\perp Y \mid Z$, if for any sets $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset \mathbb{R}$,

$$\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B} | Z \in \mathcal{C}] = \mathbb{P}[X \in \mathcal{A} | Z \in \mathcal{C}] \cdot \mathbb{P}[Y \in \mathcal{B} | Z \in \mathcal{C}].$$

Independence

Definition 1.24: Conditional Independence

In a general context, consider three random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y})|\boldsymbol{Z}$ if and only if for any $\mathcal{A} \subset \mathbb{R}^{d_x}$, $\mathcal{B} \subset \mathbb{R}^{d_y}$ and $\mathcal{C} \subset \mathbb{R}^{d_z}$,

$$\mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\} \cap \{\boldsymbol{Y} \in \mathcal{B}\}|\boldsymbol{Z} \in \mathcal{C}] = \mathbb{P}[\{\boldsymbol{X} \in \mathcal{A}\}|\boldsymbol{Z} \in \mathcal{C}] \cdot \mathbb{P}[\{\boldsymbol{Y} \in \mathcal{B}\}|\boldsymbol{Z} \in \mathcal{C}].$$

Proposition 1.10

Consider three random variables X, Y, and Z. If $X \perp Z$ and $Y \perp Z$, then $aX + bY \perp Z$, for any $a, b \in \mathbb{R}$.

Independence

Proposition 1.11: $X \perp Z,\ Y \perp Z \not\Longrightarrow \psi(X,Y) \perp Z$

Consider three random variables X, Y, and Z. If $X \perp Z$ and $Y \perp Z$, it does not imply that $\psi(X,Y) \perp Z$, for any $\psi : \mathbb{R}^2 \to \mathbb{R}$.

$$(X,Y,Z) = \begin{cases} (0,0,0) & \text{with probability } 1/4, \\ (0,1,1) & \text{with probability } 1/4, \\ (1,0,1) & \text{with probability } 1/4, \\ (1,1,0) & \text{with probability } 1/4. \end{cases}$$

Proposition 1.12

Consider a random vector $\boldsymbol{X}$ in $\mathbb{R}^k$, and a random variable Z.
$\boldsymbol{X} \perp Z$ does not imply that $\psi(\boldsymbol{X}) \perp Z$, for any $\psi : \mathbb{R}^k \to \mathbb{R}$.

Independence

Proposition 1.13

Consider three random variables X, Y, and Z. Even if $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$, it does not imply either that $\psi(X, Y) \perp Z$ or that $\psi(X, Y) \perp\!\!\!\perp Z$, for any $\psi : \mathbb{R}^2 \to \mathbb{R}$.

Proposition 1.14

Consider a random vector $\boldsymbol{X}$ in $\mathbb{R}^k$, and a random variable Z.
$\boldsymbol{X} \perp\!\!\!\perp Z$ does not imply either that $\psi(\boldsymbol{X}) \perp Z$ **or** $\psi(\boldsymbol{X}) \perp\!\!\!\perp Z$, for any $\psi : \mathbb{R}^k \to \mathbb{R}$.

Random Numbers

- In Machine Learning, simulations help estimate quantities that are hard to compute directly.
- **Monte Carlo Simulation** is a popular method for approximating complex integrals, probabilities, and expectations.
- Simulations are essential in probabilistic modeling, where we rely on random sampling to estimate model parameters and distributions.
- Example: Simulating the expected value of a random variable when analytical solutions are not feasible.
- **Monte Carlo Simulation**: A statistical technique using random sampling to approximate solutions to problems that may be deterministic in nature.
- Typical application in ML:
  - Estimating integrals, expectations, and probabilities.
  - Sampling from a probability distribution.
  - Bayesian Inference and Markov Chain Monte Carlo (MCMC) methods.

Random Numbers

- Monte Carlo methods are useful when an exact solution is computationally expensive or analytically impossible.
- **Importance Sampling**: A technique to improve the efficiency of Monte Carlo simulations.
- Instead of directly sampling from the distribution p, we sample from an easier-to-sample distribution q and re-weight the samples:

$$\mathbb{E}_p[h(x)] = \mathbb{E}_q\left[\frac{p(x)}{q(x)}h(x)\right]$$

- This is useful when the target distribution is difficult to sample from, but we can easily sample from a different distribution.
- Example: Importance sampling is used in rare-event simulation, such as estimating the tail probabilities of heavy-tailed distributions.

Random Numbers

- In Bayesian Inference, we often need to calculate posterior distributions:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) \cdot p(\theta)}{p(\mathcal{D})}$$

- Exact computation of the posterior may not be possible, especially when the posterior is high-dimensional.
- Monte Carlo methods (e.g., Markov Chain Monte Carlo (MCMC)) can be used to approximate the posterior by generating samples.
- Popular MCMC algorithms:
  - Metropolis-Hastings
  - Gibbs sampling
- MCMC helps to efficiently sample from complex distributions, critical in probabilistic modeling.

Random Numbers

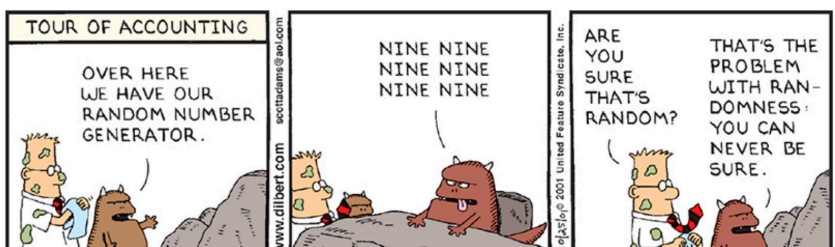
- **Bayesian Machine Learning**: Sampling from posterior distributions to make predictions or optimize hyperparameters.
- **Reinforcement Learning (RL)**: Simulating environments for agent training.
- **Uncertainty Quantification**: Simulating and modeling uncertainties in predictions, important for robustness in ML models.
- **Model Evaluation**: Using simulations to estimate model performance under various conditions (cross-validation, bootstrap).
- Example: In RL, Monte Carlo methods help simulate environments and estimate the expected reward for different actions.
- **Computational Cost**: Monte Carlo methods can be computationally expensive, especially for high-dimensional problems.
- **Convergence Rate**: In some cases, Monte Carlo estimates converge slowly, requiring a large number of samples.

Random Numbers

- **Choice of Proposal Distribution (for Importance Sampling)**: Poor choice of proposal distribution leads to inefficiency and high variance in estimates.
- **Curse of Dimensionality**: In high-dimensional spaces, the number of samples required grows exponentially.
- Despite these challenges, simulations are indispensable in modern ML for probabilistic modeling, optimization, and decision-making under uncertainty.

Random Numbers

> Random number generation is a process by which, often by means of a random number generator (RNG), a sequence of numbers or symbols is generated that cannot be reasonably predicted better than by random chance. In common understanding, "1 2 3 4 5" is not as random as "3 5 2 1 4" and certainly not as random as "47 88 1 32 41" but "we can't say authoritavely that the first sequence is not random ... it could have been generated by chance." W



Monte Carlo

> Monte Carlo methods, or Monte Carlo experiments, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. The underlying concept is to use randomness to solve problems that might be deterministic in principle. $\mathbb{W}$

The underlying idea is that if $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \cdots$ are i.i.d. with density p, from the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i) \overset{a.s.}{\to} \mathbb{E}_p[g(\boldsymbol{X})] = \int p(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Non-unique decomposition

$$\int a(\boldsymbol{x})b(\boldsymbol{x})c(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}_a[bc(\boldsymbol{X})] \text{ or } \mathbb{E}_{ab}[c(\boldsymbol{X})] \text{ or } \cdots$$

Importance sampling

if $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \cdots$ are i.i.d. with density p, from the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i) \overset{a.s.}{\to} \mathbb{E}_p[g(\boldsymbol{X})] = \int p(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

...

$$p(\boldsymbol{x})\ g(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} = q(\boldsymbol{x})\ \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\ g(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}$$

Might be easier to sample from q than from p
If $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \cdots$ are i.i.d. with density q, from the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{p(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)} g(\boldsymbol{X}_i) \overset{a.s.}{\to} \mathbb{E}_q\left[\frac{p(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)} g(\boldsymbol{X})\right] = \mathbb{E}_p[g(\boldsymbol{X})] = \int p(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Monte Carlo Markov Chain

From the law of large numbers, if $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \cdots$ are **i.i.d.** with density p,

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i) \ \to \ \mathbb{E}_p[g(\boldsymbol{X})] = \int p(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad \text{almost surely as } n \to \infty.$$

but similar results can also be obtained if if $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \cdots$ are **not i.i.d.**.

Monte Carlo Markov Chain

Definition 1.25: Markov Chains

A sequence of random variables (X_n) forms a Markov chain if

$$\mathbb{P}(X_{n+1} \in A \mid X_n, X_{n-1}, \cdots, X_0) = \mathbb{P}(X_{n+1} \in A \mid X_n).$$

If the state space is not finite, the transition probabilities are given by a Markov kernel $P(x, A)$, which satisfies:

- $P(x, A)$ is a probability measure on the state space for each fixed x.
- $P(x, A)$ is measurable in x for each measurable set A.

The stationary distribution (if it exists) π satisfies $\pi(A) = \int P(x, A)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ for all measurable sets A.

Interestingly, we have an "ergodic theorem,"

Monte Carlo Markov Chain

> A central concern of ergodic theory is the behavior of a dynamical system when it is allowed to run for a long time. W

Proposition 1.15: Ergodic Theorem

If (X_n) is an irreducible, aperiodic, and positive recurrent Markov chain with stationary distribution π, then for any function g with $\mathbb{E}_\pi[|g(X)|] < \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i) \;\to\; \mathbb{E}_\pi\big[g(\boldsymbol{X})\big] = \int \pi(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad \text{almost surely as } n \to \infty.$$

Given a distribution π, there are algorithms that can be used to create a Markov kernel $P(x, A)$ with stationary distribution π, such as Gibbs sampler and Hasting-Metropolis.

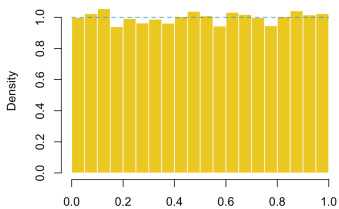
Boostrap and Resampling

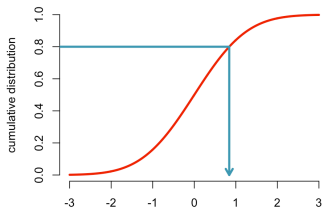
Proposition 1.16: Probability Integral Transform

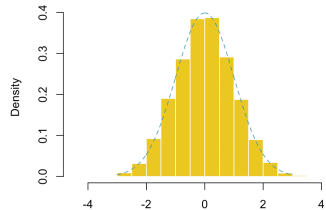
If F is a cumulative distribution function $\mathbb{R} \to [0,1]$, and if $U \sim \mathcal{U}([0,1])$, then $X := F^{-1}(U)$ has cumulative distribution function F, i.e., $X \sim F$.

Proof Let $x \in \mathbb{R}$, then $\mathbb{P}[X \leq x]$ can be written, with $F^{-1}(u) = \inf\{x \mid F(x) \geq u\}$

$$\mathbb{P}[X \leq x] = \mathbb{P}[F^{-1}(U) \leq x] = \mathbb{P}[F(F^{-1}(U)) \leq F(x)] = \mathbb{P}[U \leq F(x)] = F(x),$$







Boostrap and Resampling

```
> U <- runif(100)
  [1] 0.29 0.79 0.41 0.88 0.94 0.05 0.53 0.89 0.55 0.46 0.96 0.45 0.68
 [14] 0.57 0.10 0.90 0.25 0.04 0.33 0.95 0.89 0.69 0.64 0.99 0.66 0.71
 [27] 0.54 0.59 0.29 0.15 0.96 0.90 0.69 0.80 0.02 0.48 0.76 0.22 0.32
 [40] 0.23 0.14 0.41 0.41 0.37 0.15 0.14 0.23 0.47 0.27 0.86 0.05 0.44
 [53] 0.80 0.12 0.56 0.21 0.13 0.75 0.90 0.37 0.67 0.09 0.38 0.27 0.81
 [66] 0.45 0.81 0.81 0.79 0.44 0.75 0.63 0.71 0.00 0.48 0.22 0.38 0.61
 [79] 0.35 0.11 0.24 0.67 0.42 0.79 0.10 0.43 0.98 0.89 0.89 0.18 0.13
 [92] 0.65 0.34 0.66 0.32 0.19 0.78 0.09 0.47 0.51
```

```
> qnorm(U)
  [1] -0.56  0.80 -0.23  1.19  1.56 -1.69  0.07  1.24  0.13 -0.11  1.72
 [12] -0.12  0.46  0.18 -1.27  1.28 -0.69 -1.73 -0.45  1.69  1.22  0.50
 [23]  0.36  2.53  0.40  0.55  0.11  0.24 -0.56 -1.05  1.79  1.29  0.50
 [34]  0.83 -1.97 -0.06  0.70 -0.78 -0.47 -0.73 -1.07 -0.22 -0.22 -0.33
 [45] -1.03 -1.09 -0.73 -0.09 -0.63  1.07 -1.69 -0.15  0.84 -1.17  0.15
 [56] -0.82 -1.14  0.68  1.25 -0.32  0.43 -1.31 -0.30 -0.60  0.90 -0.13
 [67]  0.88  0.89  0.82 -0.15  0.69  0.33  0.55 -3.23 -0.06 -0.77 -0.31
```

Boostrap and Resampling

Let's consider a sample $\{x_1, \cdots, x_n\}$ i.i.d.

$$\text{of (theoretical) law} F(x) = \mathbb{P}[X_i \leq x],$$

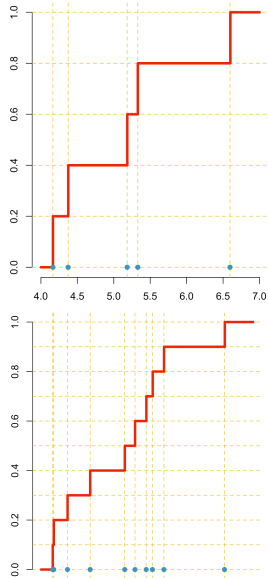
The empirical distribution function is

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \leq x), \ x \in \mathbb{R}$$

Glivenko-Cantelli theorem: $widehat F_n \to F$ when $n \to \infty$, or more precisely

$$\|\widehat{F}_n - F\|_\infty = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \longrightarrow 0$$

when $n \to \infty$.



Boostrap and Resampling

Drawing according to $\widehat{F}_n^{-1}$ means resampling in $\{x_1, \cdots, x_n\}$ with probability $1/n$ (or with replacement).

```
> x
[1] 4.164 4.374 5.184 5.330 6.595
> Qemp = Vectorize(function(u) sort(U)[ceiling(length(x)*u)])
> Qemp(U)
  [1] 4.37 5.33 5.18 6.60 6.60 4.16 5.18 6.60 5.18 5.18 6.60 5.18 5.33
 [14] 5.18 4.16 6.60 4.37 4.16 4.37 6.60 6.60 5.33 5.33 6.60 5.33 5.33
 [27] 5.18 5.18 4.37 4.16 6.60 6.60 5.33 5.33 4.16 5.18 5.33 4.37 4.37
 [40] 4.37 4.16 5.18 5.18 4.37 4.16 4.16 4.37 5.18 4.37 6.60 4.16 5.18
 [53] 5.33 4.16 5.18 4.37 4.16 5.33 6.60 4.37 5.33 4.16 4.37 4.37 6.60
 [66] 5.18 6.60 6.60 5.33 5.18 5.33 5.33 5.33 4.16 5.18 4.37 4.37 5.33
 [79] 4.37 4.16 4.37 5.33 5.18 5.33 4.16 5.18 6.60 6.60 6.60 4.16 4.16
 [92] 5.33 4.37 5.33 4.37 4.16 5.33 4.16 5.18 5.18
```

see also Davison and Hinkley (1997)

```
> sample(x, size = 100, replace = TRUE)
```


(Linear) Model & Bootstrap (1)

Dataset $\mathcal{D}_n = \{\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i)\}$, $i = 1, \cdots, n$.

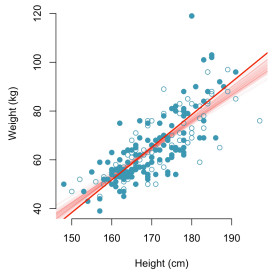
1) approach by **paired sampling**, i.e. $\{\boldsymbol{z}_1^\star, \cdots, \boldsymbol{z}_n^\star\}$.

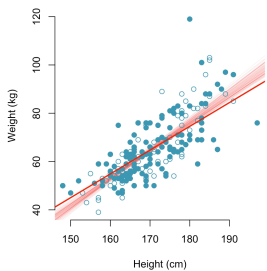
- draw $\{i_1^{(b)}, \cdots, i_n^{(b)}\}$ randomly (with replacement) in $\{1, 2, \cdots, n\}$
- consider database $(\boldsymbol{x}_i^{(b)}, y_i^{(b)}) = (\boldsymbol{x}_{i^{(b)}}, y_{i^{(b)}})$'s, and estimate a parametric model
- let $\widehat{\beta}^{(b)}$ denote the estimator of β and $\widehat{y}_{n+1}^{(b)}$ some prediction (associated with $\boldsymbol{x}_{n+1}$)

(loop over $b = 1, \cdots, B$)

(Linear) Model & Bootstrap (1)

```
> Davis = read.table("http://freakonometrics.free.fr
    /Davis.txt")
> BETA = matrix(NA,1000,2)
> for(s in 1:1000){
+  idx = sample(1:nrow(Davis),nrow(Davis),
+        replace=TRUE)
+  reg_sim = lm(weight~height, data=Davis[idx,])
+  BETA[s,] = reg_sim$coefficients
+ }
> hist(BETA[,1])
> hist(BETA[,2])
```

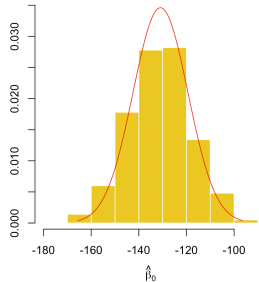


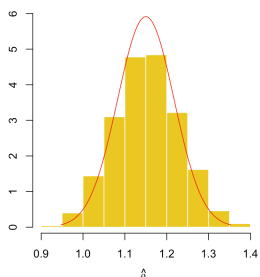


(Linear) Model & Bootstrap (1)

or use the `boot` package

```
> library(boot)
> coef = function(formula, data, indices) {
+  d = data[indices,]
+  fit = lm(formula, data=d)
+  return(coef(fit))
+ }
> results = boot(data=Davis, statistic=coef, R=1000,
+           formula=weight~height)
> plot(results, index=1)
> plot(results, index=2)
```





(Linear) Model & Bootstrap (2)

2) use **model-based resampling** (or semiparametric residual bootstrap)

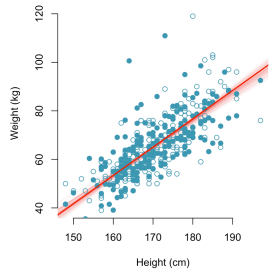
- fit a (linear) model, get predictions $\widehat{y}_i$ and residuals $\widehat{\varepsilon}_i$
- draw $\widehat{\varepsilon}_1^{(b)}, \cdots, \widehat{\varepsilon}_n^{(b)}$ from the original sample $\{\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \cdots, \widehat{\varepsilon}_n\}$
- set $y_i^{(b)} = \widehat{y}_i + \widehat{\varepsilon}_i^{(b)}$ (from the fitted model)
- from database $(\boldsymbol{x}, y^{(b)}) = (\boldsymbol{x}_i, y_i^{(b)})$, fit a (linear) parametric model
- set $\widehat{\boldsymbol{\beta}}^{(b)}$ the estimator of $\boldsymbol{\beta}$, and $\widehat{y}_{n+1}^{(b)}$ the prediction (associated to $\boldsymbol{x}_{n+1}$)

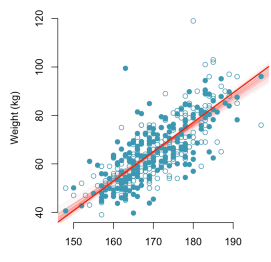
Note for a simple linear regression, $\widehat{\beta}_1^{(b)} = \dfrac{\sum[x_i - \overline{x}] \cdot y_i^{(b)}}{\sum[x_i - \overline{x}]^2} = \widehat{\beta}_1 + \dfrac{\sum[x_i - \overline{x}] \cdot \widehat{\varepsilon}_i^{(b)}}{\sum[x_i - \overline{x}]^2}$ i.e.,

$\mathbb{E}[\widehat{\beta}_1^{(b)}] = \widehat{\beta}_1$, while $\text{Var}[\widehat{\beta}_1^{(b)}] = \dfrac{\sum[x_i - \overline{x}]^2 \cdot \text{Var}[\widehat{\varepsilon}_i^{(b)}]}{\left(\sum[x_i - \overline{x}]^2\right)^2} \sim \dfrac{\sigma^2}{\sum[x_i - \overline{x}]^2}$.

(Linear) Model & Bootstrap (2)

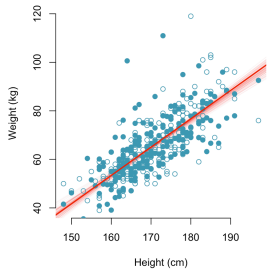
```
> BETA = matrix(NA,1000,2)
> reg = lm(weight~height, data=Davis)
> epsilon = residuals(reg)
> for(s in 1:1000){
+   eps = sample(epsilon,nrow(Davis),replace=TRUE)
+   Davis_s = data.frame(height = Davis$height,
+                        weight = predict(reg)+eps)
+   reg_sim = lm(weight~height, data=Davis_s)
+   BETA[s,] = reg_sim$coefficients
+ }
> hist(BETA[,1])
> hist(BETA[,2])
```

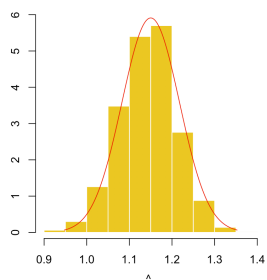




(Linear) Model & Bootstrap (2)

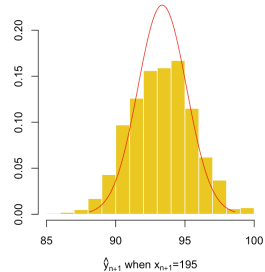
```
> BETA = matrix(NA,1000,2)
> reg = lm(weight~height, data=Davis)
> epsilon = residuals(reg)
> for(s in 1:1000){
+   eps = sample(epsilon,nrow(Davis),replace=TRUE)
+   Davis_s = data.frame(height = Davis$height,
+                        weight = predict(reg)+eps)
+   reg_sim = lm(weight~height, data=Davis_s)
+   BETA[s,] = reg_sim$coefficients
+ }
> hist(BETA[,1])
> hist(BETA[,2])
```

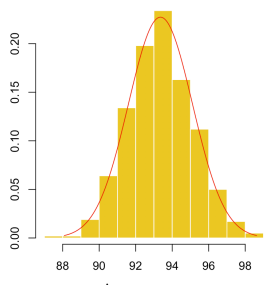




(Linear) Model & Bootstrap (1/2)

```
> PRED = matrix(NA,1000,2)
> nwDavis = data.frame(height = 195)
> for(s in 1:1000){
+   idx = sample(1:n,n,replace=TRUE)
+   reg_sim = lm(weight~height, data=Davis[idx,])
+   PRED[s,1] = predict(reg_sim, newdata=nwDavis)
+   eps = sample(epsilon,nrow(Davis),replace=TRUE)
+   Davis_s = data.frame(height = Davis$height,
+                        weight = predict(reg)+eps)
+   reg_sim = lm(weight~height, data=Davis_s)
+   PRED[s,2] = predict(reg_sim, newdata=nwDavis)
+ }
```

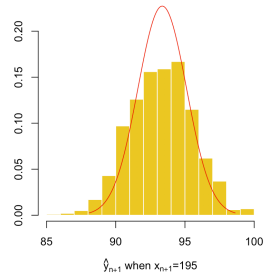


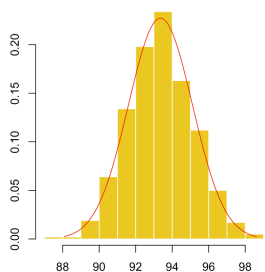


(Linear) Model & Bootstrap (2/2)

```
> apply(PRED,2,function(x) quantile(x,.025))
[1] 89.04203 89.63030
> apply(PRED,2,function(x) quantile(x,.975))
[1] 97.60345 97.02423
> predict(lm(weight~height, data=Davis),
+         newdata=nwDavis,interval="confidence",se.
    fit = TRUE)
$fit
       fit      lwr      upr
1 93.35749 89.89571 96.81927

$se.fit
[1] 1.755451
```





Introduction to Metrics and Similarities

- **Metrics** and **distances** are used to quantify the similarity or dissimilarity between data points.
- They are critical in many machine learning algorithms such as:
  - Clustering (K-means)
  - Classification (K-nearest neighbors, Support Vector Machines)
  - Dimensionality reduction (PCA)
- Commonly used metrics:
  - Euclidean distance
  - Cosine similarity
  - Manhattan distance
  - Mahalanobis distance
- Understanding the properties of different metrics helps choose the right one for a given problem.

Introduction to Metrics and Similarities

- In machine learning, **loss functions** are used to quantify how well a model's predictions match the actual values.
- Loss functions are often based on some notion of "distance" between the predicted and true values, especially in regression and classification problems.
- Key idea: The loss function can be seen as a distance metric that we minimize during model optimization.
- The relationship between distances and losses is crucial in tasks like:
  - Regression: Measuring the difference between predicted and actual values.
  - Classification: Measuring the difference between predicted class probabilities and true labels.
  - Generative Models: Comparing generated data distributions to real data distributions.

Introduction to Metrics and Similarities

- **Squared Euclidean Loss (L2 Loss)**: Measures the squared Euclidean distance between predicted and true values.

$$L_2(\hat{y}, y) = \sum_{i=1}^{n} \ell_2(\hat{y}_i, y_i) \text{ where } \ell_2(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$$

- **Mean Absolute Error (MAE, L1 Loss)**: Measures the absolute distance between predicted and actual values.

$$L_1(\hat{y}, y) = \sum_{i=1}^{n} \ell_1(\hat{y}_i, y_i) \text{ where } \ell_1(\hat{y}_i, y_i) = |\hat{y}_i - y_i|$$

- **Cross-Entropy Loss**: A loss function based on Kullback-Leibler divergence for classification tasks.

$$L_{\text{CE}}(p, q) = -\sum_{i} p_i \log(q_i)$$

where p is the true distribution and q is the predicted distribution.

Introduction to Metrics and Similarities

- **Hinge Loss**: Used for Support Vector Machines (SVM), measures how far the predictions are from the decision boundary.

$$\ell_{\text{hinge}}(y, \hat{y}) = \max(0, 1 - y \cdot \hat{y})$$

  where y is the true label and $\hat{y}$ is the predicted value.

- Many common loss functions can be interpreted as a distance metric:
  - **L2 loss** (Squared Euclidean loss) is a direct application of Euclidean distance between predicted and true values.
  - **L1 loss** (Mean Absolute Error) corresponds to Manhattan distance, measuring the absolute differences between predictions and true values.
  - **Cross-Entropy loss** can be seen as the Kullback-Leibler (KL) divergence between the true and predicted distributions.

- **Optimization Perspective:**

- Minimizing a loss function is equivalent to minimizing a certain distance between the predicted and true values or distributions.
- In most cases, we aim to minimize the distance, i.e., minimize the loss, to improve model performance.

- **Connection with Regularization:**
  - Regularization terms, such as L2 regularization (Ridge), are often added to the loss function to prevent overfitting, making the model less sensitive to small fluctuations in the data.

But one can also consider "distance between distributions"

- In many machine learning and statistical problems, we are interested in comparing probability distributions rather than just individual data points.
- This is crucial in tasks such as:
  - Probabilistic forecasting and predictive modeling.

- Generative models (e.g., GANs, VAEs) where we compare generated vs true distributions.
- Uncertainty quantification and model evaluation.

- Examples:
  - Comparing the predicted distribution of a model with the true distribution of the data.
  - Measuring the discrepancy between two probability distributions, e.g., in model calibration.
- Some key metrics used to measure the "distance" between distributions are:
  - Kullback-Leibler (KL) Divergence
  - Total Variation Distance
  - Wasserstein Distance (Earth Mover's Distance)

Introduction to Metrics and Similarities

- **KL Divergence**: Measures the "distance" between two probability distributions P and Q.

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

- The KL divergence is not a true distance because it is asymmetric (i.e., $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$).
  - It measures how much information is lost when Q is used to approximate P.
- Commonly used in:
  - Variational inference.
  - Machine learning models where we need to measure how much a learned distribution differs from the true one.
- Intuition:
  - If $P = Q$, then $D_{KL}(P \parallel Q) = 0$.
  - If P and Q are very different, the divergence becomes large.

Introduction to Metrics and Similarities

- **Wasserstein Distance** (also known as Earth Mover's Distance) measures the minimum amount of work required to transform one distribution into another.
- [•]The first Wasserstein distance between two probability distributions P and Q is defined as:

$$W_1(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]$$

  where $\Gamma(P, Q)$ is the set of all couplings between P and Q, and $\|x - y\|$ is the distance between points x and y.
- Unlike KL Divergence, the Wasserstein distance is symmetric and satisfies the triangle inequality, making it a true distance metric.
- Widely used in:
  - Generative Adversarial Networks (GANs), especially Wasserstein GANs (WGANs).

  - Comparing distributions with complex structures or in high-dimensional spaces.

- Intuition:
  - It measures the "cost" of transforming one distribution into another, which is especially useful when comparing distributions over different domains or supports.

Vectors

$$\vec{\boldsymbol{u}} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \boldsymbol{u} \in \mathbb{R}^n$$

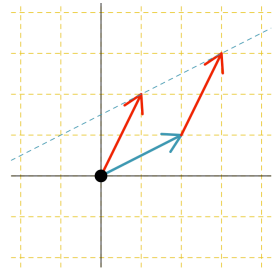
$$\boldsymbol{u} = (u_1 \cdots u_n)^\top$$

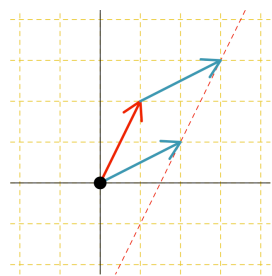
Example :
$\vec{\boldsymbol{u}}_j = (178, 185, 162, 170, \cdots, 169)^\top \in \mathbb{R}^{200}$
(vector of variables, $n = 200 =$ sample size)

$\vec{\boldsymbol{v}}_i = (178, 74, 1, 0, 14321, 1)^\top \in \mathbb{R}^p$
(vector of individuals, $p = 6 =$ features)

```
> u = [178, 185, 162, 170]
```

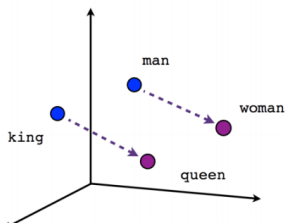
```
> u = c(178, 185, 162, 170)
```



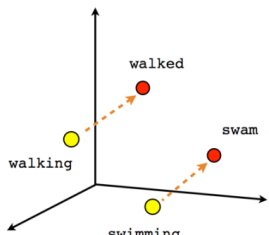


Word2Vec

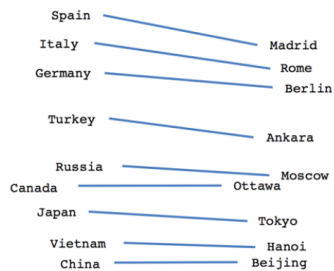
Word2vec has been one of the most popular models used to create word embeddings (words are represented in a large dimensional space... that can be projected into 2 or 3)



Male-Female



Verb tense



Country-Capital

(source https://www.tensorflow.org)

Scalar Product & Euclidean Distance

$\vec{u} + \vec{v} = (u_1 + v_1, \cdots, u_n + v_n)^\top \in \mathbb{R}^n$
Note that $\vec{u} + \vec{v} = \vec{v} + \vec{u}$

Given $\vec{u}$ and $\vec{v}$,

$$\langle \vec{u}, \vec{v} \rangle = \boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{u}^\top \boldsymbol{v} = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + \cdots + u_n v_n$$

Example Let $\mathbf{1} = (1, 1, \cdots, 1)^\top$, then $\overline{u} = \frac{1}{n} \langle \vec{\mathbf{1}}, \vec{\boldsymbol{u}} \rangle = \frac{1}{n} \mathbf{1}^\top \boldsymbol{u}$
Example $\boldsymbol{p}$ a probability vector, $\boldsymbol{x}$ a vector of outcome, $\mathbb{E}(X) = \langle \vec{\boldsymbol{p}}, \vec{\boldsymbol{x}} \rangle$

$\vec{u} \perp \vec{v}$ if and only if $\langle \vec{u}, \vec{v} \rangle = 0$

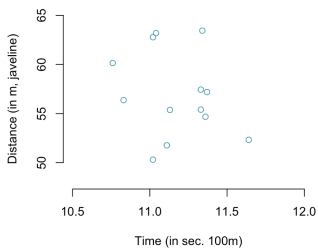
$$\|\vec{\boldsymbol{u}}\|^2 = \langle \vec{\boldsymbol{u}}, \vec{\boldsymbol{u}} \rangle = \boldsymbol{u} \cdot \boldsymbol{u} = \sum_{i=1}^{n} u_i^2 = u_1^2 + \cdots + u_n^2$$

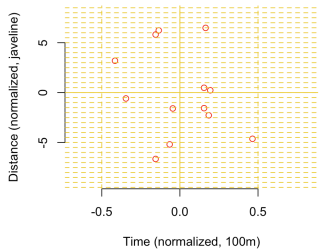
Mahalanobis Distance

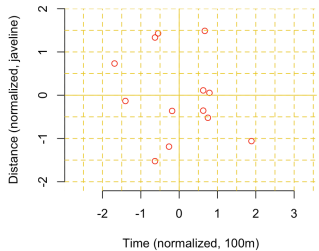
The standard Euclidean distance is $\|\vec{\boldsymbol{u}}\|_E^2 = \langle \vec{\boldsymbol{u}}, \vec{\boldsymbol{u}} \rangle = \boldsymbol{u}^\top \boldsymbol{u}$
Given some diagonal positive matrix D, and $\boldsymbol{\mu} \in \mathbb{R}^n$,

$$\|\vec{\boldsymbol{u}}\|_D^2 = (\boldsymbol{u} - \boldsymbol{\mu})^\top D^{-1} (\boldsymbol{u} - \boldsymbol{\mu}) = \|\tilde{\boldsymbol{u}}\|_E^2 \text{ where } \tilde{\boldsymbol{u}}_i = \frac{\boldsymbol{u}_i - \boldsymbol{\mu}_i}{\sqrt{D_{i,i}}}$$

(popular in statistics, when $\boldsymbol{u}_i$ is the observation of individual i)







Euclidean Distance

Define $d(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|_2 = \sqrt{(\boldsymbol{u} - \boldsymbol{v})^\top (\boldsymbol{u} - \boldsymbol{v})}$
Set of points $\boldsymbol{u}$ such that $d(\boldsymbol{0}, \boldsymbol{v}) = \|\boldsymbol{u}\|_2 = 1$

- circle (or sphere in higher dimension)

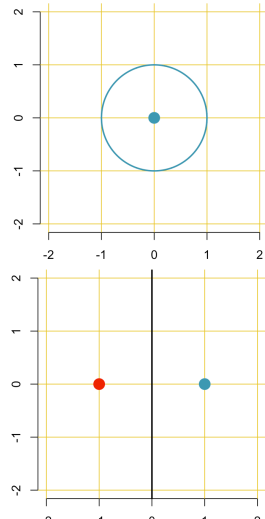
Set of points $\boldsymbol{x}$ such that $d(\boldsymbol{x}, \boldsymbol{u}) = d(\boldsymbol{x}, \boldsymbol{v})$?

- straight line (or plane in higher dimension)

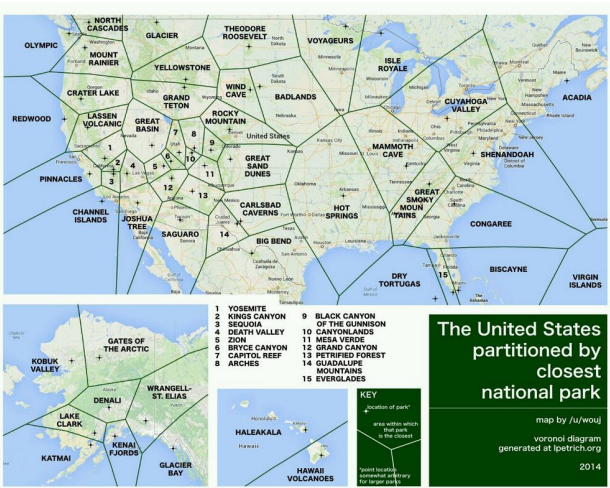
(orthogonal to $\boldsymbol{u} - \boldsymbol{v}$)

Let $\boldsymbol{X}$ be a $n \times k$ matrix, n individuals, k columns

- $\boldsymbol{X}^\top \boldsymbol{X}$ is $k \times k$, it is a "**covariance**" (related) matrix also called "Gram matrix"
- $\boldsymbol{X}\boldsymbol{X}^\top$ is $n \times n$, it is a "**similarity**" matrix

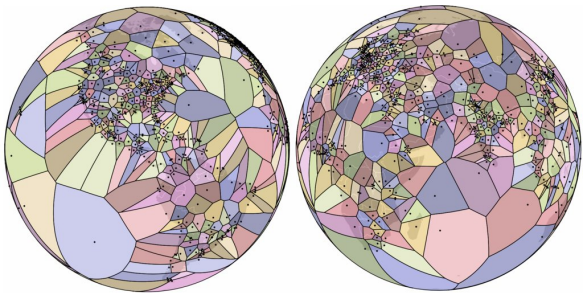


Euclidean Distance

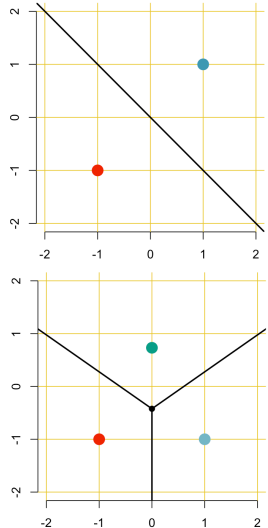


Euclidean Distance

Given a set of points $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots \boldsymbol{u}_n$
What is the set of points $\boldsymbol{x}$ the closest to $\boldsymbol{u}_i$? (see also Delaunay triangulation)



(intersections are straight linears or planes)

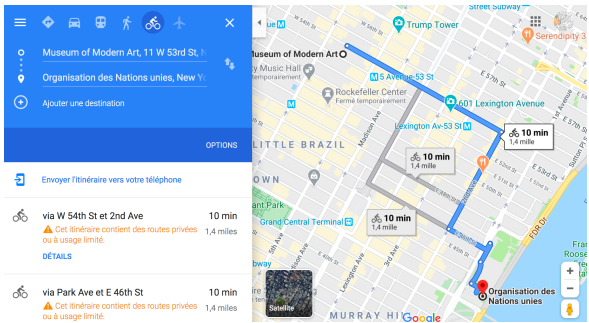


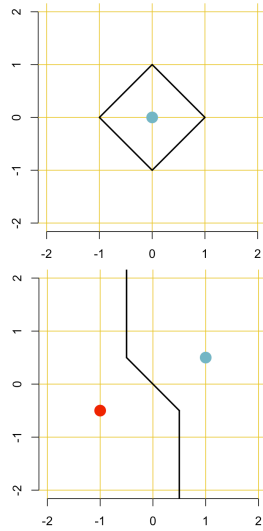
Manhattan (ℓ_1) Distance

$\|\boldsymbol{u}\|_1 = |u_1| + |u_2|$ is also a norm (Manhattan)

Set of points $\boldsymbol{u}$ such that $d(\boldsymbol{0}, \boldsymbol{v}) = \|\boldsymbol{u}\| = 1$

- square (or cube in higher dimension)





Set of points $\boldsymbol{x}$ such that $d(\boldsymbol{x}, \boldsymbol{u}) = d(\boldsymbol{x}, \boldsymbol{v})$?

- portions of straight lines (or planes)

Distance

...

A distance measure is an objective score that summarizes the relative difference between two objects in a problem domain.

Definition 1.26: Distance

A **distance** or **metric** over $\mathcal{X} \in \mathbb{R}^p$ is a map $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the following properties

- coincidence: $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,
- symmetry $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{d}(\boldsymbol{y}, \boldsymbol{x})$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,
- triangle inequality $\mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) \leq \mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) + \mathrm{d}(\boldsymbol{y}, \boldsymbol{z})$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

- non-negativity: $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,
- generalized triangle inequality $|\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) - \mathrm{d}(\boldsymbol{y}, \boldsymbol{z})| \leq \mathrm{d}(\boldsymbol{x}, \boldsymbol{y})$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

Distance

Definition 1.27: Manhattan-Taxicab distance (and ℓ_1-norm)

Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, $\mathrm{d}_1(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_1 = \sum_{i=1}^{p} |x_i - y_i|$.

Definition 1.28: Euclidean distance (and ℓ_2-norm)

Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, $\mathrm{d}_2(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y})} = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$.

Distance

Definition 1.29: Mahalanobis distance

Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ and M a positive semidefinite matrice of dimension $p \times p$,

$$\mathrm{d}_M(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^\top M(\boldsymbol{x}-\boldsymbol{y})}$$

Mahalanobis distances also satisfy

- homogeneousness: $\mathrm{d}(a\boldsymbol{x}, a\boldsymbol{y}) = |a| \cdot \mathrm{d}(\boldsymbol{x}, \boldsymbol{y})$, for every $a \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,
- translation invariance $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{d}(\boldsymbol{y}+\boldsymbol{z}, \boldsymbol{y}+\boldsymbol{z})$ for every $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{X}$,

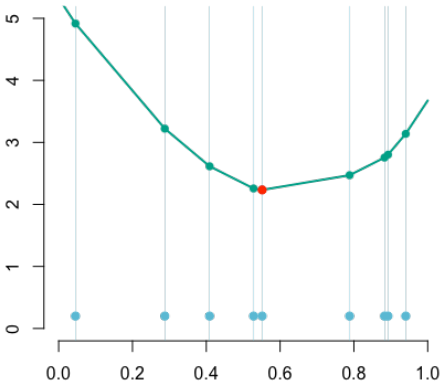
If $M = L^\top L$,

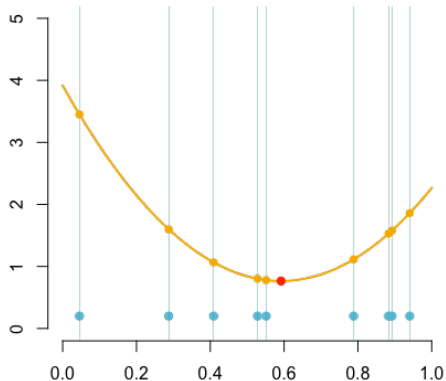
$$\mathrm{d}_M(\boldsymbol{x}, \boldsymbol{y})^2 = (\boldsymbol{x}-\boldsymbol{y})^\top M(\boldsymbol{x}-\boldsymbol{y}) = [L(\boldsymbol{x}-\boldsymbol{y})]^\top [L(\boldsymbol{x}-\boldsymbol{y})] = \|L(\boldsymbol{x}-\boldsymbol{y})\|_2^2$$

where $\|\cdot\|_2$ is the standard Euclidean norm.

Loss

Given $\boldsymbol{y} = (y_1, \cdots, y_n)$, find $x^\star \in \underset{x \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} |x - y_i| \right\}$ or $x^\star \in \underset{x \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (x - y_i)^2 \right\}$





Loss

The τth quantile ($\tau \in (0,1)$.) of Y is given by $q_\tau = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$,

$$q_Y(\tau) = \underset{q \in \mathbb{R}}{\operatorname{argmin}} \left\{ (\tau - 1) \int_{-\infty}^{q} (q - y) dF_Y(y) + \tau \int_{q}^{\infty} (y - q) dF_Y(y) \right\}.$$

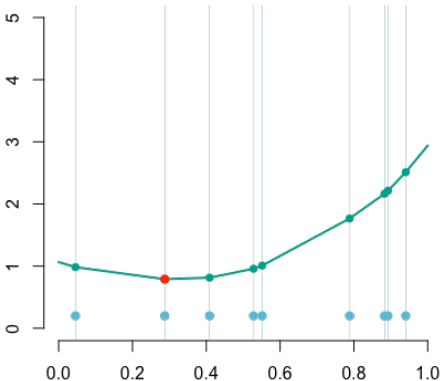
$$\hat{q}_\tau = \underset{q \in \mathbb{R}}{\operatorname{argmin}} \left\{ (\tau - 1) \sum_{y_i < q} (q - y_i) + \tau \sum_{y_i \geq q} (y_i - q) \right\},$$

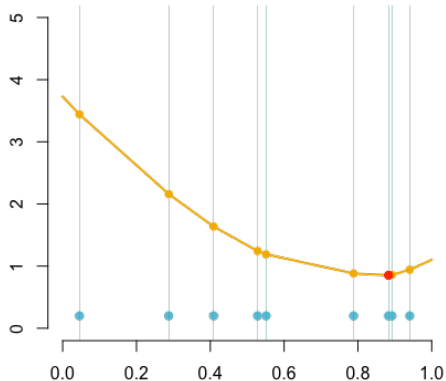
also coined pinball loss, which is asymmetric,

$$\ell_\tau(y, \hat{y}) = \begin{cases} \tau \cdot (y - \hat{y}) & \text{if } y \geq \hat{y} \\ (1 - \tau) \cdot (\hat{y} - y) & \text{if } y < \hat{y} \end{cases}$$

Loss

Given $\boldsymbol{y} = (y_1, \cdots, y_n)$, find $x_\tau^\star$ such that $x_\tau^\star \in \underset{x \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell_\tau(y_i, x) \right\}$





with $\tau = 0.2$ and $\tau = 0.7$ respectively.

Loss

```
> library(cNORM)
> set.seed(1234)
> n = 10
> x = runif(n)
> x
[1] 0.114 0.622 0.609 0.623 0.861 0.640 0.009 0.233 0.666 0.514
> tau = .7
> quantile(x,tau)
[1] 0.6284588
```

The linear program is

$$\min_{q^+,q^-,\boldsymbol{a},\boldsymbol{b}} \left\{ \sum_{i=1}^{n} \tau a_i + (1-\tau) b_i \right\}, \text{ where } y_i = q^+ - q^- + a_i - b_i, \ \forall i \in \{1, \cdots, n\}$$

and where $a_i, b_i \geq 0$ as well as $q^+, q^- \geq 0$.

Loss

Set $\boldsymbol{z} = (q^+, q^-, \boldsymbol{a}, \boldsymbol{b})^\top \in \mathbb{R}_+^{2n+2}$,

$$\min_{\boldsymbol{z}} \left\{ \boldsymbol{c}^\top \boldsymbol{z} \right\}, \text{ where } \boldsymbol{A}\boldsymbol{z} = \boldsymbol{b} \text{ and } \boldsymbol{z} \geq 0,$$

where $\boldsymbol{A}$ is $n \times (2n+2)$,

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{1}_n & -\mathbf{1}_n & \mathbb{I}_n & -\mathbb{I}_n \end{bmatrix} = \left[\begin{array}{c:c:ccc:c:ccc} 1 & -1 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ 1 & -1 & 0 & 1 & \cdots & 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & -1 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & -1 \end{array}\right]$$

$\boldsymbol{b} = \boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{c} = \begin{bmatrix} 0 & 0 & \tau\mathbf{1}_n & (1-\tau)\mathbf{1}_n \end{bmatrix}^\top \in \mathbb{R}^{2n+2}$

Loss

$$\min_{\boldsymbol{z}} \left\{ \boldsymbol{c}^\top \boldsymbol{z} \right\}, \text{ where } \boldsymbol{A}\boldsymbol{z} = \boldsymbol{b} \text{ and } \boldsymbol{z} = 0,$$

```
> library(lpSolve)
> ones = rep(1,n)
> A = cbind(ones, -ones, diag(n), -diag(n))
> b = x
> c = c(rep(0,2), tau*rep(1,n),(1-tau)*rep(1,n))
> equal_type = rep("=", n)
> lp("min", c,A,equal_type,b)$solution[1]
[1] 0.6403106
```

Deviance loss

These different variance functions can be translated to deviance loss functions by selecting the corresponding distribution within the exponential dispersion family.

- the Gaussian case translates to the square loss $\ell(y,\widehat{y}) = (y-\widehat{y})^2$
- the Poisson case translates to the Poisson deviance loss

$$\ell(y,\widehat{y}) = 2\left(\widehat{y} - y - y\log\frac{\widehat{y}}{y}\right)$$

- the Gamma case translates to the Gamma deviance loss

$$\ell(y,\widehat{y}) = 2\left(\frac{y-\widehat{y}}{\widehat{y}} + y\log\frac{\widehat{y}}{y}\right)$$

- the inverse Gaussian case translates to the inverse Gaussian deviance loss

$$\ell(y,\widehat{y}) = \frac{(y-\widehat{y})^2}{\widehat{y}^2 y}$$

Deviance loss

If $Y \sim \mathcal{P}(\lambda)$, consider

$$\widehat{y} \mapsto \mathbb{E}\big[\ell(Y,\widehat{y})\big] = 2\sum_{y=0}^{\infty}\left(\widehat{y} - y - y\log\frac{\widehat{y}}{y}\right)\frac{e^{-\lambda}\lambda^y}{y!} = 2\widehat{y} - 2\lambda - 2\lambda\log(\widehat{y}) + 2\mathbb{E}\big[Y\log Y\big]$$

thus, $\min\big\{\mathbb{E}\big[\ell(Y,\widehat{y})\big]\big\}$ is attained when $\widehat{y} = \lambda = \mathbb{E}[Y]$.

A divergence is a magnitude to measure the closeness between certain objects in a set.

Definition 1.30: Divergence

D is a divergence if it satisfies

- non negativity $\mathrm{D}(\boldsymbol{x}\|\boldsymbol{y}) \geq 0$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,
- coincidence $\mathrm{D}(\boldsymbol{x}\|\boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$.

Deviance loss

Definition 1.31: Divergence on $\mathbb{R}^n$

A divergence D on a set $E \subset \mathbb{R}^n$ is a function $E \times E \to \mathbb{R}_+$ such that

- D is separable, $\forall (\boldsymbol{x}, \boldsymbol{y}) \in E^2,\ D(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$,
- D admits development $\forall (\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\epsilon}) \in E^2,\ D(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\epsilon}) = \frac{1}{2} \sum A_{i,j}(\epsilon) \epsilon_i \epsilon_j + O(|\epsilon|^3)$, where $A(\epsilon)$ is definite positive.

A loss ℓ is consistent for mean estimation w.r.t. $\mathcal{M}$ if

$$\mathbb{E}[\ell(Y, \mathbb{E}[Y|\boldsymbol{X}])] \leq \mathbb{E}[\ell(Y, m(\boldsymbol{X}))] \text{ for all } m \in \mathcal{M}$$

Strictly consistent loss functions for mean estimation are the Bregman divergences, Savage (1971).

Deviance loss

Definition 1.32: Bregman Divergence, Bregman (1967)

Let $\psi : \mathcal{X} \to \mathbb{R}$ be a strictly convex function that is continuously differentiable. Then the Bregman divergence $D_\psi(\boldsymbol{x}, \boldsymbol{y})$ is defined as

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle \nabla\psi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.$$

All of these deviance loss functions are strictly consistent for mean estimation, but the one with the correct conditional variance behavior for the response Y, given $\boldsymbol{X}$, has the best finite sample properties (on average).

If $\psi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ (strictly convex), then $D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$.

(recall that $\nabla\|\boldsymbol{x}\|^2 = 2\boldsymbol{x}$)

Deviance loss

Proposition 1.17: Bregman Divergence

Let $\psi : \mathcal{X} \to \mathbb{R}$ be a strictly convex function that is continuously differentiable. Then Bregman divergence $D_\psi(\boldsymbol{x}, \boldsymbol{y})$ is

- strictly convex in $\boldsymbol{x}$,
- (generally) non-convex in $\boldsymbol{y}$,
- non-negative $D_\psi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$,
- separable, $D_\psi(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$,
- (generally) asymmetric.

Deviance loss

If $\mathcal{X} = \mathbb{R}^n$, and $\psi(\boldsymbol{x}) = \frac{1}{2}\sum_{ij} A_{ij}x_i x_j = \frac{1}{2}\boldsymbol{x}^\top A(\boldsymbol{x}$ for some $n \times n$ matrix A definite positive, then

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\sum_{ij} A_{ij}(x_i - y_i)(x_j - y_j) = (\boldsymbol{x} - \boldsymbol{y})^\top A(\boldsymbol{x} - \boldsymbol{y})$$

(see Mahalanobis distance).

If $\mathcal{X} = \mathbb{R}^n$, and $\psi(\boldsymbol{x}) = -\sum_i \log(x_i)$ then

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \frac{x_i}{y_i} - \log\frac{x_i}{y_i} - 1$$

See Banerjee et al. (2005) for more examples.

Deviance loss

Kullback–Leibler (KL) divergences and deviance loss functions are Bregman divergences.

Definition 1.33: Projection, Bregman (1967), Bauschke et al. (1997)

Given a strictly convex function continuously differentiable ψ and the associated Bregman divergence D_ψ, a closed closed convex $K \subset \mathcal{X}$ and a point $\boldsymbol{x} \in \mathcal{X}$. The Bregman projection of $\boldsymbol{x}$ onto K is

$$\boldsymbol{x}^\star = \underset{\boldsymbol{y} \in K}{\operatorname{argmin}} \{ D_\psi(\boldsymbol{x}, \boldsymbol{y}) \}$$

Deviance loss

Definition 1.34: Entropy or Kullback-Liebler Divergence

Kullback-Liebler divergence, between distribution f and g is

$$\mathrm{KL}(f\|g) = \mathbb{E}_f\left[\log \frac{f(X)}{g(X)}\right] = \int_{-\infty}^{\infty} \log\left(\frac{f(x)}{g(x)}\right) f(x)\mathrm{d}x,$$

when $f \ll g$, i.e., $g(x) = 0$ implies $f(x) = 0$.

$$D_{\mathrm{KL}}(\mathcal{B}(p)\|\mathcal{B}(q)) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

$$D_{\mathrm{KL}}(\mathcal{B}(n,p)\|\mathcal{B}(n,q)) = np\log\frac{p}{q} + n(1-p)\log\frac{1-p}{1-q} = nD_{\mathrm{KL}}(\mathcal{B}(p)\|\mathcal{B}(q))$$

$$D_{\mathrm{KL}}(\mathcal{U}([a_1,b_1])\|\mathcal{U}([a_2,b_2])) = \log\frac{b_2-a_2}{b_1-a_1} \text{ if } [a_1,b_1]\subset[a_2,b_2]$$

Deviance loss

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1,\sigma_1^2)\|\mathcal{N}(\mu_2,\sigma_2^2))=\frac{1}{2}\left[\frac{(\mu_1-\mu_2)^2}{\sigma_2^2}+\frac{\sigma_1^2}{\sigma_2^2}-\log\frac{\sigma_1^2}{\sigma_2^2}-1\right]$$

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\|\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))=\frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)-\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}-k\right]$$

Proposition 1.18: Divergence for Gaussian vectors

Consider two Gaussian distributions, then $D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\|\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))$ is

$$\frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)-\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}-k\right]$$

Deviance loss

Proposition 1.19: Gibbs' inequality

$D_{\mathrm{KL}}(p\|q)$ is positive and separable, i.e. $D_{\mathrm{KL}}(p\|q) \geq 0$ and $D_{\mathrm{KL}}(p\|q) = 0$ if and only if $p = q$.

Proof: $\sum_{x \in I} p(x) \log \frac{p(x)}{q(x)} \geq 0$ where I is the set of all x for which $p(x) > 0$. Recall that $\log x \leq x - 1$ (with equality only when $x = 1$), thus $\log(1/x) \geq 1 - x$, and

$$\sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} \geq \sum_{x \in I} p(x) \left(1 - \frac{q(x)}{p(x)}\right) = \underbrace{\sum_{x \in I} p(x)}_{=1} - \underbrace{\sum_{x \in I} q(x)}_{\leq 1} \geq 0.$$

Deviance loss

Proposition 1.20: Additivity for independence distributions

$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = D_{\mathrm{KL}}(p_x\|q_x) + D_{\mathrm{KL}}(p_y\|q_y)$ if $\boldsymbol{p}(x,y) = p_x(x)p_y(y)$ and $\boldsymbol{q}(x,y) = q_x(x)q_y(y)$.

Proof By definition

$$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \int_{\mathcal{X}}\int_{\mathcal{Y}} p(x,y)\cdot\log\frac{p(x,y)}{q(x,y)}\,\mathrm{d}y\,\mathrm{d}x\,.$$

and since $\boldsymbol{p}(x,y) = p_x(x)p_y(y)$ and $\boldsymbol{q}(x,y) = q_x(x)q_y(y)$,

$$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \int_{\mathcal{X}}\int_{\mathcal{Y}} p_1(x)\,p_2(y)\cdot\log\frac{p_1(x)\,p_2(y)}{q_1(x)\,q_2(y)}\,\mathrm{d}y\,\mathrm{d}x\,.$$

Deviance loss

Definition 1.35: Symetric relative entropy or Jeffrey divergence

Jeffrey divergence is some sort of "symmetric Kullback-Liebler" divergence

$$\mathrm{JF}(f,g) = \mathrm{KL}(f\|g) + \mathrm{KL}(g\|f).$$

Deviance loss

Definition 1.36: Divergence based inference

Consider some parametric family $\mathcal{Q} = \{f_\theta, \theta \in \Theta\}$. Given a divergence D, we want to find

$$\theta^\star = \underset{\theta\in\Theta}{\text{argmin}}\{D(f, f_\theta)\}$$

unknown f ; $f_\theta \in \mathcal{Q}$

or its empirical version

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\text{argmin}}\{D(\widehat{f}_n, f_\theta)\}$$

estimated $\widehat{f}_n$

Deviance loss

Definition 1.37: Hellinger distance, Hellinger (1909)

For two discrete distributions f and g, Hellinger distance is

$$d_{\mathrm{H}}(f,g)^2 = \frac{1}{2}\sum_i \left(\sqrt{f(i)} - \sqrt{g(i)}\right)^2 = 1 - \sum_i \sqrt{f(i)g(i)} \in [0,1],$$

and for absolutely continuous distributions, if f and g are densities,

$$d_{\mathrm{H}}(f,g)^2 = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \mathrm{d}x \text{ or } \frac{1}{2}\int_{\mathbb{R}^k} \left(\sqrt{f(\boldsymbol{x})} - \sqrt{g(\boldsymbol{x})}\right)^2 \mathrm{d}\boldsymbol{x}$$

See Pardo (2018).

Deviance loss

Proposition 1.21: Distance between Beta variables

Consider two Beta distribution, then $d_{\mathrm{H}}\left(\mathcal{B}(a_1, b_1), \mathcal{B}(a_2, b_2)\right)^2$ is

$$1 - \frac{1}{\sqrt{\mathrm{B}(a_1, b_1)\mathrm{B}(a_2, b_2)}} \mathrm{B}\left(\frac{a_1 + a_2}{2}, \frac{b_1 + b_2}{2}\right)$$

Proof

$$1 - \int_0^1 \sqrt{f_1(t) f_2(t)}\mathrm{d}t = 1 - \frac{1}{\sqrt{\mathrm{B}(a_1, b_1)\mathrm{B}(a_2, b_2)}} \int_0^1 t^{(a_1+a_2)/2-1}(1-t)^{(b_1+b_2)/2-1}\mathrm{d}t,$$

then use $\mathrm{B}(a, b) = \displaystyle\int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$.

Deviance loss

Proposition 1.22: Distance between Gaussian vectors

Consider two Gaussian distributions, then $d_{\mathrm{H}}\left(\mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right), \mathcal{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)\right)^2$ is

$$2 - 2\frac{\left|\boldsymbol{\Sigma}_1\right|^{\frac{1}{4}}\left|\boldsymbol{\Sigma}_2\right|^{\frac{1}{4}}}{\left|\bar{\boldsymbol{\Sigma}}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{8}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^\top \bar{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right)$$

where $\bar{\boldsymbol{\Sigma}} = \frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$.

Deviance loss

Definition 1.38: Cramér, Cramér (1928a,b) and Székely (2003)

Consider two measures on f and g on $\mathbb{R}$. Then define Cramér distance

$$C_k(f,g) = \Big(\int_{-\infty}^{\infty} |F(x) - G(x)|^k \mathrm{d}x \Big)^{1/k}, \text{ for } k \geq 1$$

Definition 1.39: Wasserstein, Wasserstein (1969)

Consider two measures on f and g on $\mathbb{R}$. Then define Wasserstein distance

$$W_k(f,g) = \Big(\int_{0}^{1} |F^{-1}(u) - G^{-1}(u)|^k \mathrm{d}u \Big)^{1/k}, \text{ for } k \geq 1$$

Deviance loss

Proposition 1.23: C_1 and W_1

Consider two measures on f and g on $\mathbb{R}$.

$$W_1(f,g) = \int_0^1 |F^{-1}(u) - G^{-1}(u)| \mathrm{d}u = \int_{-\infty}^{\infty} |F(x) - G(x)| \mathrm{d}x = C_1(f,g).$$

Proof See Prokhorov (1956), Dall'Aglio (1956)

Deviance loss

Proposition 1.24: W_2 for Gaussian / Bernoulli distributions

Consider two Gaussian distributions, then

$$W_2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \boldsymbol{\sigma}_2^2))^2 = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2,$$

and for two Bernoulli distributions, if $p_1 \leq p_2$

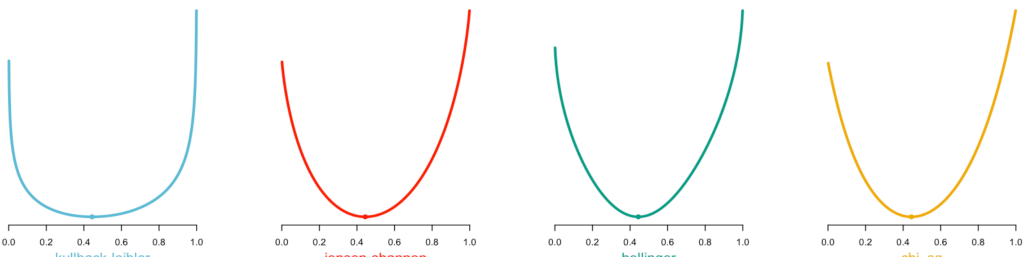
$$W_2(\mathcal{B}(p_1), \mathcal{B}(p_2)) = \sqrt{p_2 - p_1}.$$

Distribution Distance Based Inference

μ: multinomial distribution on $\{0, 1, 10\}$, with $\boldsymbol{p} = (.5, .1, .4)$

ν_θ: binomial type distribution on $\{0, 10\}$, with $\boldsymbol{q}_\theta = (1 - \theta, \theta)$

Let $\theta^\star = \text{argmin}\{d(p, q_\theta)\}$ or $\theta^\star = \text{argmin}\{d(p\|q_\theta)\}$



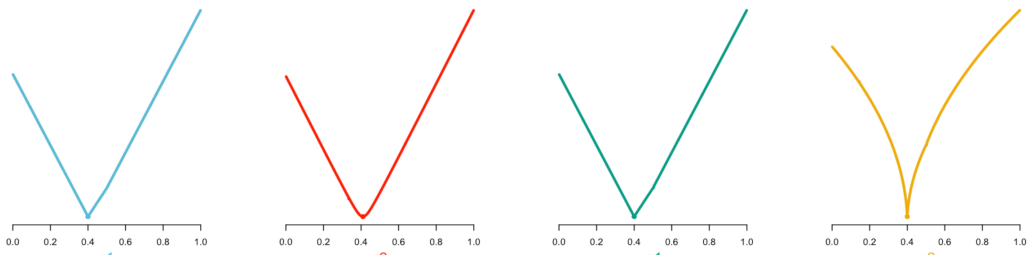
with $d_{\text{KL}}(p\|q_\theta)$, $d_{\text{JS}}(p, q_\theta)$, $d_{\text{H}}(p, q_\theta)$ and $d_{\text{H}_{\chi^2}}(p\|q_\theta)$.

Distribution Distance Based Inference

μ: multinomial distribution on $\{0, 1, 10\}$, with $\boldsymbol{p} = (.5, .1, .4)$

ν_θ: binomial type distribution on $\{0, 10\}$, with $\boldsymbol{q}_\theta = (1 - \theta, \theta)$

Let $\theta^\star = \text{argmin}\{d(p, q_\theta)\}$



with $C_1(p, q_\theta)$, $C_2(p, q_\theta)$, $W_1(p, q_\theta)$ and $W_2(p, q_\theta)$.

Optimal Transport, the Setting ⭐⭐⭐

Consider measures ν_0 and ν_1 on $\mathcal{Z}_0$ $\mathcal{Z}_1$, compact subsets of $\mathbb{R}^d$.

There exists T such that $\nu_1 = T_\sharp \nu_0$, where ν_0 is atomless ($T_\sharp \nu_0(B) = \nu_0(T^{-1}(B))$).

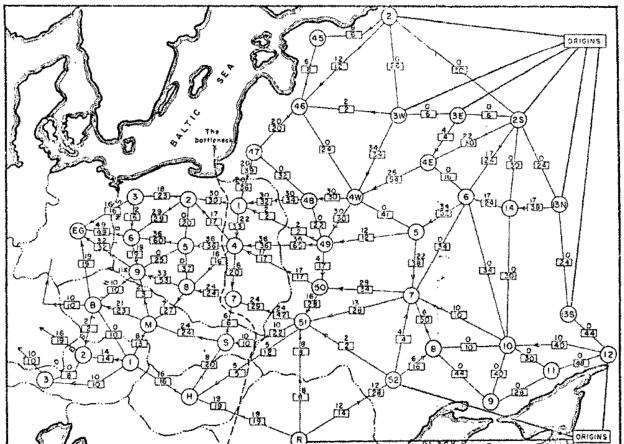
As shown in Villani (2003) and Santambrogio (2015), we can be interested in "optimal" mappings, satisfying Monge problem, from Monge (1781), i.e., solutions of

$$\inf_{T_\sharp \nu_0 = \nu_1} \int_{\mathcal{Z}_0} c(\mathbf{z}_0, T(\mathbf{z}_0)) \nu_0(\mathrm{d}\mathbf{z}_0) \,, \tag{1}$$

for some positive ground cost function $c : \mathcal{Z}_0 \times \mathcal{Z}_1 \to \mathbb{R}_+$. If ν_0 and ν_1 are not absolutely continuous (with respect to Lebesgue measure), there might not be such deterministic mapping T. This limitation motivates a relaxation of Monge's problem, as considered in Kantorovich (1942),

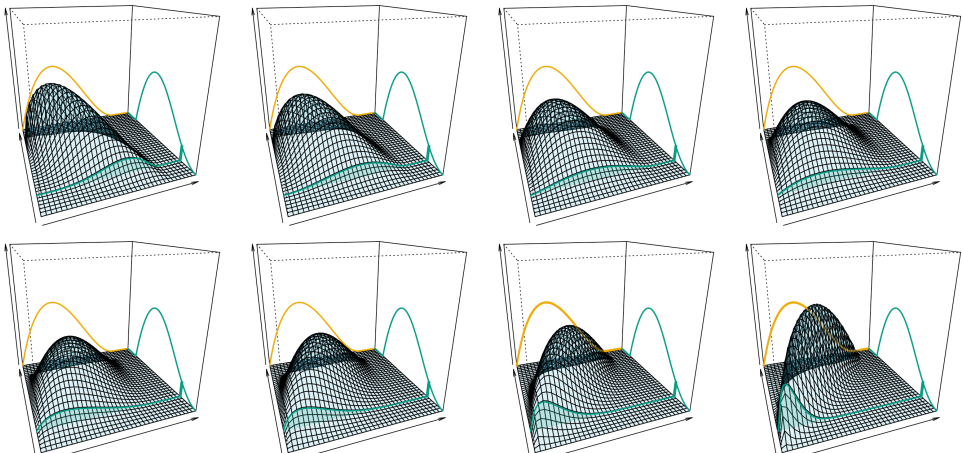
$$\inf_{\pi \in \Pi(\nu_0, \nu_1)} \int_{\mathcal{Z}_0 \times \mathcal{Z}_1} c(\mathbf{z}_0, \mathbf{z}_1) \pi(\mathrm{d}\mathbf{z}_0, \mathrm{d}\mathbf{z}_1) \,, \tag{2}$$

Optimal Transport, the Setting ⭐⭐⭐



(flow on networks, via Harris and Ross (1955))

Optimal Transport, the Setting ⭐⭐⭐



Optimal Transport, the Gaussian case ★★★

With a quadratic cost (and Euclidean distance in $\mathbb{R}^d$) and absolutely continuous measures, the optimal Monge map $T^\star$ is unique, and it is the gradient of a convex function, $T^\star = \nabla\varphi$, see Brenier (1991).
If $\nu_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$,

$$\boldsymbol{z}_1 = T^\star(\boldsymbol{z}_0) = \boldsymbol{\mu}_1 + \mathrm{M}(\boldsymbol{z}_0 - \boldsymbol{\mu}_0)\,,$$

where M is a symmetric positive matrix that satisfies $\mathrm{M}\boldsymbol{\Sigma}_0\mathrm{M} = \boldsymbol{\Sigma}_1$, which has a unique solution given by $\mathrm{M} = \boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2})^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$, where $\boldsymbol{M}^{1/2}$ denotes the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008).

In the univariate case

$$z_1 = T^\star(z_0) = \mu_1 + \sigma_1\sigma_0^{-1}(z_0 - \mu_0),$$

that is a linear non-decreasing mapping $\mathbb{R} \to \mathbb{R}$.

Optimal Transport, the discrete case ⭐⭐⭐

Consider two samples in the $\mathbb{R}^d$, $\{\boldsymbol{z}_{0,1}, \cdots, \boldsymbol{x}_{z,n_0}\}$ and $\{\boldsymbol{z}_{1,1}, \cdots, \boldsymbol{z}_{1,n_1}\}$. The discrete version of the Kantorovich problem (Equation 2) is

$$\min_{\mathrm{P} \in U(n_0,n_1)} \left\{ \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{P}_{i,j} \mathrm{C}_{i,j} \right\} = \min_{\mathrm{P} \in U(n_0,n_1)} \{\langle \mathrm{P}, \mathrm{C} \rangle \{ \tag{3}$$

where, as in Brualdi (2006), $U(n_0, n_1)$ is the set of $n_0 \times n_1$ matrices corresponding to the convex transportation polytope

$$U(n_0, n_1) = \left\{ \mathrm{P} : \mathrm{P}\mathbf{1}_{n_1} = \mathbf{1}_{n_0} \text{ and } \mathrm{P}^\top \mathbf{1}_{n_0} = \frac{n_0}{n_1} \mathbf{1}_{n_1} \right\},$$

and where C denotes the $n_0 \times n_1$ cost matrix, $\mathrm{C}_{i,j} = c(\boldsymbol{z}_i, \boldsymbol{z}_j)$, associated with cost c.

Optimal Transport, the discrete case ⭐⭐⭐

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

$$\underset{\mathrm{P}\in U(\mathbf{1}_n,\mathbf{1}_n)}{\operatorname{argmin}} \left\{ \langle \mathrm{P}, \mathrm{C} \rangle \right\}$$

over the set of **permutation matrices**.

minimal cost 0.2116

	7	8	9	10	11	12	
1	·	·	·	·	1	·	$1 \leftrightarrow 11$
2	·	1	·	·	·	·	$2 \leftrightarrow 8$
3	·	·	1	·	·	·	$3 \leftrightarrow 9$
4	1	·	·	·	·	·	$4 \leftrightarrow 7$
5	·	·	·	1	·	·	$5 \leftrightarrow 10$
6	·	·	·	·	·	1	$6 \leftrightarrow 12$

Optimal Transport, the discrete case ⭐⭐⭐

	7	8	9	10	11	12	13	14	15	16
1	0.41	0.55	0.22	0.64	0.04	0.25	0.24	0.77	0.74	0.55
2	0.28	0.24	0.73	0.22	0.64	0.80	0.76	0.76	0.12	0.10
3	0.28	0.47	0.32	0.52	0.16	0.37	0.27	0.68	0.63	0.45
4	0.28	0.62	0.81	0.25	0.64	0.85	0.58	0.32	0.51	0.48
5	0.41	0.37	0.89	0.25	0.81	0.97	0.91	0.81	0.05	0.25
6	0.66	0.76	0.21	0.89	0.22	0.14	0.33	0.96	0.99	0.79

	7	8	9	10	11	12	13	14	15	16	
1	·	·	1/5	·	3/5	·	1/5	·	·	·	$1 \leftrightarrow \{9,11,13\}$
2	·	2/5	·	·	·	·	·	·	·	3/5	$2 \leftrightarrow \{8,16\}$
3	3/5	·	·	·	·	·	2/5	·	·	·	$3 \leftrightarrow \{7,13\}$
4	·	·	·	2/5	·	·	·	3/5	·	·	$4 \leftrightarrow \{10,14\}$
5	·	1/5	·	1/5	·	·	·	·	3/5	·	$5 \leftrightarrow \{8,10,15\}$
6	·	·	2/5	·	·	3/5	·	·	·	·	$6 \leftrightarrow \{9,12\}$

Optimal Transport, the discrete case ⭐⭐⭐

$$\underset{\mathrm{P}\in U(\mathbf{1}_{n_0},\mathbf{1}_{n_1})}{\operatorname{argmin}} \Big\{ \langle \mathrm{P}, \mathrm{C} \rangle - \gamma \cdot \text{entropy}(\mathrm{P}) \Big\}$$

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

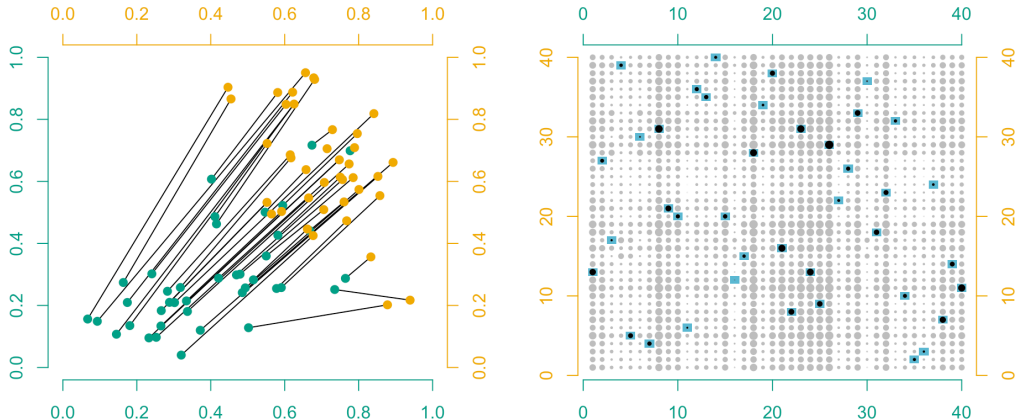
(**regularization** via Sinkhorn algorithm, associated to "Matrix Scaling Problem" in Sinkhorn (1962)) or

$$\underset{\mathrm{P}\in U(\mathbf{1}_{n_0},\mathbf{1}_{n_1})}{\operatorname{argmin}} \Big\{ \langle \mathrm{P}, \mathrm{C} \rangle + \gamma \cdot \mathrm{d_{KL}}(\mathrm{P} || \mathbf{1}_{n_0} \otimes \mathbf{1}_{n_1}) \Big\}$$

minimal cost 0.2189 (> 0.2116)

	7	8	9	10	11	12	
1	·	·	0.35	·	0.59	0.06	$1 \leftrightarrow 11 + \cdots$
2	0.12	0.82	·	0.07	·	·	$2 \leftrightarrow 8 + \cdots$
3	·	·	0.55	·	0.41	0.04	$3 \leftrightarrow 9 + \cdots$
4	0.85	·	·	0.14	·	·	$4 \leftrightarrow 7 + \cdots$
5	0.03	0.18	·	0.79	·	·	$5 \leftrightarrow 10 + \cdots$
6	·	·	0.10	·	·	0.90	$6 \leftrightarrow 12 + \cdots$

Optimal Transport, the discrete case ★★★



Discrete matching, $n_0 = n_1 = 40$ individuals in two groups, $\{z_i, s_i = 0\}$ and $\{z_i, s_i = 1\}$, with matrices C ($\mathrm{C}_{i,j} = \bullet$) and $\mathrm{P}^\star$ (■ if $\mathrm{P}^\star_{i,j} = 1$)

Optimal Transport, the univariate case ★★★

Wasserstein distance is defined as the optimal transportation cost, when c is the ℓ_k distance, for $k \geq 1$, Wasserstein (1969),

$$W_k(\nu_0, \nu_1)^k := \inf_{T_\sharp \nu_0 = \nu_1} \int_{\mathbb{R}} |z_0 - T(z_0)|^k \nu_0(\mathrm{d}z_0) = \int_0^1 |Q_1(u) - Q_0(u)|^k \mathrm{d}u$$

i.e., the optimal mapping $T^\star$ (if $k > 1$) is non-decreasing, i.e.,

$$z_1 = T^\star(z_0) = Q_1 \circ F_0(z_0) \text{ where } \begin{cases} F_j(z) = \mu_j((-\infty, z]) \\ Q_j(u) = F_j^{-1}(u) = \inf\{t \in \mathbb{R} : F_j(t) \geq u\} \end{cases}$$

$T^\star := Q_1 \circ F_0$ is a non-decreasing mapping $\mathbb{R} \to \mathbb{R}$.

In higher dimension, some multivarite extensions of quantiles can be introduced to keep this construction, see Hallin et al. (2021), Hallin and Konen (2024)

Optimal Transport, the univariate case ★★★

Given $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$ n pairs of ordered real numbers, and some supermodular function $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, for every permutation σ of $\{1, 2, \cdots, n\}$,

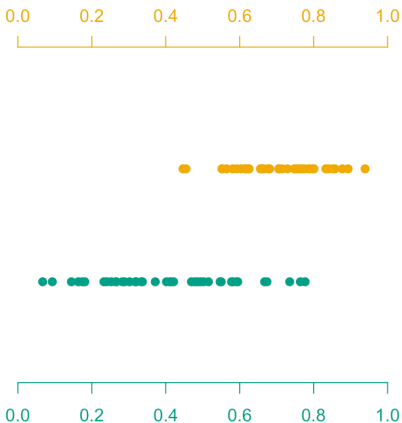
$$\sum_{i=1}^{n} \Phi(x_i, y_{n+1-i}) \leq \sum_{i=1}^{n} \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^{n} \Phi(x_i, y_i),$$

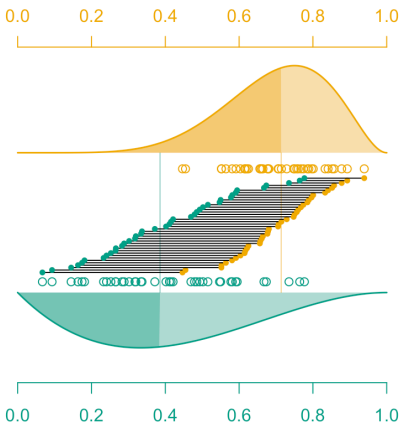
while if $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is submodular,

$$\sum_{i=1}^{n} \Phi(x_i, y_i) \leq \sum_{i=1}^{n} \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^{n} \Phi(x_i, y_{n+1-i}).$$

see Hardy et al. (1952).
See related work on risk measures, Denuit et al. (2006), Galichon (2016), Carlier et al. (2016).

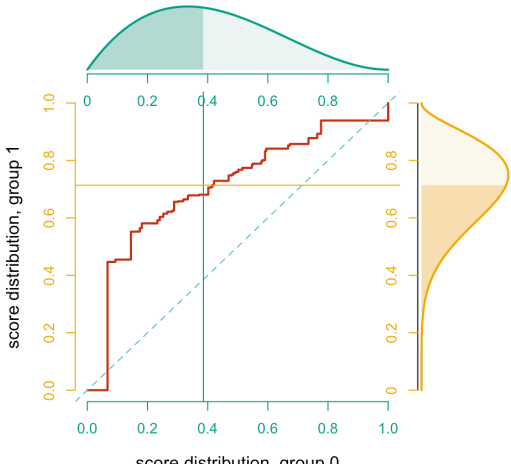
Optimal Transport with Pictures, the univariate case ★★★

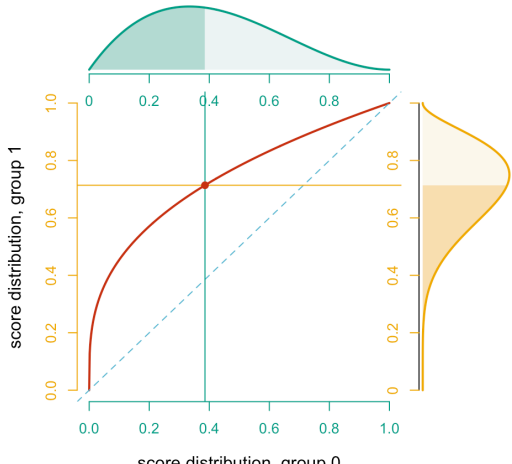




Univariate predictions for n individuals in two groups, $\{z_i, s_i = 0\}$ and $\{z_i, s_i = 1\}$, or measures ν_0 and ν_1. $T^\star(\cdot) = Q_1 \circ F_0(\cdot)$.

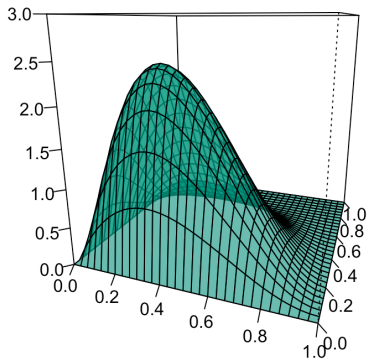
Optimal Transport with Pictures, the univariate case ⭐⭐⭐

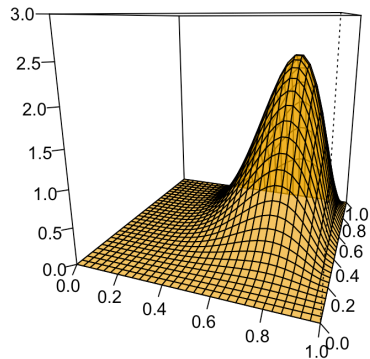




Univariate predictions for n individuals in two groups, $T^\star(\cdot) = Q_1 \circ F_0(\cdot)$.

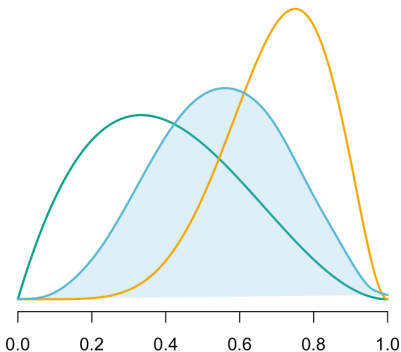
Optimal Transport with Pictures, the univariate case ⭐⭐⭐

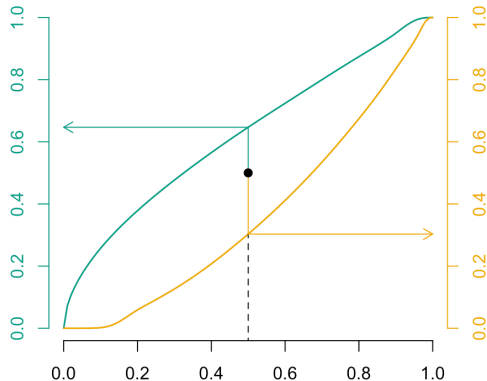




Multivariate observations for n individuals in two groups, ν_0 and ν_1 (densities).

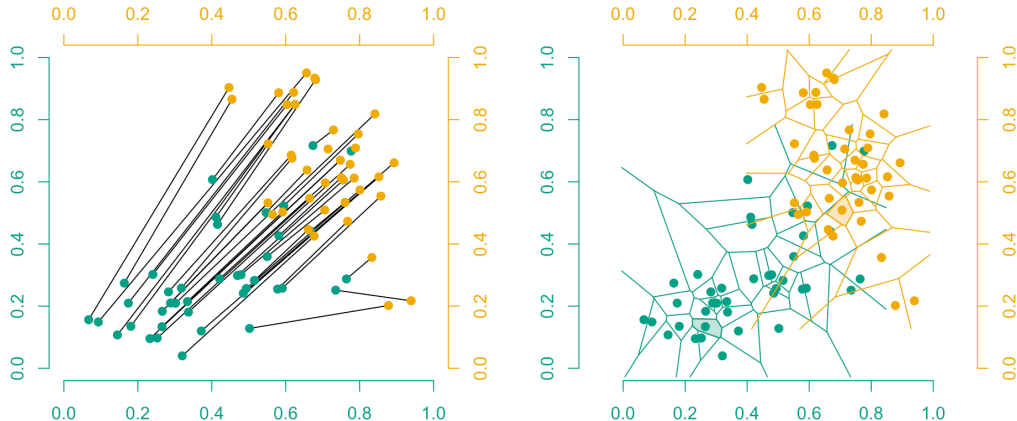
Optimal Transport with Pictures, the univariate case ⭐⭐⭐





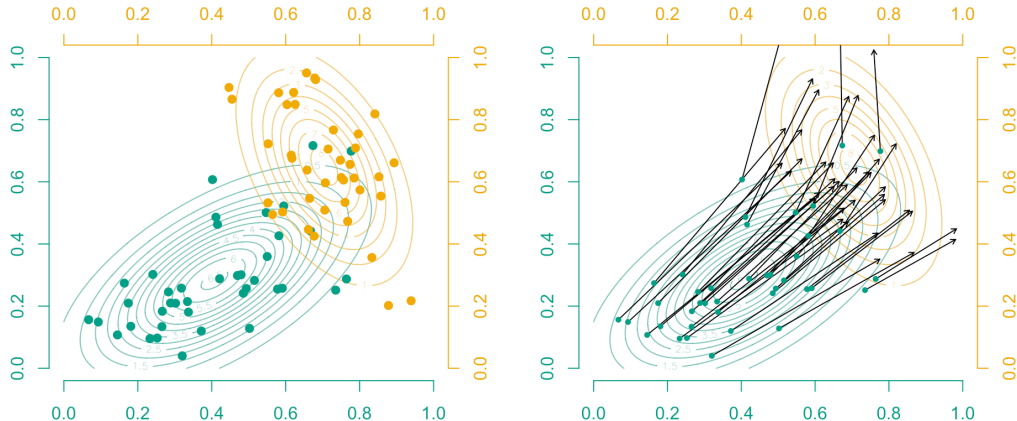
Bivariate observations for n individuals in two groups, $\{z_i, s_i = 0\}$ and $\{z_i, s_i = 1\}$.

Optimal Transport with Pictures, the univariate case ★★★



Discrete matching, n individuals in two groups, $\{z_i, s_i = 0\}$ and $\{z_i, s_i = 1\}$.

Optimal Transport with Pictures, the univariate case ⭐⭐⭐



Optimal transport with Gaussian joint distributions, $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$.

Risk

Definition 1.40: Distance

A **distance** or **metric** over $\mathcal{X} \in \mathbb{R}^p$ is a map $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the following properties

- coincidence: $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,
- symmetry $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{d}(\boldsymbol{y}, \boldsymbol{x})$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,
- triangle inequality $\mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) \leq \mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) + \mathrm{d}(\boldsymbol{y}, \boldsymbol{z})$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

Definition 1.41: Loss ℓ

A loss function ℓ is a function defined on $\mathcal{Y} \times \mathcal{Y}$ such that $\ell(y, y') \geq 0$ and $\ell(y, y) = 0$.

Risk

Standard loss functions are usually related to divergences, distances, or powers of distances, e.g.

$$\ell_1(y, y') = \mathrm{d}_1(y, y') = |y - y'| \text{ or } \ell_2(y, y') = \mathrm{d}_1(y, y')^2 = (y - y')^2.$$

Definition 1.42: Risk $\mathcal{R}$

For a fitted model $\widehat{m}$, its risk is

$$\mathcal{R}(\widehat{m}) = \mathbb{E}_{\mathbb{P}}\Big[\ell\big(Y, \widehat{m}(\boldsymbol{X})\big)\Big] = \int \ell(y, \widehat{m}(\boldsymbol{x}))\mathrm{d}\mathbb{P}(y, \boldsymbol{x}).$$

Risk

Definition 1.43: Empirical risk $\widehat{\mathcal{R}}_n$

Given a sample $\{(y_i, \boldsymbol{x}_i), i = 1, \cdots, n\}$, define the empirical risk

$$\widehat{\mathcal{R}}_n(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{m}(\boldsymbol{x}_i), y_i\right).$$

Following Vapnik (1991), the "empirical risk minimization principle" states that the learning algorithm $\widehat{m}^*$ is

$$\widehat{m}^* = \underset{\widehat{m}\in\mathcal{M}}{\operatorname{argmin}}\{\widehat{\mathcal{R}}_n(\widehat{m})\}.$$

Risk

Proposition 1.25: Consistency of the empirical risk

The empirical risk, associated with ℓ is a consistence estimator of the ℓ-risk, i.e.

$$\lim_{n\to\infty}\{\widehat{\mathcal{R}}_n(m)\}=\mathcal{R}(m),\text{ or }\lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell\left(m(\boldsymbol{X}_i),Y_i\right)\right\}=\mathbb{E}\left[\ell\left(m(\boldsymbol{X}),Y\right)\right],\text{ in probability.}$$

Proposition 1.26: Optimal Decision, "*Bayes decision rule*"

For each $\boldsymbol{x}$ choose the prediction $m_{\boldsymbol{x}}^{\star}$ that minimizes the conditional expected loss,

$$m_{\boldsymbol{x}}^{\star}\in\underset{z\in\mathcal{Y}}{\operatorname{argmin}}\left\{\int\ell(y,z)\mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})\right\}$$

Risk

It is coined "Bayes decision rule" because the conditional distribution $Y|\boldsymbol{X}$ is sometimes be referred to as the "posterior" distribution of Y given data $\boldsymbol{X}$.

Definition 1.44: Misclassification loss, $\ell_{0/1}$

If $y \in \{0,1\}$ and $\widehat{y} \in \{0,1\}$, $\ell_{0/1}(y,\widehat{y}) = \mathbf{1}(y \neq \widehat{y})$.

In the case of a binary classifier, observe that

$$\begin{aligned}
\mathcal{R}(\widehat{m}) &= \mathbb{E}[\ell(\widehat{m}(\boldsymbol{X}), Y)] = \mathbb{E}[\mathbb{E}[\ell(\widehat{m}(\boldsymbol{X}), Y) \mid \boldsymbol{X}]] \\
&= \mathbb{E}\big[\ell(\widehat{m}(\boldsymbol{X}),1) \cdot \mathbb{P}(Y=1 \mid \boldsymbol{X}) + \ell(\widehat{m}(\boldsymbol{X}),0) \cdot \mathbb{P}(Y=0 \mid \boldsymbol{X})\big] \\
&= \mathbb{E}\big[\mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] \cdot \mu(\boldsymbol{X}) + \mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 0] \cdot (1-\mu(\boldsymbol{X}))\big] \\
&= \mathbb{E}[\mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] \cdot \mu(X) + (1 - \mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1]) \cdot (1-\mu(\boldsymbol{X}))] \\
&= \mathbb{E}\big[\mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] \cdot (2\mu(\boldsymbol{X}) - 1) + 1 - \mu(\boldsymbol{X})\big].
\end{aligned}$$

Risk

Since $\widehat{m} : \mathcal{X} \to \{0, 1\}$, this expectation is minimized by choosing $\widehat{m} = m^\star$, where

$$m^\star(\boldsymbol{x}) = \mathbf{1}(\mu(\boldsymbol{x}) > 1/2) = \begin{cases} 1 \text{ if } \mu(\boldsymbol{x}) > 1/2 \\ 0 \text{ if } \mu(\boldsymbol{x}) \leq 1/2 \end{cases}$$

The optimal risk ("Bayes risk") is $\mathcal{R}(m^\star) = \inf_m \{\mathcal{R}(m)\}$.

Definition 1.45: Excess of risk of $\widehat{m}$

For any model $\widehat{m}$, the excess of risk is $\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star)$.

For a classifier

$$\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star) = \mathbb{E}\big[|2\mu(\boldsymbol{X}) - 1| \cdot \mathbf{1}(\widehat{m}(\boldsymbol{X}) \neq m^\star(\boldsymbol{X}))\big].$$

Since we do not know μ consider a classifier based on $\widehat{m}$ .....

Definition 1.46: Plug-in Estimator

Estimate $\widehat{\mu}$ and use, as a classifier, $\mathbf{1}(\widehat{\mu}(\boldsymbol{x}) > 1/2)$.

Proposition 1.27

For any model $\widehat{\mu}$, the risk of the plug-in classifier $\widehat{m}(\boldsymbol{x}) = \mathbf{1}(\widehat{\mu}(\boldsymbol{x}) > 1/2)$ satisfies

$$\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star) \leq 2\mathbb{E}|\mu(\boldsymbol{X}) - \widehat{\mu}(\boldsymbol{X})|.$$

Proof We have seen that

$$\mathcal{R}(\widehat{m}) - \mathcal{R}\left(m^*\right) = \mathbb{E}(1[\widehat{m}(\boldsymbol{X}) \neq 1] - 1\left[m^*(\boldsymbol{X}) \neq 1\right]) \cdot (2\mu(\boldsymbol{X}) - 1).$$

Risk

But

$$\begin{aligned}
&\left(\mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] - \mathbf{1}\left[m^*(\boldsymbol{X}) \neq 1\right]\right)(2\mu(\boldsymbol{X}) - 1) \\
&= \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]\left(1[\widehat{m}(\boldsymbol{X}) \neq 1] - 1\left[m^*(\boldsymbol{X}) \neq 1\right]\right)(2\mu(\boldsymbol{X}) - 1) \\
&= \begin{cases} \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right](2\mu(\boldsymbol{X}) - 1) & \text{if } 2\mu(\boldsymbol{X}) - 1 > 0, \\ \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right](-1)(2\mu(\boldsymbol{X}) - 1) & \text{if } 2\mu(\boldsymbol{X}) - 1 \leq 0. \end{cases}
\end{aligned}$$

(from the definition of m^*)

$$= \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right] \cdot |2\mu(\boldsymbol{X}) - 1|,$$

$$\mathcal{R}(\widehat{m}) - \mathcal{R}\left(m^*\right) = \mathbb{E}(\mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]) \cdot 2|\mu(\boldsymbol{X}) - 1/2|.$$

If $\widehat{m}(\boldsymbol{x}) \neq m^*(\boldsymbol{x})$, it means that $\widehat{\mu}(\boldsymbol{x})$ and $\mu(\boldsymbol{x})$ lie on opposite sides of $1/2$,

$$|\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})| = |\widehat{\mu}(\boldsymbol{x}) - 1/2| + \underbrace{|1/2 - \mu(\boldsymbol{x})|}_{\geq 0} \geq |\widehat{\mu}(\boldsymbol{x}) - 1/2|$$

Risk

i.e.

$$|\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})| \geq |\widehat{\mu}(\boldsymbol{x}) - 1/2| \cdot \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]$$

which is also valid when $\widehat{m}(\boldsymbol{x}) = m^*(\boldsymbol{x})$, thus

$$\mathcal{R}(\widehat{m}) - \mathcal{R}\left(m^*\right) = 2\mathbb{E}(\mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]) \cdot |\mu(\boldsymbol{X}) - 1/2| \leq 2\mathbb{E}\big[|\widehat{\mu}(\boldsymbol{X}) - \mu(\boldsymbol{X})|\big].$$

This $\ell_{0/1}$ loss function may be difficult to directly optimize, as shown in Bartlett et al. (2006). One could consider some surrogate loss $\tilde{\ell}$ which is easier to optimize.

Definition 1.47: Quadratic loss, ℓ_2

$\ell_2(y, \widehat{y}) = (y - \widehat{y})^2$, and the risk is then $\mathcal{R}_2(\widehat{m}) = \mathbb{E}\Big[\left(Y - \widehat{m}(\boldsymbol{X})\right)^2\Big]$.

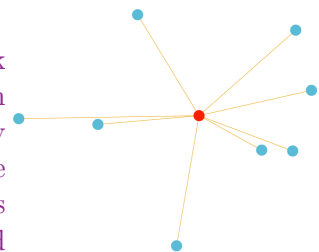
Risk

Observe that

$$\mathbb{E}[Y] = \underset{m\in\mathbb{R}}{\text{argmin}}\{\mathcal{R}_2(m)\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{\mathbb{E}\Big[\ell_2\left(Y, m\right)\Big]\Big\}.$$

The expected value is "ellicitable" (for the ℓ_2 loss).

"The elicitability of a risk measure means that the risk measure can be obtained by minimizing the expectation of a forecasting objective function. Elicitability is closely related to backtesting, whose objective is to evaluate the performance of a risk forecasting model. If a risk measure is elicitable, then the sample average forecasting error based on the objective function can be used for backtesting the risk measure," He et al. (2022).



The empirical risk minimizer is the "least-square" estimate.

Risk

Proposition 1.28: Optimal Decision, "*Bayes decision rule*", ℓ_2

For the quadratic loss ℓ_2, Bayes decision rule is the (conditional) expected value, $m_{\boldsymbol{x}}^\star = \mathbb{E}\big[Y|\boldsymbol{X}=\boldsymbol{x}\big] = \mu(\boldsymbol{x})$.

Proposition 1.29: Optimal Decision, "*Bayes decision rule*", ℓ_{0-1}

For the binary loss ℓ_{0-1}, Bayes decision rule is the majority class $m_{\boldsymbol{x}}^\star = \underset{y\in\{0,1\}}{\text{argmax}}\{\mathbb{P}[Y=y|\boldsymbol{X}=\boldsymbol{x}]\}$.

Risk

For the quadratic loss ℓ_2, Bayes decision rule is the (conditional) expected value,

$$m^\star_{\boldsymbol{x}} = \mathbb{E}\big[Y|\boldsymbol{X}=\boldsymbol{x}\big] = \mu(\boldsymbol{x}) = \underset{m\in\mathcal{M}}{\text{argmin}}\big\{\mathbb{E}\big[\ell_2(Y, m(\boldsymbol{X}))\big]\big\},$$

For the pinball loss $\ell_{q,\tau}$, Bayes decision rule is the quantile regression,

$$q^\star_{\tau,\boldsymbol{x}} = \mathsf{Q}_\tau\big[Y|\boldsymbol{X}=\boldsymbol{x}\big] = \underset{m\in\mathcal{M}}{\text{argmin}}\big\{\mathbb{E}\big[\ell_{q,\tau}(Y, m(\boldsymbol{X}))\big]\big\},$$

where $\ell_{q,\tau}(y,y') = \tau|y-y|\mathbf{1}_{y>y'} + (1-\tau)|y-y|\mathbf{1}_{y<y'}$.
For any $\alpha\in(0,1)$,

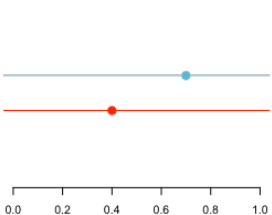
$$\mathbb{P}\Big(Y\in\Big[\mathsf{Q}_{\alpha/2}(Y|\boldsymbol{X}); \mathsf{Q}_{1-\alpha/2}(Y|\boldsymbol{X})\Big]\Big) = 1-\alpha.$$

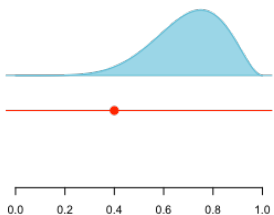
however, in general, with a finite number of observations

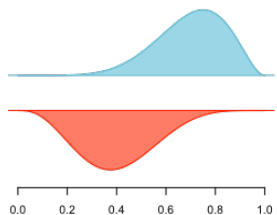
$$\mathbb{P}\Big(Y\in\Big[\widehat{\mathsf{Q}}_{\alpha/2}(Y|\boldsymbol{X}); \widehat{\mathsf{Q}}_{1-\alpha/2}(Y|\boldsymbol{X})\Big]\Big) \neq 1-\alpha.$$

(see conformal prediction).

Loss







Loss

A statistic θ is ellicitable if it can be expressed as a risk minimization for a given loss.

Definition 1.48: Elicitation, Brier et al. (1950), Good (1952)

A statistical functional $\theta = \mathcal{I}(Y)$ is said to be elicitable if it minimizes expected loss for some loss function ℓ, in the sense that

$$\mathcal{I}(Y) = \underset{y \in \mathbb{R}}{\operatorname{argmin}} \{\mathbb{E}[\ell(Y, y)]\}$$

A statistics $\widehat{\theta} = \widehat{\mathcal{I}}(\boldsymbol{y})$ (where $\boldsymbol{y} = \{y_1, \cdots, y_n\}$) is said to be elicitable if it minimizes expected loss for some loss function ℓ, in the sense that

$$\widehat{\mathcal{I}}(\boldsymbol{y}) = \underset{y \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, y) \right\}$$

Loss

E.g.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \underset{m\in\mathbb{R}}{\text{argmin}}\{\widehat{\mathcal{R}}_2(m)\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{\sum_{i=1}^{n}\ell_2(y_i, m)\Big\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{\sum_{i=1}^{n}(y_i - m)^2\Big\}.$$

for any distribution of Y_i's, or, e.g.,

$$\text{median}(y) \in \underset{m\in\mathbb{R}}{\text{argmin}}\{\widehat{\mathcal{R}}_1(m)\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{\sum_{i=1}^{n}\ell_1(y_i, m)\Big\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{\sum_{i=1}^{n}|y_i - m|\Big\}.$$

again, for any distribution of Y_i's.
But if Y_i's are i.i.d. Bernoulli $\mathcal{B}(p)$ variables,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \underset{m\in\mathbb{R}}{\text{argmin}}\{\widehat{\mathcal{R}}(m)\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{\sum_{i=1}^{n}\ell(y_i, m)\Big\}.$$

Loss

where

$$\ell(y_i, \widehat{y}_i) = \begin{cases} -2\log(1-\widehat{y}_i) & y_i = 0 \\ -2\log(\widehat{y}_i) & y_i = 1. \end{cases}$$

if Y_i's are i.i.d. Poisson $\mathcal{P}(\lambda)$ variables,

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \underset{m\in\mathbb{R}}{\text{argmin}}\{\widehat{\mathcal{R}}(m)\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{\sum_{i=1}^{n}\ell(y_i, m)\Big\}.$$

where

$$\ell(y_i, \widehat{y}_i) = \begin{cases} 2(y_i\log y_i - y_i\log\widehat{y}_i - y_i + \widehat{y}_i) & y_i > 0 \\ 2\widehat{y}_i & y_i = 0, \end{cases}$$

Loss

Definition 1.49: Scoring Rule

In decision theory, a scoring rule provides evaluation metrics for probabilistic predictions or forecasts. $\ell : (y, F_y)$ for a predictive distribution F_y

Popular in Bayesian regression

Quick Overview on Optimization

- Most ML models involve learning parameters to minimize a cost, or maximize a reward. For examples:
  - Linear regression: minimize mean squared error.
  - Logistic regression: minimize cross-entropy loss.
  - Neural networks: minimize empirical loss via backpropagation.
  - Reinforcement learning: maximize expected future rewards.
- Optimization is the engine that drives model training.
- Understanding optimization is crucial for developing efficient, robust learning algorithms.
- General form: $\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta})$, where $\mathcal{L}(\boldsymbol{\theta})$ is the loss function and $\boldsymbol{\theta}$ are the parameters.
- Often involves:
  - High-dimensional parameter spaces
  - Non-convex loss functions (e.g., deep nets)

Quick Overview on Optimization

- Large-scale datasets

- Goal: find parameters $\boldsymbol{\theta}^\star$ that minimize training loss and ideally generalize well.
- **Unconstrained vs. Constrained**

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \quad \text{vs.} \quad \min_{\boldsymbol{\theta} \in C} \mathcal{L}(\boldsymbol{\theta})$$

- **Convex vs. Non-convex**
  - Convex: single global minimum, easier to optimize.
  - Non-convex: multiple minima, may require heuristics.
- **Deterministic vs. Stochastic**
  - Full dataset vs. mini-batches (SGD)
- **Smooth vs. Non-smooth**
  - Smooth: gradients exist everywhere.
  - Non-smooth: may need subgradient or proximal methods.

Quick Overview on Optimization

- **Stochastic methods** scale to large datasets:
  - SGD, Mini-batch SGD, Adam
- **Regularization** improves generalization:

$$\min_{\theta} \mathcal{L}(\theta) + \lambda R(\theta)$$

- **Constraints and projections:** Useful for interpretability, sparsity, and physical constraints.
- **Non-smooth optimization:**
  - Lasso (L1 regularization), hinge loss in SVMs
- Meta-learning, Bayesian optimization, reinforcement learning: Advanced optimization in frontier ML tasks.
- **Convex optimization:**
  - Global minimum is also local minimum.

Quick Overview on Optimization

- Easier to analyze and solve.

- **Non-convex optimization:**
  - Many local minima and saddle points.
  - Common in deep learning.

- Visualization: think "bowl-shaped" vs "rugged terrain".
- Most ML models are trained using gradients.
- **Gradient Descent:** from x_0

$$x_{t+1} = x_t - \eta \nabla \mathcal{L}(x_t)$$

- **Variants:** Stochastic Gradient Descent
- Efficient and scalable for large datasets and high-dimensional problems.
- Not all problems are best solved with plain gradients.
- Alternatives include:

Quick Overview on Optimization

  - **Newton's Method**: uses second derivatives (Hessian).
  - **Coordinate Descent**: optimize one variable at a time.
  - **Evolutionary algorithms**: e.g., genetic algorithms.
  - **Bayesian Optimization**: for black-box functions.
- Tool choice depends on model complexity, smoothness, constraints, and noise.

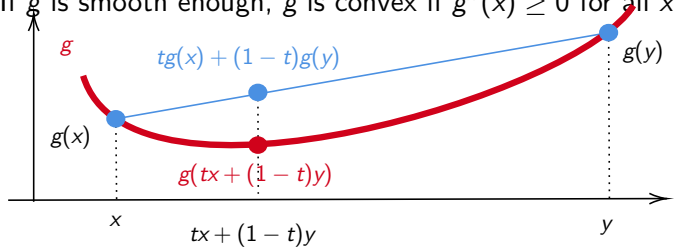
Convex

Definition 1.50: Convex Function

$g : \mathbb{R} \to \mathbb{R}$ is convex if for any $t \in [0,1]$, $g(tx + (1-t)y) \leq t\,g(x) + (1-t)\,g(y)$.

Example: $g(x) = \exp(x)$, $g(x) = |x|$ or $g(x) = x^2$

If g is smooth enough, g is convex if $g''(x) \geq 0$ for all x



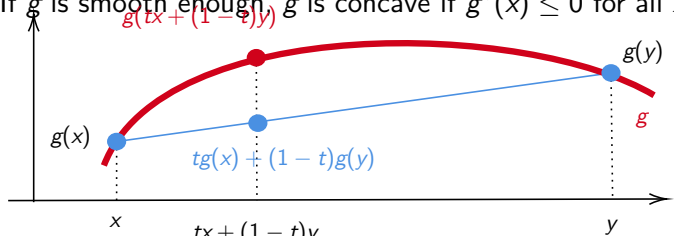
Convex and Concave

Definition 1.51: Concave Function

$g : \mathbb{R} \to \mathbb{R}$ is concave if for any $t \in [0, 1]$, $g(tx + (1-t)y) \geq t\,g(x) + (1-t)\,g(y)$.

Example: $g(x) = \log(x)$

If g is smooth enough, g is concave if $g''(x) \leq 0$ for all x



Approximation

Proposition 1.30: Taylor First Order Approximation

If g is continuously differentiable in the neighbourhood of a

$$g(x) \approx g(a) + g'(a) \cdot (x - a) \text{ or } g(a + h) \approx g(a) + g'(a) \cdot h$$

Sketch of the proof (1): definition of the derivative (in a)

$$g'(a) = \lim_{h \to 0} \frac{g(a + h) - g(a)}{h} \approx \frac{g(a + h) - g(a)}{h}$$

Sketch of the proof (2): fundamental theorem of analysis

$$g(x) = g(a) + \int_a^x g'(y)\mathrm{d}y \approx g(a) + \int_a^x g'(a)\mathrm{d}y \approx g(a) + g'(a) \cdot (x - a)$$

Approximation

since $\int_a^x g'(y)\mathrm{d}y \approx g'(a)\cdot(x-a)$ but one could be more precise

$$\int_a^x g'(y)\mathrm{d}y = \Big[-(x-y)g'(y)\Big]_a^x + \int_a^x (x-y)g''(y)\mathrm{d}y$$

$$\int_a^x g'(y)\mathrm{d}y = g'(a)(x-a) + \int_a^x (x-y)g''(y)\mathrm{d}y$$

then iterate...

$$\int_a^x (x-y)g''(y)\mathrm{d}y = \Big[-\frac{(x-y)^2}{2}g''(y)\Big]_a^x + \int_a^x \frac{(x-y)^2}{2}g^{(3)}(y)\mathrm{d}y$$

$$\int_a^x (x-y)g''(y)\mathrm{d}y \approx \frac{(x-a)^2}{2}g''(a)$$

Approximation

Proposition 1.31: Taylor Second Order Approximation

If g is twice continuously differentiable in the neighborhood of a

$$g(x) \approx g(a) + g'(a)(x-a) + \frac{g''(a)}{2}(x-a)^2$$

$$g(a+h) \approx g(a) + g'(a)h + \frac{g''(a)}{2}h^2$$

More generally, integration by parts leads to

$$\int_a^x \frac{(x-y)^{i-1}}{(i-1)!} g^{(i)}(y)\mathrm{d}y = \Big[-\frac{(x-y)^i}{(i)!} g^{(i)}(y) \Big]_a^x + \int_a^x \frac{(x-y)^i}{(i)!} f^{(i+1)}(y)\mathrm{d}y$$

Jensen Inequality

If g is twice continuously derivable in the neighborhood of a, then

$$g(x) \approx g(a) + g'(a)(x-a) + \frac{g''(a)}{2}(x-a)^2$$

Let X be a random variable, with expectation $\mathbb{E}[X] = a$, then

$$g(X) \approx g(\mathbb{E}[X]) + g'(\mathbb{E}[X])(X - \mathbb{E}[X]) + \frac{g''(\mathbb{E}[X])}{2}(X - \mathbb{E}[X])^2$$

so

$$\mathbb{E}\big[g(X)\big] \approx g(\mathbb{E}[X]) + \frac{g''(\mathbb{E}[X])}{2}\text{Var}[X]$$

Proposition 1.32: Jensen Inequality

Let g be a convex function, and X a random variable, then $\mathbb{E}\big[g(X)\big] \geq g(\mathbb{E}[X])$ (if the expected value exists).

Approximation in Higher Dimension

Proposition 1.33: Taylor Second Order Approximation, Univariate

If $g : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable in the neighborhood of a

$$g(a+h) \approx g(a) + g'(a)h + \frac{g''(a)}{2}h^2$$

Proposition 1.34: Taylor Second Order Approximation, Multivariate

If $g : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable in the neighbourhood of $\boldsymbol{a}$

$$g(\boldsymbol{a}+\boldsymbol{h}) \approx g(\boldsymbol{a}) + \nabla g(\boldsymbol{a})^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{H}g(\boldsymbol{a})\boldsymbol{h}$$

where ∇g is the gradient of g and $\boldsymbol{H}g$ is the Hessian matrix, evaluated in $\boldsymbol{a}$.

Approximation

If $g : \mathbb{R}^2 \to \mathbb{R}$, the gradient is a vector and the Hessian a 2×2 matrix

$$\nabla g = \begin{pmatrix} \dfrac{\partial g}{\partial x} \\ \dfrac{\partial g}{\partial y} \end{pmatrix} \text{ and } \boldsymbol{H}g = \begin{pmatrix} \dfrac{\partial^2 g}{\partial x^2} & \dfrac{\partial^2 g}{\partial x \partial y} \\ \dfrac{\partial^2 g}{\partial y \partial x} & \dfrac{\partial^2 g}{\partial y^2} \end{pmatrix}$$

$$g(\boldsymbol{a}+\boldsymbol{h}) \approx g(\boldsymbol{a}) + \nabla g(\boldsymbol{a})^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{H}g(\boldsymbol{a})\boldsymbol{h}$$

Example: $g(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x}$, then

$$g(\boldsymbol{a}+\boldsymbol{h}) = (\boldsymbol{a}+\boldsymbol{h})^\top(\boldsymbol{a}+\boldsymbol{h}) = \boldsymbol{a}^\top\boldsymbol{a} + \boldsymbol{a}^\top\boldsymbol{h} + \boldsymbol{h}^\top\boldsymbol{a} + \boldsymbol{h}^\top\boldsymbol{h}$$

or $\boldsymbol{a}^\top\boldsymbol{h} = \boldsymbol{h}^\top\boldsymbol{a}$ so

$$g(\boldsymbol{a}+\boldsymbol{h}) = \underbrace{\boldsymbol{a}^\top\boldsymbol{a}}_{=g(\boldsymbol{a})} + \underbrace{2\boldsymbol{a}^\top}_{=\nabla g(\boldsymbol{a})^\top}\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \underbrace{2\mathbb{I}}_{=\boldsymbol{H}g(\boldsymbol{a})}\boldsymbol{h}$$

Gradients and Differentials

- The **derivative** of a $\mathbb{R} \to \mathbb{R}$ measures how a function changes with respect to its input.
- It represents the slope or rate of change of a function.
- **Notation:** $g'(x), \dfrac{\mathrm{d}f}{\mathrm{d}x}$ or $Dg(x)$
- **Interpretation:**
  - $g'(x) > 0$ means the function is increasing.
  - $g'(x) < 0$ means the function is decreasing.
  - $g'(x) = 0$ may indicate a maximum, minimum, or inflection point.
- **Example:** If $g(x) = x^2$, then $g'(x) = 2x$
- A **differential** represents a small change in a variable.
- $\mathrm{d}g \approx g'(x)\,\mathrm{d}x$ gives a linear approximation of change in g.
- Useful for approximating function values near a point.

Gradients and Differentials

- **Example:** If $g(x) = x^2$, then
$$\mathrm{d}g = 2x\,\mathrm{d}x$$
- This tells us how a small change $\mathrm{d}x$ affects $g(x)$.
- A **gradient** generalizes the derivative to multivariable functions $\mathbb{R}^d \to \mathbb{R}$.
- For $g(x, y)$, the gradient is a vector:
$$\nabla g = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$
- Points in the direction of greatest increase of the function.
- Magnitude of ∇f shows how steep the increase is.
- **Example:** $g(x, y) = x^2 + y^2 \Rightarrow \nabla f = (2x, 2y)$
- **Example:** $g(x, y) = x + y \Rightarrow \nabla f = (1, 1)$ everywhere, $\Pi_k = \{(x, y) \ : \ g(x, y) = k\}$ it is a plane Π (linear), and $\vec{u} = \nabla f \perp \Pi_k$

Gradients and Differentials

- The **chain rule** handles compositions of functions.
- If $y = g(u)$ and $u = g(x)$, then:

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u} \cdot \frac{\partial u}{\partial x} \text{ or } \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x}$$

- **Example:** If

$$y = \log\left(\sqrt{1 - x^2}\right)$$

- Let $u = \sqrt{1 - x^2}$, then:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x} = \frac{1}{\sqrt{1 - x^2}} \cdot \left(\frac{-x}{\sqrt{1 - x^2}}\right) = \frac{-x}{1 - x^2}$$

- So,

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\log\left(\sqrt{1 - x^2}\right)\right] = \frac{-x}{1 - x^2}$$

- In higher dimensions: use Jacobians or total derivatives.

Optimization

Machine learning has a lot to do with optimization.
And a lot of concepts can be related to optimization.
E.g.

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \underset{m\in\mathbb{R}}{\text{argmin}}\{\widehat{\mathcal{R}}_2(m)\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{ \sum_{i=1}^{n} \ell_2(y_i, m) \Big\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{ \sum_{i=1}^{n} (y_i - m)^2 \Big\}.$$

or, e.g.,

$$\text{median}(y) \in \underset{m\in\mathbb{R}}{\text{argmin}}\{\widehat{\mathcal{R}}_1(m)\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{ \sum_{i=1}^{n} \ell_1(y_i, m) \Big\} = \underset{m\in\mathbb{R}}{\text{argmin}}\Big\{ \sum_{i=1}^{n} |y_i - m| \Big\}.$$

Optimisation: continuous (differentiable) case

The problem is to solve $\min_{y \in \mathbb{R}} \{g(y)\}$

Note: $\min_{y \in \mathbb{R}} \{g(y)\} = \max_{y \in \mathbb{R}} \{-g(y)\}$

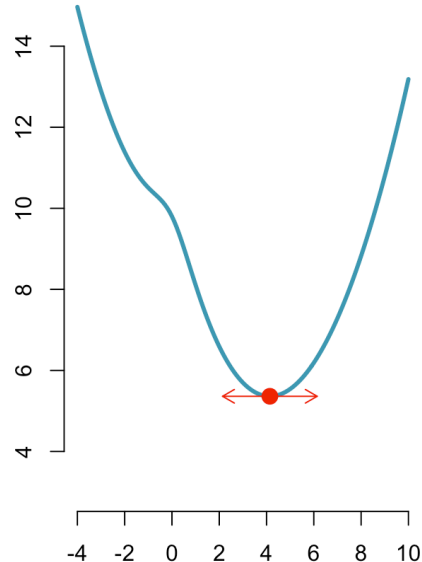
Note: $y^\star \in \operatorname{argmin}_{y \in \mathbb{R}} \{g(y)\}$

and $\min_{y \in \mathbb{R}} \{g(y)\} = g(y^\star)$.

First order condition

$$g'(y^\star) = \left.\frac{\partial g(y)}{\partial y}\right|_{y=y^\star} = 0$$

(necessary condition)



Optimisation: continuous (differentiable) case

First order condition

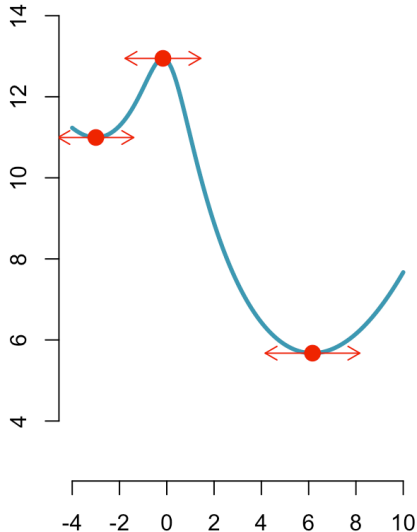
$$g'(y^\star) = \left.\frac{\partial g(y)}{\partial y}\right|_{y=y^\star} = 0$$

might be not sufficient

$$g''(y^\star) = \left.\frac{\partial^2 f}{\partial y^2}\right|_{y=y^\star} > 0 : \text{ minimum}$$

$$g''(y^\star) = \left.\frac{\partial^2 f}{\partial y^2}\right|_{y=y^\star} < 0 : \text{ maximum}$$

can be a local minimum...



Optimisation: continuous (differentiable) case

Example : $\{y_1, \cdots, y_n\}$ in $\mathbb{R}$, let

$$g(y) = \sum_{i=1}^{n} (y_i - y)^2$$

$$\frac{\partial g(y)}{\partial y} = \frac{\partial}{\partial y} \sum_{i=1}^{n} (y_i - y)^2 = \sum_{i=1}^{n} \frac{\partial (y_i - y)^2}{\partial y} = \sum_{i=1}^{n} -2(y_i - y)$$

so

$$\left.\frac{\partial g(y)}{\partial y}\right|_{y=y^\star} = 0 \text{ if and only if } \sum_{i=1}^{n} (y_i - y^\star) = 0 \text{ or } \sum_{i=1}^{n} y_i = n y^\star$$

i.e. $y^\star = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$.

Optimisation: continuous (differentiable) case

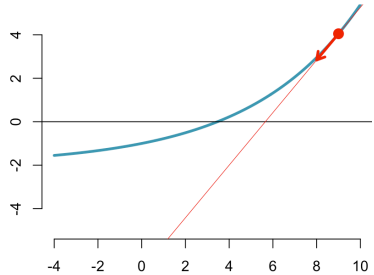
Solving $g'(y^\star) = 0$ numerically
Newton's method: solve $g(y^\star) = 0$

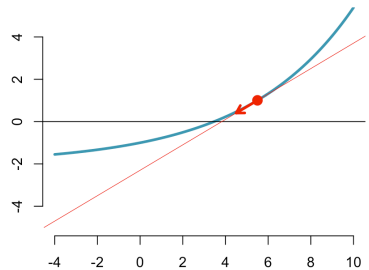
$$g(y) \simeq g(y_0) + g'(y_0)(y - y_0)$$

If $g(y) \simeq 0$, $g(y_0) + g'(y_0)(y - y_0) \simeq 0$

Start from y_0, then

$$y_{k+1} = y_k - \frac{g(y_k)}{g'(y_k)}$$



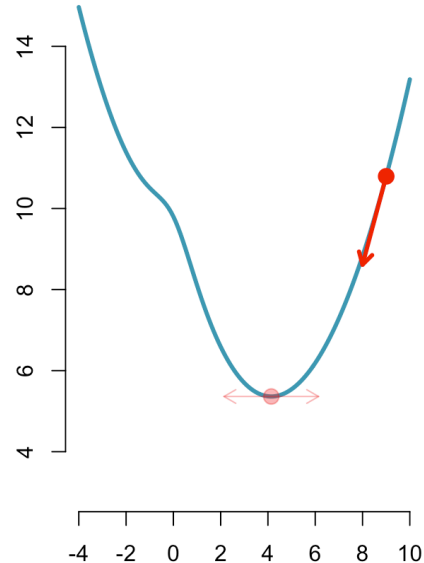


Optimisation: continuous (differentiable) case

To solve $g'(y^\star) = 0$ numerically
Start from y_0, then

$$y_{k+1} = y_k - \frac{g'(y_k)}{g''(y_k)}$$

$g'(y_k)$ gives the direction
$g''(y_k)$ gives the speed of convergence
(close to the minimum $g''(y_k) > 0$)



Optimisation: continuous (differentiable) case

In python

```
> import statistics as stat
> v=[0.89367,-1.04729,1.97133,-0.38363,1.65414]
> stat.mean(v)
0.617644
> import numpy as np
> def f0(x):
...   return np.sum((np.array(v)-x)**2)
> f = np.vectorize(f0)
> from scipy.optimize import fminbound
> fminbound(f, -1, 1)
0.6176439999999992
```

or

Optimisation: continuous (differentiable) case

```
> from scipy.optimize import minimize
> minimize(f, 0, method='nelder-mead')
 final_simplex: (array([[0.617625 ],
       [0.6176875]]), array([6.75753495, 6.75753495]))
           fun: 6.757534946524999
       message: 'Optimization terminated successfully.'
```

or

```
> def g(x):
...   s=[0]*len(v)
...   for i in range(len(v)):
...     s[i]=((v[i]-x)**2)
...   return sum(s)
> fminbound(f, -1, 1)
0.6176439999999992
```

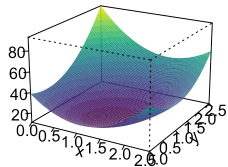
Optimisation: continuous (differentiable) case

and in R

```
> v = c(0.89367,-1.04729,1.97133,-0.38363,1.65414)
> mean(v)
[1] 0.617644
> f = function(x) sum((v-x)^2)
> optim(0, f)
$par
[1] 0.6175781
$value
[1] 6.757535
```

Optimisation: continuous (differentiable) case

The problem is $\min_{\boldsymbol{y}\in\mathbb{R}^p}\{g(\boldsymbol{y})\}$ or $\min_{(y_1,\cdots,y_p)\in\mathbb{R}^p}\{g(y_1,\cdots,y_p)\}$



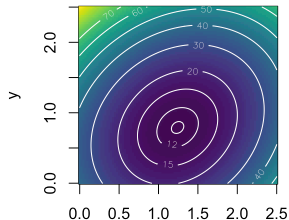
First order conditions: $\nabla g(\boldsymbol{y}^\star)=\boldsymbol{0}$,

$$\left.\frac{\partial g(y_1,y_2,\cdots,y_p)}{\partial y_1}\right|_{\boldsymbol{y}=\boldsymbol{y}^\star}=0$$

$$\left.\frac{\partial g(y_1,y_2,\cdots,y_p)}{\partial y_2}\right|_{\boldsymbol{y}=\boldsymbol{y}^\star}=0$$

$$\vdots$$

$$\left.\frac{\partial g(y_1,y_2,\cdots,y_p)}{\partial y_p}\right|_{\boldsymbol{y}=\boldsymbol{y}^\star}=0$$



Optimisation: continuous (differentiable) case

Example : $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ in $\mathbb{R}^2$, let

$$g(a, b) = \sum_{i=1}^{n} (y_i - [a + bx_i])^2$$

$$\frac{\partial g(a, b)}{\partial a} = -2 \sum_{i=1}^{n} (y_i - [a + bx_i]) = -2(n\bar{y} - [a + bn\bar{y}])$$

$$\frac{\partial g(a, b)}{\partial b} = -2 \sum_{i=1}^{n} (y_i - [a + bx_i]) x_i$$

$$\left.\frac{\partial g(a, b)}{\partial a}\right|_{(a,b)=(a^\star,b^\star)} = 0 \text{ means that } \bar{y} = a^\star + b^\star \bar{x},$$

$$\left.\frac{\partial g(a, b)}{\partial b}\right|_{(a,b)=(a^\star,b^\star)} = 0 \text{ means that } \widehat{\varepsilon} \perp \boldsymbol{x},\ \widehat{\varepsilon}_i = y_i - [a^\star + b^\star x_i],$$

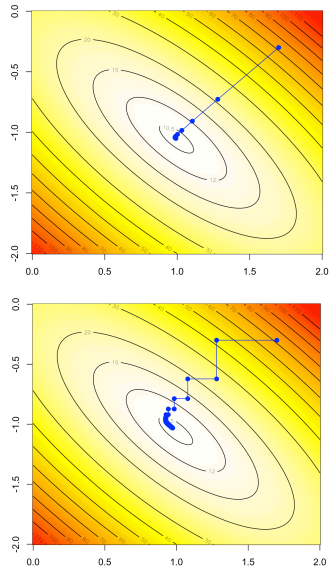
Optimisation: continuous (differentiable) case

To solve $\nabla g(\boldsymbol{y}^\star) = \mathbf{0}$ numerically
Start from $\boldsymbol{y}_0$, then

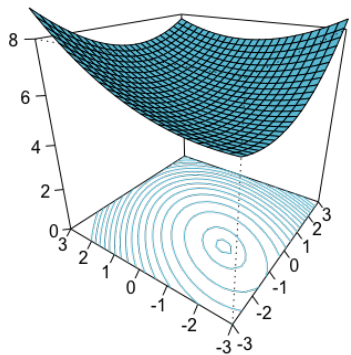
$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \boldsymbol{H}_k^{-1} \nabla g(\boldsymbol{y}_k)$$

$\nabla g(\boldsymbol{y}_k)$ gives the direction
$\boldsymbol{H}_k^{-1}$ gives the speed of convergence
$\boldsymbol{H}_k^{-1}$ is the inverse of the Hessian matrix

One can also consider some numerical tricks,
see coordinate descent
where we iterate on the dimension
(univariate optimisation problems)



Optimisation: continuous (differentiable) case

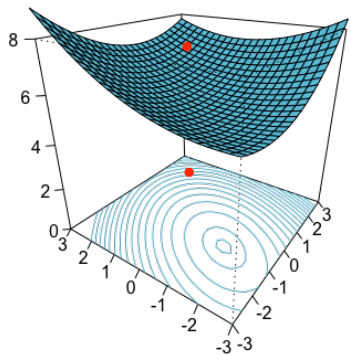


The problem is to find $\boldsymbol{y}^\star = \underset{\boldsymbol{y}\in\mathbb{R}^2}{\text{argmin}}\{g(\boldsymbol{y})\}$,

for some convex function g

corresponding to $\nabla g(\boldsymbol{y}^\star) = \mathbf{0}$

Optimisation: continuous (differentiable) case

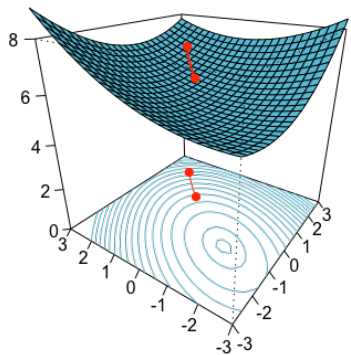


The problem is to find $\boldsymbol{y}^\star = \underset{\boldsymbol{y}\in\mathbb{R}^2}{\text{argmin}}\{g(\boldsymbol{y})\}$,

for some convex function g
corresponding to $\nabla g(\boldsymbol{y}^\star) = \mathbf{0}$

Start from $\boldsymbol{y}_0$,

Optimisation: continuous (differentiable) case



The problem is to find $\boldsymbol{y}^\star = \underset{\boldsymbol{y}\in\mathbb{R}^2}{\text{argmin}}\{g(\boldsymbol{y})\}$,

for some convex function g
corresponding to $\nabla g(\boldsymbol{y}^\star) = \boldsymbol{0}$

Start from $\boldsymbol{y}_0$, then

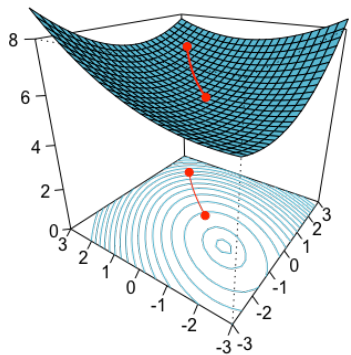
$$\boldsymbol{y}_1 = \boldsymbol{y}_0 - \boldsymbol{H}_0^{-1}\nabla g(\boldsymbol{y}_0)$$

where $\nabla g(\boldsymbol{y}_k)$ and $\boldsymbol{H}_k$ are

$$\begin{pmatrix} \dfrac{\partial g(y_1,y_2)}{\partial y_1} \\ \dfrac{\partial g(y_1,y_2)}{\partial y_2} \end{pmatrix}\Bigg|_{\boldsymbol{y}=\boldsymbol{y}_0} \quad \begin{pmatrix} \dfrac{\partial^2 g(y_1,y_2)}{\partial y_1^2} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_1 \partial y_2} \\ \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2 \partial y_1} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2^2} \end{pmatrix}$$

that are either know, or approximated

Optimisation: continuous (differentiable) case



The problem is to find $\boldsymbol{y}^\star = \underset{\boldsymbol{y}\in\mathbb{R}^2}{\text{argmin}}\{g(\boldsymbol{y})\}$,

for some convex function g
corresponding to $\nabla g(\boldsymbol{y}^\star) = \boldsymbol{0}$

Start from $\boldsymbol{y}_0$, then

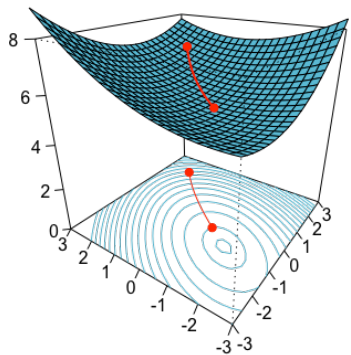
$$\boldsymbol{y}_2 = \boldsymbol{y}_1 - \boldsymbol{H}_1^{-1}\nabla g(\boldsymbol{y}_1)$$

where $\nabla g(\boldsymbol{y}_1)$ and $\boldsymbol{H}_1$ are

$$\left(\begin{array}{c}\dfrac{\partial g(y_1,y_2)}{\partial y_1}\\ \dfrac{\partial g(y_1,y_2)}{\partial y_2}\end{array}\right)\Bigg|_{\boldsymbol{y}=\boldsymbol{y}_1} \quad \left(\begin{array}{cc}\dfrac{\partial^2 g(y_1,y_2)}{\partial y_1^2} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_1\partial y_2}\\ \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2\partial y_1} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2^2}\end{array}\right)$$

that are either know, or approximated

Optimisation: continuous (differentiable) case



The problem is to find $\boldsymbol{y}^\star = \underset{\boldsymbol{y}\in\mathbb{R}^2}{\text{argmin}}\{g(\boldsymbol{y})\}$,

for some convex function g
corresponding to $\nabla g(\boldsymbol{y}^\star) = \boldsymbol{0}$

Start from $\boldsymbol{y}_0$, then iterate

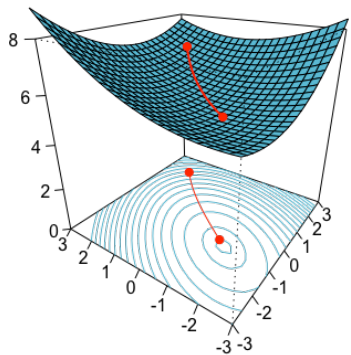
$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \boldsymbol{H}_k^{-1}\nabla g(\boldsymbol{y}_k)$$

where $\nabla g(\boldsymbol{y}_k)$ and $\boldsymbol{H}_k$ are

$$\left(\begin{array}{c}\dfrac{\partial g(y_1,y_2)}{\partial y_1}\\ \dfrac{\partial g(y_1,y_2)}{\partial y_2}\end{array}\right)\Bigg|_{\boldsymbol{y}=\boldsymbol{y}_k} \quad \left(\begin{array}{cc}\dfrac{\partial^2 g(y_1,y_2)}{\partial y_1^2} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_1\partial y_2}\\ \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2\partial y_1} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2^2}\end{array}\right)$$

that are either know, or approximated

Optimisation: continuous (differentiable) case



The problem is to find $\boldsymbol{y}^\star = \underset{\boldsymbol{y}\in\mathbb{R}^2}{\text{argmin}}\{g(\boldsymbol{y})\}$,

for some convex function g
corresponding to $\nabla g(\boldsymbol{y}^\star) = \boldsymbol{0}$

Start from $\boldsymbol{y}_0$, then iterate

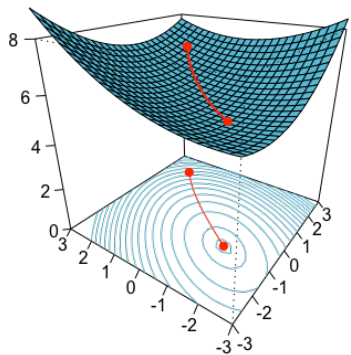
$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \boldsymbol{H}_k^{-1}\nabla g(\boldsymbol{y}_k)$$

where $\nabla g(\boldsymbol{y}_k)$ and $\boldsymbol{H}_k$ are

$$\begin{pmatrix} \dfrac{\partial g(y_1,y_2)}{\partial y_1} \\ \dfrac{\partial g(y_1,y_2)}{\partial y_2} \end{pmatrix}\Bigg|_{\boldsymbol{y}=\boldsymbol{y}_k} \quad \begin{pmatrix} \dfrac{\partial^2 g(y_1,y_2)}{\partial y_1^2} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_1 \partial y_2} \\ \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2 \partial y_1} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2^2} \end{pmatrix}$$

that are either know, or approximated

Optimisation: continuous (differentiable) case



The problem is to find $\boldsymbol{y}^\star = \underset{\boldsymbol{y}\in\mathbb{R}^2}{\text{argmin}}\{g(\boldsymbol{y})\}$,

for some convex function g
corresponding to $\nabla g(\boldsymbol{y}^\star) = \boldsymbol{0}$

Start from $\boldsymbol{y}_0$, then

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \boldsymbol{H}_k^{-1}\nabla g(\boldsymbol{y}_k)$$

where $\nabla g(\boldsymbol{y}_k)$ and $\boldsymbol{H}_k$ are

$$\left(\begin{array}{c}\dfrac{\partial g(y_1,y_2)}{\partial y_1}\\ \dfrac{\partial g(y_1,y_2)}{\partial y_2}\end{array}\right)\Bigg|_{\boldsymbol{y}=\boldsymbol{y}_k} \quad \left(\begin{array}{cc}\dfrac{\partial^2 g(y_1,y_2)}{\partial y_1^2} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_1\partial y_2}\\ \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2\partial y_1} & \dfrac{\partial^2 g(y_1,y_2)}{\partial y_2^2}\end{array}\right)$$

that are either know, or approximated

Constrained Optimisation: continuous case

Constrained optimization seeks to find the best solution to an objective function subject to constraints:

$$\max_x \{f(x)\} \quad \text{subject to} \quad \begin{cases} g_i(x) \leq 0, \ \forall i \\ h_j(x) = 0, \ \forall j, \end{cases}$$

where:

- $f(x)$ is the objective function.
- $g_i(x)$ are inequality constraints.
- $h_j(x)$ are equality constraints.

The Lagrangian function transforms a constrained problem into an unconstrained one:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x)$$

Constrained Optimisation: continuous case

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x)$$

where:

- λ_i are Lagrange multipliers for inequality constraints.
- μ_j are Lagrange multipliers for equality constraints.

Necessary conditions for a local optimum in constrained optimization:

- Stationarity: $\nabla f(x) + \sum_i \lambda_i \nabla g_i(x) + \sum_j \mu_j \nabla h_j(x) = 0$
- Primal feasibility: $g_i(x) \leq 0, \quad h_j(x) = 0$
- Dual feasibility: $\lambda_i \geq 0$
- Complementary slackness: $\lambda_i g_i(x) = 0$

so-called **Karush-Kuhn-Tucker** (KKT) conditions.

Constrained Optimisation: continuous case

For equality constraints only:

- Construct the Lagrangian:

$$\mathcal{L}(x, \mu) = f(x) + \sum_j \mu_j h_j(x)$$

- Solve $\nabla \mathcal{L} = 0$ to find x and μ.

For inequality constraints:

$$\lambda_i g_i(x) = 0$$

This means:

- If $g_i(x) < 0$, then $\lambda_i = 0$ (inactive constraint).
- If $g_i(x) = 0$, then $\lambda_i \geq 0$ (active constraint).

Constrained Optimisation: continuous case

The dual function:

$$q(\lambda, \mu) = \max_x \mathcal{L}(x, \lambda, \mu)$$

gives a lower bound on the optimal value:

$$q(\lambda, \mu) \leq f^*$$

- Weak Duality: $q(\lambda, \mu) \leq f^*$
- Strong Duality: $q(\lambda^*, \mu^*) = f^*$ (holds if Slater's condition is satisfied)

Slater's condition ensures strong duality for convex optimization problems. It states that if the problem:

$$\min g(x) \quad \text{subject to } g_i(x) \leq 0$$

is convex and there exists a strictly feasible point x_0 such that:

$$g_i(x_0) < 0 \quad \forall i$$

Constrained Optimisation: continuous case

then strong duality holds, meaning:

$$\max_{\lambda \geq 0} \min_{x} \{\mathcal{L}(x, \lambda)\} = \min_{x} \{f(x)\}$$

This guarantees that the primal and dual problems have the same optimal value.

Constrained Optimisation: continuous case

Quadratic Programming corresponds to minimizing

$$f(x) = \frac{1}{2}x^T Qx + c^T x$$

subject to:

$$Ax \leq b$$

Solution obtained using KKT conditions or interior-point methods.

- Constrained optimization incorporates constraints into objective minimization.
- The Lagrangian method helps transform constrained problems into unconstrained ones.
- KKT conditions provide necessary conditions for optimality.
- Duality theory helps analyze convex problems.

Convex Optimization Problem: Ridge

$$\mathcal{L}(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(x_i^\top\beta - y_i\right)^2 + \frac{\lambda}{2}\|\beta\|^2 = \frac{1}{2}(\boldsymbol{X}\beta - \boldsymbol{y})^\top(\boldsymbol{X}\beta - \boldsymbol{y}) + \frac{1}{2}\beta^\top\beta$$

Primal problem

$$\min_{\beta}\left\{\frac{1}{2}\beta^\top(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I}_p)\beta - \beta^\top\boldsymbol{X}^\top\boldsymbol{y}\right\}$$

The solution of the unconstrained quadratic program is $\widehat{\beta} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$

Dual problem

$$\min_{\alpha}\left\{\frac{1}{2}\alpha^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\mathbb{I}_n)\alpha - \lambda\alpha^\top\boldsymbol{y}\right\}$$

The solution is $\widehat{\alpha} = \lambda(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\mathbb{I}_n)^{-1}\boldsymbol{y}$, i.e. $\widehat{\beta} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\mathbb{I}_n)^{-1}\boldsymbol{y}$

Differentials and Gradients

Let $f : \mathbb{R}^n \to \mathbb{R}^k$, the "derivative" is $\nabla f(\boldsymbol{x})$, a $n \times k$ matrix, and

$$f(\boldsymbol{x} + \boldsymbol{u}) \sim f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{u} = f(\boldsymbol{x}) + \mathrm{d}f(\boldsymbol{x}; \boldsymbol{u})$$

where the differential $\mathrm{d}f$ is linear in $\boldsymbol{u}$.

Chain rule for differentials

If $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$

$$\mathrm{d}h(\boldsymbol{x}; \boldsymbol{u}) = \mathrm{d}g(f(\boldsymbol{x}); \mathrm{d}f(\boldsymbol{x}; \boldsymbol{u})) \ .$$

From chain rule to back-propagation

Differentials and Gradients

If $h(\boldsymbol{x}) = f_n(\cdots f_2(f_1(\boldsymbol{x}, \boldsymbol{\theta}_1), \boldsymbol{\theta}_2), \cdots, \boldsymbol{\theta}_n)$. Let

$$\boldsymbol{y}_0 = \boldsymbol{x}, \ \boldsymbol{y}_i = f_i(\boldsymbol{y}_{i-1}, \boldsymbol{\theta}_i), \text{ and } \boldsymbol{y}_n = h(\boldsymbol{x}).$$

Step 1

$$\frac{\partial h}{\partial \boldsymbol{y}_{n-1}} = \frac{\partial f_n(\boldsymbol{y}_{n-1}, \boldsymbol{\theta}_n)}{\partial \boldsymbol{y}_{n-1}} \text{ and } \frac{\partial h}{\partial \boldsymbol{\theta}_n} = \frac{\partial f_n(\boldsymbol{y}_{n-1}, \boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n}$$

then loop.

Back-propagation proceeds by obtaining the derivatives with respect to $\boldsymbol{\theta}_i$ and $\boldsymbol{y}_{i-1}$ from the derivative with respect to $\boldsymbol{y}_i$

$$\frac{\partial h}{\partial \boldsymbol{\theta}_i} = \frac{\partial h}{\partial \boldsymbol{y}_i} \cdot \frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{\theta}_i} = \frac{\partial h}{\partial \boldsymbol{y}_i} \cdot \frac{\partial f_i(\boldsymbol{y}_{i-1}, \boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i}$$

$$\frac{\partial h}{\partial \boldsymbol{y}_{i-1}} = \frac{\partial h}{\partial \boldsymbol{y}_i} \cdot \frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{y}_{i-1}} = \frac{\partial h}{\partial \boldsymbol{y}_i} \cdot \frac{\partial f_i(\boldsymbol{y}_{i-1}, \boldsymbol{\theta}_i)}{\partial \boldsymbol{y}_{i-1}}$$

(etc)

Differentials and Gradients

For **back-propagation**, compute

$$\frac{\partial h}{\partial \theta_i} = \left(\left(\cdots \left(\frac{\partial h}{\partial y_{n-1}} \times \frac{\partial y_{n-1}}{\partial y_{n-2}} \right) \times \cdots \right) \times \frac{\partial y_{i+1}}{\partial y_i} \right) \times \frac{\partial y_i}{\partial \theta_i} \quad ,$$

For **forward-propagation**, compute

$$\frac{\partial h}{\partial \theta_i} = \frac{\partial h}{\partial y_{n-1}} \times \left(\frac{\partial y_{n-1}}{\partial y_{n-2}} \times \left(\cdots \times \left(\frac{\partial y_{i+1}}{\partial y_i} \times \frac{\partial y_i}{\partial \theta_i} \right) \cdots \right) \right)$$

Convex Optimization Algorithms

$$\min_{\boldsymbol{x}}\{g(\boldsymbol{x})\}$$

with g convex, and differentiable.

Algorithm 1: Gradient Descent

initialization : $\boldsymbol{x}^{(0)}$;
for t=1,2,... **do**
  $\boldsymbol{x}^{(t)} \leftarrow \boldsymbol{x}^{(t-1)} - h_t \nabla g(\boldsymbol{x}^{(t-1)})$

Heuristics: Taylor expansion

$$g(\boldsymbol{y}) \sim g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2h}\|\boldsymbol{y}-\boldsymbol{x}\|^2$$

From Gradient Descent to Newton's Method

$$\min_{\boldsymbol{x}}\{g(\boldsymbol{x})\}$$

with g convex, twice differentiable.

Algorithm 2: Newton's Method

initialization : $\boldsymbol{x}^{(0)}$;
for $t=1,2,...$ **do**
  $\boldsymbol{H}_t \leftarrow \nabla^2 g(\boldsymbol{x}^{(t-1)})$;
  $\boldsymbol{x}^{(t)} \leftarrow \boldsymbol{x}^{(t-1)} - \boldsymbol{H}_t^{-1} g\ \nabla g(\boldsymbol{x}^{(t-1)})$

Use a better quadratic approximation in Taylor expansion – $\frac{1}{h}\mathbb{I} \ \rightarrow \ H$,

$$g(\boldsymbol{y}) \sim g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^\top \boldsymbol{H} g(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$$

Convex Optimization Problem

$$\min_{\boldsymbol{x}}\{g(\boldsymbol{x})\}$$

with g convex, but non-differentiable.

Algorithm 3: Subgradient 'Descent'

initialization : $\boldsymbol{x}^{(0)}$;
for t=1,2,... **do**
  $\boldsymbol{g}^{(t-1)} \in \partial g(\boldsymbol{x}^{(t-1)})$;
  $\boldsymbol{x}^{(t)} \leftarrow \boldsymbol{x}^{(t-1)} - h_t\, \boldsymbol{g}^{(t-1)}$

Note that it is not necessarily a descent, so pick

$$\boldsymbol{x}^\star = \text{argmin}\{g(\boldsymbol{x}^{(0)}), g(\boldsymbol{x}^{(1)}), g(\boldsymbol{x}^{(2)}), \cdots\}$$

Convex Optimization Problem

$$\min_{\boldsymbol{x}}\{g(\boldsymbol{x})\} \text{ or } \min_{\boldsymbol{x}}\{f_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\}$$

with f_1 and f_2 convex, but f_2 non-differentiable.

Algorithm 4: Proximal Gradient 'Descent'

initialization : $\boldsymbol{x}^{(0)}$;

for $t=1,2,...$ **do**

$$\gamma_h(x) = \frac{1}{h}(\boldsymbol{x} - \text{proximal}_{h,f_2}(\boldsymbol{x} - h\nabla f_1(\boldsymbol{x}));$$

$$\boldsymbol{x}^{(t)} \leftarrow \boldsymbol{x}^{(t-1)} - h_t\, \gamma_{h_{t-1}}(\boldsymbol{x}^{(t-1)}) = \text{proximal}_{h,f_2}(\boldsymbol{x}^{(t-1)} - h\nabla f_1(\boldsymbol{x}^{(t-1)})$$

γ_h is the "generalized gradient of g".
The trick is that $\text{proximal}_{h,f_2}(\cdot)$ usually has a closed form in most applications.

Constrained Convex Optimization

Let $C \subset \mathbb{R}^n$ denote a convex set, with g convex, and differentiable.

$$\min_{x \in C}\{g(x)\} \iff \min_{x}\{g(x) + I_C(x)\} \text{ where } I_C(x) = \begin{cases} 0 \text{ if } x \in C \\ \infty \text{ if } x \notin C \end{cases}$$

$$\text{proximal}_h(x) = \underset{z}{\text{argmin}} \left\{ \frac{1}{2h}\|x - z\|_2^2 + \lambda I_C(z) \right\} = \underset{z \in C}{\text{argmin}} \left\{ \frac{1}{2h}\|x - z\|_2^2 \right\}$$

i.e. $\text{proximal}_h(x)$ is the projection operator onto C, $\Pi_C(x)$

Algorithm 5: Projected Gradient Descent

initialization : $x^{(0)}$;
for t=1,2,... **do**
  $x^{(t)} \leftarrow \Pi_C(x^{(t-1)} - h_t \nabla f(x^{(t-1)}))$

Coordinate Descent

Let $\{\vec{e}_1, \cdots, \vec{e}_n\}$ denote the standard basis in $\mathbb{R}^n$,

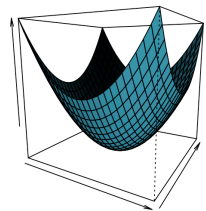
$$\vec{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0) \in \mathbb{R}^n$$

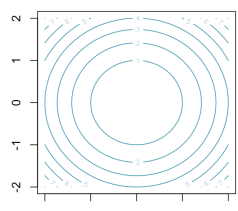
Proposition 1.35

If $g : \mathbb{R}^n \to \mathbb{R}$ is convex, differentiable,

$$g(\boldsymbol{x}) \leq g(\boldsymbol{x} + \delta \vec{e}_i),\ \forall i \Longrightarrow g(\boldsymbol{x}) = \min\{g\}$$

i.e. if we are at a point $\boldsymbol{x}$ such that $g(\boldsymbol{x})$ is minimized along each coordinate axis, then we have found a global minimizer.





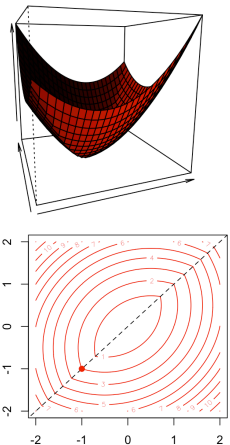
Coordinate Descent

Proposition 1.36

If $g : \mathbb{R}^n \to \mathbb{R}$ is convex, but not differentiable,

$$g(\boldsymbol{x}) \leq g(\boldsymbol{x} + \delta \vec{\boldsymbol{e}}_i),\ \forall i \Longrightarrow g(\boldsymbol{x}) = \min\{g\} = g^\star$$

i.e. if we are at a point $\boldsymbol{x}$ such that $g(\boldsymbol{x})$ is minimized along each coordinate axis, then we have not found a global minimizer.



Coordinate Descent

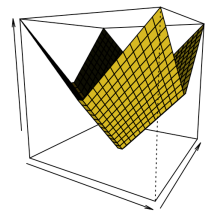
Proposition 1.37

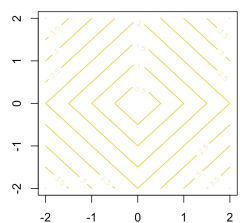
If $g : \mathbb{R}^n \to \mathbb{R}$ can be written, for f_0, f_i convex,

$$g(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \underbrace{\sum_{i=1}^{n} f_i(\boldsymbol{x}_i)}_{\text{separable}}, \quad \text{where} \begin{cases} f_0 \text{ differentiable} \\ f_i \text{ non-differentiable} \end{cases}$$

$$g(\boldsymbol{x}) \leq g(\boldsymbol{x} + \delta \vec{\boldsymbol{e}}_i), \ \forall i \Longrightarrow g(\boldsymbol{x}) = \min\{g\} = g^\star$$

i.e. if we are at a point $\boldsymbol{x}$ such that $g(\boldsymbol{x})$ is minimized along each coordinate axis, then we have found a global minimizer.





Coordinate Descent

If we want to solve $\min\{g(\boldsymbol{x})\}$ for some $g : \mathbb{R}^n \to \mathbb{R}$ such that

$$g(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \underbrace{\sum_{i=1}^{n} f_i(\boldsymbol{x}_i)}_{\text{separable}}, \quad \text{where} \begin{cases} f_0 \text{ convex and differentiable} \\ f_i \text{ convex and non-differentiable} \end{cases}$$

we can use a coordinate descent algorithm

Algorithm 6: Coordinate Descent

initialization : $\boldsymbol{x}^{(0)}$;
for $t=1,2,...$ **do**
  for $j=1,2,...,n$ **do**
    $\boldsymbol{x}_j^{(t)} \leftarrow \text{argmin}\{g(\boldsymbol{x}_1^{(t)}, \cdots, \boldsymbol{x}_{j-1}^{(t)}, x_j, \boldsymbol{x}_{j+1}^{(t-1)}, \cdots, \boldsymbol{x}_n^{(t-1)})\}$

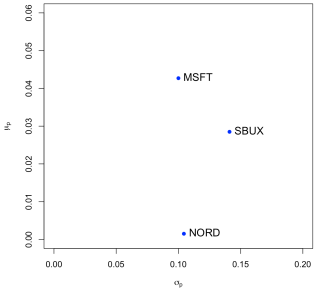
Portfolio Optimization

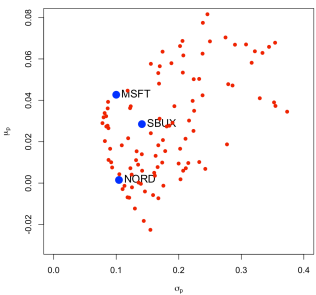
Consider n assets, with expected return $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

```
> asset.names = c("MSFT", "NORD", "SBUX")
> mu.vec = c(0.0427, 0.0015, 0.0285)
> names(mu.vec) = asset.names
> sigma.mat = matrix(c(0.0100, 0.0018,
    0.0113, 0.0018, 0.0109, 0.0026, 0.0113,
    0.0026, 0.0199), nrow=3, ncol=3)
> dimnames(sigma.mat) = list(asset.names,
    asset.names)
```

A portfolio is $\boldsymbol{\omega} \in \mathbb{R}^n$ with $\boldsymbol{\omega}^\top \mathbf{1} = 1$,

$$\mathbb{E}(P) = \boldsymbol{\omega}^\top \boldsymbol{\mu}, \ \text{Var}(P) = \boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega}$$





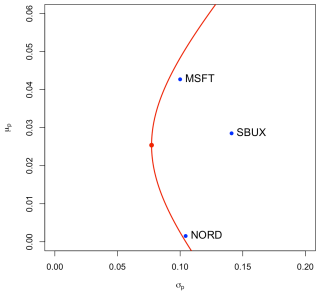
Portfolio Optimization – Markowitz (1952)

Efficient portfolio: $\omega_r^\star$,

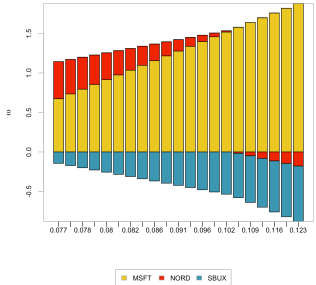
$$\omega_r^\star = \text{argmin}\{\omega^\top \boldsymbol{\Sigma} \omega\} \text{ s.t. } \begin{cases} \omega^\top \mathbf{1} = 1 \\ \omega^\top \boldsymbol{\mu} = r \end{cases}$$

or minimum variance portfolio

$$\omega^\star = \text{argmin}\{\omega^\top \boldsymbol{\Sigma} \omega\} \text{ s.t. } \omega^\top \mathbf{1} = 1$$




```
> one.vec = rep(1, 3)
> sigma.inv.mat = solve(sigma.mat)
> top.mat = sigma.inv.mat%*%one.vec
> bot.val = as.numeric((t(one.vec)%*%sigma.
    inv.mat%*%one.vec))
> m.mat = top.mat/bot.val
> m.mat[, 1]
[1]   0.674   0.470 -0.144
```



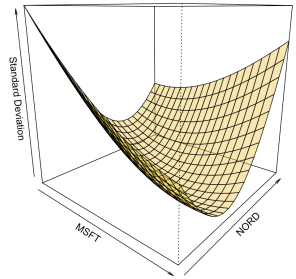
Possible only if short sells are possible,

Minimum Variance Portfolio – Markowitz (1952)

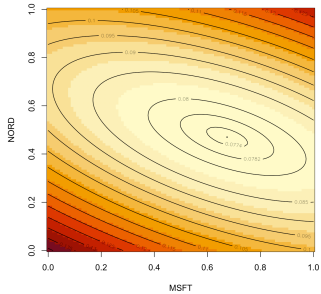
minimum variance portfolio

$$\boldsymbol{\omega}^\star = \text{argmin}\{\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega}\} \text{ s.t. } \boldsymbol{\omega}^\top \mathbf{1} = 1$$

$$\begin{pmatrix} \omega_M \\ \omega_N \\ 1-\omega_M-\omega_N \end{pmatrix}^\top \boldsymbol{\Sigma} \begin{pmatrix} \omega_M \\ \omega_N \\ 1-\omega_M-\omega_N \end{pmatrix}$$




```
> var_poids = function(z){
+   z.vec = cbind(z[1],z[2],1-z[1]-z[2])
+   sqrt((z.vec)%*%sigma.mat%*%t(z.vec))
+ }
> optim(par=c(1,1),var_poids)
$par
[1] 0.675 0.470
```

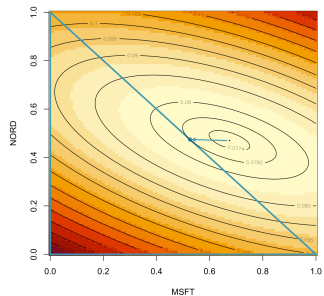


i.e. $\boldsymbol{\omega}^\star = (0.675, 0.470, -0.145)$

Minimum Variance Portfolio – Markowitz (1952)

with no-short sales,

$$\omega^\star_{NSS} = \text{argmin}\{\omega^\top \mathbf{\Sigma} \omega\} \text{ s.t. } \begin{cases} \omega^\top \mathbf{1} = 1 \\ \omega \geq \mathbf{0} \\ \omega \leq \mathbf{1} \end{cases}$$



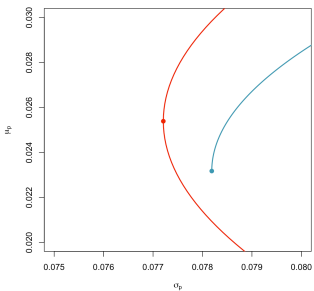

```
> grd_var_poids= function(z){
+   h=1e-5
+   c((var_poids(c(z[1]+h,z[2]))-var_poids(c
    (z[1]-h,z[2])))/(2*h), (var_poids(c(z
    [1],z[2]+h))-var_poids(c(z[1],z[2]-h)))
    /(2*h)) }
```

```
> constrOptim(c(.2,.2),var_poids,grd_var_poids, ui=matrix(c
    (1,0,0,1,-1,0,0,-1,-1,-1),5,2,byrow = TRUE), ci=c(0,0,-1,-1,-1))
$par
[1] 0.526 0.474
```

Portfolio Optimization – Markowitz (1952)

Efficient portfolio: $\boldsymbol{\omega}^\star_{r:NSS}$,

$$\boldsymbol{\omega}^\star_{r:NSS} = \text{argmin}\{\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega}\} \text{ s.t. } \begin{cases} \boldsymbol{\omega}^\top \mathbf{1} = 1 \\ \boldsymbol{\omega}^\top \boldsymbol{\mu} = r \\ \boldsymbol{\omega} \geq \mathbf{0} \\ \boldsymbol{\omega} \leq \mathbf{1} \end{cases}$$



Support Vector Machines

SVMs are a classification method that finds the optimal separating hyperplane:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

subject to:

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, \ \forall i$$

where $\boldsymbol{w}$ is the weight vector, and b is the bias.
Using Lagrange multipliers α_i, we define the Lagrangian:

$$\mathcal{L}(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_i \alpha_i [y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1]$$

By solving for $\boldsymbol{w}$ and b, we get the dual problem:

$$\max_{\boldsymbol{\alpha}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{w}_i^\top \boldsymbol{w}_j$$

Support Vector Machines

subject to:

$$\sum_i \alpha_i y_i = 0, \quad \alpha_i \geq 0$$

Support vectors are the data points for which $\alpha_i > 0$. They determine the decision boundary:

$$\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$$

For non-linearly separable data, we use kernel functions $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to transform the input space:

$$\max_{\boldsymbol{\alpha}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Common kernels:

- Polynomial: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j + c)^d$
- Gaussian (RBF): $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)$

Support Vector Machines

- SVMs rely on constrained optimization to find an optimal hyperplane.
- The Lagrangian formulation leads to the dual problem, reducing complexity.
- The kernel trick extends SVMs to non-linear classification problems.

Linear Algebra and Matrices

- A vector is an ordered list of numbers (or entries). It is represented as a column or as a row

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \text{ or } \boldsymbol{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

- Vectors are used to represent features in machine learning models (e.g., data points in multi-dimensional space).
- Common operations on vectors:
  - **Addition**: Adding two vectors element-wise.
  - **Dot product**: Computes a scalar value. $\boldsymbol{v} \cdot \boldsymbol{w} = \sum_{i=1}^{n} v_i w_i$.
  - **Scalar multiplication**: Multiplying a vector by a scalar.

Linear Algebra and Matrices

- In machine learning, vectors represent data points, and operations on vectors allow us to compute distances, projections, and transformations.
- A matrix is a two-dimensional array of numbers. It has rows and columns.

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}$$

- In machine learning, matrices represent datasets, where each row corresponds to an observation, and each column corresponds to a feature.
- Key operations on matrices:
  - **Matrix multiplication**: $\boldsymbol{AB}$, multiplying rows of $\boldsymbol{A}$ with columns of $\boldsymbol{B}$.
  - **Transpose**: The transpose of matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}^\top$, swaps its rows and columns.

- **Inverse**: If $\boldsymbol{A}$ is square, its inverse $\boldsymbol{A}^{-1}$ satisfies $\boldsymbol{A}\boldsymbol{A}^{-1} = \mathbb{I}$, where $\mathbb{I}$ is the identity matrix.

- Matrices are essential for representing data transformations and for operations in machine learning models.
- A **linear transformation** is a function that maps vectors to other vectors while preserving vector addition and scalar multiplication.
- Matrices can represent linear transformations. For example, applying matrix $\boldsymbol{A}$ to a vector $\boldsymbol{x}$ transforms the vector $\boldsymbol{x}$ into a new vector $\boldsymbol{y}$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

- In machine learning, linear transformations are used to:
  - Map data points from one space to another (e.g., in neural networks, during the forward pass).
  - Transform features in PCA or other dimensionality reduction techniques.

Linear Algebra and Matrices

- Key properties of linear transformations:
  - Preserves vector addition: $\boldsymbol{T}(\boldsymbol{x}+\boldsymbol{y}) = \boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{T}(\boldsymbol{y})$.
  - Preserves scalar multiplication: $\boldsymbol{T}(c\boldsymbol{x}) = c \cdot \boldsymbol{T}(\mathbf{x})$.
- An **eigenvector** $\boldsymbol{v}$ of a matrix $\boldsymbol{A}$ is a non-zero vector that satisfies the equation:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

  where λ is the corresponding **eigenvalue**.
- In machine learning, eigenvectors and eigenvalues are used in:
  - Principal Component Analysis (PCA) for dimensionality reduction.
  - Neural networks for understanding weight updates and activations.
- The matrix $\boldsymbol{A}$ acts by scaling the eigenvector $\boldsymbol{v}$ by the eigenvalue λ, without changing its direction.
- **Geometric interpretation**: Eigenvectors point in the direction of maximum variance or stretching, and eigenvalues measure the scaling factor.

Linear Algebra and Matrices

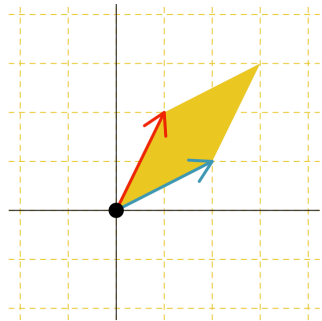
- **Matrix Decomposition** refers to breaking down a matrix into simpler components to facilitate computations and analysis.
- Common matrix decompositions include:
  - **Singular Value Decomposition (SVD)**: Decomposes a matrix into three matrices $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, used in PCA, recommendation systems, and low-rank approximations.
  - **QR Decomposition**: Decomposes a matrix into an orthogonal matrix $\boldsymbol{Q}$ and an upper triangular matrix $\boldsymbol{R}$, useful in solving linear systems and optimization.
  - **LU Decomposition**: Decomposes a matrix into a lower triangular matrix $\boldsymbol{L}$ and an upper triangular matrix $\boldsymbol{U}$, used in solving systems of linear equations.
- Applications in machine learning:
  - Dimensionality reduction via PCA (using SVD).
  - Solving linear systems in regression analysis.
  - Optimization algorithms for deep learning.

Determinant of square matrices (2×2)

Let $\mathbf{M} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$, then

$$\det(\mathbf{M}) = |\mathbf{M}| = \begin{vmatrix} a & c \\ b & d \end{vmatrix} = ad - bc$$

Note that $|\det(\mathbf{M})|$ is the area of the parallelogram, $\vec{\boldsymbol{u}} = \begin{pmatrix} a \\ b \end{pmatrix}$ and $\vec{\boldsymbol{v}} = \begin{pmatrix} c \\ d \end{pmatrix}$



Thus, $\det(\mathbf{M}) = 0$ if and only if $\vec{\boldsymbol{u}} /\!/ \vec{\boldsymbol{v}}$
Let θ denote the angle $(\vec{\boldsymbol{u}}, \vec{\boldsymbol{v}})$,

$$\cos(\theta) = \frac{\langle \vec{\boldsymbol{u}}, \vec{\boldsymbol{v}} \rangle}{\|\vec{\boldsymbol{u}}\| \cdot \|\vec{\boldsymbol{v}}\|} \text{ and } \sin(\theta) = \frac{\det(\mathbf{M})}{\|\vec{\boldsymbol{u}}\| \cdot \|\vec{\boldsymbol{v}}\|}, \text{ where } M = (\vec{\boldsymbol{u}}\ \vec{\boldsymbol{v}})$$

Inverse of square matrices (2×2)

If $\det(\mathbf{M}) \neq 0$, then $\mathbf{M}$ has an inverse, denoted $\mathbf{M}^{-1}$,

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{M}^{-1}\mathbf{M} = \mathbb{I}$$

Example: $\mathbf{M} = \begin{pmatrix} 1 & 2 \\ -2 & 0 \end{pmatrix}$ then $\mathbf{M}^{-1} = \frac{1}{4}\begin{pmatrix} 0 & -1 \\ 2 & 1 \end{pmatrix}$

Important for linear systems: find $\boldsymbol{x}$ such that $\mathbf{M}\boldsymbol{x} = \boldsymbol{a}$,
then the unique solution is $\boldsymbol{x} = \mathbf{M}^{-1}\boldsymbol{a}$

Example: $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ then $\mathbf{\Sigma}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$, if $|\rho| < 1$.
(here $\mathbf{\Sigma}$ is a variance matrix)

A projection matrix cannot be inverted

Eigenvalues and Eigenvectors of square matrices

Let $\mathbf{M} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$, $\vec{\boldsymbol{u}} = \begin{pmatrix} x \\ y \end{pmatrix}$ and λ such that $\mathbf{M}\vec{\boldsymbol{u}} = \lambda\vec{\boldsymbol{u}}$,

$\vec{\boldsymbol{u}}$ is the eigenvector associated with eigenvalue λ.

$$\begin{pmatrix} a-\lambda & c \\ b & d-\lambda \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ implies } \det\begin{pmatrix} a-\lambda & c \\ b & d-\lambda \end{pmatrix} = 0.$$

i.e. solve (in λ) $\det(M - \lambda\mathbb{I}) = 0$.
This can be extended in higher dimension

If $\det(\mathbf{M}) = 0$ then $\lambda = 0$ is an eigenvalue.

If $\mathbf{M}$ is symmetric, eigenvectors are orthogonal.

$\mathbf{M}$ is said to be positive if all eigenvalues are positive, and then $\boldsymbol{z}^\top\mathbf{M}\boldsymbol{z} \geq 0$, $\forall \boldsymbol{z}$.

Example: Let $\boldsymbol{X}$ be a $n \times m$ matrix, then $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^\top$ are symmetric and positive matrices (think of $\boldsymbol{X}^\top\boldsymbol{X}$ as a covariance matrix).

Quadratic Forms

Consider $\mathbf{M} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$,

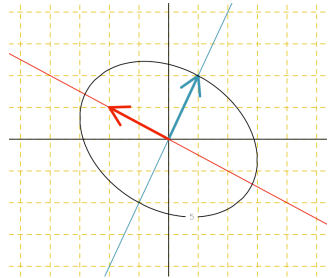
and function $\boldsymbol{z} \mapsto \boldsymbol{z}^\top \mathbf{M} \boldsymbol{z}$, i.e.

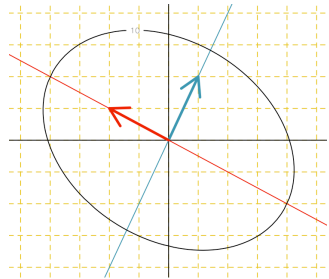
$$f : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

or $ax^2 + 2bxy + cy^2$ is a quadratic form.

If $\det(\mathbf{M}) > 0$, points $\boldsymbol{z} = (x, y)$ such that $\boldsymbol{z}^\top \mathbf{M} \boldsymbol{z} = \gamma$, for some $\gamma > 0$, are on an ellipse (centered on $\mathbf{0}$)

Let $\lambda_1 \geq \lambda_2 > 0$ denote the eigenvalues of $\mathbf{M}$ and $\vec{\boldsymbol{v}}_1$ and $\vec{\boldsymbol{v}}_2$ denote the eigenvectors.





Quadratic Forms

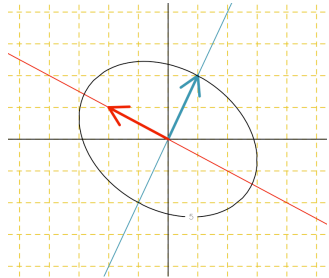
On the picture, $\mathbf{M} = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.9 \end{pmatrix}$

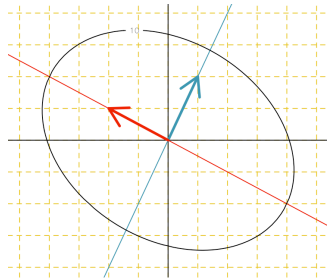
in Python

```
> import numpy as np
> from numpy import linalg
> M = np.array([[.6,.2], [.2, .9]])
> l, v = linalg.eig(M)
> print(l)
[ 0.5  1. ]
> print(v)
[[-0.89442719 -0.4472136 ]
 [ 0.4472136  -0.89442719]]
```

i.e. $\lambda_1 = 1/2$ and $\lambda_2 = 1$, and

$$\vec{v}_1 = \begin{pmatrix} -2\sqrt{5} \\ \sqrt{5} \end{pmatrix} = \sqrt{5}\begin{pmatrix} -2 \\ 1 \end{pmatrix}, \ \vec{v}_2 = \sqrt{5}\begin{pmatrix} -1 \\ -2 \end{pmatrix}$$





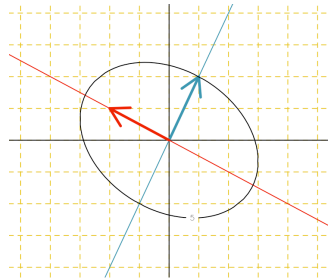
Quadratic Forms

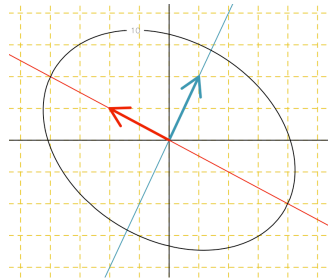
On the picture, $\mathbf{M} = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.9 \end{pmatrix}$

in R

```
> M=matrix(c(.6,.2,.2,.9),2,2)
> eigen(M)
eigen() decomposition
$values
[1] 1.0 0.5
$vectors
          [,1]       [,2]
[1,] 0.4472136 -0.8944272
[2,] 0.8944272  0.4472136
```

Note that $\|\vec{v}_1\| = \|\vec{v}_2\| = 1$ and $\vec{v}_1 \perp \vec{v}_2$





Spectral & Singular Value Decomposition

Let $\mathbf{M}$ denote a symmetric $d \times d$, with eigenvalues $\lambda_1, \cdots, \lambda_d$ and eigenvectors $\vec{\boldsymbol{u}}_1, \cdots, \vec{\boldsymbol{u}}_d$ with $\|\vec{\boldsymbol{u}}_j\| = 1$,

$$\underset{d\times d}{\mathbf{M}} = \underset{d\times d}{\begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_d \\ | & | & & | \end{pmatrix}} \underset{d\times d}{\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}} \underset{d\times d}{\begin{pmatrix} -\boldsymbol{u}_1^\top - \\ -\boldsymbol{u}_2^\top - \\ \vdots \\ -\boldsymbol{u}_d^\top - \end{pmatrix}}$$

i.e. $\mathbf{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. Let $\mathbf{M}$ denote $p \times q$, with $p \leq q$, $\mathbf{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, with $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p \in \mathbb{R}^p$ and $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_p \in \mathbb{R}^q$

$$\underset{p\times q}{\mathbf{M}} = \underset{p\times p}{\begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_p \\ | & | & & | \end{pmatrix}} \underset{p\times p}{\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}} \underset{p\times q}{\begin{pmatrix} -\boldsymbol{v}_1^\top - \\ -\boldsymbol{v}_2^\top - \\ \vdots \\ -\boldsymbol{v}_p^\top - \end{pmatrix}}$$

Trace of square matrices

Definition 1.52: Trace of square matrix

The trace of a square matrix $\boldsymbol{A}$, denoted $\text{trace}(\boldsymbol{A})$ is the sum of the elements on its main diagonal,

$$\text{trace}(\boldsymbol{A}) = A_{11} + A_{22} + \cdots + A_{nn}.$$

The trace of a matrix is the sum of its eigenvalues.
The trace of a projection matrix is the dimension of the target space.

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\top} \implies \text{trace}\left(\mathbf{P}_{\mathbf{X}}\right) = \text{rank}(\mathbf{X})$$

Definition 1.53: Froebenius inner product and norm

$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \text{trace}(\boldsymbol{A}^{\top}\boldsymbol{B})$ the Frobenius inner product of $\boldsymbol{A}$ and $\boldsymbol{B}$

Approximations

Consider $\mathbf{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ et $\tilde{\mathbf{M}} = \boldsymbol{U}\tilde{\boldsymbol{\Lambda}}\boldsymbol{V}^\top$ where $\tilde{\boldsymbol{\Lambda}} = \text{diag}(\lambda_1, \lambda_2, \mathbf{0})$.

$$\begin{pmatrix} 91 & 107 & 136 & 4 \\ 91 & 82 & 86 & 81 \\ 31 & 33 & -52 & 300 \\ 105 & 86 & 97 & 28 \\ 84 & 80 & 78 & 76 \\ 96 & 97 & 124 & 15 \end{pmatrix} \simeq \begin{pmatrix} 103.5 & 101.4 & 130.5 & 2.2 \\ 87.9 & 85.7 & 85.6 & 80.9 \\ 33.5 & 31.7 & -53.0 & 299.7 \\ 91.1 & 89.1 & 105.6 & 30.8 \\ 82.0 & 80.0 & 79.6 & 76.5 \\ 99.3 & 97.2 & 121.3 & 14.1 \end{pmatrix}$$

The old dataset $\mathbf{M}$ is $n = 6$ observations, in dimension 4
The new dataset $\tilde{\mathbf{M}}$ is $n = 6$ observations, in (real) dimension 2

Réduction de dimension, cas linéaire

- If $\mathbf{M}$ is a matrix $n \times n$, $\vec{\boldsymbol{v}}$ is a proper vector for $\mathbf{M}$ (associated to λ) if $\mathbf{M}\vec{\boldsymbol{v}} = \lambda\vec{\boldsymbol{v}}$
- If $\mathbf{M}$ is diagonalize, $\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ where $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, et $\mathbf{V} = [\vec{\boldsymbol{v}}_1, \cdots, \vec{\boldsymbol{v}}_n]$.
- If $\mathbf{M}$ is invertible, $\mathbf{M}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top$ where $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, et $\mathbf{V} = [\vec{\boldsymbol{v}}_1, \cdots, \vec{\boldsymbol{v}}_n]$.
- **Eckart-Young-Mirsky** theorem (Eckart and Young (1936), Mirsky (1960)):

$$\mathbf{M}^\star \in \underset{\mathbf{H}:n\times n}{\text{argmin}}\left\{\|\mathbf{M}-\mathbf{H}\|_F \text{ such that } \text{rank}(\mathbf{H}) \leq r\right\}$$

If $\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top = \begin{bmatrix}\mathbf{V}_r & \mathbf{V}_{n-r}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_r & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_{n-r}\end{bmatrix}\begin{bmatrix}\mathbf{V}_r & \mathbf{V}_{n-r}\end{bmatrix}^\top$, $\mathbf{M}^\star = \mathbf{V}_r\boldsymbol{\Lambda}_r\mathbf{V}_r^\top$

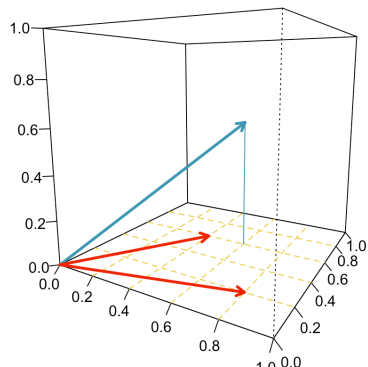
(which will be unique if $\lambda_r > \lambda_{r+1}$)

- The approximation error verifies $\|\mathbf{M}-\mathbf{M}^\star\|_F = \sqrt{\lambda_{r+1}^2 + \cdots + \lambda_n^2}$.

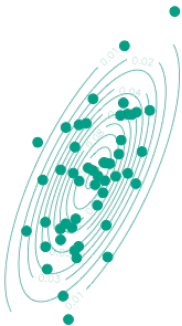
Projection

Consider the projection (in $\mathbb{R}^3$) on $\{\vec{x}_1, \vec{x}_2\}$. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$.
$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is the (orthogonal) projection on $\{\vec{x}_1, \vec{x}_2\}$

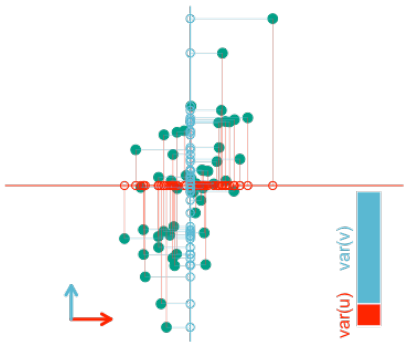
```
> X=cbind(c(.8,.2,0),c(.4,.6,0))
> P=X %*% solve(t(X)%*%X) %*% t(X)
> P
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    0
> P %*% c(.6,.6,0.6)
      [,1]
[1,]  0.6
[2,]  0.6
[3,]  0.0
```

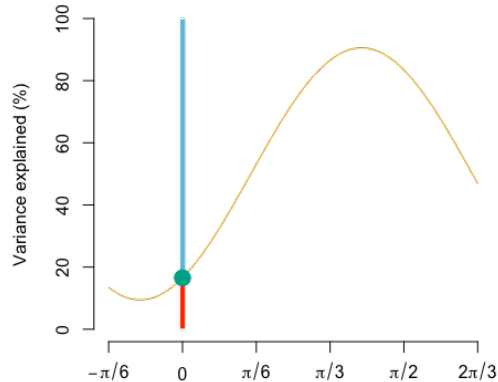


Projection and Variance

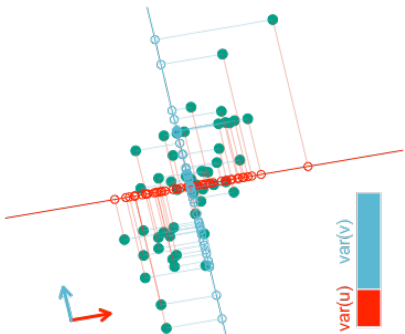


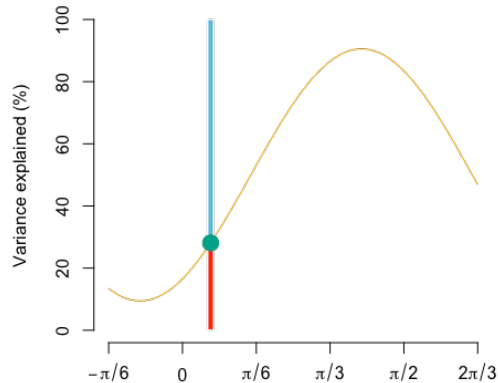
Projection and Variance



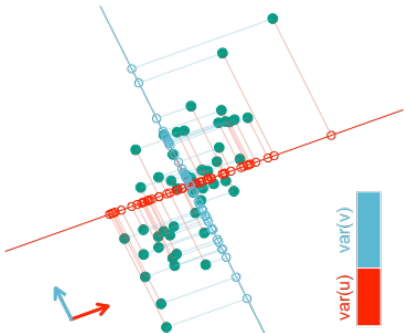


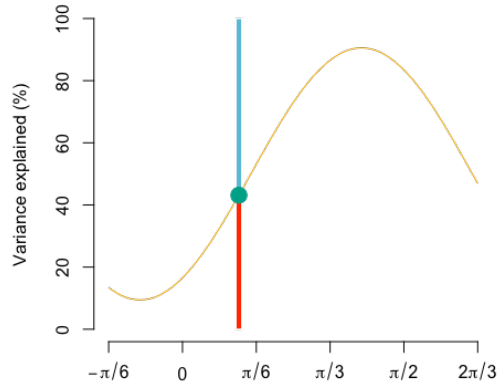
Projection and Variance



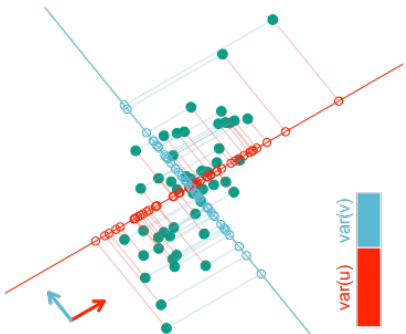


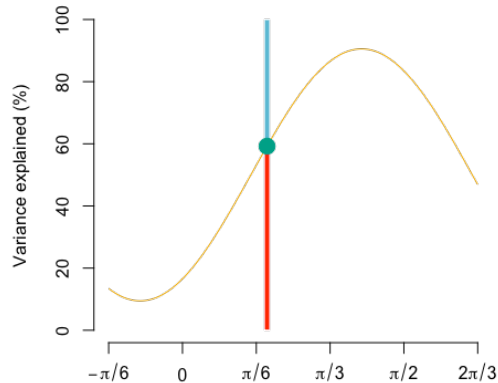
Projection and Variance



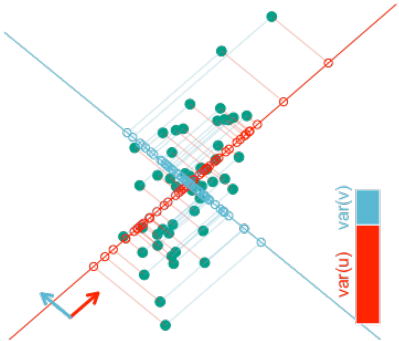


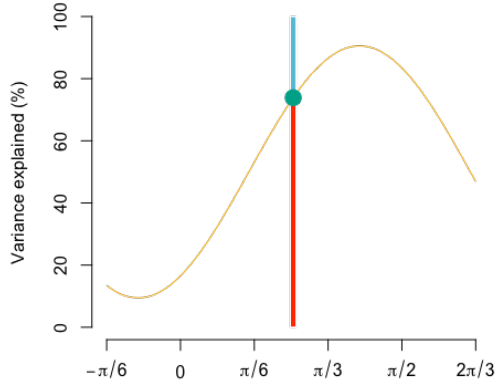
Projection and Variance



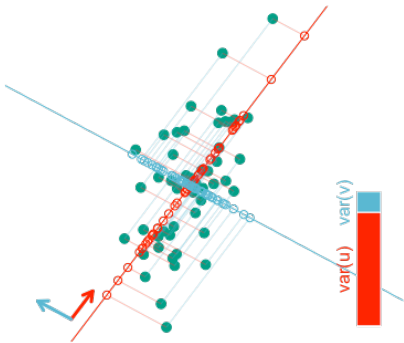


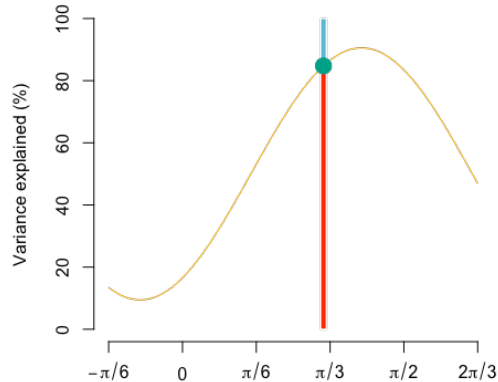
Projection and Variance



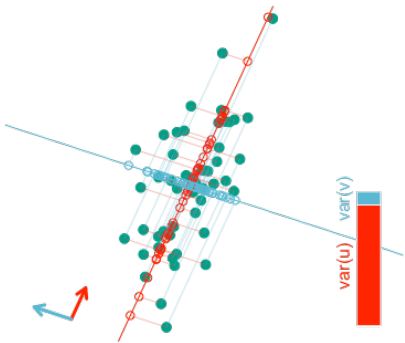


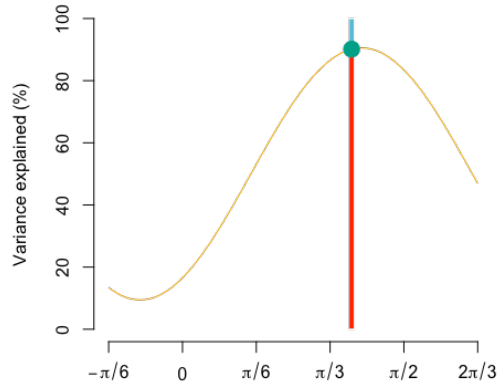
Projection and Variance





Projection and Variance





Projection and Variance

Proposition 1.38: Eigenvectors, optimization and projection

iven a symmetric $n \times n$ matrix $\boldsymbol{M}$, the eigenvector $\vec{\boldsymbol{v}}_1$ corresponding to the largest eigenvalue λ_1 can be found as the solution to the following optimization problem

$$\max_{\vec{\boldsymbol{v}}:\ \|\vec{\boldsymbol{v}}\|=1} \left\{ \boldsymbol{v}^\top \boldsymbol{M} \boldsymbol{v} \right\}.$$

Rotation & Orthogonal Matrices

> **Orthogonal matrix**
> In linear algebra, an orthogonal matrix, or orthonormal matrix, is a real square matrix whose columns and rows are orthonormal vectors. One way to express this is $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\top = \mathbb{I}$. W

One can use matrices to transform vectors, e.g. $\vec{\boldsymbol{y}} = \boldsymbol{A}\vec{\boldsymbol{x}}$, with $\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}} \in \mathbb{R}^n$, and $\boldsymbol{A}$ is some $n \times n$ matrix.

Example:

$$\boldsymbol{A}\vec{\boldsymbol{x}} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{bmatrix}.$$

If $\boldsymbol{A} = R_{\boldsymbol{0}}(\theta)$, $\boldsymbol{A}^\top = R_{\boldsymbol{0}}(-\theta) = \boldsymbol{A}^{-1}$

Rotation & Orthogonal Matrices

Example: $\boldsymbol{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = R_{\mathbf{0}}(\theta)$ and $\boldsymbol{B} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} = R_{\mathbf{0}}(\phi)$ then

$$\boldsymbol{AB} = \begin{bmatrix} \cos\theta\cos\phi - \sin\theta\sin\phi & -\cos\theta\sin\phi - \sin\theta\cos\phi \\ \sin\theta\cos\phi + \cos\theta\sin\phi & -\sin\theta\sin\phi + \cos\theta\cos\phi \end{bmatrix} \text{ i.e.}$$

$$\boldsymbol{AB} = \begin{bmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{bmatrix} = R_{\mathbf{0}}(\theta+\phi)$$

In higher dimension n, a $n \times n$ matrix $\boldsymbol{A}$ is orthogonal if its columns and rows are orthogonal unit vectors, i.e.

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \mathbb{I}$$

or equivalently, $\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$.

In dimension 2,

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \text{rotation}, \quad \text{and} \quad \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} = \text{reflection}$$

Other Important Matrices

- A row-stochastic matrix is a square matrix where each row sums to 1, and all entries are non-negative.

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \quad \text{with} \quad \sum_{j=1}^{n} A_{ij} = 1 \quad \forall i$$

- Key properties:
  - Non-negative entries ($A_{ij} \geq 0$).
  - Row sums equal to 1: $\sum_{j} A_{ij} = 1$ for each row.

- **Application in Machine Learning**:
  - **Markov Chains**: Transition probability matrices are row-stochastic matrices. Each row represents the probability of transitioning from one state to another.

Other Important Matrices

- A doubly stochastic matrix is a square matrix where:
  - Each row sums to 1 (row-stochastic).
  - Each column sums to 1 (column-stochastic).

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \quad \text{with} \quad \sum_{j=1}^{n} A_{ij} = 1 \quad \forall i, \quad \sum_{i=1}^{m} A_{ij} = 1 \quad \forall j$$

- Key properties:
  - Non-negative entries.
  - Both row sums and column sums equal to 1.
- Application in Machine Learning:
  - Sinkhorn Algorithm: Used in optimal transport problems, where doubly stochastic matrices are used to balance marginal distributions.

Other Important Matrices

- Softmax: In the softmax function used in classification tasks, the output probabilities can be represented by a doubly stochastic matrix.

- A permutation matrix is a square matrix that has exactly one entry of 1 in each row and column, with all other entries being 0.

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

- Key properties:
  - The rows and columns are permutations of the standard basis vectors.
  - $\boldsymbol{P}^\top = \boldsymbol{P}^{-1}$, meaning the inverse of a permutation matrix is its transpose.
- Application in Machine Learning:
  - Data Shuffling: Permutation matrices are used in data preprocessing to shuffle datasets randomly while preserving the structure of the data.

Other Important Matrices

- Attention Mechanisms: In transformer models, permutation matrices are used to shuffle input sequences or to reorder elements based on attention scores.

- A projection matrix projects vectors onto a subspace.

$$\boldsymbol{P}^2 = \boldsymbol{P}, \quad \boldsymbol{P} = \boldsymbol{P}^\top \quad \text{(for an orthogonal projection matrix)}$$

- Key properties:
  - $\boldsymbol{P}$ is idempotent: $\boldsymbol{P}^2 = \boldsymbol{P}$.
  - For orthogonal projections, $\boldsymbol{P}$ is symmetric: $\boldsymbol{P} = \boldsymbol{P}^T$.
- Application in Machine Learning:
  - Linear Regression: The projection matrix in ordinary least squares (OLS) regression projects the target vector onto the subspace spanned by the predictors.
  - PCA: The matrix that projects data onto principal components is a projection matrix.

Other Important Matrices

- Both row-stochastic and doubly stochastic matrices are used to represent graph-based data:
  - A row-stochastic matrix can represent the transition probabilities in a Markov chain on a graph.
  - A doubly stochastic matrix can be used to represent balanced flows in networks, where both rows and columns represent marginal distributions.
- Application in Machine Learning:
  - Graph Neural Networks (GNNs): Markov chains and transition matrices (row-stochastic) are foundational in GNNs for message passing.
  - Optimal Transport: The Sinkhorn algorithm uses doubly stochastic matrices to solve optimal transport problems, which is applied in generative models and domain adaptation.

References

Adjemian, S. and Pelgrin, F. (2008). Un regard bayésien sur les modèles dynamiques de la macroéconomie. *Economie prevision*, (2):127–152.

Anscombe, F. J., Aumann, R. J., et al. (1963). A definition of subjective probability. *Annals of mathematical statistics*, 34(1):199–205.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(10).

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bauschke, H. H., Borwein, J. M., et al. (1997). Legendre functions and the method of random bregman projections. *Journal of convex analysis*, 4(1):27–67.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the Royal Society of London*, (53):370–418.

References

Bell, E. T. (1945). *The development of mathematics*. Courier Corporation.

Bernoulli, J. (1713). *Ars conjectandi: opus posthumum: accedit Tractatus de seriebus infinitis; et Epistola gallice scripta de ludo pilae reticularis*. Impensis Thurnisiorum.

Black, F. and Litterman, R. (1990). Asset allocation: combining investor views with market equilibrium. *Goldman Sachs Fixed Income Research*, 115.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial analysts journal*, 48(5):28–43.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Brualdi, R. A. (2006). *Combinatorial matrix classes*, volume 13. Cambridge University Press.

References

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin: The Journal of the IAA*, 4(3):199–207.

Cardano, G. (1564). *Liber de ludo aleae*. Franco Angeli.

Carlier, G., Chernozhukov, V., and Galichon, A. (2016). Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44.

Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités*. Hachette.

Cramér, H. (1928a). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Cramér, H. (1928b). On the composition of elementary errors: second paper: statistical applications. *Scandinavian Actuarial Journal*, 1928(1):141–180.

Dall'Aglio, G. (1956). Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 10(1-2):35–74.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge university press.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

References

Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. (2006). *Actuarial theory for dependent risks: measures, orders and models*. John Wiley & Sons.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Fenton, N. and Neil, M. (2018). *Risk assessment and decision analysis with Bayesian networks*. CRC Press.

Fenton, N., Neil, M., and Berger, D. (2016). Bayes and the law. *Annual Review of Statistics and Its Application*, 3:51.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Fisher, R. (1921). Studies in crop variation. i. an examination of the yield of dressed grain from broadbalk. *The Journal of Agricultural Science*, 11(2):107–135.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A*, 222(594-604):309–368.

Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.

References

Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Verlag.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005). Kernel constrained covariance for dependence measurement. In *International Workshop on Artificial Intelligence and Statistics*, pages 112–119. PMLR.

Hájek, A. (2002). Interpretations of probability. *Stanford Encyclopedia of Philosophy*.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165.

References

Hallin, M. and Konen, D. (2024). Multivariate quantiles: Geometric and measure-transportation-based contours.

Hannan, E. J. (1961). The general theory of canonical correlation and its relation to functional analysis. *Journal of the Australian Mathematical Society*, 2(2):229–242.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1952). *Inequalities*. Cambridge university press.

Harris, T. and Ross, F. (1955). Fundamentals of a method for evaluating rail net capacities. Technical report.

He, X. D., Kou, S., and Peng, X. (2022). Risk measures: robustness, elicitability, and backtesting. *Annual Review of Statistics and Its Application*, 9:141–166.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. SIAM.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524.

Hunt, I. and Mostyn, J. (2020). Probability reasoning in judicial fact-finding. *The international Journal of evidence & Proof*, 24(1):75–94.

References

Jeffrey, R. (2004). *Subjective probability: The real thing*. Cambridge University Press.

Jensen, D. and Mayer, L. (1977). Some variational results and their applications in multiple inference. *The Annals of Statistics*, pages 922–931.

Jonakait, R. N. (1983). When blood is their argument: probabilities in criminal cases, genetic markers, and, once again, bayes' theorem. *University of Illinois Law Review*, page 369.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454.

Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201.

Kimeldorf, G., May, J. H., and Sampson, A. R. (1982). Concordant and discordant monotone correlations and their evaluation by nonlinear optimization. *Studies in the Management Sciences*, 19:117–130.

Kimeldorf, G. and Sampson, A. R. (1978). Monotone dependence. *The Annals of Statistics*, pages 895–903.

Klugman, S. A. (1991). *Bayesian statistics in actuarial science: with emphasis on credibility*, volume 15. Springer Science & Business Media.

References

Knott, M. and Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49.

Kolmogorov, A. (1933). Grundbegriffe der wahrscheinlichkeitsrechnung.

Kremer, W. (2014). Do doctors understand test results. *BBC World Service*.

Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, 44(1/2):289–292.

Lancaster, H. O. (1958). The Structure of Bivariate Distributions. *The Annals of Mathematical Statistics*, 29(3):719 – 736.

Laplace, P. S. (1774). Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*.

Lebesgue, H. (1918). Remarques sur les théories de la mesure et de l'intégration. *Annales scientifiques de l'École Normale Supérieure*, 35:191–250.

Lin, P.-E. (1987). Measures of association between vectors. *Communications in Statistics-Theory and Methods*, 16(2):321–338.

Loève, M. (1977). *Probability Theory*. Springer.

Longley-Cook, L. H. (1962). An introduction to credibility theory. Casualty Actuarial Society.

References

Martin, T. (2009). la probabilité, un concept pluriel. *Pour la science)*, (385):46–50.

Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Mourier, E. (1953). Eléments aléatoires dans un espace de banach. In *Annales de l'institut Henri Poincaré*, volume 13, pages 161–244.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, pages 97–131.

Orbanz, P. and Teh, Y. W. (2010). Bayesian nonparametric models. *Encyclopedia of machine learning*, 1.

Pardo, L. (2018). *Statistical inference based on divergence measures*. CRC press.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the royal society of London*, 58(347-352):240–242.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Popper, K. R. (1959). The propensity interpretation of probability. *The British journal for the philosophy of science*, 10(37):25–42.

References

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214.

Reichenbach, H. (1971). *The theory of probability*. University of California Press.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451.

Saini, A. (2011). A formula for justice. *The Guardian*, October 2nd.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*, volume 87. Springer.

Sarmanov, O. (1958a). Maximum correlation coefficient (non-symmetrical case). *Doklady Akademii Nauk SSSR*, 121(1):52–55.

Sarmanov, O. V. (1958b). The maximum correlation coefficient (symmetrical case). *Doklady Akademii Nauk SSSR*, 120(4):715–718.

Satchell, S. and Scowcroft, A. (2000). A demystification of the black–litterman model: Managing quantitative and traditional portfolio construction. *Journal of Asset Management*, 1(2):138–150.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

References

Sinkhorn, R. (1962). On the factor spaces of the complex doubly stochastic matrices. *Notices of the American Mathematical Society*, 9:334–335.

Székely, G. J. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18.

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Society.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer.

von Mises, R. (1939). *Probability, statistics and truth*. Macmillan.

Wasserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

Whitney, A. W. (1918). Theory of experience rating. In *Proceedings of the Casualty Actuarial society*, volume 1, pages 274–292.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62.