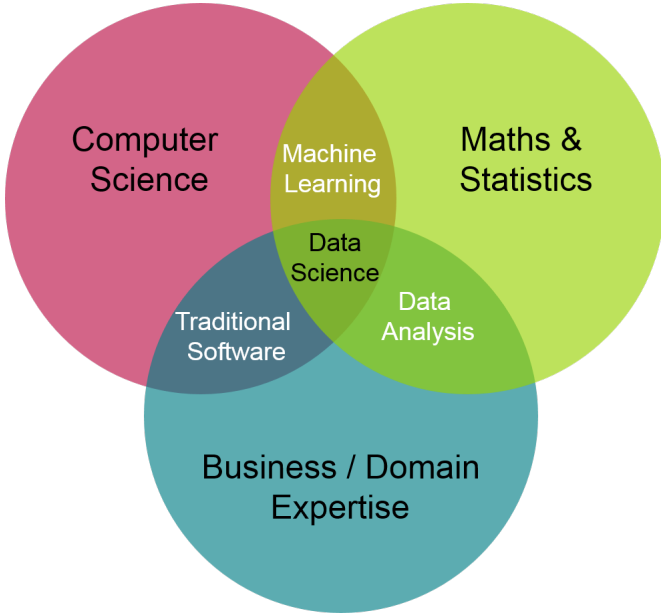
Highdimensional Statistics

Marie Düker
Winter term 2023/24

October 17, 2023



[0]https://thedatascientist.com/data-science-without-programming/

Syllabus

- Review: Statistics and Probability
- Motivation: Why high-dimensional statistics? (MW Chapter 1)
- Concentration inequalities (MW Chapter 2 and RV Chapter 2)
- Sparse linear models (MW Chapter 7 and BG Chapter 11)
- Random matrices and covariance estimation (MW Chapter 6)
- Covariance estimation and thresholding (MW Chapter 6 and BL)
- Inverse Covariance estimation (RBLZ)
- Principal component analysis in high dimensions (MW Chapter 8)
- Reproducing kernel Hilbert spaces (MW Chapter 12)
- Review Session

Relevant Literature

- MW: High-Dimensional Statistics: A Non-Asymptotic Viewpoint, by Martin J. Wainwright
- RV: High-Dimensional Probability, by Roman Vershynin
- BG: Statistics for High-Dimensional Data: Methods, Theory and Applications, by Peter Bühlmann and Sara van de Geer
- BL: Covariance regularization by thresholding, by Peter Bickel and Elizaveta Levina
- RBLZ: Sparse permutation invariant covariance estimation, by Adam Rothmann, Peter Bickel, Elizaveta Levina and Ji Zhu

Part 0

Review of some basic concepts in statistics and probability

Outline

Basics

Experiment A repeatable procedure

1. random
2. deterministic

- Toss a coin twice
- H head, T tail

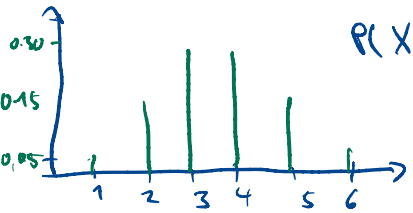
Sample space Set of all possible outcomes → $\Omega = \{HH, TT, HT, TH\}$

Event Subset of the sample space

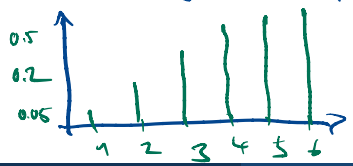
Probability function Gives probability for each outcome

$P(\text{at least one tail}) = P(\{TT, HT, TH\}) = \frac{|\{TT, HT, TH\}|}{|\Omega|} = \frac{3}{4}$

- probability fct.
- 100 students, grades 1–6



$P(X = 3) = 0.3$ probability distribution



Discrete distributions

Discrete random variable Set of possible outcomes is discrete.

Probability mass function The probability mass function (abbreviated pmf) p_X for the random variable X is defined by $p_X(k) = P(X = k)$.

Expected value Given a discrete random variable X which takes values in a set $A = \{x_1, x_2, x_3, \dots\}$, the expected value of X is denoted $\mathrm{E}(X)$ or μ_X, and is given by multiplying each possible value of X by its probability.

$$\mathrm{E}(X) = \sum_{x \in A} x \cdot \mathrm{P}(X = x) = \sum_{x \in A} x \cdot p_X(x).$$

Expected value of $g(X)$

$$\mathrm{E}(g(X)) = \sum_{x \in A} g(x) \cdot \mathrm{P}(X = x) = \sum_{x \in A} g(x) \cdot p_X(x).$$

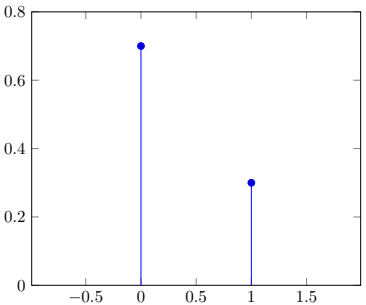
Variance of a discrete random variable X is denoted $\mathrm{Var}(X)$, and defined by $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}\,X)^2]$.
Standard deviation of X is then defined by $\sqrt{\mathrm{Var}(X)}$.

Discrete distributions

Bernoulli The random variable X has a Bernoulli distribution with parameter p, written $X \sim Bernoulli(p)$, if

$$p_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, X takes the value 1 with probability p, and 0 with probability $1 - p$.



Discrete distributions

If $X \sim Bernoulli(p)$, then $\mathrm{E}(X) = p$ and $\mathrm{Var}(X) = p(1-p)$.

$EX = 1 \cdot P(X=1) + 0 \cdot P(X=0) = 1 \cdot p = p$

$\mathrm{Var}(X) = E(X - EX)^2 = E(X-p)^2$

$= EX^2 - 2E(Xp) + p^2$

$= pEX$

$= EX^2 - p^2$

$*$

$= p - p^2 = p(1-p)$

$* \quad EX^2 = \sum x^2 P(X=x)$

$= 1^2 P(X=1) + 0^2 P(X=0)$

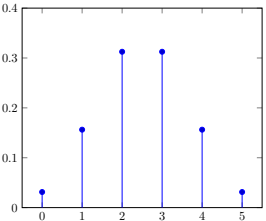
$= p$

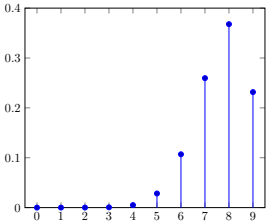
Discrete distributions

Binomial Distribution The random variable X has a binomial distribution with parameters n and p, written $X \sim Binomial(n, p)$, if

success probability

$$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$





X counts the number of successes after n trials

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ $\quad n! = n \cdot (n-1) \cdot (n-2) \cdots 1$ $\quad$ Exp: $3! = 3 \cdot 2 \cdot 1 = 6$

Exp: S: success, F: failure, success prob. $p = 0.9$, 7 attempts

Prob to succeed 4 out of 7 times

$P(X=4) = \binom{7}{4} 0.9^4 (1-0.9)^3$, $\quad P(X \leq 4) = \sum_{k=0}^{4} P(X=k)$

Discrete distributions

If $X \sim Binomial(n, p)$, then $E(X) = np$ and $\text{Var}(X) = np(1-p)$.

$$EX = \sum x P(X=x) = \sum_{k=0}^{n} k P(X=k)$$

Calculate $\text{Var}(X)$!

$$= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \sum_{k=1}^{n} \frac{n(n-1)!}{(k-1)!(n-1-(k-1))!} p^{k-1} (1-p)^{n-1-(k-1)} p$$

$$= \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \, np = np$$

$\sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} = 1$

Discrete distributions

Poisson Distribution The random variable X has a Poisson distribution with parameter λ, written $X \sim Pois(\lambda)$, if

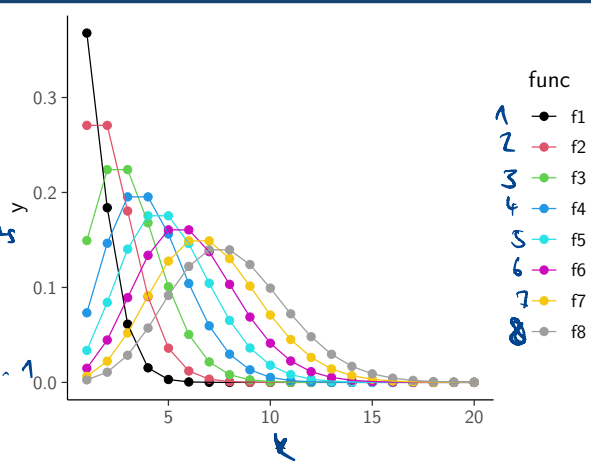
$$p_X(k) = \mathrm{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, 3, \ldots$$

λ: average number of events

k: number of times an event occurs

$k! = k \cdot (k-1) \cdots 1$

$$P(X=3) = \frac{2^3 e^{-2}}{3!}$$



Discrete distributions

If $X \sim Pois(\lambda)$, then $E(X) = \lambda$ and $\mathrm{Var}(X) = \lambda$.

$$EX = \sum_{k=0}^{\infty} k\, P(X=k)$$

$$= \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!}$$

$k!$: $k \cdot (k-1)(k-2)\cdots 1$

$$= \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!}\lambda \overset{k-1=j}{=} \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!}\lambda$$

$\sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} = 1$

$$= \lambda$$

Continuous distributions

Continuous random variable Takes values in an interval of real numbers.

Probability density function If X is a continuous random variable with cdf F_X, the probability density function of X is denoted f_X and is defined by $f_X(x) = \frac{d}{dx} F_X(x)$. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

Expected value If X is a continuous random variable, the expected value of X is denoted $E(X)$ or μ_X, and defined by the integral

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx.$$

Expected value of $g(X)$

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

Variance of a continuous random variable X is denoted $\mathrm{Var}(X)$ and defined by $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}\,X)^2]$.
Standard deviation of X is then defined by $\sqrt{\mathrm{Var}(X)}$.

Some useful properties

For any constants $a, b \in \mathbb{R}$,

1. $E(aX + b) = aE(X) + b$
2. $E(X + Y) = E(X) + E(Y)$.
3. $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$
4. If X and Y are independent, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
5. $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$

1) $E(aX+b) = \int (ax+b) f(x)\,dx = a \underbrace{\int x f(x)\,dx}_{= EX} + b \underbrace{\int f(x)\,dx}_{= 1} = aEX + b$

3) $\mathrm{Var}(X) = E(X - EX)^2 = EX^2 - 2E(XEY) + (EY)^2 = EX^2 - 2(EX)^2 + (EX)^2 = EX^2 - (EX)^2$

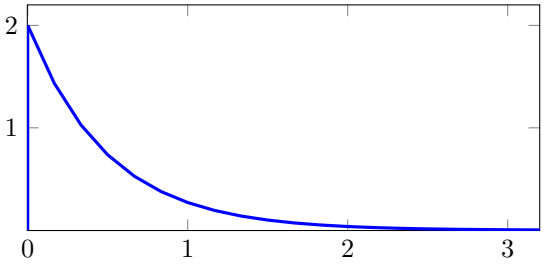
Exercise: Show 2, 4 and 5

Continuous distributions

Exponential Distribution The random variable X is exponentially distributed with parameter $\lambda > 0$ (often called the rate parameter), written $X \sim \textit{Exponential}(\lambda)$, if its pdf is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$X \sim Exponential(2)$



Continuous distributions

If $X \sim Exponential(\lambda)$, then $E(X) = \frac{1}{\lambda}$ and $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.

$$EX = \int_0^\infty x f(x)\,dx$$

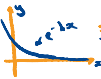
$$= \int_0^\infty x \lambda e^{-\lambda x}\,dx$$

$$\overset{*}{=} \lambda \left[x\left(-\tfrac{1}{\lambda}\right)e^{-\lambda x}\Big|_0^\infty - \int_0^\infty 1\left(-\tfrac{1}{\lambda}\right)e^{-\lambda x}\,dx \right]$$

$$\overset{*}{=} \lambda \left[\tfrac{1}{\lambda} \int_0^\infty e^{-\lambda x}\,dx \right]$$

$$= -\tfrac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty$$

$$= \tfrac{1}{\lambda}$$



(*) Integration by parts

$$\int_0^\infty g(x) z'(x)\,dx = g(x) z(x)\Big|_0^\infty - \int_0^\infty g'(x) z(x)\,dx$$

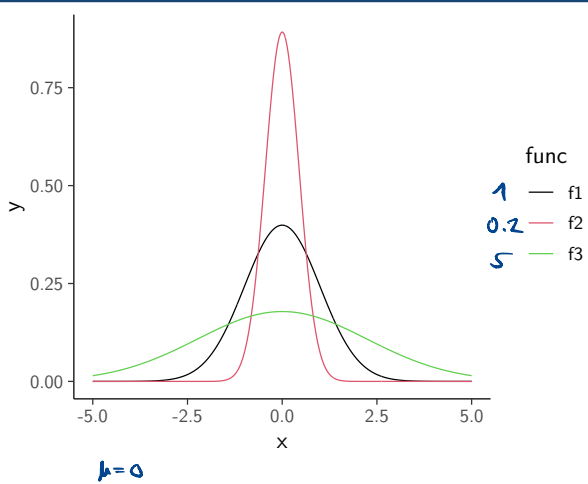
with $g(x) = x$ and $z(x) = e^{-\lambda x}$

Exercise: $\mathrm{Var}(X) =$

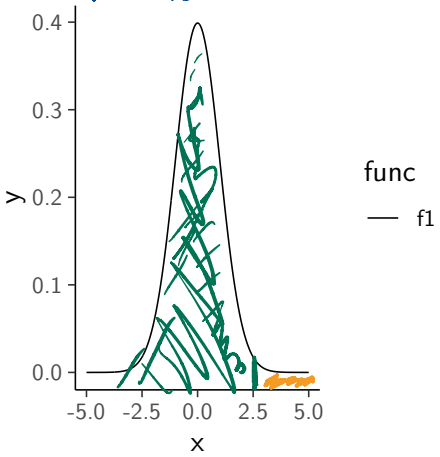
Continuous distributions

Normal/Gaussian distribution The random variable X is normally distributed with parameters μ and $\sigma > 0$, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if

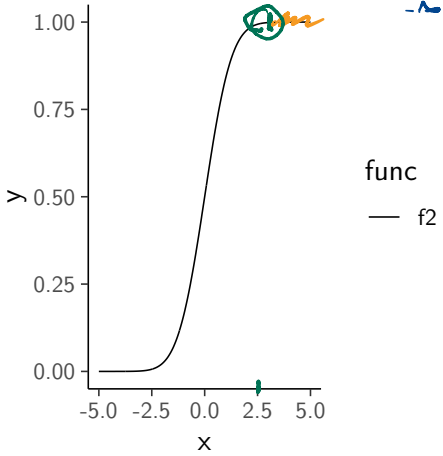
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Gaussian distribution



density



distribution

Gaussian distribution

Example

- heights of adults is normally distributed.
- heights of adults are normally distributed $N(\mu = 175, \sigma^2 = 2.75)$
- Find probability that randomly picked adult is less or equal to 168

$$P(X \leq 168) = \int_{-\infty}^{168} \frac{1}{\sqrt{2\pi \cdot 2.75}} \exp\left(-\frac{1}{2}\frac{(x-175)^2}{2.75}\right) dx$$

Some nice property

$X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ and X_1, X_2 independent

$\Rightarrow X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Exp: X_1: height of male adults,
X_2: height of female adults

Estimators

Estimator Rule for estimating a quantity based on observed data.

Bias The bias of an estimator is the difference between its expected value and the actual value of the parameter being estimated.

$$Bias(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

$Bias(\hat{\theta}) > 0$: $E(\hat{\theta})$ is overestimate of θ, $Bias(\hat{\theta}) < 0$: $E(\hat{\theta})$ underestimate, θ, $E(\hat{\theta}) = \theta$: unbiased estimator

Mean-squared error The mean-squared error of an estimator is the expected value of the squared difference between the estimator and the parameter it estimates.

$$\mathrm{MSE}(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2].$$

Example: $\{X_1, \dots X_n\}$ Estimator for the mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$X_i \sim N(\mu, \sigma^2)$, Bias $E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{E(X_i)}_{=\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$

$\Rightarrow$ $\hat{\mu}$ is an unbiased estimator

Lemma

$MSE(\widehat{\theta}) = Bias^2(\widehat{\theta}) + \mathrm{Var}(\widehat{\theta})$, *where* $\mathrm{Var}(\widehat{\theta}) = E[(\widehat{\theta} - E(\widehat{\theta}))^2]$

Proof

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2]$$

$$= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2]$$

$$= Var(\hat{\theta}) + 2E[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] + E[(E(\hat{\theta}) - \theta)^2]$$

(linearity of E; $E(\hat{\theta}) - \theta$ deterministic; $(E(\hat{\theta}) - \theta)^2$ deterministic)

$$E[\hat{\theta} - E(\hat{\theta})](E(\hat{\theta}) - \theta) = E(\hat{\theta}) - E(\hat{\theta}) = 0$$

$$(E(\hat{\theta}) - \theta)^2 = Bias^2(\hat{\theta})$$

$$= Var(\hat{\theta}) + Bias^2(\hat{\theta})$$

Convergence

Convergence in distribution
A sequence $\{X_n\}$ converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_n(x) = F(x).$$

Convergence in probability
A sequence $\{X_n\}$ converges in probability to a random variable X if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}\left(|X_n - X| > \varepsilon\right) = 0.$$

Convergence almost surely
A sequence $\{X_n\}$ converges almost surely to a random variable X if

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Probabilistic O-notation

For random variables X_n and a corresponding set of constants a_n,
Big O: stochastic boundedness the notation

$$X_n = O_p(a_n)$$

means that for any $\varepsilon > 0$, there exists a finite $M > 0$ and a finite $N > 0$ such that

$$\mathrm{P}\left(|\frac{X_n}{a_n}| > M\right) < \varepsilon \quad \text{for all} \quad n > N.$$

Small o: convergence in probability the notation

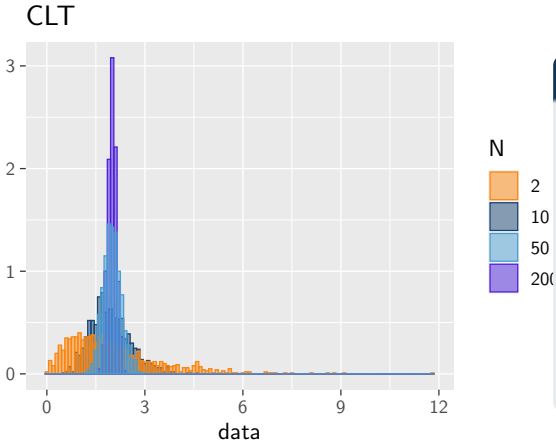
$$X_n = o_p(a_n)$$

means that the set of values $\frac{X_n}{a_n}$ converges to zero in probability. That is,

$$\lim_{n\to\infty} P\left[\left|\frac{X_n}{a_n}\right| > \varepsilon\right] = 0 \quad \text{for all} \quad \varepsilon > 0.$$

Example

CLT



Lemma (CLT)

Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Consider the sum

$$S_N = X_1 + \cdots + X_N \text{ and } Z_N = \frac{S_N - \mathrm{E}\, S_N}{\sqrt{\mathrm{Var}(S_N)}}.$$

Then, as $N \to \infty$,

$$Z_N \to \mathcal{N}(0,1) \text{ in distribution.}$$

The weak Law of Large Numbers

Lemma (Weak law of large numbers)

Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean μ. Consider the sum

$$S_N = X_1 + \cdots + X_N.$$

Then, as $N \to \infty$,

$$S_N \to \mu \quad \text{in probability.}$$

The Law of Large Numbers

Lemma (Strong law of large numbers)

Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean μ. Consider the sum

$$S_N = X_1 + \cdots + X_N.$$

Then, as $N \to \infty$,

$$S_N \to \mu \quad \text{almost surely.}$$

Multivariate random variables

- The multivariate normal distribution is by far the most important multivariate distribution in statistics.
- It's important for all the reasons that the one-dimensional Gaussian distribution is important, but even more so in higher dimensions because many distributions that are useful in one dimension do not easily extend to the multivariate case
- requires some important results from linear algebra

Linear algebra

Inverse The inverse of an $p \times p$ matrix A (if it exists), denoted A^{-1} is the matrix satisfying $AA^{-1} = I_p$.

- Only square matrices have inverses.
- Matrices which are not invertible are called singular

Positive definite A symmetric $p \times p$ matrix A is said to be positive (semi)definite if for all $x \in \mathbb{R}$, $x \neq 0$

$$x^T Ax > (\geq) 0.$$

Rank The rank of a matrix is the dimension of its largest nonsingular submatrix.

- Note that a nonsingular $p \times p$ matrix has rank p, and is said to be full rank

Expectation and variance

Let A be a matrix of constants and X a random vector with mean μ and variance Σ,

- $\mathrm{E}(A^T X) = A^T \mu$
- $\mathrm{Var}(A^T X) = A^T \Sigma A$
- $\mathrm{E}(X^T A X) = \mu^T A \mu + \mathrm{tr}(A\Sigma)$,

Linear algebra

Some useful facts about traces:

- $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$
- $\mathrm{tr}(cA) = c\,\mathrm{tr}(A)$
- $\mathrm{tr}(A) = \mathrm{rk}(A)$ if $AA = A$

Some useful facts about the rank:

- A and B with appropriate dimensions, $\mathrm{rk}(AB) \leq \mathrm{rk}(A)$ and $\mathrm{rk}(AB) \leq \mathrm{rk}(B)$.
- $\mathrm{rk}(A^T A) = \mathrm{rk}(AA^T) = \mathrm{rk}(A)$.

Standard normal

Standard normal A real random vector $Z = (Z_1, \ldots, Z_p)^T$ is called p-variate standard normal random vector if $Z_1, \ldots, Z_p$ are mutually independent and each follows a standard normal distribution.

We write $X \sim \mathcal{N}_p(0, I)$ and Z has the density:

$$f_{\mathrm{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k}} \exp\left(-\frac{1}{2}\mathrm{x}^{\mathrm{T}}\mathrm{x}\right).$$

Multivariate normal

Multivariate normal A real random vector $X = (X_1, \ldots, X_p)^T$ is called a normal random vector if there exists a random p-vector Z, which is a standard normal random vector, a p-vector μ, and a matrix A, such that $X = A^{\mathrm{T}} Z + \mu$.

Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$ and that Σ is full rank; then X has a density:

$$f_X(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (x - \boldsymbol{\mu})\right)$$

where Σ denotes the determinant of Σ.

Multivariate normal

Properties

- connection between covariance and independence in the multivariate normal distribution
- Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$. If $\Sigma_{ij} = 0$, i.e. the off-diagonal corresponding to X_i and X_j is zero, then X_i and X_j are independent.
- linear combinations are also normally distributed
- Let b be a $p \times 1$ vector of constants, B a $k \times d$ matrix of constants, and $X \sim \mathcal{N}_p(\mu, \Sigma)$. Then

$$b + BX \sim \mathcal{N}_p(B\mu + b, B\Sigma B')$$

Summary

- What is the difference between discrete and continuous distributions?
- How does one calculate the expected value and the variance of a random variable?
- What are some basic properties of the expected value and the variance?
- How is convergence in probability defined?
- How is convergence in distribution defined?
- What do the CLT and LLN tell us?
- What is the probabilistic big-O notation?
- How to calculate mean and covariance of multivariate normal distribution?
- How is the rank of a matrix defined?