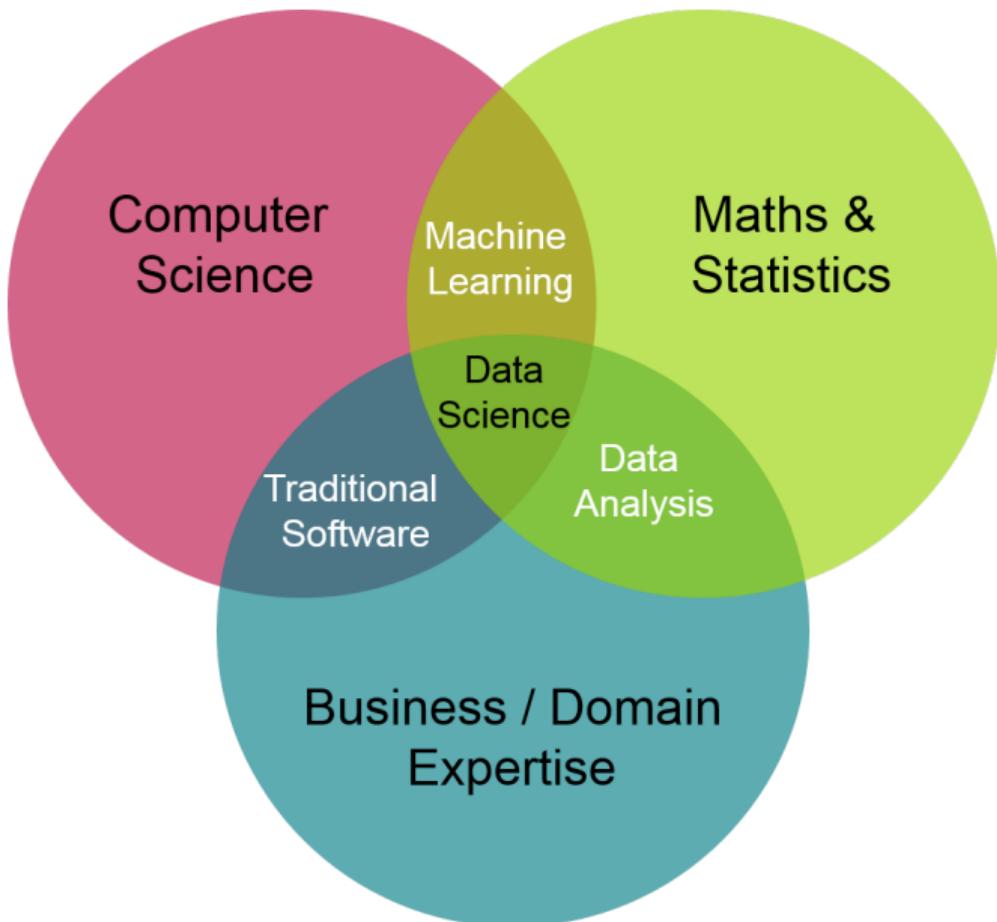


# Highdimensional Statistics

**Marie Düker**

Winter term 2023/24

November 7, 2023



<sup>0</sup><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

- ① Basics
- ② Discrete distributions
- ③ Continuous distributions
- ④ Convergence
- ⑤ Estimators
  - ① General concepts
  - ② The law of large numbers
  - ③ The Central Limit Theorem
- ⑥ Some linear algebra
- ⑦ The multivariate normal distribution

# 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

$P(\text{at least one tail})$

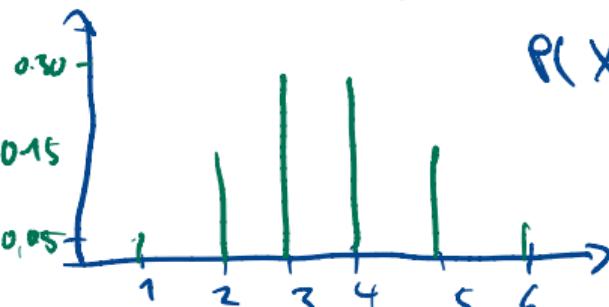
$$= P(\{TT, HT, TH\})$$

$$= \frac{1+TT, HT, TH\}}{4} = \frac{3}{4}$$

**Probability function** Gives probability for each outcome

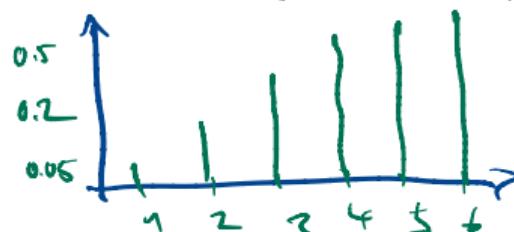
• probability fct.

• 100 students, grades 1-6



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

probability distribution  $|\Omega|$



# 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  $E(X)$  or  $\mu_X$ , and is given by multiplying each possible value of  $X$  by its probability.

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

Expected value of  $g(X)$

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

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

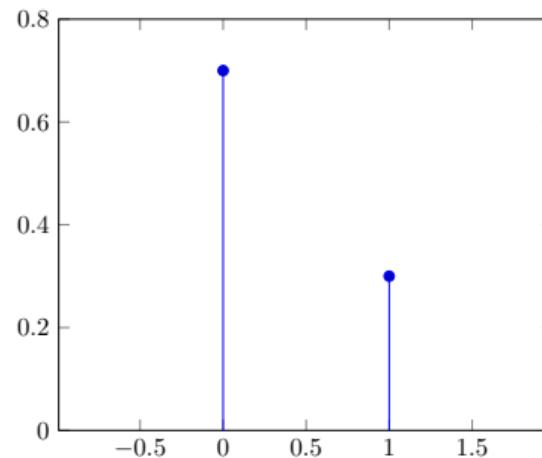
**Standard deviation** of  $X$  is then defined by  $\sqrt{\text{Var}(X)}$ .

# Discrete distributions

**Bernoulli** The random variable  $X$  has a Bernoulli distribution with parameter  $p$ , written  $X \sim \text{Bernoulli}(p)$ , 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  $X \sim \text{Bernoulli}(p)$ , then  $E(X) = p$  and  $\text{Var}(X) = p(1 - p)$ .

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

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

$$= EX^2 - 2EXp + p^2$$

$$\stackrel{\text{def}}{=} pEX$$

$$= EX^2 - p^2$$

\*

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

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

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

$$= p$$

# Discrete distributions

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

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

*success probability*

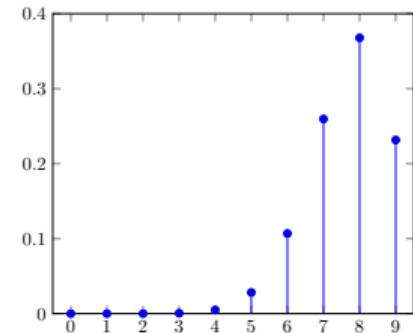
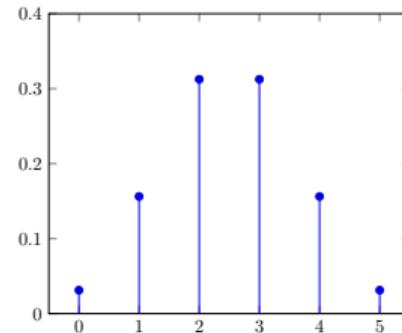
$X$  counts the number of successes after  $n$  trials

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

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

$$\begin{aligned} E(X) &= \sum_{x=0}^n x P(X=x) = \sum_{k=0}^n k P(X=k) \quad \text{Calculate } \text{Var}(X) ! \\ &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n \frac{n(n-1)!}{k!(n-k)!} (1-p)^{n-k} p^{k-n} \\ &= \sum_{j=0}^{n-n} \binom{n-n}{j} p^j (1-p)^{n-n-j} \quad np = np \\ &\quad \underbrace{\qquad\qquad}_{=1} \end{aligned}$$

# Discrete distributions

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

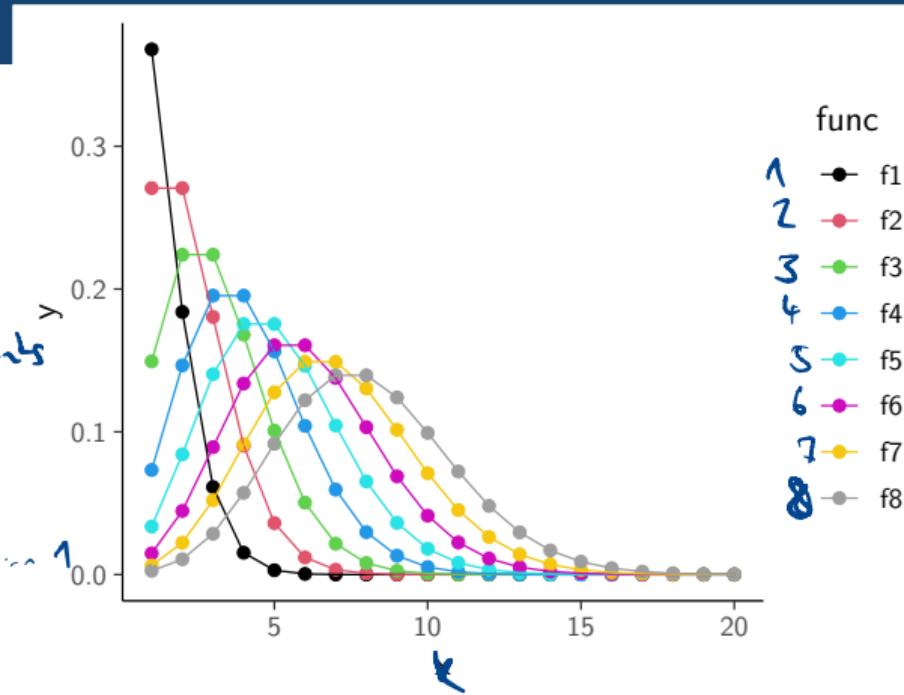
*average number of events*

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

*number of times  
an event occurs*

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

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



# Discrete distributions

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

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k P(X=k) \\ &= \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^k}{(k-1)!} \\ &\quad \underbrace{k(k-1)(k-2)\dots 1}_{k-1=j} \\ &= \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^{k-1} \lambda}{(k-1)!} \stackrel{j=0}{=} \sum_{j=0}^{\infty} \underbrace{\frac{e^{-\lambda} \lambda^j}{j!}}_{=1} \lambda \\ &= \lambda \end{aligned}$$

# 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  $\mu_X$ , and defined by the integral

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

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  $\text{Var}(X)$  and defined by  $\text{Var}(X) = E[(X - E X)^2]$ .  
**Standard deviation** of  $X$  is then defined by  $\sqrt{\text{Var}(X)}$ .

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

4 • If  $X$  and  $Y$  are independent,  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .

5 •  $\text{Var}(aX + b) = a^2 \text{Var}(X)$

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

3)  $\text{Var}(X) = E(X-EX)^2 = EX^2 - 2E(XEX) + (EX)^2 = EX^2 - 2(EX)^2 + (EX)^2 = EX^2 - (EX)^2$

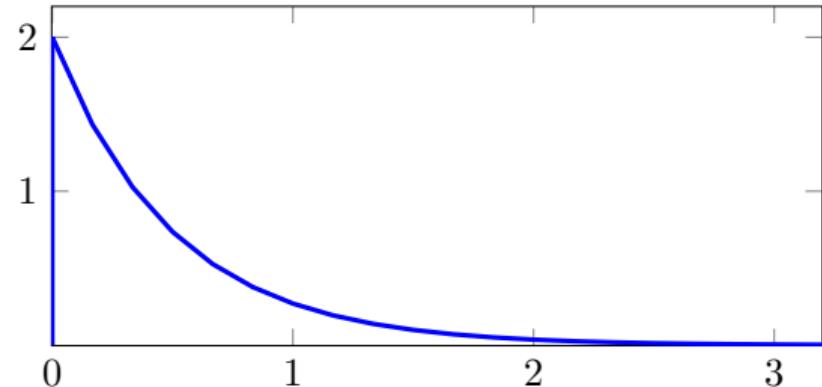
Exercise: Show 2, 4 and 5

# Continuous distributions

$$X \sim \text{Exponential}(2)$$

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

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



# Continuous distributions

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

$$\begin{aligned} E(X) &= \int_0^\infty x f(x) dx \\ &= \int_0^\infty x \lambda e^{-\lambda x} dx \\ &\stackrel{*}{=} \lambda \left[ x \left( -\frac{1}{\lambda} \right) e^{-\lambda x} \right]_0^\infty - \int_0^\infty 1 \left( -\frac{1}{\lambda} \right) e^{-\lambda x} dx \\ &\stackrel{*}{=} \lambda \left[ \left\{ \int_0^\infty e^{-\lambda x} dx \right\} \right] \\ &= -\frac{1}{\lambda} e^{-\lambda x} \Big|_0^\infty \\ &= \frac{1}{\lambda} \end{aligned}$$

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



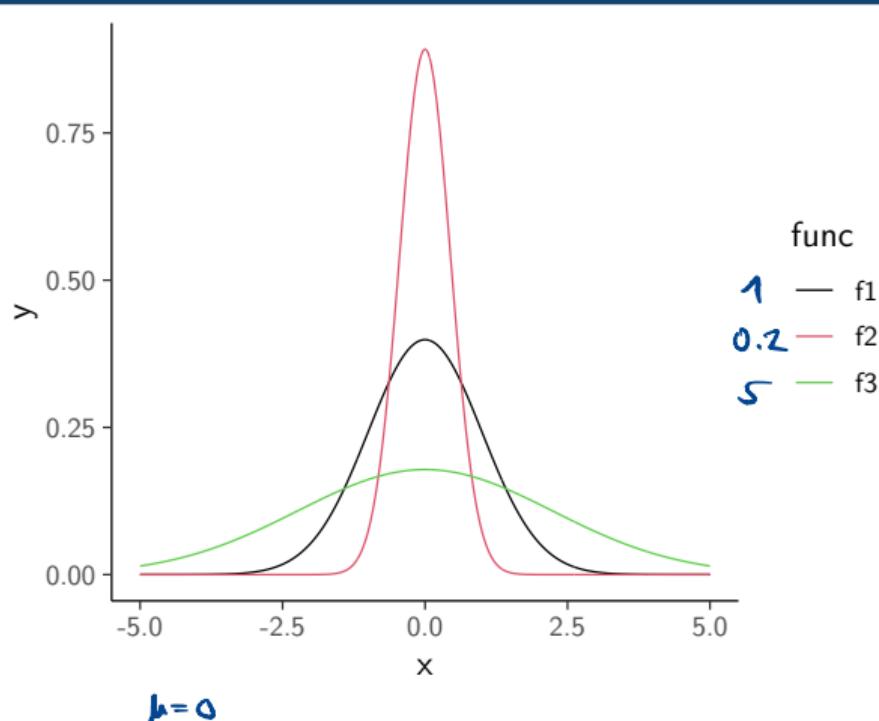
(\*) Integration by parts

$$\begin{aligned} \int_0^\infty g(x) z'(x) dx &= g(x) z(x) \Big|_0^\infty \\ &\quad - \int_0^\infty g'(x) z(x) dx \\ \text{with } g(x) &= x \text{ and } z(x) = e^{-\lambda x} \end{aligned}$$

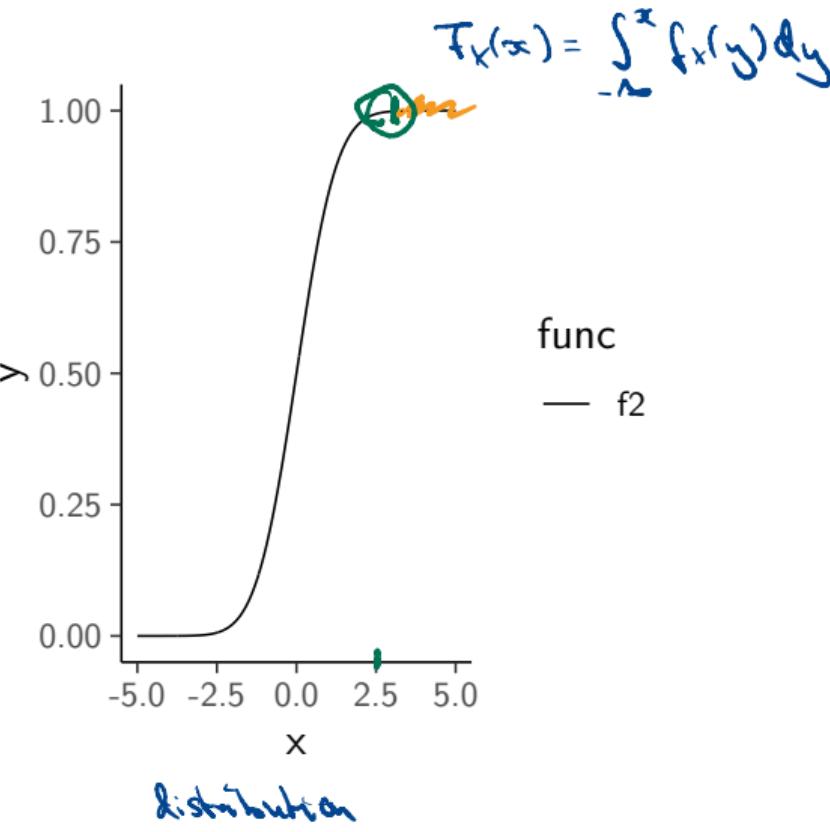
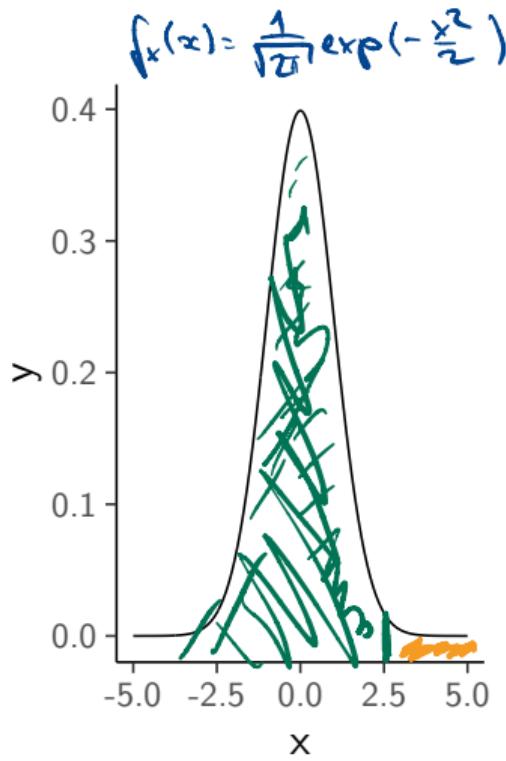
# Continuous distributions

**Normal/Gaussian distribution** The random variable  $X$  is normally distributed with parameters  $\mu$  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



# 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

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

Some nice property

$$X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2) \text{ and } X_1, X_2 \text{ independent}$$
$$\Rightarrow X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Ex:  $X_1$ : 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.

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

$\text{Bias}(\hat{\theta}) > 0$  :  $E(\hat{\theta})$  is overestimate of  $\theta$ ,  $\text{Bias}(\hat{\theta}) < 0$  :  $E(\hat{\theta})$  underestimates  $\theta$ ,  $E(\hat{\theta}) = \theta$  : 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.

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

Example:  $\{\hat{x}_1, \dots, \hat{x}_n\}$  Estimator for the mean  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \hat{x}_i$ :

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

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

## Lemma

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

## Proof

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\ &\stackrel{\text{Linearity } E}{=} \text{Var}(\hat{\theta}) + 2E[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] + E[(E(\hat{\theta}) - \theta)^2] \\ &\quad \underbrace{E[\hat{\theta} - E(\hat{\theta})]}_{\text{deterministic}} \underbrace{(E(\hat{\theta}) - \theta)}_{\text{deterministic}} \quad \underbrace{= (E(\hat{\theta}) - \theta)^2}_{= \text{Bias}^2(\hat{\theta})} \\ &\quad = E(\hat{\theta}) - E(\hat{\theta}) = 0 \\ &= \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) \end{aligned}$$

# Convergence

## Convergence in distribution

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

$$\lim_{n \rightarrow \infty} \underbrace{F_n(x)}_{P(X_n \leq x)} = \overbrace{F(x)}^{P(X \leq x)}.$$

## Convergence in probability

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

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0. \Leftrightarrow \lim_{n \rightarrow \infty} P(|X_n - X| \leq \varepsilon) = 1$$

## Convergence almost surely

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

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

# Probabilistic O-notation

For random variables  $X_n$  and a corresponding set of constants  $a_n$ ,

**Big O: stochastic boundedness** the notation

$$X_n = O_p(a_n)$$

*deterministic*

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

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

*$P(|X_n| \leq M |a_n|) \geq \varepsilon$*

**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,

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

## Example

$$\{X_n\}_{n \geq 1} \quad \mu_n = E X_n, \quad \sigma_n^2 = \text{Var}(X_n) \quad \leftarrow \infty$$

$$X_n - \mu_n = O_p(\sigma_n)$$

$$P(|X_n - \mu_n|/\sigma_n > n^{-\frac{1}{2}})$$

$$\begin{aligned} &\stackrel{\text{chebyshev}}{\leq} \underbrace{E|X_n - \mu_n|^2}_{=\sigma_n^2} / \sigma_n^2 \cdot \frac{1}{n^{-1}} \\ &= n \end{aligned}$$

$$X_n - \mu_n = O_p(s_n) \quad s_n^{-2} \text{Var}(X_n) = \frac{\sigma_n^2}{s_n^2} \rightarrow 0$$

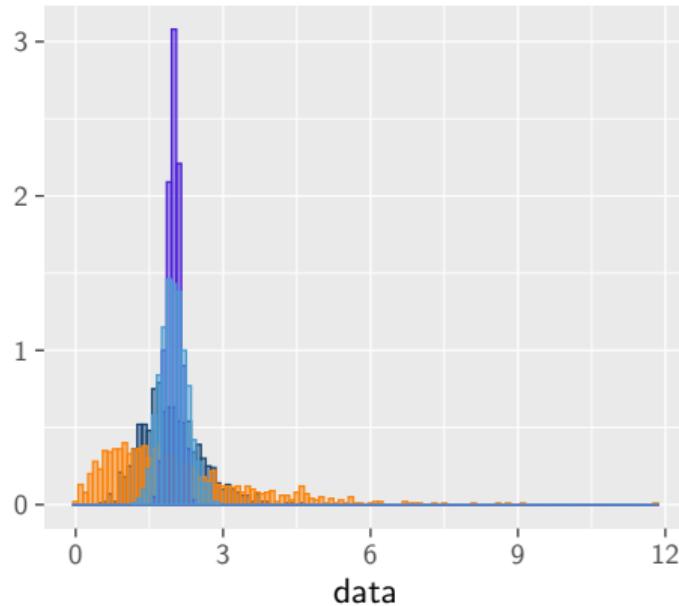
$$P(|X_n - \mu_n| \frac{1}{s_n} > \varepsilon)$$

$$\leq E|X_n - \mu_n|^2 \frac{1}{s_n^2 \varepsilon^2}$$

$$= \frac{\sigma_n^2}{s_n^2} \cdot \frac{1}{\varepsilon^2}$$

$$\rightarrow 0$$

## CLT



## Lemma (CLT)

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

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

Then, as  $N \rightarrow \infty$ ,

$Z_N \rightarrow \mathcal{N}(0, 1)$  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  $\mu$ . Consider the sum

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

Then, as  $N \rightarrow \infty$ ,

$$\hat{\mathbb{P}}_N S_N \rightarrow \mu \text{ in probability.}$$

$$\forall \varepsilon > 0 \quad \lim_{N \rightarrow \infty} \mathbb{P}(|\frac{1}{N} S_N - \mu| > \varepsilon) = 0$$

# 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  $\mu$ . Consider the sum

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

Then, as  $N \rightarrow \infty$ ,

$$S_N \rightarrow \mu \text{ almost surely.}$$

# 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  $x \in \mathbb{R}^p$ ,  $x \neq 0$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \quad x^T Ax > (\geq) 0.$$

**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

$$A = \begin{pmatrix} 1 & 2 & | & 3 \\ 4 & 5 & | & 6 \\ \hline 7 & 8 & | & 9 \end{pmatrix}$$
$$\det(A) > 0$$
$$\det\begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix} = 5 - 8 = -3 < 0$$

# Expectation and Variance

$X$  is random vector  $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$  \* decorrelation:  $A =$   
Expectation and variance

$$EX = \begin{pmatrix} EX_1 \\ \vdots \\ EX_p \end{pmatrix}, \quad Cov(X) = E[(X - EX)(X - EX)^T]$$

$$= \left( E[(X_i - EX_i)(X_j - EX_j)] \right)_{i,j=1,\dots,p}$$

Let  $A$  be a matrix of constants and  $X$  a random vector with mean  $\mu$  and variance  $\Sigma$ ,

- $E(A^T X) = A^T \mu$   $EATX = A^T EX = A^T \mu$
- \* •  $\text{Var}(A^T X) = A^T \Sigma A$   $\text{Cov}(A^T X) = E[(A^T X - E(A^T X))(A^T X - E(A^T X))^T]$   
 $= A^T E[(X - \mu)(X - \mu)^T] A = (X - \mu)^T A^T A$
- $E(X^T A X) = \mu^T A \mu + \text{tr}(A \Sigma)$ ,  
 $= A^T E[(X - \mu)(X - \mu)^T] A = (X - \mu)^T A^T A$

# Linear algebra

$$\text{tr}(A) = \sum_{i=1}^p a_{ii} \quad A = (a_{ij})_{i,j=1,\dots,p}$$

Some useful facts about traces:

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

Some useful facts about the rank:

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

## Standard normal

$$Z_i \sim N(0, 1) \quad f_{Z_i}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

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

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

$$EX = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{p-dim vector}$$

$$f_X(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p}} \exp\left(-\frac{1}{2} \underbrace{x^T x}_{\text{sum of squares}}\right).$$

$$(x_1, \dots, x_p) \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = x_1^2 + \dots + x_p^2$$

# Multivariate normal

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

$\underbrace{p \times p}_{p \times p}$   $\underbrace{p \text{-vector}}$

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

$$f_X(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

where  $\Sigma$  denotes the determinant of  $\Sigma$ .

$$E\mathbf{X} = \boldsymbol{\mu}$$

$$\text{Cov}(X) = A^T \underbrace{\text{Cov}(Z)}_{I_p} A = A^T A \stackrel{!}{=} \Sigma$$

# 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  $p \times p$  matrix of constants, and  $X \sim \mathcal{N}_p(\mu, \Sigma)$ . Then

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

$$- E(b + BX) = E(b) + E(BX) = b + E(BX) = b + BEX = b + B\mu$$

$$- E[(b + BX - E(b + BX))(b + BX - E(b + BX))^T]$$

$$= E[(b + BX - (b + B\mu))(b + BX - (b + B\mu))^T]$$

$$= E[(BX - B\mu)(BX - B\mu)^T] = B E[(X - \mu)(X - \mu)^T] B^T = B \Sigma B^T$$

# 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?

# Part I

Why high-dimensional statistics?

# Outline

- ① Why high-dimensional statistics?
- ② What can go wrong in high dimensions?
  - ① Example 1: Covariance estimation
  - ② Example 2: Linear regression
- ③ What can help?
- ④ Summary

# Why high-dimensional statistics?

## Motivation

### ① Intensive data collection

- More and more features are measured per individual.
- Biology (specifically genetics) where millions of (combinations of) genes are measured for a single individual.
- High resolution imaging.
- Finance (stock indices).
- Climate studies.

### ② Number of measured features possibly exceeds the number of observations.

### ③ Not all measured features are relevant to answer a given question.

## Classical theory

Statements that apply to a fixed class of models, parameterized by an index  $N$  (sample size) that is allowed to increase ("large  $N$ , fixed  $p$ ").

## High-dimensional theory

Classical methods can break down. Consider either "large  $N$ , large  $p$ " or entirely non-asymptotic approaches.

## ① Classical theory

- ①  $N \gg p$ .  $N$  is "much larger" than  $p$
- ② Asymptotic assumption:  $p$  is fixed and  $N \rightarrow \infty$ .  $\frac{p}{N} \rightarrow 0$
- ③ Basic tools: LLN and CLT.

## ② High-dimensional theory

- ①  $N \sim p$ , e.g.  $\frac{N}{p} \rightarrow \alpha > 0$   $\text{Exp } \rho = b_1 N + b_2$
- ②  $p \gg N$ , e.g.  $p \sim e^N$
- ③ non-asymptotic

① e.g. 104 genes with only 50 samples.

② Classical methods fail. E.g., Linear regression, covariance estimation. Why?

# Example: Covariance estimation

Setup:

$\mathbb{R}^p$

- given i.i.d. samples  $X_i \sim \mathcal{N}(0, \Sigma)$ , for  $i = 1, 2, \dots, N$
- want to estimate a covariance matrix  $\Sigma \in \mathbb{R}^{p \times p}$

Classical approach: Estimate  $\Sigma$  via sample covariance matrix:

$$\hat{\Sigma}_N = \frac{1}{N} \sum_{i=1}^N \underbrace{X_i X_i'}_{\mathbb{R}^{p \times p}} \quad \text{mean} = 0$$

rank( $X_i X_i'$ ) = rank( $X_i' X_i$ ) = 1

Reasonable properties: (p fixed, N increasing)

- Unbiased:  $E \hat{\Sigma}_N = \Sigma$
- Consistent:  $\hat{\Sigma}_N \rightarrow \Sigma$  as  $N \rightarrow \infty$ .
- Asymptotic distributional properties available

$$\|A\|_2 = \sqrt{\lambda_{\max}(A)} \quad \begin{array}{l} A \in \mathbb{R}^{p \times p} \\ A \text{ symmetric, pos. def} \end{array}$$

$$\|\hat{\Sigma}_N - \Sigma\|_2 \rightarrow 0 \quad \text{with high probability as } N \rightarrow \infty$$

# Example: Covariance estimation

**Reasonable properties:** ( $p$  fixed,  $N$  increasing)

- Unbiased:  $E \hat{\Sigma}_N = \Sigma$
- Consistent:  $\hat{\Sigma}_N \xrightarrow{a.s.} \Sigma$  as  $N \rightarrow \infty$ .
- Asymptotic distributional properties available

**An alternative experiment:**

- Fix some  $\alpha > 0$
- Study behavior over sequences with  $\frac{p}{N} \rightarrow \alpha > 0$
- Does  $\hat{\Sigma}_{N(p)}$  converge to anything reasonable?

$$\hat{\Sigma}_{N(p)} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i'$$

## Example: Covariance estimation

$$\begin{pmatrix} 1 & \cdot & \cdot & \cdot \\ \cdot & \ddots & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & 1 \end{pmatrix}$$

Suppose we have a sample  $\{X_1, \dots, X_N\}$  with  $X_i \in \mathbb{R}^p$  and  $X_i \sim \mathcal{N}(0, \mathbf{I})$ . Then,

$$\widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^N X_i X_i'$$

is a consistent estimator for the covariance matrix  $\Sigma$ . Distance between estimator and population quantity can be measured through the so-called operator norm.

Consider the eigenvalues of  $\widehat{\Sigma}$ :

$$\lambda_{\max}(\widehat{\Sigma}) = \lambda_1(\widehat{\Sigma}) \geq \dots \geq \lambda_p(\widehat{\Sigma}) = \lambda_{\min}(\widehat{\Sigma}) \geq 0.$$

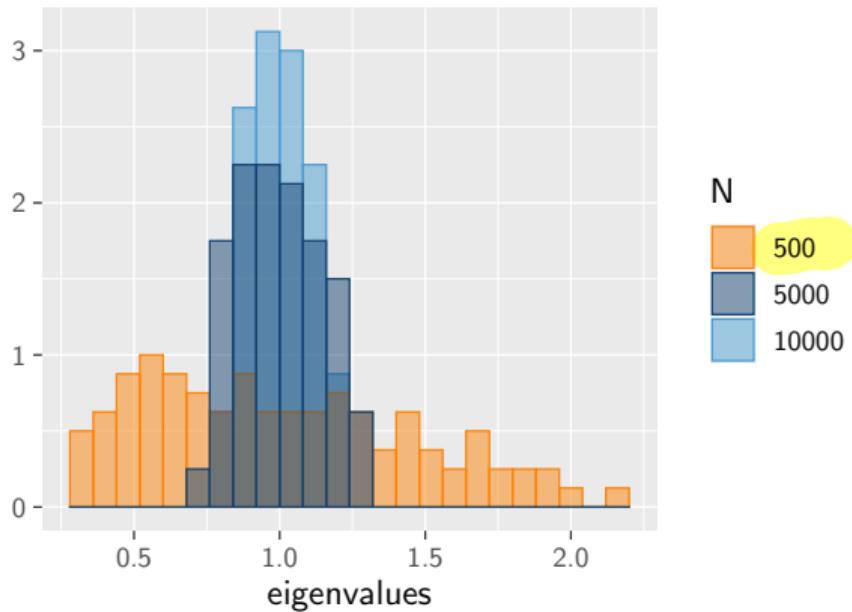
## Example: Covariance estimation

Histograms of eigenvalues

$$X_1, \dots, X_N \sim N(0, I_p)$$

$$\lambda_1(\hat{\Sigma}), \dots, \lambda_p(\hat{\Sigma})$$

fixed dimension  $p=100$



What we expect:

- all eigenvalues of true  $\Sigma = I_p$  are 1
- plot will peak at 1 with increasing  $N$

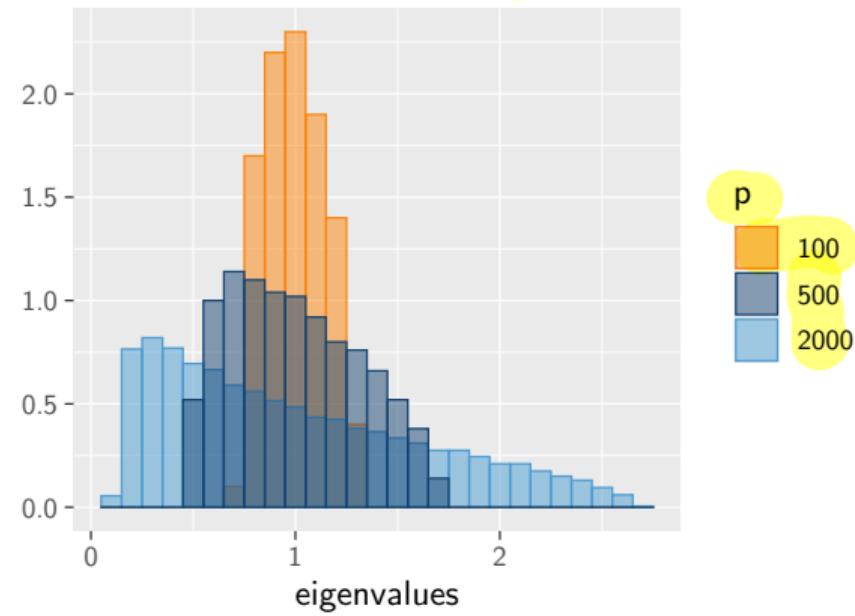
# Example

What can go wrong in high dimensions?

What happens in high-dimensional setting:

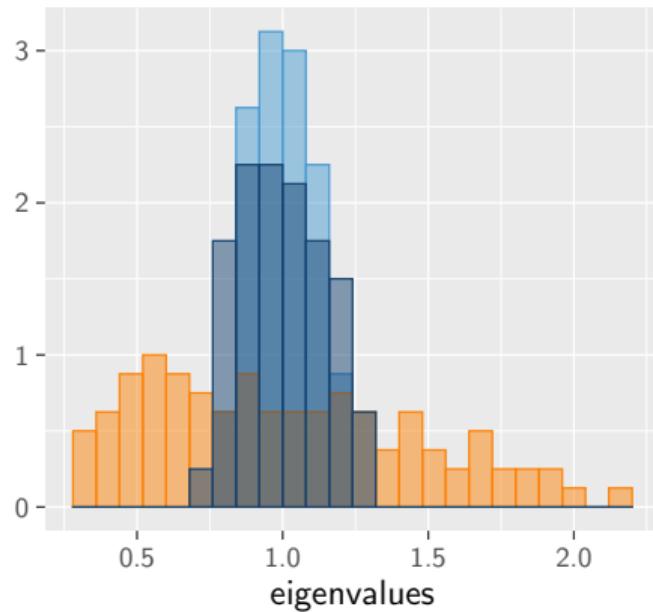
- doesn't center around 1 with increasing dimension

fixed sample size  $N=5000$

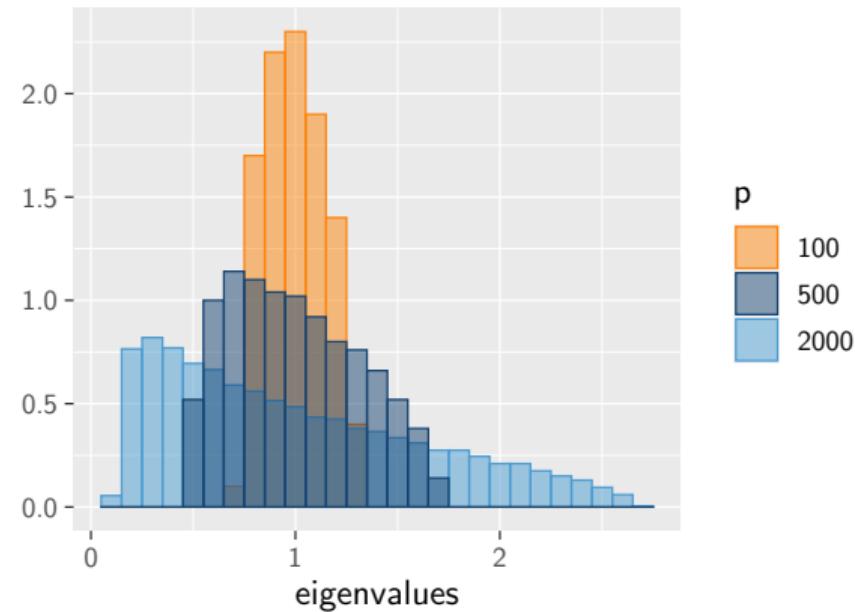


# Example

fixed dimension  $p=100$



fixed sample size  $N=5000$



# Example: linear models

What can go wrong in high dimensions?

## Setup:

- Linear regression
- We aim to recover the unknown true  $\beta_0$ .

$$y = X\beta + \varepsilon \quad \rightarrow \quad \varepsilon = y - X\beta$$

$\underbrace{y}_{\mathbb{R}^n} \quad \underbrace{X\beta}_{\mathbb{R}^{n \times p}} \quad \underbrace{\varepsilon}_{\mathbb{R}^n}$

## Classical approach:

- Given the observations  $(y, X)$  we get OLS  $\hat{\beta} = \underbrace{(X^T X)^{-1} X^T y}_{\text{not invertible}}$ .

$$\mathcal{E}' \mathcal{E} = (y - X\beta)' (y - X\beta)$$

not invertible

since we average over rank 1  
matrices with  $p < N$

$$\xrightarrow{\text{HW}} \hat{\beta}$$

Reasonable properties: (p fixed, N increasing)

- Unbiased:  $E \hat{\beta} = \beta_0$
- Consistent:  $\hat{\beta} \rightarrow \beta$  as  $N \rightarrow \infty$ .
- Asymptotic distributional properties available

## Example: linear models

regular regime of fixed n large

$$y = x \beta +$$

# Example: linear models

What can go wrong in high dimensions?

## Reasonable properties: ( $p$ fixed, $N$ increasing)

- Unbiased:  $E \hat{\beta} = \beta_0$
- Consistent:  $\hat{\beta} \rightarrow \beta_0$  as  $N \rightarrow \infty$ .
- Asymptotic distributional properties available

## An alternative experiment:

- $p \gg N$
- Important prior : many extracted feature in  $X$  are irrelevant
- many coefficients in  $\beta_0$  are “exactly zero”.

**Example:**  $Y$  is the size of a tumor, it might be reasonable to suppose that it can be expressed as a linear combination of genetic information in the genome described in  $X$ . BUT most components of  $X$  will be zero and most genes will be unimportant to predict  $Y$  : We are looking for meaningful few genes.

## Example 3: linear models

OLS cannot be calculated

$$\begin{array}{ccccccc} y & = & X & \beta & + & \epsilon \\ & = & \begin{matrix} \text{colorful matrix} \end{matrix} & \begin{matrix} \text{colorful vector} \end{matrix} & + & \text{noise} \\ n \times 1 & & n \times p & p \times 1 & & \end{array}$$

# Summary

## What are the different view points?

- Classical asymptotics.
- High-dimensional asymptotics.
- Non-asymptotic bounds.

## What can go wrong in high dimensions?

- no consistent estimator
- low rank matrices, not invertible

## What can help?

- Finding or imposing lower dimensional structure
- sparsity
- low rank
- graphical structure

Key questions: What embedded low-dimensional structures are present in data?

How can they be exploited?

# Part II

## Concentration bounds

# Outline

- ① Concentration bounds: Classical examples
- ② Sub-Gaussian Random variables
  - ① Tail bound
  - ② Sum of sub-Gaussian RVs
  - ③ Hoeffding
  - ④ Chernoff
- ③ Equivalent characterizations of sub-Gaussianity
- ④ Sub-exponential concentration

# Concentration bounds

- Main tools in dealing with highdimensional randomness
- Non-asymptotic versions of CLT
- General form

- Classical examples:

- Markov's inequality: Assume  $X \geq 0$

$\text{HW: Proof Markov's =}$

$$\mathbb{1}_{\{X \geq t\}} \leq \frac{X}{t}$$

- Chebyshev's: Assume  $E X^2 < \infty$

$$P[|X - E[X]| > t] \leq \varepsilon(t) \quad \begin{matrix} \downarrow \\ \text{nonasymptotic} \end{matrix} \quad \begin{matrix} \varepsilon(t) \text{ is small} \\ \text{if } t \text{ is large} \end{matrix}$$
$$P(|X - EX| \leq t) \geq 1 - \varepsilon(t)$$

$$E[X] < \infty$$

$$P[X > t] \leq \frac{E[X]}{t}, \quad \forall t > 0. \quad O(\frac{1}{t})$$

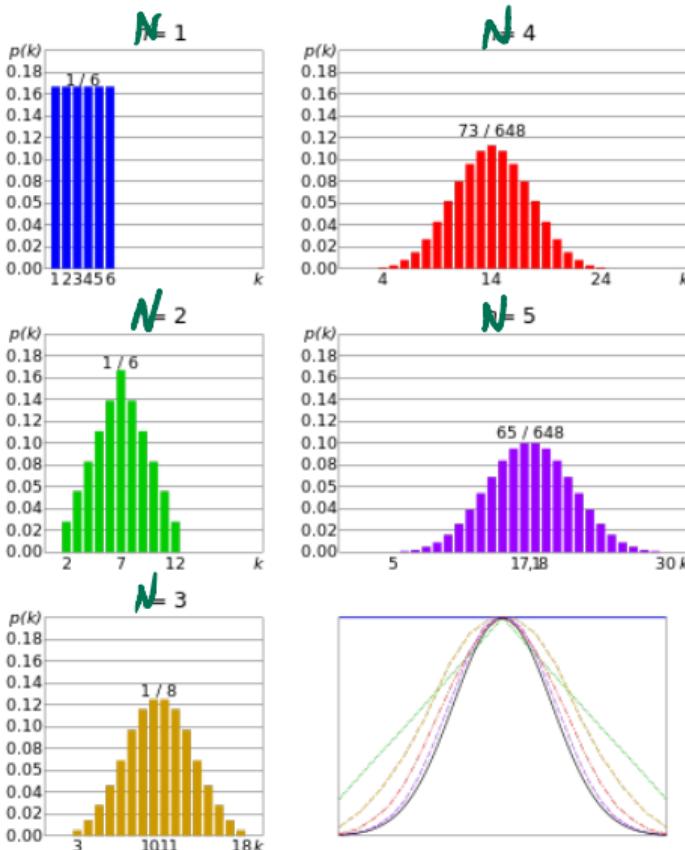
$$\mathbb{1}_{\{X \geq t\}} = \begin{cases} 1 & X \geq t \\ 0 & \text{otherwise} \end{cases}$$

$$\underbrace{P[|X - E[X]| > t]}_{= P(|X - EX|^2 > t^2)} \leq \frac{\text{Var}(X)}{t^2}, \quad \forall t > 0. \quad O(\frac{1}{t^2})$$

- Stronger assumption  $E |X|^k < \infty$

$$P[|X - E[X]| > t] \leq t^{-k} E[|X - E[X]|^k]. \Rightarrow$$
$$P(|X - EX|^k > t^k) \leq \inf_{n \geq 1} t^{-k} E|X - EX|^k$$

# Recall CLT



## Lemma (CLT)

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

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

Then, as  $N \rightarrow \infty$ ,

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

# Some Motivation

$$E S_n = \frac{n}{2} \quad \text{Var}(S_n) = \frac{n}{4}$$

- Let  $X_1, \dots, X_n \sim \text{Ber}(1/2)$  and  $S_n = \sum_{i=1}^n X_i$ . Then, by CLT

$$Z_n := \frac{S_n - n/2}{\sqrt{n/4}} \xrightarrow{D} \mathcal{N}(0, 1)$$

Goal: avoid  $n \rightarrow \infty$

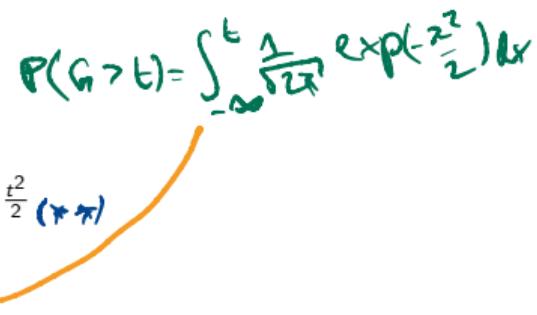
- Let  $G \sim \mathcal{N}(0, 1)$ ,

$$P(Z_n > t) = P\left(\frac{S_n - n/2}{\sqrt{n/4}} > t\right) = P\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right) \xrightarrow[n \rightarrow \infty]{\substack{\downarrow \\ CLT}} G \sim \mathcal{N}(0, 1)$$

- Letting  $t = \alpha\sqrt{n}$

$$P\left(S_n > \frac{n}{2}(1 + \alpha)\right) \lesssim \frac{1}{2} e^{-\frac{n\alpha^2}{2}}$$

- Problem: Approximation is not tight in general.



## Theorem (Berry-Esseen CLT)

Under the assumption of CLT, with  $\delta = E|X_1 - \mu|^3/\sigma^3$ ,

$$G \sim \mathcal{N}(0,1)$$

$$Z_n = \frac{S_n - E S_n}{\sqrt{\text{Var}(S_n)}}, \quad S_n = \sum_{i=1}^n X_i$$

$$|P(Z_n > t) - P(G > t)| \leq \frac{\delta}{\sqrt{n}}. \quad \underset{n \rightarrow \infty}{\longrightarrow} 0 \quad (\times)$$

- Bound is tight since for Bernoulli example  $P(S_n = n/2) = \frac{1}{2^n} \binom{n}{n/2} \sim \frac{1}{\sqrt{n}}$
- Conclusion, the approximation error is  $O(\frac{1}{\sqrt{n}})$  which is larger than the exponential bound  $O(e^{-\frac{n\alpha^2}{2}})$  that we want to establish.
- Solution: Directly obtain the concentration inequalities,
- Often using Chernoff bounding technique: for any  $\lambda > 0$ ,

$$\leq |P(Z_n > t) - P(G > t)| + P(G > t)$$

$$\leq \frac{\delta}{\sqrt{n}} + \frac{1}{2} e^{-\lambda^2 t^2} \quad t = \delta \sqrt{n}$$

$$P(Z_n > t) = P(\exp(\lambda Z_n) > \exp(\lambda t)) \leq \frac{E e^{\lambda Z_n}}{e^{\lambda t}}, \quad t \in \mathbb{R}$$

$\exp(\lambda t) \nearrow 0$   
 $\uparrow$   
 Markov's  $\neq$

$$= \frac{\delta}{\sqrt{n}} + \frac{1}{2} e^{-\lambda^2 t^2}$$

$$\leq \frac{\delta}{\sqrt{n}} = O\left(\frac{1}{\sqrt{n}}\right)$$

- Leads to the study of the MGF of random variables.

Moment generating function

## Definition (Sub-Gaussian Random Variables)

A random variable  $X$  with finite mean  $\mu$  is *sub-Gaussian* with parameter  $\sigma > 0$  if

$$E \left[ e^{\lambda(X-\mu)} \right] \leq e^{\sigma^2 \lambda^2 / 2}, \quad \forall \lambda \in \mathbb{R}.$$

*MGF*

We say that  $X$  is  $\sigma$ -sub-Gaussian and say it has *variance proxy*  $\sigma^2$ .

- $X \sim \mathcal{N}(0, 1)$  is sub-Gaussian with equality.  $E[e^{\lambda X}] = e^{\lambda^2/2}$
- A *Rademacher random variable*  $R$  takes a value of 1 with probability 1/2 and a value of -1 with probability 1/2.  $R$  is 1-sub-Gaussian.
- Suppose a random variable  $X$  satisfies  $|X - E[X]| \leq M$  almost surely for some constant  $M$ . Then  $X$  is  $M$ -sub-Gaussian.

$X \sim \mathcal{N}(0, 1)$  is sub-Gaussian with equality.

$X$  has density  $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$

$$\begin{aligned} E[e^{\lambda x}] &= \int_{-\infty}^{\infty} e^{\lambda x} f(x) dx \\ &= \int_{-\infty}^{\infty} e^{\lambda x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(\lambda x - \frac{x^2}{2}) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\frac{1}{2}(x^2 - 2\lambda x)) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\frac{1}{2}(x^2 - 2\lambda x + \lambda^2) + \frac{1}{2}\lambda^2) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\frac{1}{2}(x - \lambda)^2) \exp(+\frac{1}{2}\lambda^2) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x - \lambda)^2) dx \exp(\frac{1}{2}\lambda^2) = \exp(\frac{1}{2}\lambda^2) \end{aligned}$$

density of Gaussian RV

## Theorem (Tail bound for sub-Gaussian random variables)

If a random variable  $X$  with finite mean  $\mu$  is  $\sigma$ -sub-Gaussian, then

$$P[|X - \mu| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall t \in \mathbb{R}.$$

Proof: We know  $X$  is sub-Gaussian, that means  $E e^{X(X-\mu)} \leq e^{\sigma^2 \lambda^2 / 2}$

$$P(X - \mu \geq t) = P(\underbrace{\exp(\lambda(X-\mu))}_{\text{Markov's}} \geq \exp(\lambda t)) \quad (*)$$

$$\stackrel{*}{\leq} E(\exp(\lambda(X-\mu))) \frac{1}{\exp(\lambda t)}$$

$$\stackrel{*}{\leq} \exp(\sigma^2 \lambda^2 / 2) \frac{1}{\exp(\lambda t)}$$

$$= \exp(\sigma^2 \lambda^2 / 2 - \lambda t)$$

$$\frac{d}{d\lambda} [\sigma^2 \lambda^2 / 2 - \lambda t]$$

$$= \sigma^2 \lambda - t = 0$$

$$\Rightarrow \lambda = \frac{t}{\sigma^2}$$

$$P(X - \mu \geq t) \leq \inf_{\lambda > 0} \exp(\sigma^2 \lambda^2 / 2 - \lambda t) = \exp(\sigma^2 \frac{t^2}{\sigma^2} / 2 - \frac{t^2}{\sigma^2}) \\ = \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

$$\{|X - \mu| \geq t\} = \{X - \mu \geq t\} \cup \{-(X - \mu) \geq t\}$$

## Proposition (Sum of sub-Gaussian random variables is sub-Gaussian)

If  $X_1, \dots, X_n$  are independent sub-Gaussian random variables with variance proxies  $\sigma_1^2, \dots, \sigma_n^2$ , then  $Z = \sum_{i=1}^n X_i$  is sub-Gaussian with variance proxy  $\sum_{i=1}^n \sigma_i^2$ .

Proof:  $X_i$   $\sigma_i$ -sub-Gaussian, that means:  $E(\exp(\lambda(X_i - EX_i))) \leq \exp(\sigma_i^2 \lambda^2 / 2)$  (\*)

We have to show:  $E(\exp(\lambda(Z - EZ))) \leq \exp(\sum_{i=1}^n \sigma_i^2 \lambda^2 / 2)$   
 $= \sum_{i=1}^n E(X_i)$

$$\begin{aligned} E[\exp(\lambda(Z - EZ))] &= E[\exp(\lambda(\sum_{i=1}^n X_i - \sum_{i=1}^n EX_i))] \\ &= E[\exp(\sum_{i=1}^n \lambda(X_i - EX_i))] \\ &= E[\prod_{i=1}^n \exp(\lambda(X_i - EX_i))] \\ &\stackrel{X_i \text{ indep.}}{=} \prod_{i=1}^n E[\exp(\lambda(X_i - EX_i))] \\ &\leq \prod_{i=1}^n \exp(\sigma_i^2 \lambda^2 / 2) = \exp(\sum_{i=1}^n \sigma_i^2 \lambda^2 / 2) \end{aligned}$$

(\*)  $X_i$  sub-Gaussian  $\downarrow$

## Theorem (Hoeffding)

Let  $X_1, \dots, X_n$  be independent sub-Gaussian random variables with variance proxies  $\sigma_1^2, \dots, \sigma_n^2$ , then  $Z = \sum_{i=1}^n X_i$  satisfies the tail bound

$$P[|Z - E[Z]| \geq t] \leq 2 \exp\left(-\frac{t^2}{2 \sum_{i=1}^n \sigma_i^2}\right), \quad \forall t \in \mathbb{R}.$$

Proof:  $X_1, \dots, X_n$  sub-Gaussian  $\Rightarrow Z = \sum_{i=1}^n X_i$  sub-Gaussian with variance proxy  $\sum_{i=1}^n \sigma_i^2$   
 $\Rightarrow P(|Z - EZ| \geq t) \leq 2 \exp\left(-\frac{t^2}{2 \sum_{i=1}^n \sigma_i^2}\right)$

Example :  $X_i \sim N(\mu, 1)$  ,  $X_1, \dots, X_n$

$\mu$  can be estimated through  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$

$$P\left(|\sum_{i=1}^n X_i - E(\sum_{i=1}^n X_i)| \geq t\right) \leq 2 \exp(-nt^2/2)$$

$$t = \left(\frac{\log(n)}{n}\right)^{\frac{1}{2}}$$

$$\begin{aligned} \Rightarrow P\left(|\sum_{i=1}^n X_i - \underbrace{E(\sum_{i=1}^n X_i)}_{n\mu}| \geq \left(\frac{\log(n)}{n}\right)^{\frac{1}{2}}\right) &\leq 2 \exp\left(-\log(n)/2\right) \\ &= 2 \exp(-\log(n^{1/2})) \\ &= 2^{-n^{-\frac{1}{2}}} \xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

## Lemma (Chernoff's inequality)

Let  $X_i$  be independent Bernoulli random variables with parameters  $p_i$ . Consider their sum  $S_N = \sum_{i=1}^N X_i$  and denote its mean by  $\mu = E S_N$ . Then, for any  $t > \mu$ , we have

$$P(S_N \geq t) \leq \exp(-\mu) \left( \frac{\exp(1)\mu}{t} \right)^t.$$

Proof:  $P(S_N \geq t)$

$$= P\left(\sum_{i=1}^N X_i \geq t\right)$$

$$= P\left(\exp(\lambda \sum_{i=1}^N X_i) \geq \exp(\lambda t)\right)$$

$$\leq E\left[\exp(\lambda \sum_{i=1}^N X_i)\right] \frac{1}{\exp(\lambda t)}$$

$$= E\left[\prod_{i=1}^N \exp(\lambda X_i)\right] \frac{1}{\exp(\lambda t)}$$



# Equivalent characterizations of sub-Gaussianity

For a RV  $X$ , the following are equivalent:

- ① The tails of  $X$  satisfy

$$P(|X| > t) \leq 2 \exp(-t^2/K_1^2), \text{ for all } t \geq 0.$$

- ② The moments of  $X$  satisfy

$$(E|X|^p)^{\frac{1}{p}} \leq K_2 \sqrt{p}, \text{ for all } p \geq 1.$$

- ③ The MGF of  $X^2$  satisfies

$$E \exp(\lambda^2 X^2) \leq \exp(K_3^2 \lambda^2), \text{ for all } |\lambda| \leq \frac{1}{K_3}.$$

- ④ The MGF of  $X^2$  is bounded at some point,

$$E \exp(X^2/K_4^2) \leq 2.$$

# Equivalent characterizations of sub-Gaussianity

Proof Sketch:

# Sub-exponential concentration

## Definition

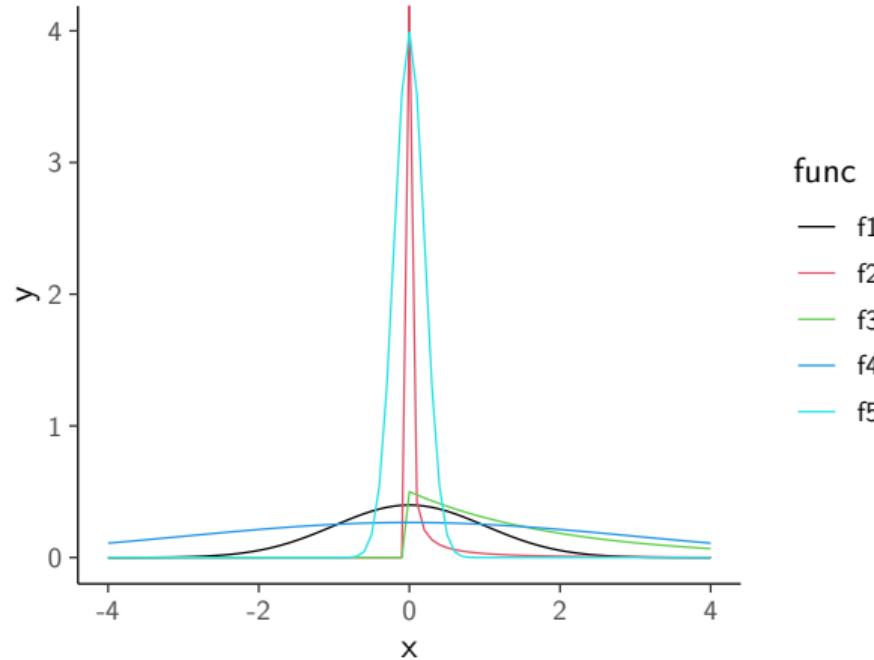
A zero-mean random variable  $X$  is sub-exponential if for some  $\nu, \alpha > 0$ .

$$E e^{\lambda X} \leq \exp(\nu^2 \lambda^2 / 2), \quad \text{for all } |\lambda| < \frac{1}{\alpha}.$$

A general random variable is sub-exponential if  $X - E X$  is sub-exponential.

- If  $Z \sim \mathcal{N}(0, 1)$ , then  $Z^2$  is sub-exponential with parameters  $(2, 4)$ .
- Tails of  $Z^2 - 1$  are heavier than a Gaussian.

# Illustration of tail behavior



func

- $f_1$
- $f_2$
- $f_3$
- $f_4$
- $f_5$

# Sub-exponential concentration

## Proposition

Assume that  $X$  is zero-mean sub-exponential with parameters  $(\nu, \alpha)$ . Then,

$$\mathbb{P}(X > t) \leq \exp\left(-\frac{1}{2} \min\left\{\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right\}\right).$$

# Maximum of sub-Gaussian vectors

For any vector  $X = (X_1, \dots, X_n)' \in \mathbb{R}^n$ , the max-norm is defined as  $\|X\|_{\max} = \max_{i=1, \dots, n} |X_i|$ .

## Lemma

Let  $X = (X_1, \dots, X_n) \in \mathbb{R}^n$  be a random vector with zero-mean, sub-Gaussian coordinates  $X_i$  with parameter  $\sigma_i > 0$ . Then, for any  $\gamma \geq 0$ ,

$$P(\|X\|_{\max} > \sigma \sqrt{2(1 + \gamma) \log(n)}) \leq 2n^{-\gamma}$$

where  $\sigma = \max_{i=1, \dots, n} \sigma_i$ .

Proof:



# Lipschitz functions of standard Gaussian vector

## Theorem

Suppose  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz with respect to Euclidean distance, and let  $X = (X_1, \dots, X_n)$ , where  $X_1, \dots, X_n \sim \mathcal{N}(0, 1)$  i.i.d. Then for all  $t \in \mathbb{R}$ ,

$$P[|f(X) - E[f(X)]| \geq t] \leq 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

In particular,  $f(X)$  is sub-Gaussian.

- In other words,  $f(X)$  is sub-Gaussian with parameter  $L$ .
- Deep result, no easy proof!
- Has far-reaching consequences.
- One-sided bounds holds with factor 2 removed.

## Example: Singular values

- Consider a matrix  $X \in \mathbb{R}^{n \times p}$  where  $n > p$ .
- Let  $\sigma_1(X) \geq \sigma_2(X) \geq \dots \geq \sigma_k(X)$  be (ordered) singular values of  $X$ .
- By Weyl's theorem, for any  $X, Y \in \mathbb{R}^{n \times d}$ :  $|\sigma_k(X) - \sigma_k(Y)| \leq \|X - Y\|_{op} \leq \|X - Y\|_F$ .
- Thus,  $X \rightarrow \sigma_k(X)$  is 1-Lipschitz:

### Proposition

Let  $X \in \mathbb{R}^{n \times d}$  be a random matrix with iid  $\mathcal{N}(0, 1)$  entries. Then,

$$\mathbb{P}(|\sigma_k(X) - \mathbb{E} \sigma_k(X)| \geq \delta) \leq 2 \exp(-\delta^2/2)$$

# Summary

## What is concentration?

- non-asymptotic bound on probability to control deviations
- possible deviations of interest: RV from mean
- deviation between estimator and true quantity

## Important definitions:

- tails of a sub-Gaussian distribution are dominated by the tails of a Gaussian
- distributions with heavy tails are not sub-Gaussian
- tails of distributions with heavy tails might be sub-exponential

**Main technique:** Let  $Z$  be zero-mean random variable. Then,

$$P(Z > t) = P(\exp(\lambda Z) > \exp(\lambda t)) \leq \frac{E e^{\lambda Z}}{e^{\lambda t}}, t \in \mathbb{R}$$

- Apply  $\exp(\lambda \cdot)$  to both sides.
- Apply Markov's inequality.
- Calculate MGF.
- Minimize over  $\lambda$ .

# Part III

## Sparse linear models

# Outline

- ① Linear regression setup
- ② Recall what goes wrong in large dimensions
- ③ Ridge Regression
- ④ Lasso Regression
- ⑤ Comparison