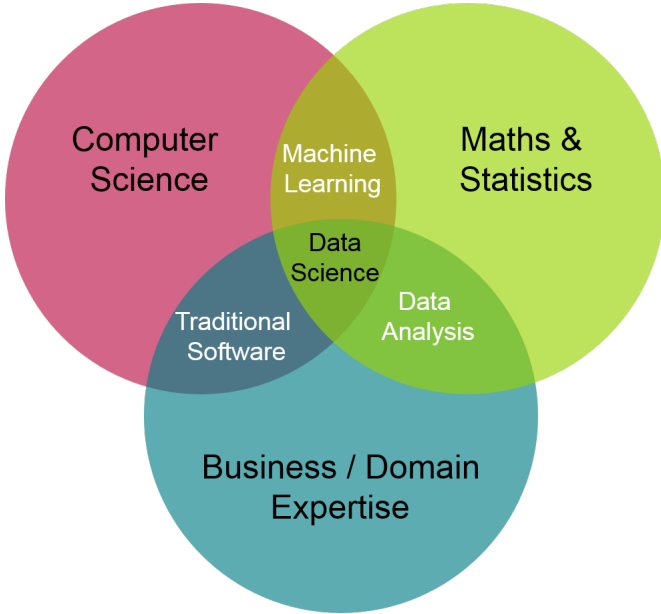
Highdimensional Statistics

Marie Düker
Winter term 2023/24

November 14, 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

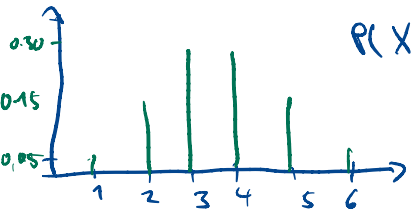
Sample space Set of all possible outcomes → $\Omega = \{HH, TT, HT, TH\}$

Event Subset of the sample space

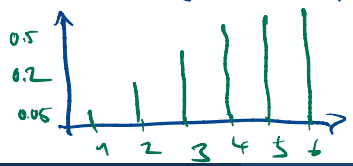
Probability function Gives probability for each outcome

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

- 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 $g(X)$

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

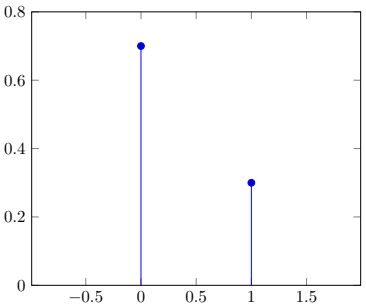
Variance of a discrete random variable X is denoted $\mathrm{Var}(X)$, and defined by $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}\,X)^2]$.
Standard deviation of X is then defined by $\sqrt{\mathrm{Var}(X)}$.

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 $\mathrm{Var}(X) = p(1-p)$.

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

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

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

$= pEX$

$= EX^2 - p^2$

$*$

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

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

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

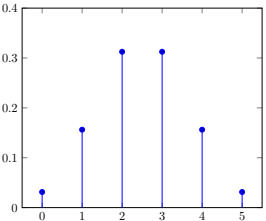
$= p$

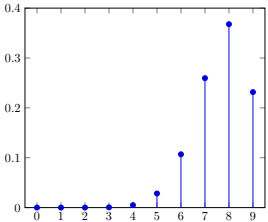
Discrete distributions

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

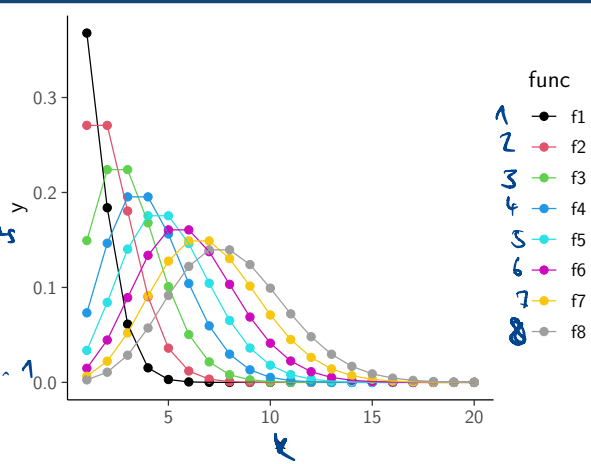
$$p_X(k) = \mathrm{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, 3, \ldots$$

λ: average number of events

k: number of times an event occurs

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

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



Discrete distributions

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

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

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

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

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

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

$$= 1$$

$$= \lambda$$

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

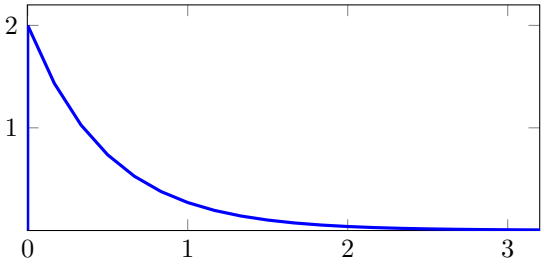
Exercise: Show 2, 4 and 5

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

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

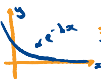
$$= \int_0^\infty x \lambda e^{-\lambda x}\,dx$$

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

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

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

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



(*) Integration by parts

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

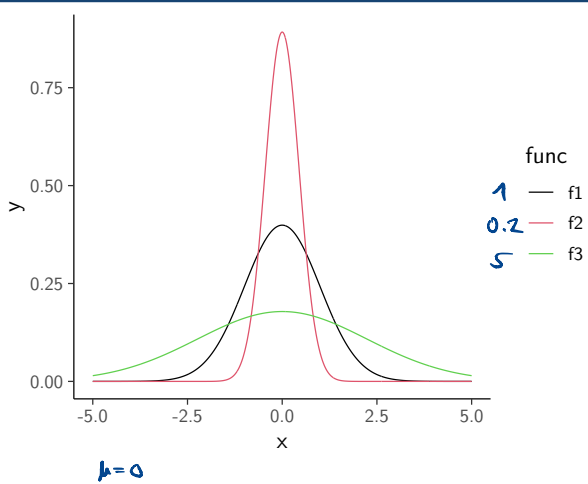
with $g(x) = x$ and $z(x) = e^{-\lambda x}$

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

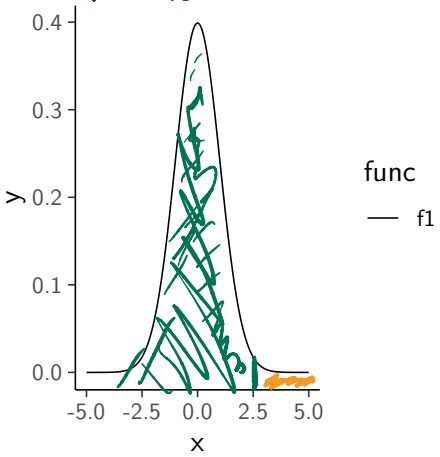
Continuous distributions

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

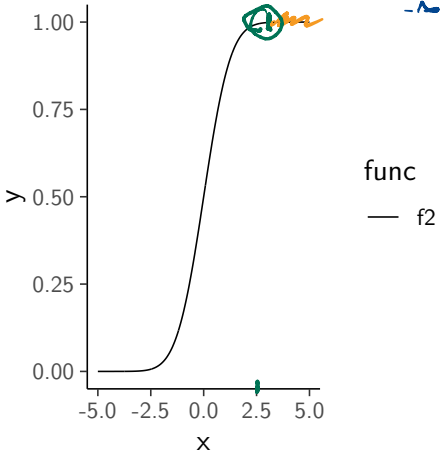
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



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

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

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

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

deterministic: $E[\hat{\theta} - E(\hat{\theta})](E(\hat{\theta}) - \theta)$

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

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

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

Convergence

Convergence in distribution

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

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

$F_n(x) = P(X_n \le x)$, $F(x) = P(X \le x)$

Convergence in probability

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

$$\lim_{n\to\infty} \mathrm{P}\left(|X_n - X| > \varepsilon\right) = 0. \quad \Leftrightarrow \quad \lim_{n\to\infty} P(|X_n - X| \le \varepsilon) = 1$$

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

deterministic

$$X_n = O_p(a_n)$$

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

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

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,

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

Example

$\{X_n\}_{n\geq 1}$ $\mu_n = EX_n$, $\sigma_n^2 = \mathrm{Var}(X_n) < \infty$

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

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

Chebyshev $\leq$

$\leq \underbrace{E|X_n - \mu_n|^2}_{=\sigma_n^2} / \sigma_n^2 \; \frac{1}{\eta^{-1}}$

$= \eta$

$X_n - \mu_n = o_p(s_n)$ $s_n^{-2}\mathrm{Var}(X_n) = \frac{\sigma_n^2}{s_n^2} \to 0$

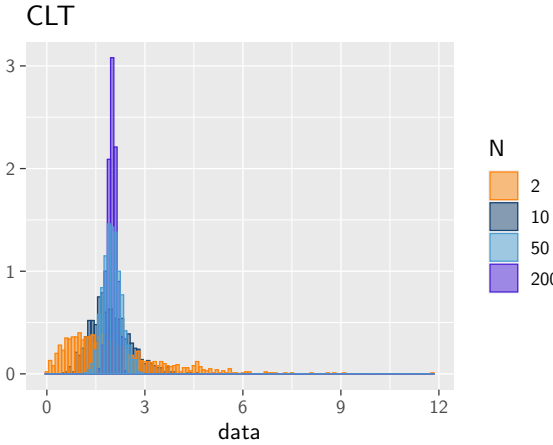
$P(|X_n - \mu_n| \frac{1}{s_n} > \varepsilon)$

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

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

$\to 0$

CLT



Lemma (CLT)

Let $X_1, X_2, \ldots$ 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 $N \to \infty$,

$$\frac{1}{N} S_N \to \mu \quad \text{in probability.}$$

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

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 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 $x \in \mathbb{R}$, $x \neq 0$

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

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

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

$\det(1) > 0$

$\det\begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix} = 5 - 8 = -3 < 0$

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 μ and variance Σ,

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

$$= A^T E[(X-\mu)(X-\mu)^T] A = (X-\mu)^T A^T$$

Linear algebra

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

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

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

Standard normal A real random vector $Z = (Z_1, \ldots, Z_p)^T$ 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 $X \sim \mathcal{N}_p(0, I_p)$ and Z has the density:

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

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

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

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 $X = A^{\mathrm{T}} Z + \mu$.

$p \times p$ $\quad$ p-vector

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

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

where Σ denotes the determinant of Σ.

$EX = \mu$

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

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

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. $N >> p$. N is "much larger" than p $\frac{p}{N} \to 0$
   2. Asymptotic assumption: p is fixed and $N \to \infty$.
   3. Basic tools: LLN and CLT.
2. High-dimensional theory
   1. $N \sim p$, e.g. $\frac{N}{p} \to \alpha > 0$ Ex $p = b_1 N + b_2$
   2. $p >> N$, 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:

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

Classical approach: Estimate Σ via sample covariance matrix:

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

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

- Unbiased: $\mathrm{E}\,\widehat{\Sigma}_N = \Sigma$
- Consistent: $\widehat{\Sigma}_N \to \Sigma$ as $N \to \infty$.
- Asymptotic distributional properties available

Example: Covariance estimation

Reasonable properties: (p fixed, N increasing)

- Unbiased: $\mathrm{E}\,\widehat{\Sigma}_N = \Sigma$
- Consistent: $\widehat{\Sigma}_N \xrightarrow{a.s.} \Sigma$ as $N \to \infty$.
- Asymptotic distributional properties available

An alternative experiment:

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

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

Example: Covariance estimation

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 Σ. Distance between estimator and population quantity can be measured through the so-called operator norm.

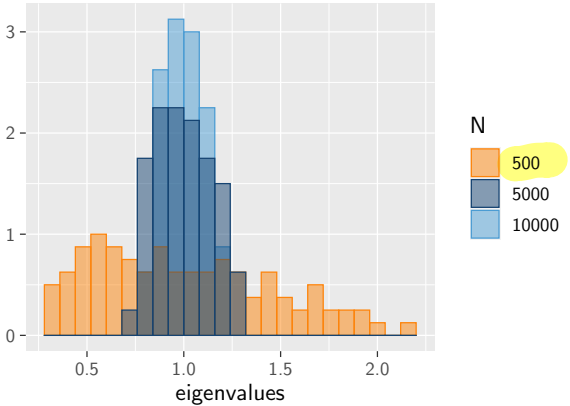
Consider the eigenvalues of $\widehat{\Sigma}$:

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

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

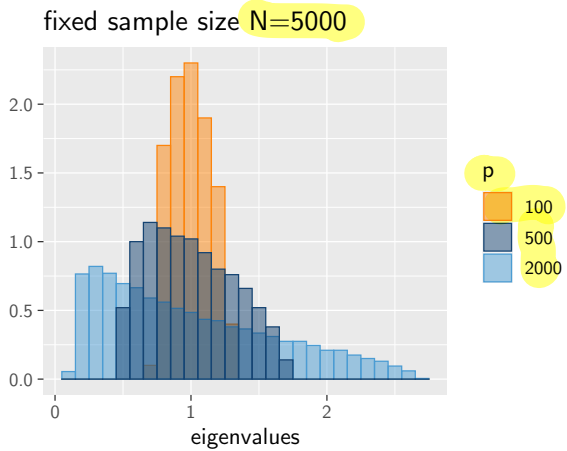


What we expect:

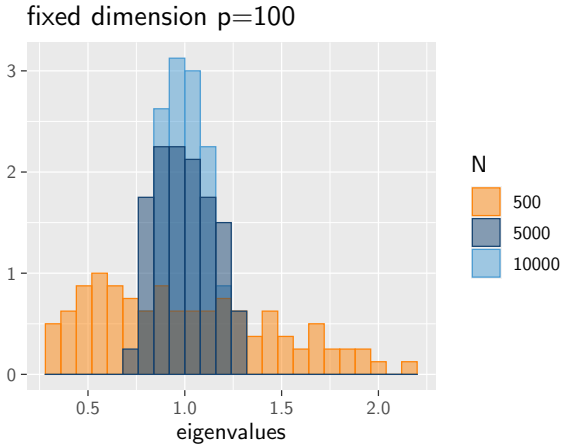
- all eigenvalues of true $\Sigma = I_p$ are 1
- plot will peak at 1 with increasing N

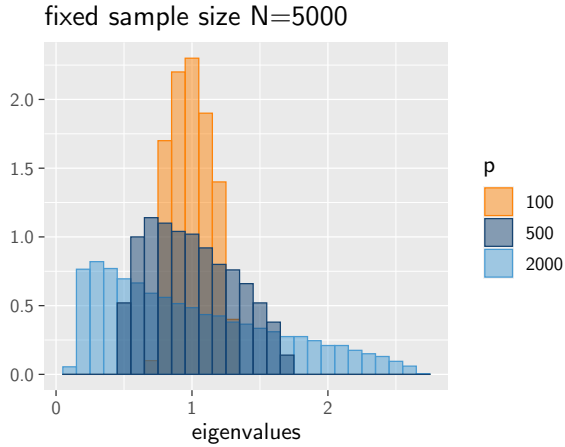
What happens in high-dimensional setting:

- doesn't center around 1 with increasing dimension



Example


fixed dimension p=100
3
2
1
0
0.5
1.0
1.5
2.0
eigenvalues
N
500
5000
10000


fixed sample size N=5000
2.0
1.5
1.0
0.5
0.0
0
1
2
eigenvalues
p
100
500
2000

Example: linear models

What can go wrong in high dimensions?

Setup:

- Linear regression

$$y = X\beta + \varepsilon$$

- We aim to recover the unknown true β_0 .

Classical approach:

- Given the observations (y, X) we get OLS $\widehat{\beta} = (X^T X)^{-1} X^T y$.

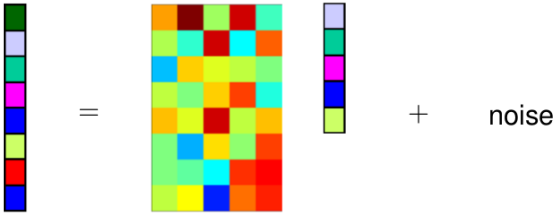
Reasonable properties: (*p* **fixed,** *N* **increasing**)

- Unbiased: $\mathrm{E}\,\widehat{\beta} = \beta_0$
- Consistent: $\widehat{\beta} \to \beta$ as $N \to \infty$.
- Asymptotic distributional properties available

Example: linear models

regular regime p fixed n large

$$y = X \beta + \epsilon$$



$n \times 1$ $n \times p$ $p \times 1$ $n \times 1$

Example: linear models

What can go wrong in high dimensions?

Reasonable properties: (p fixed, N increasing)

- Unbiased: $\mathrm{E}\,\widehat{\beta} = \beta_0$
- Consistent: $\widehat{\beta} \to \beta_0$ as $N \to \infty$.
- Asymptotic distributional properties available

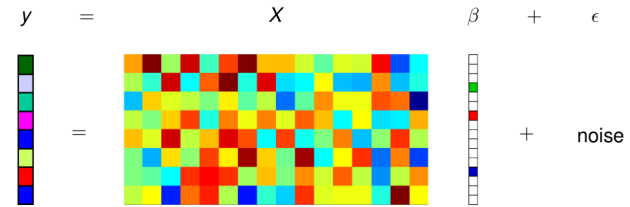
An alternative experiment:

- $p >> N$
- Important prior : many extracted feature in X are irrelevant
- many coefficients in β_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



Summary

What are the different view points?

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

What can go wrong in highdimensions?

- 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

$$\mathrm{P}[|X - \mathrm{E}[X]| > t] \leq \varepsilon(t)$$

nonasymptotic

$\varepsilon(t)$ is small

$\mathrm{P}(|X - \mathrm{E}X| \leq t) \geq 1 - \varepsilon(t)$

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

    HW: Proof Markov's ≠

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

    $\mathrm{E}|X| < \infty$

    $$\mathrm{P}[X > t] \leq \frac{\mathrm{E}[X]}{t}, \quad \forall t > 0.$$

    $O(\frac{1}{t})$

    $\mathbb{1}_{\{X \geq t\}} = \begin{cases} 1 & X \geq t \\ 0 & \text{otherwise} \end{cases}$
  - Chebyshev's: Assume $\mathrm{E}\,X^2 < \infty$

$$\mathrm{P}[|X - \mathrm{E}[X]| > t] \leq \frac{\mathrm{Var}(X)}{t^2}, \quad \forall t > 0.$$

$O(\frac{1}{t^2})$

$= \mathrm{P}(|X - \mathrm{E}X|^2 > t^2)$

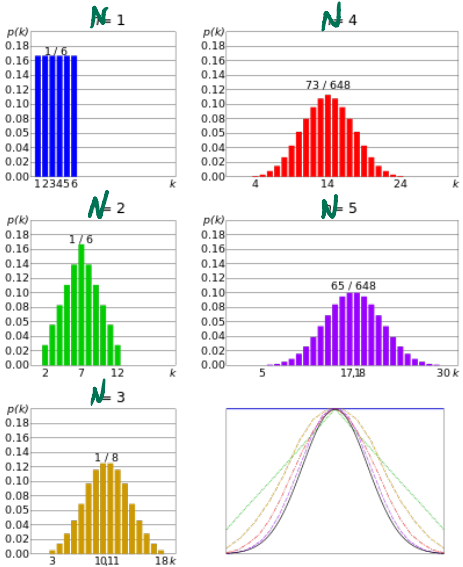
- Stronger assumption $\mathrm{E}\,|X|^k < \infty$

$$\mathrm{P}[|X - \mathrm{E}[X]| > t] \leq t^{-k}\,\mathrm{E}[|X - \mathrm{E}[X]|^k].$$

$\mathrm{P}(|X - \mathrm{E}X|^k > t^k)$

$\mathrm{P}(|X - \mathrm{E}X| > t) \leq \inf_{k \geq 1} t^{-k}\,\mathrm{E}|X - \mathrm{E}X|^k$

Recall 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 $N \to \infty$,

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

Some Motivation

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

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

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

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

Goal: avoid $n \to \infty$

$G \sim N(0,1)$

$P(Z_n > t) = P\left(\frac{S_n - n/2}{\sqrt{n/4}} > t\right) = \mathrm{P}\left(S_n > \frac{n}{2} + \sqrt{\frac{n}{4}}t\right) \overset{\text{CLT}}{\approx} P(G > t) \overset{\text{HW}}{\leq} \frac{1}{2}e^{-\frac{t^2}{2}}$ (*)

$P(G > t) = \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})\,dx$

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

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

- Problem: Approximation is not tight in general.

Theorem (Berry-Esseen CLT)

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

$$Z_n = \frac{S_n - \mathrm{E}S_n}{\sqrt{\mathrm{Var}(S_n)}}, \quad S_n = \sum_{i=1}^n X_i \qquad |\,\mathrm{P}(Z_n > t) - \mathrm{P}(G > t)| \le \frac{\delta}{\sqrt{n}}. \quad (*)$$

- Bound is tight since for Bernoulli example $\mathrm{P}(S_n = n/2) = \frac{1}{2^n}\binom{n}{n/2} \sim \frac{1}{\sqrt{n}}$
- 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 $\lambda > 0$,

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

- 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 μ is *sub-Gaussian* with parameter $\sigma > 0$ if

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

MGF

We say that X is σ-sub-Gaussian and say it has *variance proxy* σ^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 $|X - \mathrm{E}|X]| \leq M$ 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 $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

$$\mathbb{E}[e^{\lambda X}] = \int_{-\infty}^{\infty} e^{\lambda x} f(x)\, dx$$

$$= \int_{-\infty}^{\infty} e^{\lambda x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\lambda x - \frac{x^2}{2}\right) dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x^2 - 2\lambda x)\right) dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x^2 - 2\lambda x + \lambda^2) + \frac{1}{2}\lambda^2\right) dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-\lambda)^2\right) \exp\left(+\frac{1}{2}\lambda^2\right) dx$$

$$= \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\lambda)^2\right) dx}_{\text{density of Gaussian RV}} \exp\left(\frac{1}{2}\lambda^2\right) = \exp\left(\frac{1}{2}\lambda^2\right)$$

Theorem (Tail bound for sub-Gaussian random variables)

If a random variable X with finite mean μ is σ-sub-Gaussian, then

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

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

$$P(X-\mu \geq t) = P(\exp(\lambda(X-\mu)) \geq \exp(\lambda t))$$

$\exp(\lambda(X-\mu)) \geq 0$

Markov's

$$\overset{+}{\leq} \mathbb{E}(\exp(\lambda(X-\mu))) \frac{1}{\exp(\lambda t)}$$

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

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

$$P(X-\mu \geq t) \leq \inf \exp(\sigma^2\lambda^2/2 - \lambda t) = \exp\left(\sigma^2 \frac{t^2}{\sigma^4}/2 - \frac{t^2}{\sigma^2}\right)$$

$$= \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

$$\frac{d}{d\lambda}\left[\sigma^2\lambda^2/2 - \lambda t\right] = \sigma^2\lambda - t \overset{!}{=} 0 \Rightarrow \lambda = \frac{t}{\sigma^2}$$

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

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 σ_i-sub-Gaussian, that means: $\mathbb{E}(\exp(\lambda(X_i - \mathbb{E}X_i)) \le \exp(\sigma_i^2\lambda^2/2)$ (*)

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

$Z = \sum_{i=1}^n X_i$

$$\mathbb{E}[\exp(\lambda(Z-\mathbb{E}Z))] = \mathbb{E}[\exp(\lambda(\sum_{i=1}^n X_i - \sum_{i=1}^n \mathbb{E}X_i))]$$
$$= \mathbb{E}[\exp(\sum_{i=1}^n \lambda(X_i - \mathbb{E}X_i))]$$
$$= \mathbb{E}[\prod_{i=1}^n \exp(\lambda(X_i - \mathbb{E}X_i))]$$
X_i indep. $=$ $\prod_{i=1}^n \mathbb{E}(\exp(\lambda(X_i - \mathbb{E}X_i)))$
$\le \prod_{i=1}^n \exp(\sigma_i^2\lambda^2/2) = \exp(\sum_{i=1}^n \sigma_i^2\lambda^2/2)$

(*) X_i sub-Gaussian ✓

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

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

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

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

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

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

$$\Rightarrow P\left(\left|\sum_{i=1}^{n} X_i - \underbrace{E\left(\sum_{i=1}^{n} X_i\right)}_{n\mu}\right| \geq \left(\frac{\log(n)}{n}\right)^{\frac{1}{2}}\right) \leq 2\exp\left(-\log(n)/2\right)$$

$$= 2\exp\left(-\log(n^{1/2})\right)$$

$$= 2\, n^{-\frac{1}{2}} \xrightarrow[n\to\infty]{} 0$$

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 $\mu = \mathrm{E}\, S_N$. Then, for any $t > \mu$, we have

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

Proof:

$$\begin{aligned}
&\mathrm{P}(S_N \geq t) \\
&= \mathrm{P}\Big(\sum_{i=1}^N X_i \geq t \Big) \\
&= \mathrm{P}\big(\exp\big(\lambda \textstyle\sum_{i=1}^N X_i\big) \geq \exp(\lambda t) \big) \\
&\leq \mathrm{E}\big[\exp\big(\lambda \textstyle\sum_{i=1}^N X_i\big) \big] \,/\, \exp(\lambda t) \\
&= \mathrm{E}\Big[\prod_{i=1}^N \exp(\lambda X_i) \Big] \,/\, \exp(\lambda t)
\end{aligned}$$

Equivalent characterizations of sub-Gaussianity

For a RV X, the following are equivalent:

1. The tails of X satisfy
$$\mathrm{P}(|X| > t) \leq 2\exp(-t^2/K_1^2), \text{ for all } t \geq 0.$$
2. The moments of X satisfy
$$(\mathrm{E}\,|X|^p)^{\frac{1}{p}} \leq K_2\sqrt{p}, \text{ for all } p \geq 1.$$
3. The MGF of X^2 satisfies
$$\mathrm{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2\lambda^2), \text{ for all } |\lambda| \leq \frac{1}{K_3}.$$
4. The MGF of X^2 is bounded at some point,
$$\mathrm{E}\exp(X^2/K_4^2) \leq 2.$$

Sub-exponential concentration

Definition

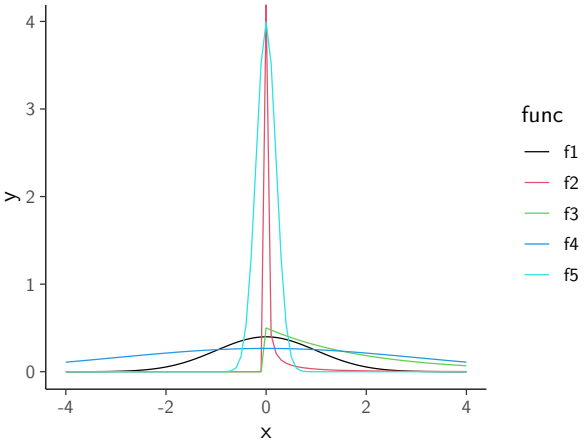
A zero-mean random variable X is sub-exponential if for some $\nu, \alpha > 0$.

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

A general random variable is sub-exponential if $X - \mathrm{E}\, X$ 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



Sub-exponential concentration

Proposition

Assume that X is zero-mean sub-exponential with parameters (ν, α). Then,

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

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 $\gamma \geq 0$,

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

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

Proof:

Lipschitz functions of standard Gaussian vector

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ 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 $t \in \mathbb{R}$,

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

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 $\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_k(X)$ 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, $X \to \sigma_k(X)$ is 1-Lipschitz:

Proposition

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

$$\mathrm{P}\left(|\sigma_k(X) - \mathrm{E}\,\sigma_k(X)| \geq \delta\right) \leq 2\exp(-\delta^2/2)$$

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,

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

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

Part III

Sparse linear models

Outline

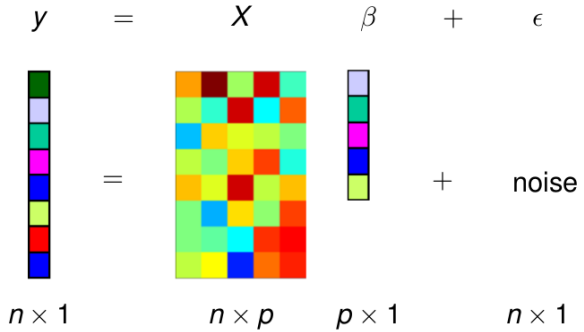
1. Linear regression setup
2. Recall what goes wrong in large dimensions
3. Ridge Regression
4. Lasso Regression
5. Comparison

Linear regression setup I

Recall the linear regression setup in low dimension

- The data is (y, X) where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$, and the model

$$y = X\beta + \varepsilon$$



Linear regression setup II

The data is (y, X) where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$, and the model

$$y = X\beta + \varepsilon$$

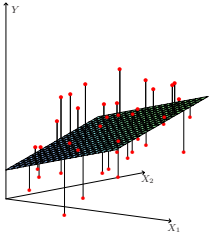


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

What can go wrong in high dimensions?

The data is (y, X) where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$, and the model

$$y = X\beta + \varepsilon$$

Suppose $p > n$

Ridge Regression

Ridge estimator: For any $\lambda > 0$, set

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

with $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$. For any $\lambda > 0$ the solution to the minimization problem is

$$\widehat{\beta} = (X'X + \lambda I_{p \times p})^{-1} X'y.$$

- When $\lambda = 0$, we get the linear regression estimate.

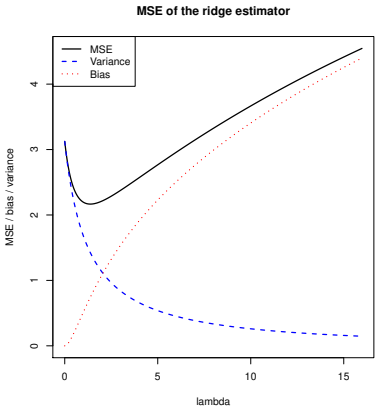
Some properties:

- $\text{Bias}(\widehat{\beta}) = \mathrm{E}\,\widehat{\beta} - \beta_0 = -\lambda(X'X + \lambda I)^{-1}\beta$
- $\text{Var}\,\widehat{\beta} = \mathrm{E}[(\widehat{\beta} - E(\widehat{\beta}))^2] = (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1}\sigma^2$
- $\text{MSE}(\widehat{\beta}) = \text{Bias}^2(\widehat{\theta}) + \text{Var}(\widehat{\theta}) = \text{tr}((X'X + \lambda I)^{-2}(\lambda^2 \beta\beta' + \sigma^2 X'X))$

Questions:

- What is the bias at $\lambda = 0$?
- What is the variance at $\lambda = \infty$?

RIDGE: Bias-variance tradeoff

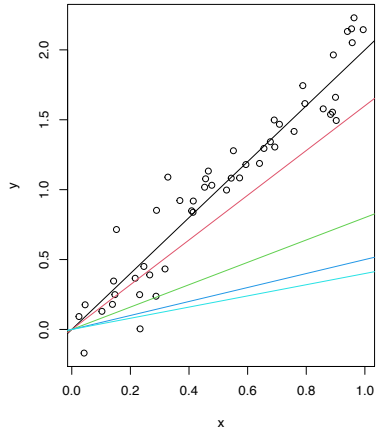


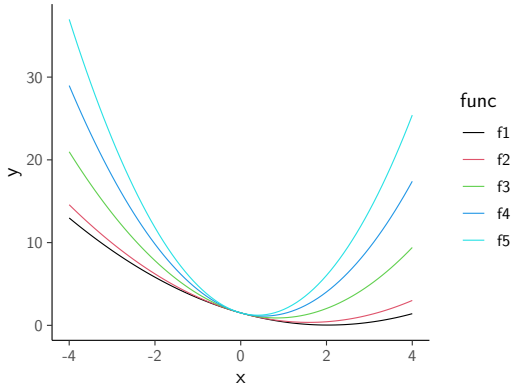
Observations

- The bias increases as λ (amount of shrinkage) increases.
- The variance decreases as λ (amount of shrinkage) increases.
- Recall our earlier questions: What is the bias at $\lambda = 0$? The variance at $\lambda = \infty$?

Observations: Ridge shrinks components of its estimate toward zero, but never sets these components to be zero exactly (unless $\lambda = \infty$, in which case all components are zero). So strictly speaking, ridge regression does not perform variable selection.

Example: One-dimensional case: $\text{argmin}_\alpha \frac{1}{2n}\|y - \alpha x\|_2^2 + \lambda|\alpha|^2$

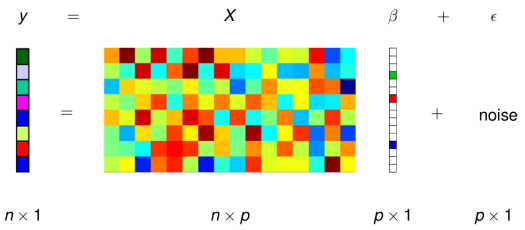




Sparsity

An introductory example:

- In many applications, $p > n$ but...
- Important prior: many extracted features in X are irrelevant
- In an equivalent way: many coefficients in β_0 are exactly zero.
- For example, if 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.



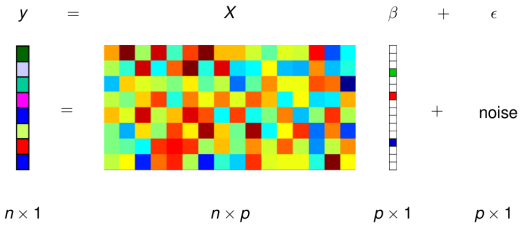
- Sparsity: Assumption that the unknown β_0 has entries which are exactly equal to zero. In other words, only s of the $p$entries are non-zero and determine the important features.
- Sparsity assumption: $|S| = s$ with $S = \text{supp}(\beta_0) = \{i \in \{1, \dots, p\} \mid \beta_{0,i} \neq 0\}$.
- It permits to reduce the effective dimension of the problem.

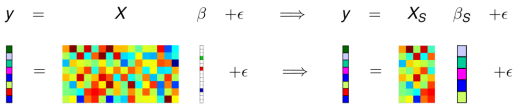
Sparsity

What does this dimension reduction look like?

Sparsity assumption: $|S| = s$ with $s << n$ and $S = \text{supp}(\beta_0) = \{i \in \{1, \dots, p\} \mid \beta_{0,i} \neq 0\}$.

Sparsity assumption allows to reduce the effective dimension of the problem.





Hope: $X_S' X_S$ has full rank and lienar model can be applied.

Problem: This only works if the support of β_0 is known!!!

Sparsity models

What would be a natural relaxation of the problem?
Subset selection

$$\min_{\beta \in \mathbb{R}^q} \|y - X\beta\|_2^2 \text{ subject to } \|\beta\|_0 \leq s$$

with

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}},$$

Sparsity: Assume β_0 is a s–sparse, i.e. $\|\beta_0\|_0 = \sum_{j=1}^{p} \|\beta\|_0 = s$.

- **Problem:** $\|\beta\|_0$ is not convex.
- **Idea:** use convex relaxation of the l_0 norm: instead of considering a variable $z \in \{0, 1\}$, imagine that $z \in [0, 1]$.

LASSO estimator:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{N} \|y - X\beta\|_2^2 + \lambda_N \|\beta\|_1 \tag{1}$$

- A large value of λ leads to a very sparse solution, with bias.
- A low value of λ yields overfitting with no penalization (too much variance).

LASSO Regression

LASSO estimator: For any $\lambda > 0$, set

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{N} \|y - X\beta\|_2^2 + \lambda_N \|\beta\|_1$$

with $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$.

LASSO does not have an explicit solution in general!

But we can find an explicit solution in the case of orthogonal design matrix, e.g. $X'X = I$. Then:

$$\widehat{\beta}_i = \max\{\widehat{\beta}_i^{OLS} - \frac{\lambda}{2}, 0\} \text{ for } \widehat{\beta}_i^{OLS} > 0$$

$$\widehat{\beta}_i = \min\{\widehat{\beta}_i^{OLS} + \frac{\lambda}{2}, 0\} \text{ for } \widehat{\beta}_i^{OLS} \leq 0$$

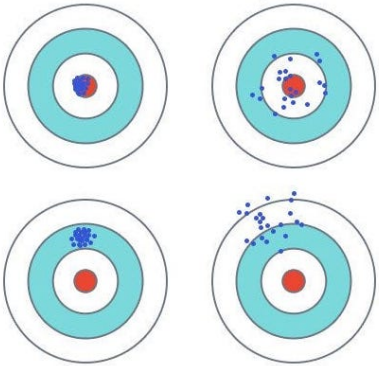
LASSO Regression

For orthogonal design matrix, e.g. $X'X = I$, the Lasso has the solution

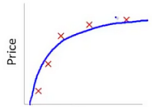
$$\widehat{\beta}_i = \max\{\widehat{\beta}_i^{OLS} - \frac{\lambda}{2}, 0\} \text{ for } \widehat{\beta}_i^{OLS} > 0$$

$$\widehat{\beta}_i = \min\{\widehat{\beta}_i^{OLS} + \frac{\lambda}{2}, 0\} \text{ for } \widehat{\beta}_i^{OLS} \leq 0$$

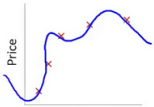
Illustrations bias-variance



$\theta_0 + \theta_1 x$



$\theta_0 + \theta_1 x + \theta_2 x^2$



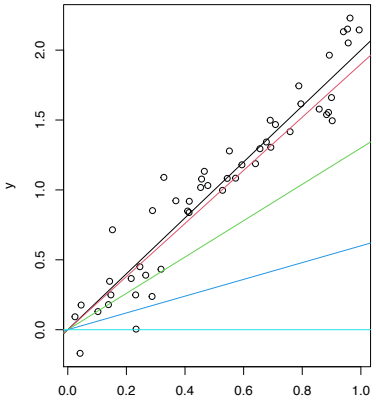
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

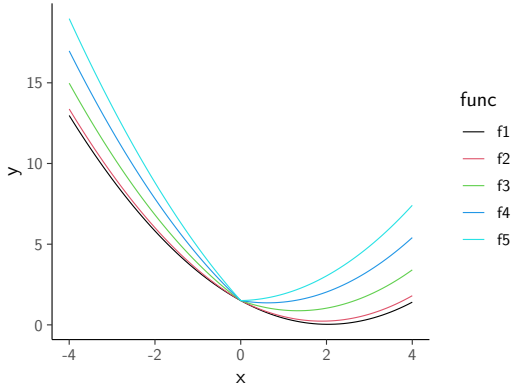
LASSO: Bias-variance tradeoff: What are the implications for LASSO?

- The bias increases as λ (amount of shrinkage) increases.
- The variance decreases as λ (amount of shrinkage) increases.

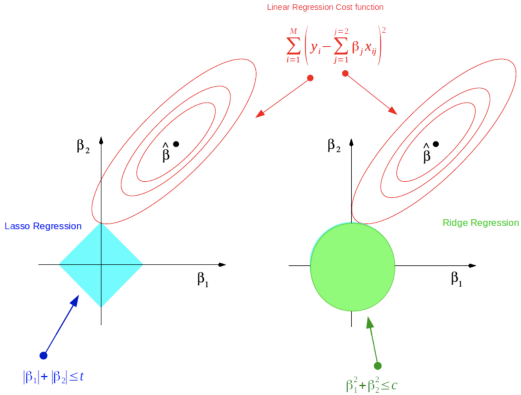
LASSO

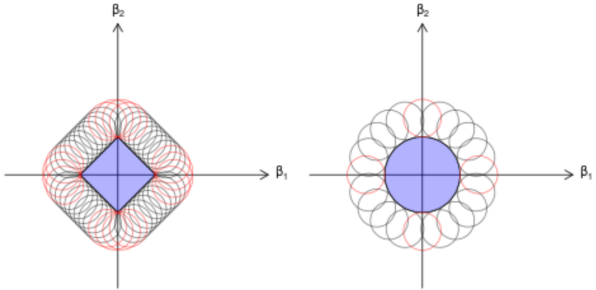
Observations: A large value of λ leads to a very sparse solution, with bias. A low value of λ yields overfitting with no penalization (too much variance).
Example: One-dimensional case: $\text{argmin}_\alpha \frac{1}{2n}\|y - \alpha x\|_2^2 + \lambda|\alpha|$





Comparision of penalties: Ridge vs LASSO





Major difference between ridge and lasso:

- the sharp, non-differentiable corners of the l_1-ball produce parsimonious models for sufficiently large values of λ
- the lasso lacks an analytic solution, making both computation and theoretical results more difficult

Summary of penalizations

Three canonical choices of penalization: the ℓ_0, ℓ_1 and ℓ_2 (norms):

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2 \right)^{\frac{1}{2}}$$

Penalized forms:

- Best subset selection

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_0$$

- Lasso regression

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

- Ridge regression

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Sparsity: Assume β_0 is a s–sparse, i.e. $\|\beta_0\|_0 = \sum_{j=1}^{p} \|\beta\|_0 = s$.

Summary

What are possible penalization and their advantages and disadvantages?

- Ridge:
  - penalizes with l_2-norm $\|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2\right)^{\frac{1}{2}}$
  - explicit representation
  - shrinks towards small values
  - does not do model selection
- Best subset selection:
  - penalizes with l_0-norm $\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$,
  - not convex
- LASSO:
  - penalizes with l_1-norm $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$
  - shrinks exactly to zero
  - convex

What is the bias-variance trade-off?

- The bias increases as λ (amount of shrinkage) increases.
- The variance decreases as λ (amount of shrinkage) increases.