High-Dimensional Statistics

I. Introduction

high-dimensional (model): more parameters than sample size

no. par. $= p$; n

$p \gg n$

Notation: asymptotic viewpoint

$p = p_n$, $p_n \to \infty$ $(n \to \infty)$, e.g. $p_n = n^\alpha$ $\alpha > 0$!

$p \gg n \iff \frac{p}{n} = \frac{p_n}{n} \to \infty$ $(n \to \infty)$

$$p = O(a_n) \iff \frac{p}{a_n} = \frac{p_n}{a_n} \le M < \infty \quad \forall n$$

$$p = o(a_n) \iff \frac{p}{a_n} = \frac{p_n}{a_n} \to 0 \quad (n \to \infty)$$

$$p \asymp a_n \iff p_n = O(a_n), \; a_n = O(p_n)$$

$$p \sim a_n \iff p_n / a_n \to 1 \quad (n \to \infty)$$

$p \gg n$ is common in nonparametric statistics

example: $Y_i = f(X_i) + \varepsilon_i$, $X_i \in \mathbb{R}^1$, $Y_i \in \mathbb{R}^1$

$f(\cdot)$ smooth, $E[\varepsilon_i] = 0$

$f(\cdot)$ is an infinite-dimensional parameter

e.g. $f(x) = \sum_{j=1}^{\infty} b_j \phi_j(x)$ ← basis functions; par $= b_1, b_2, \ldots$

assuming smoothness of $f(\cdot)$ is typically required
to get a good estimator $\hat{f}(\cdot)$ for $f(\cdot)$

for example: smoothing splines with $p \asymp n$ which need to
be estimated

new view: sparsity assumption (in statistics: Donoho, Johnstone ≈ 1990)

that is: p parameters but many of them are zero
(approx. zero)

this is another assumption which guarantees good estimators
in the $p \gg n$ setting

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\beta + \varepsilon$$ linear model

$n = 115$

$p = 4088$ $\rightsquigarrow$ ordinary least squares "does not work"

I. 1. High-dimensional linear model

a) a prime example

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\,\underset{p\times 1}{\beta^0} + \underset{n\times 1}{\varepsilon}, \qquad X \text{ fixed (or random and conditioning on } X)$$

$$\mathbb{E}[\varepsilon] = 0$$

$$S^0 = \mathrm{supp}(\beta^0) = \{j;\ \beta^0_j \neq 0\}$$

sparsity index $s_0 = |S^0|$. assume $s_0 = |S^0| \ll n$

in many practical applications: assuming sparsity is "reasonable"

example: biomarker discovery

Riboflavin production data

⟶ "hope" that only a few genes are relevant

for Y = log-production rate

$$Y_i = \sum_{j=1}^{p} \beta_j^0 X_i^{(j)} + \varepsilon_i$$

Y_i: log-prod. rate

$X_i^{(j)}$: expr. of gene j

II. Lasso for linear models

II.1. Introduction

$$Y = X\beta^* + \varepsilon$$

in practice: centered and scaled variables,

$$Y_i \leftarrow Y_i - \bar{Y}$$

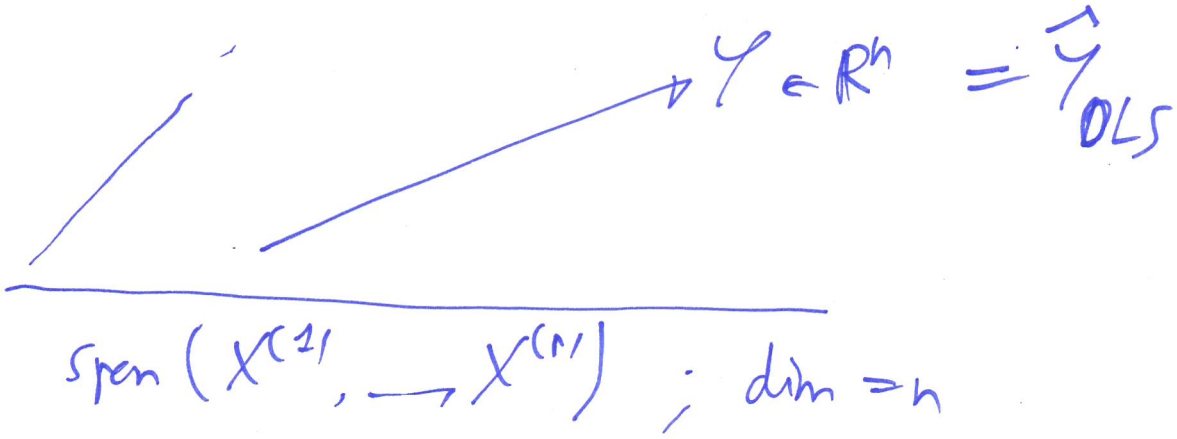
$$X_i^{(j)} \leftarrow \frac{X_i^{(j)} - \bar{X}^{(j)}}{\hat{\sigma}_j}, \quad \hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i^{(j)} - \bar{X}^{(j)})^2$$

II. 2. The Lasso estimator

most popular estimator for high-dim. lin. model
(Tibshirani, 1996)

for $p > n$: $\operatorname{rank}(X) < p$

⟶ ordinary least squares is not unique:
it overfits and produces residual sum of squares $= 0$



$\rightsquigarrow$ need complexity regularization

$\sum_{j=1}^{p} |\beta_j|$

Lasso:

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\|Y - X\beta\|_2^2 / n + \lambda \|\beta\|_1 \right)$$

$\lambda \geq 0$ penalty parameter

different from **Ridge regression** (Tikhonov regularization)

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

$\lambda \geq 0$

First properties:

(1) Lasso is sparse estimator

$\hat{\beta}_j(\lambda) = 0$ for "many" j's (depending on λ)

"Lasso is doing variable selection"

LASSO = Least Absolute Shrinkage and Selection Operator

not true for Ridge estimator

(2) Lasso involves convex optimization: "easy"

every local minimum is a global one

but in general not unique

(under additional conditions $\rightsquigarrow$ uniqueness, see later)

Explanation for (1): due to convexity

$$\hat{\beta}(\lambda) = \operatorname{argmin}_{\beta} \left(\|Y - X\beta\|_2^2 / n + \lambda \|\beta\|_1 \right) \quad \text{"Lagrangian form"}$$

$\Longleftrightarrow$

primal problem:

$$\hat{\beta}_{\text{primal}}(R) = \operatorname*{argmin}_{\beta : \|\beta\|_1 \le R} \|Y - X\beta\|_2^2 / n$$

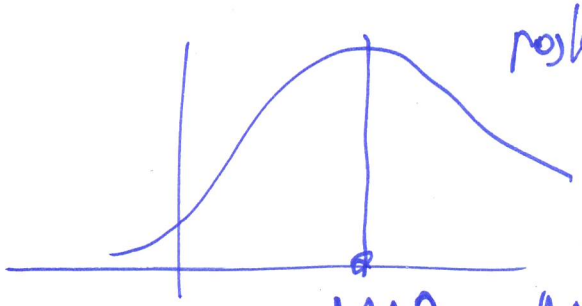
equivalent optimization with a one-to-one correspondence between λ and R (not explicit; depends on data $(Y_i, X_i)_{i=1,\dots,n}$)

software: glmnet in R
including CV for choosing λ

Bayesian :

$\beta_1, \dots, \beta_p$ i.i.d. $\sim \mathcal{N}(0, \tau^2)$ prior

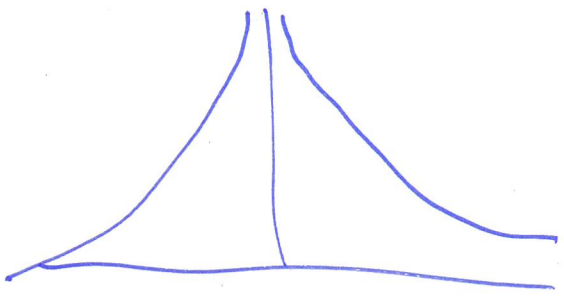
$\rightsquigarrow$ posterior assuming $\varepsilon \sim$ Gaussian



$\rightsquigarrow = \hat{\beta}_{Ridge}$

MAP = Maximum A Posteriori mode

$\beta_1, \ldots, \beta_p$ i.i.d. $\sim$ Double-Exp. (Laplace) mean 0 variance τ^2



$\varepsilon \sim$ Gaussian

$\longrightarrow$ MAP = Lasso