High-Dimensional Statistics

I. Introduction

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

no. parameters: p

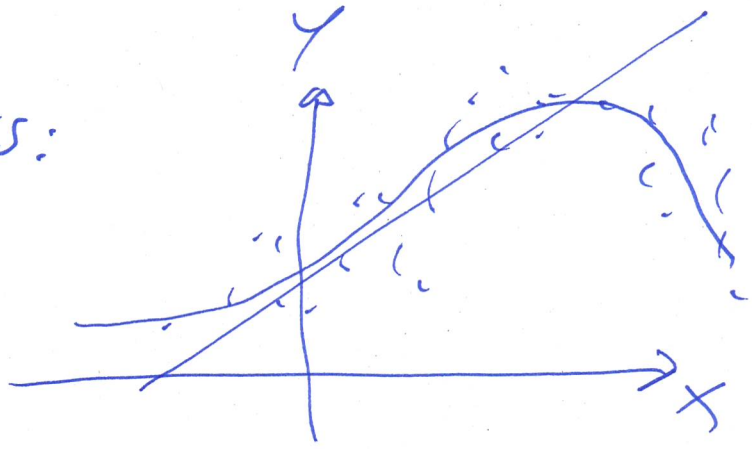
n

$p \gg n$

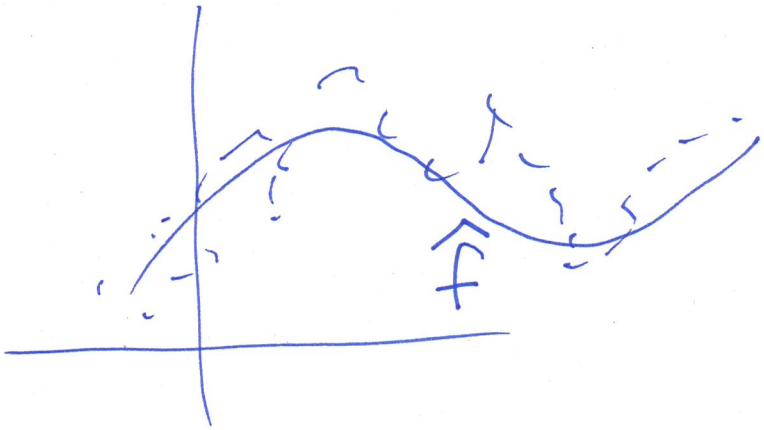
Very common in nonparametric statistics:

e.g. $Y = f(X) + \varepsilon$

nonparametric smooth function $\mathbb{R} \to \mathbb{R}$: ∞-dim. object



for smoothing spline method:



$\hat{f}$ parameterized (with splines) with $p \gtrsim n$ parameters

"order"

and smoothness of f is crucial to deal with this problem

new view: sparsity instead of smoothness

p parameters but many of them are zero
(but unknown which of them are ~~zero~~)

prime example:

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\,\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \qquad p \gg n$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_1^{(1)} & \cdots & X_1^{(p)} \\ \vdots & & \vdots \\ X_n^{(1)} & & X_n^{(p)} \end{pmatrix}$$

high-dimensional linear model

p different features, covariates

sparsity: $S = \text{supp}(\beta) = \{j;\ \beta_j \neq 0\}$

active set of covariates

cardinality of S: $|S| \ll n$

$$Y = \underset{n \times p}{X} \underset{p \times 1}{\beta} + \varepsilon = \underset{n \times |S|}{X_S} \underset{|S| \times 1}{\beta_S} + \varepsilon$$

practical implication: "throw in many potentially useless covariates — so that we reduce the chance to miss an important one"

What is the price of collecting too many covariates ?

(besides the benefit to catch the important ones)

II. The Lasso for linear models (Ch. 2 in Bühlmann & van de Geer)

II.1. Introduction

data are realizations of

$(X_1, Y_1), \ldots, (X_n, Y_n)$

random

fixed or random

$X_i \in \mathcal{X} \subseteq \mathbb{R}^p$; $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$

useful linear model:

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \qquad (i = 1, \dots, n)$$

$$X_i = \left(X_i^{(1)}, \dots, X_i^{(p)}\right)^T, \quad X^{(j)} = \left(X_1^{(j)}, \dots, X_n^{(j)}\right)^T$$

$$\varepsilon_1, \dots, \varepsilon_n \text{ i.i.d.}, \quad \mathbb{E}[\varepsilon_i] = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2$$

independent of $X_1, \dots, X_n$

often in practice: centered and scaled

$$Y_i \leftarrow Y_i - \bar{Y}; \quad X_i^{(j)} \leftarrow \frac{X_i^{(j)} - \bar{X}^{(j)}}{\hat{\sigma}^{(j)}}$$

$$\left(\hat{\sigma}^{(j)}\right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i^{(j)} - \bar{X}^{(j)}\right)^2$$

notation: $Y = X\beta + \varepsilon$

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

(if model is assumed correct: "true" parameter is β^0)

II. 2. The Lasso estimator

For $p > n$: OLS is not unique
will overfit the data

⟶ complexity regularization

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

$\lambda > 0$ regularization parameter

makes only sense if $\hat{\sigma}^{(j)}$ are (roughly) equal

different from Ridge Regression:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left(\|Y - X\beta\|_2^2 / n + \lambda \|\beta\|_2^2 \right)$$

fast properties of Lasso:

(1) sparse estimator

$\hat{\beta}_j(\lambda) = 0$ depending on j and λ

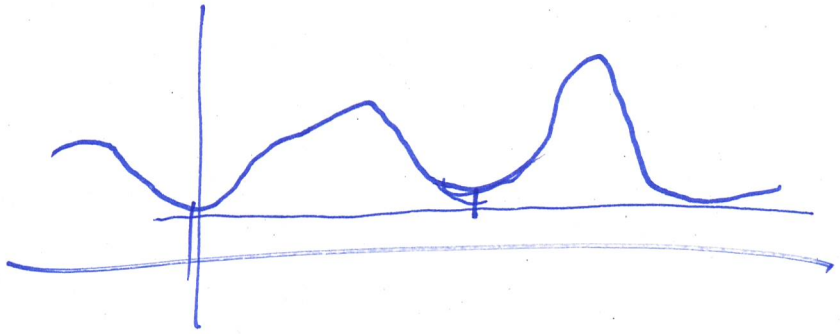
"Lasso is doing variable selection"

in contrast to Ridge regression

(2) Lasso involves convex optimization:

every local minimum is a global minimum

there are typically many of them
all of them being global



but under further assumptions: uniqueness

Explanation for (1):

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\| \cdots \|_2^2 / n + \lambda \|\beta\|_1 \right)$$ Lagrangian form

primal problem:

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

convex constraints

the two optimization problems are equivalent with a one-to-one correspondence between λ and R (depending on $(X_i, Y_i)_{i=1}^n$)

For Ridge:

$$\hat{\beta}_{primal}(R) = \underset{\beta:\ \|\beta\|_2 \le R}{\operatorname{argmin}} \|y - X\beta\|_2^2$$

Lasso: Tibshirani (1996)

Least Absolute Selection and Shrinkage Operator

Sparsity

II.3 Orthonormal design

X with $n^{-1} X^T X = I_{p \times p}$ $(\rightarrow p \leq n)$

Then: Lasso

$$\hat{\beta}_j(\lambda) = \text{sign}(z_j)\,(|z_j| - \lambda/2)_+ \qquad j = 1, \ldots, p$$

$$z_j = (X^T Y)_j / n = \hat{\beta}_{OLS;j}$$