Sylvain Le Corff

Introduction to statistical learning

Contents

Notations

For all $n \geq 1$, I_n is the identity matrix of size $n \times n$ and $\mathbf{1}_n$ is the vector of size n with all entries equal to 1. For all matrix $A \in \mathbb{R}^{n\times d}$, $A^\top$ is the transpose of A. For all $1 \leqslant i \leqslant n$, $A_{i,\cdot}$ is the column vector given by the i-th row of A and for all $1 \leqslant j \leqslant d$, $A_{\cdot,j}$ is the column vector given by the j-th column of A. For all $(a_1,\dots,a_p) \in \mathbb{R}^p$, $\mathrm{diag}(a_1,\dots,a_p)$ is the diagonal matrix of size $p \times p$ with diagonal given by $(a_1,\dots,a_p)$.

Chapter 1

Principal component analysis

Contents

Keywords 1.1 *Principal components; singular value decomposition.*

Principal component analysis is a multivariate technique which aims at analyzing the statistical structure of high dimensional dependent observations by representing data using orthogonal variables called *principal components*. Its origin may be traced back to [Hotelling, 1933] who first introduced the principal components as a way to reduce the dimensionality of the data. Reducing the dimensionality of the data is motivated by several practical reasons such as improving computational complexity.

Let $(X_i)_{1\leqslant i\leqslant n}$ be i.i.d. random variables in $\mathbb{R}^d$ and consider the matrix $X \in \mathbb{R}^{n\times d}$ such that the i-th row of X is the observation $X_i^\top$. In this chapter, it is assumed that data are preprocessed so that the columns of X are centered (X is replaced by $X - n^{-1}\mathbf{1}_n^\top \mathbf{1}_n X$ if the columns of X are not centered). This means that for all $1 \leqslant j \leqslant d$, $\sum_{i=1}^n X_{i,j} = 0$. Let Σ_n be the empirical covariance matrix:

$$\Sigma_n = n^{-1}\sum_{i=1}^n X_i X_i^\top \,.$$

Principal Component Analysis aims at reducing the dimensionality of the observations $(X_i)_{1\leqslant i\leqslant n}$ using a *compression* matrix $W \in \mathbb{R}^{p\times d}$ with $1 \leqslant p \leqslant d$ so that for each $1 \leqslant i \leqslant n$, WX_i is a low dimensional representation of X_i. The original observation may then be partially recovered using another matrix $U \in \mathbb{R}^{d\times p}$. Principal Component Analysis computes U and W using the least squares approach:

$$(U_\star, W_\star) \in \underset{(U,W)\in\mathbb{R}^{d\times p}\times\mathbb{R}^{p\times d}}{\operatorname{argmin}} \sum_{i=1}^n \|X_i - UWX_i\|_2^2 \,.$$

1.1 Principal Component Analysis as a singular value decomposition problem

1.1.1 Singular value decomposition

Proposition 1.1 *For all $\mathbb{R}^{n\times d}$ matrix A with rank r, there exist $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ such that*

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^\top \,,$$

where $\{u_1,\ldots,u_r\} \in (\mathbb{R}^n)^r$ and $\{v_1,\ldots,v_r\} \in (\mathbb{R}^d)^r$ are two orthonormal families. The vectors $\{\sigma_1,\ldots,\sigma_r\}$ are called singular values of A and $\{u_1,\ldots,u_r\}$ (resp. $\{v_1,\ldots,v_r\}$) are the left-singular (resp. right-singular) vectors of A.

Remark 1.2 *If U denotes the $\mathbb{R}^{n\times r}$ matrix with columns given by $\{u_1,\ldots,u_r\}$ and V denotes the $\mathbb{R}^{p\times r}$ matrix with columns given by $\{v_1,\ldots,v_r\}$, then the singular value decomposition of A may also be written as*

$$A = UD_rV^\top \,,$$

where $D_r = \mathrm{diag}(\sigma_1,\ldots,\sigma_r)$.

Remark 1.3 *The singular value decomposition is closely related to the spectral theorem for symmetric semipositive definite matrices. In the framework of Proposition 7.11, $A^\top A$ and $AA^\top$ are positive semidefinite such that*

$$A^\top A = VD_r^2V^\top \quad \text{and} \quad AA^\top = UD_r^2U^\top \,.$$

PROOF. Since the matrix $AA^\top$ is positive semidefinite, its spectral decomposition is given by

$$AA^\top = \sum_{k=1}^{r} \lambda_k u_k u_k^\top \,,$$

where $\lambda_1 \geqslant \ldots \geqslant \lambda_r > 0$ are the nonzero eigenvalues of $AA^\top$ and $\{u_1,\ldots,u_r\}$ is an orthonormal family of $\mathbb{R}^n$. For all $1 \leqslant k \leqslant r$, define $v_k = \lambda_k^{-1/2}A^\top u_k$ so that

$$\|v_k\|^2 = \lambda_k^{-1}\langle A^\top u_k; A^\top u_k\rangle = \lambda_k^{-1}u_k^\top AA^\top u_k = 1 \,,$$
$$A^\top A v_k = \lambda_k^{-1/2}A^\top AA^\top u_k = \lambda_k v_k \,.$$

On the other hand, for all $1 \leqslant k \neq j \leqslant r$, $\langle v_k; v_j\rangle = \lambda_k^{-1/2}\lambda_j^{-1/2}u_k^\top AA^\top u_j = \lambda_k^{-1/2}\lambda_j^{1/2}u_k^\top u_j = 0$. Therefore, $\{v_1,\ldots,v_r\}$ is an orthonormal family of eigenvector of $A^\top A$ associated with the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_r > 0$. Define, for all $1 \leqslant k \leqslant r$, $\sigma_k = \lambda_k^{1/2}$ which yields

$$\sum_{k=1}^{r} \sigma_k u_k v_k^\top = \sum_{k=1}^{r} u_k u_k^\top A = \left(\sum_{k=1}^{r} u_k u_k^\top\right) A \,.$$

As $\{u_1,\ldots,u_r\}$ is an orthonormal family, by Lemma 7.9, $UU^\top = \sum_{k=1}^{r} u_k u_k^\top$ is the orthogonal projection onto the range$(AA^\top) =$ range(A) which implies

$$\sum_{k=1}^{r} \sigma_k u_k v_k^\top = \left(\sum_{k=1}^{r} u_k u_k^\top\right) A = A \,.$$

■

An illustration of SVD is given in Figure 1.1. In this setting, the matrix A is a 848 × 1280 matrix of grayscale pixels.

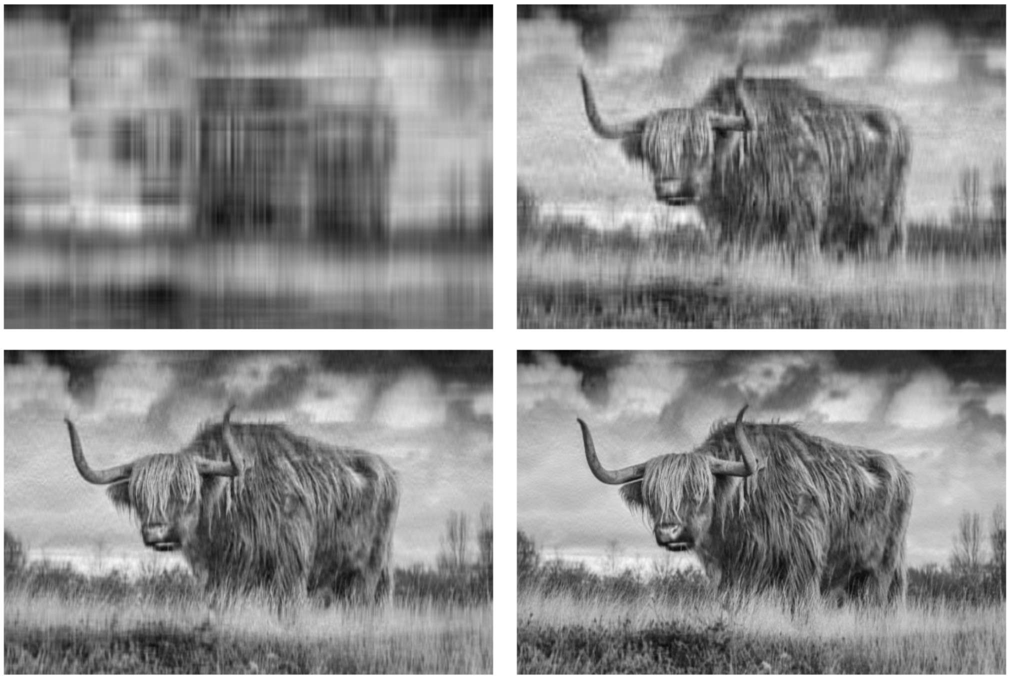

Fig. 1.1 Image reconstruction using the largest singular values of the SVD. The original grayscale image (from `https://pixabay.com/`) is given by a matrix ox pixels of size 848 × 1280. Reconstruction given with the top 5 (top left), 25 (top right), 50 (bottom left) and 100 (bottom right) singular values.

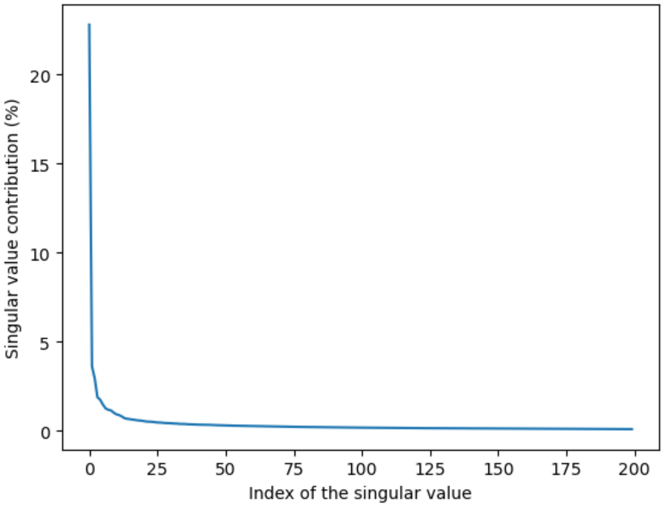

Fig. 1.2 Singular value contributions: for all $1 \leqslant i \leqslant 200$, the contribution of σ_i is given by $\sigma_i / (\sum_{j=1}^{200} \sigma_j)$.

1.1.2 Application to Principal Component Analysis

As mentioned in the introduction, Principal Component Analysis aims at solving the following optimization problem:

$$(U_\star, W_\star) \in \operatorname*{argmin}_{(U,W)\in\mathbb{R}^{d\times p}\times\mathbb{R}^{p\times d}} \sum_{i=1}^n \|X_i - UWX_i\|_2^2 \,. \tag{1.1}$$

Lemma 1.4 *Let $(U_\star, W_\star) \in \mathbb{R}^{d\times p}\times\mathbb{R}^{p\times d}$ be a solution to* (1.1)*. Then, the columns of $U_\star$ are orthonormal and $W_\star = U_\star^\top$.*

Let $U \in \mathbb{R}^{d\times p}$ be such that $U^\top U = \mathrm{I}_p$. Then,

$$\begin{aligned}
\sum_{i=1}^n \|X_i - UU^\top X_i\|_2^2 &= \sum_{i=1}^n \|X_i\|_2^2 + \sum_{i=1}^n \|UU^\top X_i\|_2^2 - 2\sum_{i=1}^n \langle X_i; UU^\top X_i\rangle \,,\\
&= \sum_{i=1}^n \|X_i\|_2^2 + \sum_{i=1}^n X_i^\top UU^\top X_i - 2\sum_{i=1}^n X_i^\top UU^\top X_i \,,\\
&= \sum_{i=1}^n \|X_i\|_2^2 - \sum_{i=1}^n X_i^\top UU^\top X_i \,,\\
&= \sum_{i=1}^n \|X_i\|_2^2 - \sum_{i=1}^n \mathrm{trace}(U^\top X_i X_i^\top U) \,.
\end{aligned}$$

Therefore, by Lemma 1.4, solving (1.1) boils down to computing

$$U_\star \in \operatorname*{argmax}_{U\in\mathbb{R}^{d\times p}\,,\, U^\top U = \mathrm{I}_n} \{\mathrm{trace}(U^\top \Sigma_n U)\} \,. \tag{1.2}$$

Proposition 1.5 *Let $\{\vartheta_1,\ldots,\vartheta_d\}$ be orthonormal eigenvectors associated with the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_d$ of Σ_n. Then a solution to* (1.1) *is given by the matrix $U_\star$ with columns $\{\vartheta_1,\ldots,\vartheta_p\}$ and $W_\star = U_\star^\top$.*

PROOF. By Lemma 1.4 and (1.2), the proof is equivalent to prove that $U_\star$ is a solution to (1.2). Let $\Sigma_n = VD_nV^\top$ be the spectral decomposition of Σ_n where $D_n = \mathrm{diag}(\lambda_1,\ldots,\lambda_d)$ and $V \in \mathbb{R}^{d\times d}$ is a matrix with columns $\{\vartheta_1,\ldots,\vartheta_d\}$. For all matrix $U \in \mathbb{R}^{d\times p}$ with orthonormal colums define $B = V^\top U$ so that, as $V \in \mathbb{R}^{d\times d}$ is an orthogonal matrix,

$$VB = VV^\top U = U \quad \text{and} \quad U^\top \Sigma_n U = B^\top V^\top V D_n V^\top V B = B^\top D_n B \,.$$

Therefore,

$$\mathrm{Trace}(U^\top \Sigma_n U) = \mathrm{Trace}(B^\top D_n B) = \sum_{i=1}^d \lambda_i \sum_{j=1}^p b_{i,j}^2 \,. \tag{1.3}$$

On the other hand,

$$B^\top B = U^\top V V^\top U = U^\top U = \mathrm{I}_p \,,$$

so that the columns of B are orthonormal and

$$\sum_{i=1}^d \sum_{j=1}^p b_{i,j}^2 = p \,.$$

By (1.3),

$$\mathrm{Trace}(U^\top \Sigma_n U) = \sum_{i=1}^d \alpha_i \lambda_i \,,$$

with, for all $1 \leqslant i \leqslant d$, $\alpha_i \in [0,1]$ and $\sum_{i=1}^d \alpha_i = p$. As $\lambda_1 \geqslant \lambda_2 \geqslant \ldots, \lambda_d$,

$$\mathrm{Trace}(U^\top \Sigma_n U) \leqslant \sum_{i=1}^{p} \lambda_i \,.$$

As the columns of $U_\star$ are$\{\vartheta_1,\ldots,\vartheta_p\}$, for all $1 \leqslant i \leqslant d$ and $1 \leqslant j \leqslant p$, $b_{i,j} = \langle \vartheta_i ; \vartheta_j \rangle = \delta_{i,j}$. Therefore, for all $1 \leqslant i \leqslant d$, $\sum_{j=1}^{p} b_{i,j}^2 = 1$ and by (1.3),

$$\mathrm{Trace}(U_\star^\top \Sigma_n U_\star) = \sum_{i=1}^{p} \lambda_i \,,$$

which completes the proof. ■

1.2 Principal Component Analysis: projection onto a subspace

For any dimension $1 \leqslant p \leqslant d$, let $\mathscr{F}_d^p$ be the set of all vector suspaces of $\mathbb{R}^d$ with dimension p. In this section, it is proved that Principal Component Analysis computes a linear span V_d such as

$$V_p \in \underset{V \in \mathscr{F}_d^p}{\mathrm{argmin}} \sum_{i=1}^{n} \| X_i - \pi_V(X_i) \|_2^2 \,, \tag{1.4}$$

where π_V is the orthogonal projection onto the linear span V. Assume first that $p = 1$ and write $V_1 = \mathrm{span}\{v_1\}$ for $v_1 \in \mathbb{R}^d$ such that $\|v_1\|_2 = 1$. Then,

$$\begin{aligned}
\sum_{i=1}^{n} \| X_i - \pi_{V_1}(X_i) \|_2^2 &= \sum_{i=1}^{n} \| X_i - \langle X_i ; v_1 \rangle v_1 \|_2^2 \,, \\
&= \sum_{i=1}^{n} \left(\|X_i\|_2^2 - 2\langle X_i ; \langle X_i ; v_1 \rangle v_1 \rangle + \| \langle X_i ; v_1 \rangle v_1 \|_2^2 \right) \,, \\
&= \sum_{i=1}^{n} \left(\|X_i\|_2^2 - \langle X_i ; v_1 \rangle^2 \right) .
\end{aligned}$$

Consequently, V_1 is a solution to (1.4) if and only if v_1 is solution to:

$$v_1 \in \underset{v \in \mathbb{R}^d \,;\, \|v\|_2 = 1}{\mathrm{argmax}} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \,.$$

For all $2 \leqslant p \leqslant d$, following the same steps, it can be proved that a solution to (1.4) is given by $V_p = \mathrm{span}\{v_1,\ldots,v_p\}$ where

$$v_1 \in \underset{v \in \mathbb{R}^d \,;\, \|v\|_2 = 1}{\mathrm{argmax}} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \quad \text{and for all } 2 \leqslant k \leqslant p \,, \;\; v_k \in \underset{\substack{v \in \mathbb{R}^d \,;\, \|v\|_2 = 1 \,; \\ v \perp v_1, \ldots, v \perp v_{k-1}}}{\mathrm{argmax}} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \,. \tag{1.5}$$

It remains to prove that the vectors $\{v_1,\ldots,v_k\}$ defined by (1.5) can be chosen as the orthonormal eigenvectors associated with the k largest eigenvalues of the empirical covariance matrix Σ_n. Note that for all $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$,

$$\frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 = \frac{1}{n} \sum_{i=1}^{n} (v^\top X_i)(X_i^\top v) = v^\top \Sigma_n v \,.$$

As $(\vartheta_i)_{1 \leqslant i \leqslant d}$ are the orthonormal eigenvectors associated with the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_d \geqslant 0$ of Σ_n. Then,

$$\frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 = v^\top \left(\sum_{i=1}^{d} \lambda_i \vartheta_i \vartheta_i^\top \right) v = \sum_{i=1}^{d} \lambda_i \langle v, \vartheta_i \rangle^2 \leqslant \lambda_1 \sum_{i=1}^{d} \langle v, \vartheta_i \rangle^2$$

and, as $(\vartheta_i)_{1\leqslant i\leqslant d}$ is an orthonormal basis of $\mathbb{R}^d$, $\sum_{i=1}^d \langle v,\vartheta_i\rangle^2 = \|v\|_2^2 = 1$. Therefore,

$$\frac{1}{n}\sum_{i=1}^n \langle X_i, v\rangle^2 \leqslant \lambda_1 \,.$$

On the other hand, for all $2\leqslant i\leqslant d$, $\langle \vartheta_1,\vartheta_i\rangle = 0$ and $\langle \vartheta_1,\vartheta_1\rangle = 1$ so that $\sum_{i=1}^d \lambda_i\langle \vartheta_1,\vartheta_i\rangle^2 = \lambda_1$ which proves that ϑ_1 is solution to (1.5).

Assume now that $v\in\mathbb{R}^d$ is such that $\|v\|_2 = 1$ and for all $1\leqslant j\leqslant k-1$, $\langle v;\vartheta_j\rangle = 0$ and write

$$\frac{1}{n}\sum_{i=1}^n \langle X_i, v\rangle^2 = \sum_{i=1}^d \lambda_i\langle v,\vartheta_i\rangle^2 \leq \lambda_k\sum_{i=k}^d \langle v,\vartheta_i\rangle^2 \leq \lambda_k \,,$$

since, as $(\vartheta_i)_{1\leqslant i\leqslant d}$ is an orthonormal basis of $\mathbb{R}^d$, $\sum_{i=1}^d \langle v,\vartheta_i\rangle^2 = \sum_{j=k}^d \langle v,\vartheta_i\rangle^2 = \|v\|_2^2 = 1$. On the other hand, for all $1\leqslant i\leqslant d$, $i\neq k$, $\langle \vartheta_k,\vartheta_i\rangle = 0$ and $\langle \vartheta_k,\vartheta_k\rangle = 1$ so that $\sum_{i=1}^d \lambda_i\langle \vartheta_k,\vartheta_i\rangle^2 = \lambda_k$ which proves that ϑ_k is solution to (1.5).

Therefore, $V_p = \mathrm{span}\{\vartheta_1,\ldots\vartheta_p\}$ is a solution to (1.5) and, as $(\vartheta_i)_{1\leqslant i\leqslant p}$ is an orthonormal family, the projection matrix onto V_p is given by $U_\star U_\star^\top$ where $U_\star$ is a $\mathbb{R}^{d\times p}$ matrix with columns $\{\vartheta_1,\ldots\vartheta_p\}$.

1.3 Interpretation of the Principal Component Analysis

1.3.1 Principal components

The orthonormal eigenvectors associated with the eigenvalues of Σ_n allow to define the principal components. As $V_d = \mathrm{span}\{\vartheta_1,\ldots,\vartheta_d\}$, for all $1\leqslant i\leqslant n$,

$$\pi_{V_d}(X_i) = \sum_{k=1}^d \langle X_i,\vartheta_k\rangle\vartheta_k = \sum_{k=1}^d (X_i^\top\vartheta_k)\vartheta_k = \sum_{k=1}^d c_k(i)\vartheta_k \,,$$

where for all $1\leqslant k\leqslant d$, the k-th principal component is defined as $c_k = X\vartheta_k$. Therefore the k-th principal component is the vector whose components are the coordinate are the coordinates of each X_i, $1\leqslant i\leqslant n$, relative to the vector ϑ_k of the basis $\{\vartheta_1,\ldots,\vartheta_d\}$ of V_d. For all $1\leqslant i\neq j\leqslant d$,

$$\langle c_i,c_j\rangle = \vartheta_i^\top X^\top X\vartheta_j = \vartheta_i^\top(n\Sigma_n)\vartheta_j = n\lambda_j\vartheta_i^\top\vartheta_j = 0 \,, \tag{1.6}$$

as $\{\vartheta_1,\ldots,\vartheta_d\}$ is an orthonormal family. Let W_d be the vector subspace of $\mathbb{R}^n$ generated by $\{c_1,\ldots,c_d\}$. Since $(c_j)_{1\leqslant j\leqslant d}$ form a orthogonal basis of W_d, for all $1\leqslant j\leqslant d$,

$$\pi_{W_d}(X_{.,j}) = \sum_{\ell=1}^d \frac{\langle c_\ell, X_{.,j}\rangle}{\|c_\ell\|_2^2}c_\ell \,.$$

By (1.6), for all $1\leqslant \ell\leqslant d$, $\|c_\ell\|_2^2 = n\lambda_\ell$ and

$$\langle c_\ell, X_{.,j}\rangle = \langle X\vartheta_\ell, X_{.,j}\rangle = X_{.,j}^\top X\vartheta_\ell = (X^\top X\vartheta_\ell)_j = (n\Sigma_n\vartheta_\ell)_j = n\lambda_\ell\vartheta_\ell(j) \,.$$

This yields, for all $1\leqslant j\leqslant d$,

$$\pi_{W_d}(X_{.,j}) = \sum_{\ell=1}^d \vartheta_\ell(j)c_\ell \,.$$

1.3.2 Explained variance

The percentage of variance explained by the first p dimensions is:

$$\alpha_p = \frac{n^{-1}\sum_{i=1}^{n}\|\pi_{V_p}(X_i)\|_2^2}{n^{-1}\sum_{i=1}^{n}\|X_i\|_2^2} = \frac{n^{-1}\sum_{i=1}^{n}\|\pi_{V_p}(X_i)\|_2^2}{\mathrm{trace}(\Sigma_n)} = \frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{d}\lambda_i}\,.$$

The sum of the variance of all the variables is equal to the number of variables when the variables are scaled. Figure 1.3 displays a toy example in which $n = 200$ and $(X_i)_{1\leq i\leq n}$ are i.i.d. Gaussian random variables

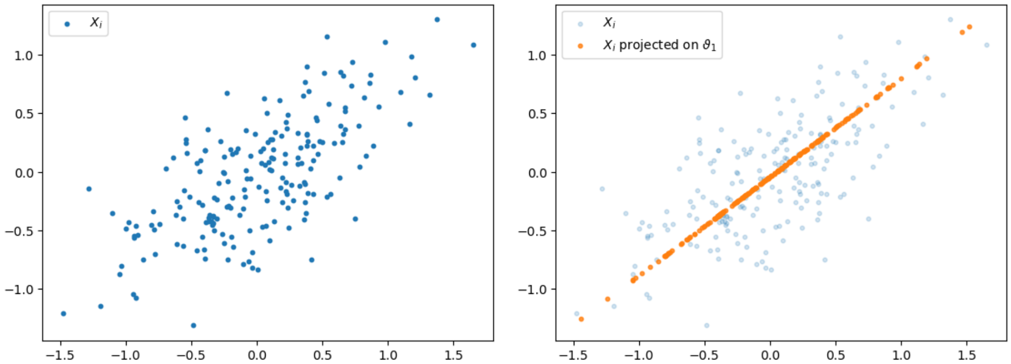

Fig. 1.3 PCA on the first component of $(X_i)_{1\leq i\leq 200}$ i.i.d. Gaussian random variables in $\mathbb{R}^2$. The first vector ϑ_1 provides 46% of the explained variance. (Right) The coordinate of X_i along ϑ is given by $c_1(i)$ (orange dots).

1.4 PCA visualization

Using Scikit-learn (`https://scikit-learn.org/`) we can display comprehensive visualizations of dimensionality reduction with PCA. For a simple illustration we use the Iris dataset[1] in which $n = 150$ and each X_i, $1 \leqslant i \leqslant n$, is the observation of an iris (Setosa, Versicolour, or Virginica) described by $d = 4$ features: Sepal Length, Sepal Width, Petal Length and Petal Width. Figure 1.4 displays the dataset and Figure 1.5 the result of a PCA on the 4 components. The two first components provide 97.8% of the explained variance. The illustrations are obtained following `https://plotly.com/python/pca-visualization/`.

[1] `https://scikit-learn.org/1.5/auto_examples/datasets/plot_iris_dataset.html`

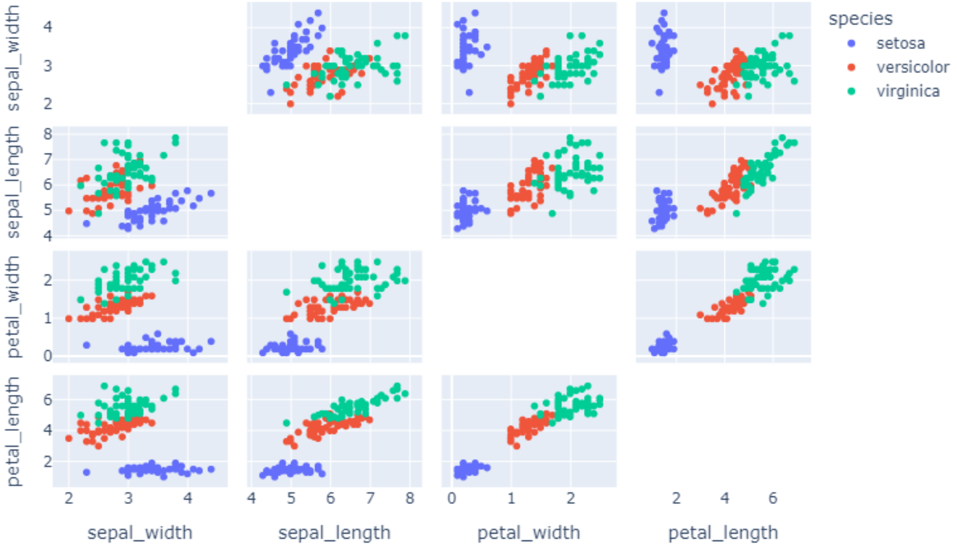

Fig. 1.4 Iris dataset.

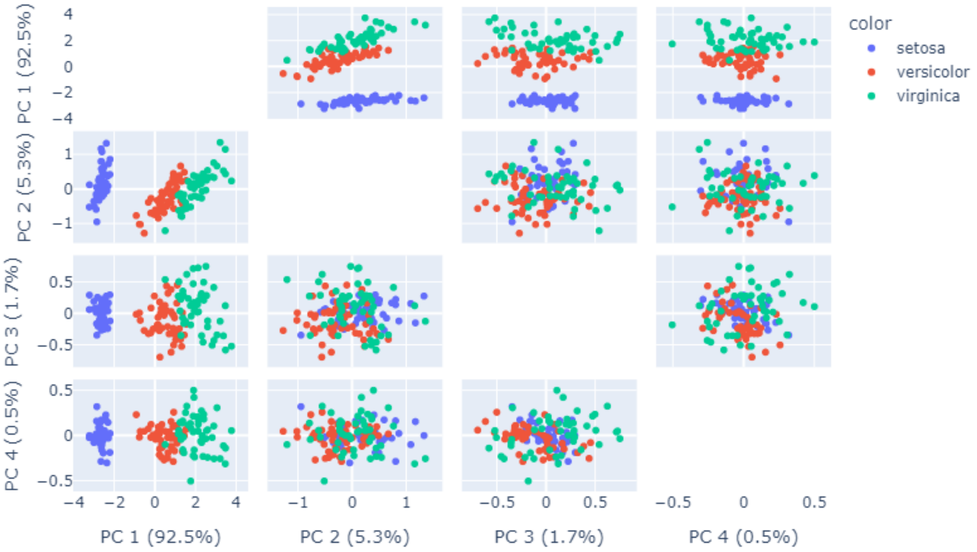

Fig. 1.5 PCA on the four components of the iris dataset. The graph in row i and column j displays the projection of each X_i on the subspace generated by $\{\vartheta_i, \vartheta_j\}$ i.e. the two dimensional projection related to principal components c_i and c_j. Illustration from `https://scikit-learn.org/1.5/auto_examples/datasets/plot_iris_dataset.html`.

Chapter 2

Supervised learning

Contents

Keywords 2.1 *Bayes classifier, empirical risk, oracle inequality, linear discriminant analysis, logistic regression.*

In a supervised learning framework, a set $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ of input data (also referred to as *features*) $X_i \in \mathscr{X}$ and output data $Y_i \in \mathscr{Y}$ (also referred to as *observations*), for $1 \leqslant i \leqslant n$, is available, where $\mathscr{X}$ is a general feature space and $\mathscr{Y}$ is a general observation space. In a supervised classification setting, the problem is to learn wether an individual from a given state space $\mathscr{X}$ belongs to some class, so that $\mathscr{Y} = \{1,\ldots,M\}$ for some $M \geqslant 1$. In a regression framework, the observation set $\mathscr{Y}$ is usually a subset of $\mathbb{R}^m$. The state space $\mathscr{X}$ is usually a subset of $\mathbb{R}^d$ and an element of $\mathscr{X}$ contains all the features the observation prediction is based on.

One of the main goals of supervised learning is to design an automatic procedure, based on the *training dataset* $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$, to predict the observation $y \in \mathscr{Y}$ associated with an input $x \in \mathscr{X}$ which is not in the training dataset.

The simulations presented in these notes can be found at `https://sylvainlc.github.io/`. Most elementary numerical solutions are based on scikit-learn, the website `https://scikit-learn.org/stable/supervised_learning.html` provides many helpful comments and advices.

2.1 Losses and risks

In these notes, we consider that $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ are independent and identically distributed (i.i.d.) with the same distribution as a couple of random variables (X,Y) defined on a measured space $(\Omega,\mathscr{F},\mathbb{P})$ and taking values in $\mathscr{X}\times\mathscr{Y}$. The joint distribution of (X,Y) is unknown. A loss function is used to evaluate the prediction of the observations: $\ell : \mathscr{Y}\times\mathscr{Y} \to \mathbb{R}$.

- In a classification setting where $\mathscr{Y} = \{1,\ldots,M\}$ for some $M \geqslant 1$, a common loss function is $\ell : (y,y') \mapsto \mathbb{1}_{y\neq y'}$. This 0-1 loss outputs 1 if the prediction $y' \in \mathscr{Y}$ is different from the true class $y \in \mathscr{Y}$.
- In a regression setting, common loss functions are $\ell : (y,y') \mapsto \|y-y'\|_2^2$ and $\ell : (y,y') \mapsto \|y-y'\|_1$.

Once the loss function is chosen, the expected risk allows to evaluate all predictors $f : \mathscr{X} \to \mathscr{Y}$. It is defined as the expected loss between the observation Y and the prediction $f(X)$:

$$\mathsf{R}(f) = \mathbb{E}\left[\ell(Y, f(X))\right] .$$

- In a classification setting using the 0-1 loss, the risk function is $\mathsf{R}(f) = \mathbb{E}\left[\mathbb{1}_{Y \neq h(X)}\right] = \mathbb{P}(Y \neq f(X))$. It is also known as the *misclassification loss*.
- In a regression setting using the square loss, the risk function is $\mathsf{R}(f) = \mathbb{E}[\|Y - f(X)\|_2^2]$.

This risk is typically unknown as the joint distribution of (X,Y) is unknown, i.e. the expectation cannot be computed explicitly. Therefore, a classical surrogate is given by the empirical risk:

$$\mathsf{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) .$$

In empirical risk minimization, we often define a parameterized family of predictors $\{f_\theta\}_{\theta\in\Theta}$ where $\Theta \in \mathbb{R}^q$ and for all $\theta \in \Theta$, $f_\theta : \mathscr{X} \to \mathscr{Y}$ and seek to minimize the empirical risk over this parameterized family:

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta\in\Theta} \left\{ \mathsf{R}_n(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)) \right\} .$$

Cross-validation (see also Section 4.1.3)

In practice, several parameterized families of predictors can be considered to minimize the empirical risk. Each parameterized family provides an empirical best predictor. Cross-validation (CV) provides appealing strategies for algorithm selection. Cross-validation proposes to split the data to estimate the risk of each algorithm: part of data is used to train each algorithm (or minimize the empirical risk over each parameterized family), and the remaining part is used to estimate the risk of the estimated predictor. In [Arlot and Celisse, 2010], the authors provide a survey of the most common cross-validation techniques and a few guidelines to choose the best technique depending on the statistical learning setting.

The most widespread technique is probably the k-fold cross-validation approach, see [Geisser, 1975]. In this case, the training dataset is first randomly partitioned into k subset $\mathscr{D}_i$, $1 \leqslant i \leqslant k$. Each subset $\mathscr{D}_i$, $1 \leqslant i \leqslant k$, is iteratively removed from the dataset before training the algorithm. Then, the empirical risk of the trained algorithm is computed over $\mathscr{D}_i$. The estimated risk of the algorithm is given by the empirical mean of the risks obtained when each $\mathscr{D}_i$, $1 \leqslant i \leqslant k$, is removed from the training dataset and used to evaluate the risk. Additional details and illustrations can be found for instance here: `https://scikit-learn.org/stable/modules/cross_validation.html`.

2.2 Bayes classifier

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Assume that (X,Y) is a couple of random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $\mathscr{X} \times \{-1,1\}$ where $\mathscr{X}$ is a given state space, which means that the focus is set on a two-class classification problem. One aim of supervised classification is to define a function $h : \mathscr{X} \to \{-1,1\}$, called *classifier*, such that $h(X)$ is the best prediction of Y in a given context. For instance, the risk of misclassification of h is

$$\mathsf{R}_{\mathrm{miss}}(h) = \mathbb{E}\left[\mathbb{1}_{Y\neq h(X)}\right] = \mathbb{P}\left(Y \neq h(X)\right) .$$

Note that $\mathbb{E}[Y|X]$ is a random variable measurable with respect to the σ-algebra $\sigma(X)$. Therefore, there exists a function $\eta : \mathscr{X} \to [-1,1]$ so that $\mathbb{E}[Y|X] = \eta(X)$ almost surely.

Lemma 2.1 *The classifier $h_\star$, defined for all $x \in \mathscr{X}$, by*

$$h_\star(x) = \begin{cases} 1 & \text{if } \eta(x) > 0 \,, \\ -1 & \text{otherwise} \,, \end{cases}$$

is such that

$$h_\star = \operatorname*{arg\,min}_{h:\mathscr{X}\to\{-1,1\}} \mathsf{R}_{\mathrm{miss}}(h) \,.$$

PROOF. For all $u, v \in \{-1,1\}$, $\mathbb{1}\{u \neq v\} = \mathbb{1}\{uv = -1\} = (1-uv)/2$. Since Y and $h(X)$ take values in $\{-1,1\}$, this implies

$$\mathsf{R}_{\mathrm{miss}}(h) = \mathbb{P}(Y \neq h(X)) = (1 - \mathbb{E}[Yh(X)])/2 \,. \tag{2.1}$$

Using sucessively the tower property, the equality $|u| = u \times \operatorname{sgn}(u)$, and the tower property again yields

$$\mathbb{E}[Yh(X)] = \mathbb{E}[\mathbb{E}[Y|X]h(X)] \leq \mathbb{E}[|\mathbb{E}[Y|X]||h(X)|] = \mathbb{E}[\mathbb{E}[Y|X]\operatorname{sgn}(\mathbb{E}[Y|X])] = \mathbb{E}[Yh_\star(X)] \,.$$

Plugging this into (2.1) yields $\mathsf{R}_{\mathrm{miss}}(h) \geq \mathsf{R}_{\mathrm{miss}}(h_\star)$, which concludes the proof. ■

Note that

$$\mathbb{E}[Y|X] = \mathbb{P}(Y=1|X) - \mathbb{P}(Y=-1|X) = 2\mathbb{P}(Y=1|X) - 1 \,,$$

which motivates this alternative definition of $h_\star$.

Definition 2.2. The classifier $h_\star$ is called the Bayes classifier. It may also be written as follows:

$$h_\star(X) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X) > 1/2 \text{ i.e. if } \mathbb{P}(Y=1|X) > \mathbb{P}(Y=-1|X) \,, \\ -1 & \text{otherwise} \,. \end{cases}$$

The Bayes classifier is the optimal choice to minimize the probability of misclassification $\mathsf{R}_{\mathrm{miss}}$. However, as the conditional distribution of Y given X is usually unknown, it cannot be computed explicitly. Supervised classification aims at designing an approximate classifier $\widehat{h}_n$ using independent observations $(X_i, Y_i)_{1\leqslant i\leqslant n}$ with the same distribution as (X,Y) so that the error $\mathsf{R}_{\mathrm{miss}}(\widehat{h}_n) - \mathsf{R}_{\mathrm{miss}}(h_\star)$ may be controlled.

2.3 Parametric and semiparametric classifiers

2.3.1 Mixture of Gaussian distributions

In this first example, we consider a *parametric model*, that is, we assume that the joint distribution of (X,Y) belongs to a family of distributions parametrized by a vector θ with real components. For $k \in \{-1,1\}$, write $\pi_k = \mathbb{P}(Y=k)$. Assume that $\mathscr{X} = \mathbb{R}^d$ and that, conditionally on the event $\{Y=k\}$, X has a Gaussian distribution with mean $\mu_k \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, whose density is denoted g_k. In this case, the parameter $\theta = (\pi_1, \mu_1, \mu_{-1}, \Sigma)$ belongs to the set $\Theta = [0,1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$. The parameter π_{-1} is not part of the components of θ since $\pi_{-1} = 1 - \pi_1$. In this case, the parameter $\theta = (\pi_1, \mu_1, \mu_{-1}, \Sigma)$ belongs to the set $\Theta = [0,1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$. The parameter π_{-1} is not part of the components of θ since $\pi_{-1} = 1 - \pi_1$. The explicit computation of $\mathbb{P}(Y=1|X)$ writes

$$\mathbb{P}(Y=1|X) = \frac{\pi_1 g_1(X)}{\pi_1 g_1(X) + \pi_{-1} g_{-1}(X)} = \frac{1}{1 + \frac{\pi_{-1} g_{-1}(X)}{\pi_1 g_1(X)}} = \sigma\left(\log(\pi_1/\pi_{-1}) + \log(g_1(X)/g_{-1}(X))\right) \,,$$

where $\sigma : x \mapsto (1 + \mathrm{e}^{-x})^{-1}$ is the sigmoid function. Then,

$$\mathbb{P}\left(Y=1|X\right) = \sigma\left(X^\top \omega + b\right) \,, \tag{2.2}$$

where

$$\omega = \Sigma^{-1}\left(\mu_1 - \mu_{-1}\right) , b = \log(\pi_1/\pi_{-1}) + \frac{1}{2}\left(\mu_1 + \mu_{-1}\right)^\top \Sigma^{-1}\left(\mu_{-1} - \mu_1\right) \,.$$

Since

$$\mathbb{E}\left[Y|X\right] = \mathbb{P}\left(Y=1|X\right) - \mathbb{P}\left(Y=-1|X\right) \,,$$

the Bayes classifier is such that for all $x \in \mathscr{X}$,

$$\begin{aligned} h_\star(x) = 1 &\Leftrightarrow \pi_1 \exp\left\{-\frac{1}{2}\left(x-\mu_1\right)^\top \Sigma^{-1}\left(x-\mu_1\right)\right\} > \pi_{-1}\exp\left\{-\frac{1}{2}\left(x-\mu_{-1}\right)^\top \Sigma^{-1}\left(x-\mu_{-1}\right)\right\} \,, \\ &\Leftrightarrow \log\left(\frac{\pi_1}{\pi_{-1}}\right) > -\frac{1}{2}\left(x-\mu_{-1}\right)^\top \Sigma^{-1}\left(x-\mu_{-1}\right) + \frac{1}{2}\left(x-\mu_1\right)^\top \Sigma^{-1}\left(x-\mu_1\right) \,, \\ &\Leftrightarrow \log\left(\frac{\pi_1}{\pi_{-1}}\right) > x^\top \Sigma^{-1}\left(\mu_{-1}-\mu_1\right) + \frac{1}{2}\left(\mu_1+\mu_{-1}\right)^\top \Sigma^{-1}\left(\mu_1-\mu_{-1}\right) \,. \end{aligned}$$

In this case, the Bayes classifier is given by

$$h_\star : x \mapsto \begin{cases} 1 & \text{if } \left\langle \Sigma^{-1}\left(\mu_1-\mu_{-1}\right) ; x - \frac{\mu_1+\mu_{-1}}{2}\right\rangle + \log\left(\frac{\pi_1}{\pi_{-1}}\right) > 0 \,, \\ -1 & \text{otherwise} \,, \end{cases}$$

Additional numerical considerations can be found for instance here `https://scikit-learn.org/`

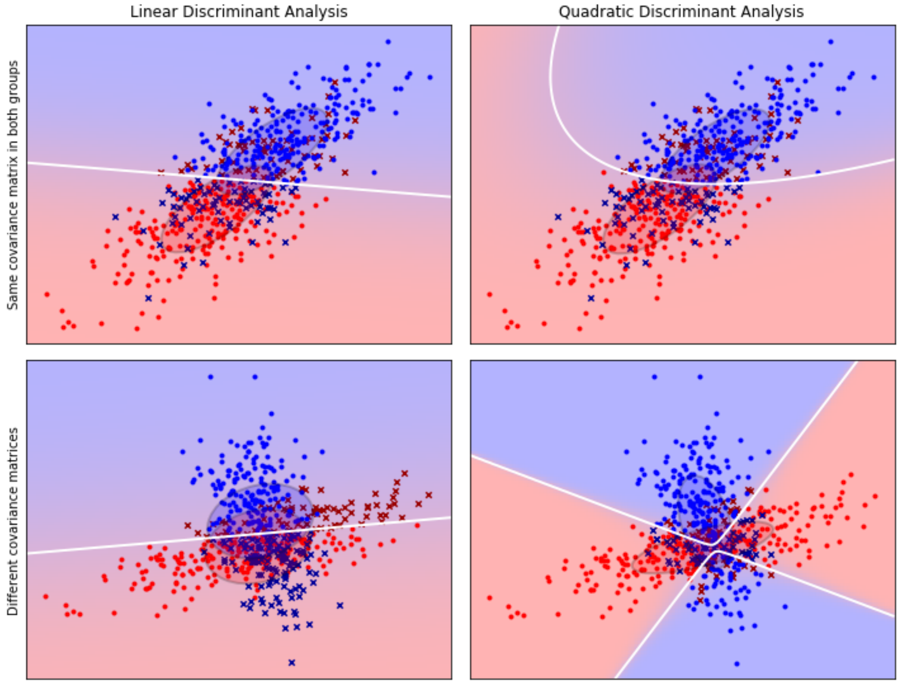

Fig. 2.1 (Top) Data are generated with the same covraiance matrix in each group. (Bottom) Data are generated with different covraiance matrices in the two groups. (Left) Classification boundary obtained with LDA, assuming the covariance matrix is the same in each group. (Right) Classification boundary obtained with QDA, assuming the covariance matrices are different in each group. Crosses are all false positives i.e. all data wrongly classified by the discriminant analysis. Simulations are inspired by `https://scikit-learn.org/stable/modules/lda_qda.html#mathematical-formulation-of-the-lda-and-qda-classifiers` and can be found here `https://sylvainlc.github.io/`.

`stable/modules/lda_qda.html#mathematical-formulation-of-the-lda-and-qda-classifiers`.

When Σ and μ_1 and μ_{-1} are unknown, this classifier cannot be computed explicitely. We will approximate it using the observations. Assume that $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ are independent observations with the same distribution as (X, Y). The loglikelihood of these observations is given by

$$\begin{aligned}
\log \mathbb{P}_\theta(X_{1:n}, Y_{1:n}) &= \sum_{i=1}^n \log \mathbb{P}_\theta(X_i, Y_i) \;, \\
&= -\frac{nd}{2}\log(2\pi) + \sum_{i=1}^n \sum_{k \in \{-1,1\}} \mathbb{1}_{Y_i = k} \left(\log \pi_k - \frac{\log \det \Sigma}{2} - \frac{1}{2}(X_i - \mu_k)^\top \Sigma^{-1} (X_i - \mu_k) \right) \;, \\
&= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^n \mathbb{1}_{Y_i=1}\right)\log \pi_1 + \left(\sum_{i=1}^n \mathbb{1}_{Y_i=-1}\right)\log(1-\pi_1) \\
&\quad - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=1}(X_i - \mu_1)^\top \Sigma^{-1}(X_i - \mu_1) - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=-1}(X_i - \mu_{-1})^\top \Sigma^{-1}(X_i - \mu_{-1}) \;.
\end{aligned}$$

By Lemma 7.15, the gradient of $\log \mathbb{P}_\theta(X_{1:n}, Y_{1:n})$ with respect to θ is therefore given by

$$\begin{aligned}
\frac{\partial \log \mathbb{P}_\theta(X_{1:n}, Y_{1:n})}{\partial \pi_1} &= \left(\sum_{i=1}^n \mathbb{1}_{Y_i=1}\right)\frac{1}{\pi_1} - \left(\sum_{i=1}^n \mathbb{1}_{Y_i=-1}\right)\frac{1}{1-\pi_1} \;, \\
\frac{\partial \log \mathbb{P}_\theta(X_{1:n}, Y_{1:n})}{\partial \mu_1} &= \sum_{i=1}^n \mathbb{1}_{Y_i=1}\left(2\Sigma^{-1}X_i - 2\Sigma^{-1}\mu_1\right) \;, \\
\frac{\partial \log \mathbb{P}_\theta(X_{1:n}, Y_{1:n})}{\partial \mu_{-1}} &= \sum_{i=1}^n \mathbb{1}_{Y_i=-1}\left(2\Sigma^{-1}X_i - 2\Sigma^{-1}\mu_{-1}\right) \;, \\
\frac{\partial \log \mathbb{P}_\theta(X_{1:n}, Y_{1:n})}{\partial \Sigma^{-1}} &= \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=1}(X_i - \mu_1)(X_i - \mu_1)^\top - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=-1}(X_i - \mu_{-1})(X_i - \mu_{-1})^\top \;.
\end{aligned}$$

The maximum likelihood estimator is defined as the only parameter $\widehat{\theta}^n$ such that all these equations are set to 0. For $k \in \{-1, 1\}$, it is given by

$$\begin{aligned}
\widehat{\pi}_k^n &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i=k} \;, \\
\widehat{\mu}_k^n &= \frac{1}{\sum_{i=1}^n \mathbb{1}_{Y_i=k}} \sum_{i=1}^n \mathbb{1}_{Y_i=k} X_i \;, \\
\widehat{\Sigma}^n &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \widehat{\mu}_{Y_i}^n\right)\left(X_i - \widehat{\mu}_{Y_i}^n\right)^\top \;.
\end{aligned}$$

Therefore, a natural surrogate for the bayes classifier is

$$\widehat{h}_n : x \mapsto \begin{cases} 1 \quad \text{if } \left\langle \widehat{\Omega}^n\left(\widehat{\mu}_1^n - \widehat{\mu}_{-1}^n\right); x - \frac{\widehat{\mu}_1^n + \widehat{\mu}_{-1}^n}{2} \right\rangle + \log\left(\frac{\widehat{\pi}_1^n}{\widehat{\pi}_{-1}^n}\right) > 0 \,, \\ -1 \text{ otherwise} \,, \end{cases}$$

where $\widehat{\Omega}^n = (\widehat{\Sigma}^n)^{-1}$. From the asymptotic properties of the Maximum Likelihood Estimator as n goes to infinity, this classifier converges almost surely to the Bayes classifier as the number of observations n tends to infinity.

2.3.2 Logistic regression

In some situations, it may be too restrictive to assume that the joint distribution of (X,Y) belongs to a parametric family. Instead, since the Bayes classifier defined in Lemma 2.1 only depends on the conditional distribution of Y given X, we only assume that this *conditional distribution* depends on a parameter. The model is said to be *semiparametric* instead of parametric. In the case where $\mathscr{X} = \mathbb{R}^d$, one of the most widely spread model for this conditional distribution is the *logistic regression* which is defined by

$$\mathbb{P}(Y=1|X) = \sigma(\alpha + \beta^T X)\,, \tag{2.3}$$

where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^d$ and σ is the sigmoid function. The parameter θ is thus $\theta = (\alpha,\beta) \in \mathbb{R}\times\mathbb{R}^d$. Note that for all $x \in \mathscr{X}$,

$$\begin{aligned} \sigma(\alpha+\beta^T x) &= \frac{1}{1+\mathrm{e}^{-\alpha-\langle\beta;x\rangle}}\,,\\ 1-\sigma(\alpha+\beta^T x) &= \frac{1}{1+\mathrm{e}^{\alpha+\langle\beta;x\rangle}}\,,\\ \log\left(\frac{\sigma(\alpha+\beta^T x)}{1-\sigma(\alpha+\beta^T x)}\right) &= \alpha+\langle\beta;x\rangle\,. \end{aligned}$$

The Bayes classifier is then given by

$$h_\star : x \mapsto \begin{cases} 1 & \text{if } \alpha+\langle\beta;x\rangle > 0\,,\\ -1 & \text{otherwise}\,. \end{cases}$$

When α and β are unknown, this classifier cannot be computed explicitely and is approximated using the observations. Assume that $(X_i,Y_i)_{1\leqslant i\leqslant n}$ are independent observations with the same distribution as (X,Y). The conditional likelihood of the observations $Y_{1:n}$ given $X_{1:n}$ is:

$$\begin{aligned} \mathbb{P}_\theta\left(Y_{1:n}|X_{1:n}\right) &= \prod_{i=1}^n \mathbb{P}_\theta\left(Y_i|X_i\right)\,,\\ &= \prod_{i=1}^n \left(\sigma_{\alpha,\beta}\right)^{(1+Y_i)/2}(X_i)\left(1-\sigma_{\alpha,\beta}(X_i)\right)^{(1-Y_i)/2}\,,\\ &= \prod_{i=1}^n \left(\frac{\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}{1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}\right)^{(1+Y_i)/2}\left(\frac{1}{1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}\right)^{(1-Y_i)/2}\,. \end{aligned}$$

The associated conditional loglikelihood is therefore

$$\begin{aligned} \log\mathbb{P}_\theta\left(Y_{1:n}|X_{1:n}\right) &= \sum_{i=1}^n \left\{\frac{1+Y_i}{2}\log\left(\frac{\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}{1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}\right) + \frac{1-Y_i}{2}\log\left(\frac{1}{1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}}\right)\right\}\,,\\ &= \sum_{i=1}^n \left\{\frac{1+Y_i}{2}\left(\alpha+\langle\beta;X_i\rangle\right) - \log\left(1+\mathrm{e}^{\alpha+\langle\beta;X_i\rangle}\right)\right\}\,. \end{aligned}$$

This conditional loglikelihood function cannot be maximized explictly yet numerous numerical optimization methods are available to maximize $(\alpha,\beta)\mapsto\log\mathbb{P}_\theta\left(Y_{1:n}|X_{1:n}\right)$. If $(\widehat{\alpha}_n,\widehat{\beta}_n)$ is an approximate solution to the optimization problem:

$$(\widehat{\alpha}_n,\widehat{\beta}_n) \in \underset{\theta=(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^d}{\arg\max}\ \log\mathbb{P}_\theta(Y_{1:n}|X_{1:n})\,, \tag{2.4}$$

then the associated logistic regression classifier is given by

$$\widehat{h}_n : x \mapsto \begin{cases} 1 & \text{if } \widehat{\alpha}_n + \langle \widehat{\beta}_n ; x \rangle > 0 \,, \\ -1 & \text{otherwise} \,, \end{cases}$$

Even though, the model is semiparametric (and not parametric), it can be shown that, specifically for logistic regression model, the approximated classifier almost surely tends to the Bayes classifier as the number of observations n tends to infinity.

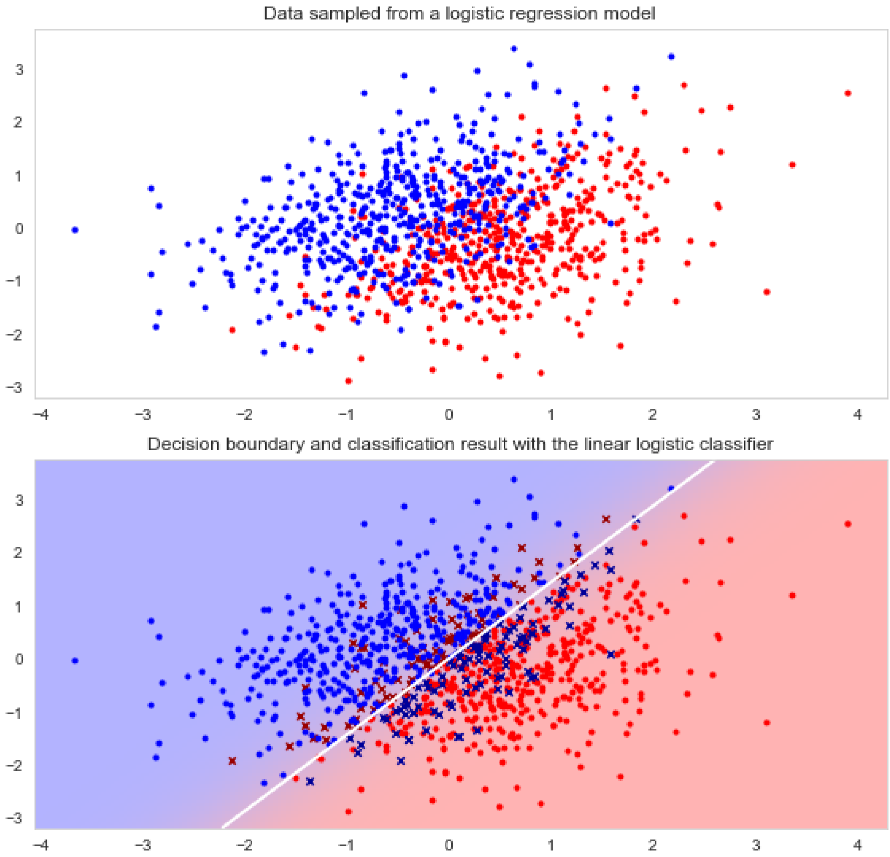

Fig. 2.2 (Top) Data are generated with a logistic regression model. The input data are Gaussian vectors in dimension $d = 2$ and the class of each data is chosen randomly according to (2.3) with a fixed weight w. (Bottom) Classification boundary obtained with the logistic classifier. Crosses are all false positives i.e. all data wrongly classified by the classifier (eventhough the weight w is known in this case). Simulations can be found here `https://sylvainlc.github.io/`.

Additional numerical considerations can be found for instance here `https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression`.

2.4 Nonparametric Bayes classifier

In the case of *nonparametric* models, it is not assumed anymore that the joint law of (X,Y) belongs to any parametric or semiparametric family of models. The assumption on the distribution of (X,Y) is relaxed but instead, we will make some restrictions on the set of classifiers on which the optimisation occurs.

More precisely, we consider that the optimization of classifiers holds on a specific set $\mathscr{H}$ of classifiers (often called the *dictionary*), which may possibly not contain the Bayes classifier. Moreover, since in most cases, the classification risk R_{miss} cannot be computed nor minimized, it is instead estimated by the empirical classification risk defined as

$$\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq h(X_i)} ,$$

where $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ are independent observations with the same distribution as (X,Y). The classification problem then boilds down to solving

$$\widehat{h}^n_{\mathscr{H}} \in \underset{h \in \mathscr{H}}{\arg\min}\, \widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) . \tag{2.5}$$

In this context several practical and theoretical challenges arise from the minimization of the empirical classification risk. The choice of $\mathscr{H}$ is pivotal in designing an efficient classification procedure. Note that choosing $\mathscr{H}$ as all possible classifiers is meaningless, in this case, $\widehat{h}^n_{\mathscr{H}}$ is such that $\widehat{h}^n_{\mathscr{H}}(X_i) = Y_i$ for all $1 \leqslant i \leqslant n$ and $\widehat{h}^n_{\mathscr{H}}(x)$ is any element of $\{-1,1\}$ for all $x \notin \{X_1, \ldots, X_n\}$. Although $\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h^n_{\mathscr{H}}) = 0$, is likely to be a poor approximation of $\mathsf{R}_{\mathrm{miss}}(h^n_{\mathscr{H}})$. To understand this, the excess misclassification risk may be decomposed as follows

$$\mathsf{R}_{\mathrm{miss}}\left(\widehat{h}^n_{\mathscr{H}}\right) - \mathsf{R}_{\mathrm{miss}}(h_\star) = \mathsf{R}_{\mathrm{miss}}\left(\widehat{h}^n_{\mathscr{H}}\right) - \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) + \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}(h_\star) \geqslant 0 .$$

The first term of the decomposition $\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h)$ is a **stochastic error** which is likely to grow when the size of $\mathscr{H}$ grows while $\min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}(h_\star)$ is **deterministic** and likely to decrease as the size of $\mathscr{H}$ grows.

Lemma 2.3 *For all set $\mathscr{H}$ of classifiers and all $n \geqslant 1$,*

$$\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) \leqslant 2 \sup_{h \in \mathscr{H}} \left| \widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}(h) \right| . \tag{2.6}$$

PROOF. By definition of $\widehat{h}^n_{\mathscr{H}}$, for any $h \in \mathscr{H}$,

$$\begin{aligned}\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) &= \mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) - \widehat{\mathsf{R}}^n_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) + \widehat{\mathsf{R}}^n_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) ,\\ &\leqslant \mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) - \widehat{\mathsf{R}}^n_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) + \widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) .\end{aligned}$$

For all $\varepsilon > 0$ there exists $h_\varepsilon \in \mathscr{H}$ such that $\mathsf{R}_{\mathrm{miss}}(h_\varepsilon) < \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) + \varepsilon$ so that

$$\begin{aligned}\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) - \min_{h \in \mathscr{H}} \mathsf{R}_{\mathrm{miss}}(h) &\leqslant \mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) - \widehat{\mathsf{R}}^n_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) + \widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_\varepsilon) - \mathsf{R}_{\mathrm{miss}}(h_\varepsilon) + \varepsilon ,\\ &\leqslant 2 \sup_{h \in \mathscr{H}} \left| \widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}(h) \right| + \varepsilon ,\end{aligned}$$

which concludes the proof. ■

Oracle inequality when $\mathscr{H}$ is finite

This section considers the simple case where the dictionary is finite, i.e., $\mathscr{H} = \{h_1, \ldots, h_M\}$ where $M \geqslant 1$ and for all $1 \leqslant j \leqslant M$, $h_j : \mathscr{X} \to \{-1,1\}$ is a given classifier.

Proposition 2.4 *Assume that $\mathscr{H} = \{h_1, \ldots, h_M\}$, then, for all $\delta > 0$,*

$$\mathbb{P}\left(\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) \leqslant \min_{1\leqslant j\leqslant M} \mathsf{R}_{\mathrm{miss}}(h_j) + \sqrt{\frac{2}{n}\log\left(\frac{2M}{\delta}\right)}\right) \geqslant 1-\delta\,.$$

PROOF. By Lemma 2.3, for all $u > 0$,

$$\mathbb{P}\left(\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) > \min_{1\leqslant j\leqslant M} \mathsf{R}_{\mathrm{miss}}(h_j) + u\right) \leqslant \mathbb{P}\left(\sup_{h\in\mathscr{H}}\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}(h)\right| > \frac{u}{2}\right) \leqslant \sum_{j=1}^{M}\mathbb{P}\left(\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right| > \frac{u}{2}\right)\,.$$

By Hoeffding's inequality, see Theorem 7.4,

$$\mathbb{P}\left(\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}}) > \min_{1\leqslant j\leqslant M} \mathsf{R}_{\mathrm{miss}}(h_j) + u\right) \leqslant 2M\mathrm{e}^{-nu^2/2}\,,$$

which concludes the proof by choosing

$$u = \sqrt{\frac{2}{n}\log\left(\frac{2M}{\delta}\right)}\,.$$

■

Proposition 2.5 *Assume that $\mathscr{H} = \{h_1,\ldots,h_M\}$, then,*

$$\mathbb{E}\left[\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}})\right] \leqslant \min_{1\leqslant j\leqslant M} \mathsf{R}_{\mathrm{miss}}(h_j) + \sqrt{\frac{2\log(2M)}{n}}\,.$$

PROOF. By Lemma 2.3,

$$\mathbb{E}\left[\mathsf{R}_{\mathrm{miss}}(\widehat{h}^n_{\mathscr{H}})\right] - \min_{1\leqslant j\leqslant M} \mathsf{R}_{\mathrm{miss}}(h_j) \leqslant 2\mathbb{E}\left[\sup_{h\in\mathscr{H}}\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h) - \mathsf{R}_{\mathrm{miss}}(h)\right|\right] = \frac{2}{n}\mathbb{E}\left[\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right]\,.$$

Note that

$$n\left\{\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right\} = \sum_{i=1}^{n}\left\{\mathbb{1}_{Y_i\neq h_j(X_i)} - \mathsf{R}_{\mathrm{miss}}(h_j)\right\}\,,$$

where the random variables $(\mathbb{1}_{Y_i\neq h_j(X_i)})_{1\leqslant i\leqslant n}$ are independent Bernoulli random variables with mean $\mathsf{R}_{\mathrm{miss}}(h_j)$. By Lemma 7.5, for all $t > 0$,

$$\mathbb{E}\left[\exp\left\{t\sum_{i=1}^{n}\left\{\mathbb{1}_{Y_i\neq h_j(X_i)} - \mathsf{R}_{\mathrm{miss}}(h_j)\right\}\right\}\right] = \prod_{i=1}^{n}\mathbb{E}\left[\exp\left\{t\left(\mathbb{1}_{Y_i\neq h_j(X_i)} - \mathsf{R}_{\mathrm{miss}}(h_j)\right)\right\}\right] \leqslant \mathrm{e}^{nt^2/8}$$

and similarly

$$\mathbb{E}\left[\exp\left\{-t\sum_{i=1}^{n}\left\{\mathbb{1}_{Y_i\neq h_j(X_i)} - \mathsf{R}_{\mathrm{miss}}(h_j)\right\}\right\}\right] \leqslant \mathrm{e}^{nt^2/8}\,.$$

Then, for all $t > 0$, by Jensen's inequality,

$$\begin{aligned}\exp\left\{t\mathbb{E}\left[\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right]\right\} &\leqslant \mathbb{E}\left[\exp\left\{t\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right\}\right]\,,\\ &\leqslant 2M\mathrm{e}^{nt^2/8}\,,\end{aligned}$$

which yields

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right] \leqslant \frac{\log(2M)}{t} + \frac{nt}{8}\,.$$

Choosing $t = \sqrt{8\log(2M)/n}$,

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant M}\left\{n\left|\widehat{\mathsf{R}}^n_{\mathrm{miss}}(h_j) - \mathsf{R}_{\mathrm{miss}}(h_j)\right|\right\}\right] \leqslant \sqrt{n\log(2M)/2}\,,$$

which concludes the proof. ■

Chapter 3

Multivariate regression

Contents

Keywords 3.1

3.1 Gaussian vectors

Definition 3.1. A random variable $X \in \mathbb{R}^d$ is a Gaussian vector if and only if, for all $a \in \mathbb{R}^d$, the random variable $\langle a\,;\,X\rangle$ is a Gaussian random variable.

For all random variable $X \in \mathbb{R}^d$, $X \sim \mathscr{N}(\mu, \Sigma)$ means that X is a Gaussian vector with mean $\mathbb{E}[X] = \mu \in \mathbb{R}^n$ and covariance matrix $\mathbb{V}[X] = \Sigma \in \mathbb{R}^{n\times n}$. The characteristic function of X is given (see exercises), for all $t \in \mathbb{R}^n$, by

$$\mathbb{E}[\mathrm{e}^{i\langle t\,;\,X\rangle}] = \mathrm{e}^{i\langle t\,;\,\mu\rangle - t^\top \Sigma t/2}\,.$$

Therefore, the law of a Gaussian vector is uniquely defined by its mean vector and its covariance matrix. If the covariance matrix Σ is nonsingular, then the law of X has a probability density with respect to the Lebesgue measure on $\mathbb{R}^n$ given by :

$$x \mapsto \det(2\pi\Sigma)^{-1/2} \exp\left\{-(x-\mu)^\top \Sigma^{-1}(x-\mu)/2\right\}\,,$$

where $\mu = \mathbb{E}[X]$.

Proposition 3.2 *Let $X \in \mathbb{R}^d$ be a Gaussian vector. Let $\{i_1, \ldots, i_p\}$ be a subset of $\{1, \ldots, d\}$, $p \geqslant 1$. If for all $1 \leqslant k \neq j \leqslant p$, $\mathrm{Cov}(X_{i_k}, X_{i_j}) = 0$, then $(X_{i_1}, \ldots, X_{i_p})$ are independent.*

PROOF. The random vector $(X_{i_1},\ldots,X_{i_p})^\top$ is a Gaussian vector with mean $(\mathbb{E}[X_{i_1}],\ldots,\mathbb{E}[X_{i_p}])^\top$ and diagonal covariance matrix $\mathrm{diag}(\mathbb{V}[X_{i_1}],\ldots,\mathbb{V}[X_{i_p}])$. Consider $(\xi_{i_1},\ldots,\xi_{i_p})$ i.i.d. random variables with distribution $\mathcal{N}(0,1)$ and define, for all $1 \leqslant j \leqslant p$,

$$Z_{i_j} = \mathbb{E}[X_{i_j}] + \sqrt{\mathbb{V}[X_{i_j}]}\xi_{i_j} \,.$$

Then, the random vector $(Z_{i_1},\ldots,Z_{i_p})^\top$ is a Gaussian vector with the same mean and the same covariance matrix as $(X_{i_1},\ldots,X_{i_p})^\top$. The two vectors have therefore the same characteristic function and the same law and $(X_{i_1},\ldots,X_{i_p})$ are independent as $(\xi_{i_1},\ldots,\xi_{i_p})$ are independent. ■

Theorem 3.3 (Cochran). *Let $X \sim \mathcal{N}(0,I_d)$ be a Gaussian vector in $\mathbb{R}^d$, F be a vector subspace of $\mathbb{R}^d$ and $F^\perp$ its orthogonal. Denote by $\pi_F(X)$ (resp. $\pi_{F^\perp}(X)$) the orthogonal projection of X on F (resp. on $F^\perp$). Then, $\pi_F(X)$ and $\pi_{F^\perp}(X)$ are independent, $\|\pi_F(X)\|_2^2 \sim \chi^2(p)$ and $\|\pi_{F^\perp}(X)\|_2^2 \sim \chi^2(d-p)$, where p is the dimension of F.*

PROOF. Let $(u_1,\ldots,u_d)$ be an orthonormal basis of $\mathbb{R}^d$ where $(u_1,\ldots,u_p)$ is an orthonormal basis of F and $(u_{p+1},\ldots,u_d)$ and orthonormal basis of $F^\perp$. Consider the matrix $U \in \mathbb{R}^{d\times d}$ such that for all $1 \leqslant i \leqslant d$, the i-th column of U is u_i and $U_{(p)}$ (reps. $U^\perp_{(d-p)}$) the matrix made of the first p (resp. last $d-p$) columns of U. Note that

$$\pi_F(X) = \sum_{i=1}^{p} \langle X \,;\, u_i\rangle u_i \,,$$

which can be written $\pi_F(X) = U_{(p)}U_{(p)}^\top X$. Similarly, $\pi_{F^\perp}(X) = U^\perp_{(d-p)}(U^\perp_{(d-p)})^\top X$ Therefore,

$$\begin{pmatrix} \pi_F(X) \\ \pi_{F^\perp}(X) \end{pmatrix} = \begin{pmatrix} U_{(p)}U_{(p)}^\top \\ U^\perp_{(d-p)}(U^\perp_{(d-p)})^\top \end{pmatrix} X$$

is a centered Gaussian vector with covariance matrix given by

$$\begin{pmatrix} U_{(p)}U_{(p)}^\top & 0 \\ 0 & U^\perp_{(d-p)}(U^\perp_{(d-p)})^\top \end{pmatrix} \,.$$

By Proposition 3.2, $\pi_F(X)$ and $\pi_{F^\perp}(X)$ are independent. On the other hand,

$$\|\pi_F(X)\|_2^2 = \sum_{i=1}^{p} \langle X \,;\, u_i\rangle^2 \quad \text{and} \quad \|\pi_{F^\perp}(X)\|_2^2 = \sum_{i=p+1}^{d} \langle X \,;\, u_i\rangle^2 \,.$$

The random vector $(\langle X \,;\, u_i\rangle)_{1\leqslant i\leqslant d}$ is given by U^TX: it is a Gaussian random vector with mean 0 and covariance matrix I_d. The random variables $(\langle X \,;\, u_i\rangle)_{1\leqslant i\leqslant d}$ are therefore i.i.d. with distribution $\mathcal{N}(0,1)$, which concludes the proof. ■

3.2 Full rank multivariate regression

3.2.1 Preliminaries

In a supervised learning framework, a set $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ of input data (also referred to as *features*) $X_i \in \mathscr{X}$ and output data $Y_i \in \mathscr{Y}$ (also referred to as *observations*), $1 \leqslant i \leqslant n$, is available, where $\mathscr{X}$ is a general feture space and $\mathscr{Y}$ is a general observation space. For instance, in a supervised classification framework, the problem is to learn wether an individual from a given state space $\mathscr{X}$ belongs to some class in $\mathscr{Y} = \{1,\ldots,M\}$. The state space $\mathscr{X}$ is usually a subset of $\mathbb{R}^d$ and an element of $\mathscr{X}$ contains all the features ued to predict the associated observation. In a regression framework, the observation set $\mathscr{Y}$ is usually a subset of $\mathbb{R}^m$.

Our aim is to introduce a regression function using the training dataset $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ employed to predict the observations associated with new features in a test dataset. In these lecture notes, we focus on empirical risk minimization. We consider a parameter set Θ and a family of regression functions $\{f_\theta\}_{\theta\in\Theta}$ where for all $\theta\in\Theta$, $f_\theta:\mathscr{X}\to\mathscr{Y}$. Considering first that $\mathscr{Y}=\mathbb{R}$ and $\mathscr{X}=\mathbb{R}^d$ we focus on solving the following optimization problem:

$$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\theta(X_i))^2 \;.$$

The components of the vector $\widehat{\theta}_n$ are often referred to as the *weights* or the *regression coefficients*. Each component $\widehat{\theta}_n(j)$, $1\leqslant j\leqslant d$, specifies the expected change in the output when the input $X(j)$ is changed by one unit.

3.2.2 Least squares estimator

In a linear regression setting, we assume that $\Theta=\mathbb{R}^d$ and that for all $\theta\in\Theta$, $f_\theta : x\mapsto\theta^\top x$. Let $Y\in\mathbb{R}^d$ be the random (column) vector such that for all $1\leqslant i\leqslant n$, the i-th component of Y is Y_i and $X\in\mathbb{R}^{n\times d}$ the matrix with line i equal to $X_i^\top$. The least squares estimate is defined as a solution to

$$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \|Y - X\theta\|_2^2 \;.$$

In a *well-specified setting*, we assumed that for all $1\leqslant i\leqslant n$, $Y_i = X_i^\top\theta_\star + \varepsilon_i$ for some unknown $\theta_\star\in\mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. random variables in $\mathbb{R}$. Let $\varepsilon\in\mathbb{R}^n$ be the random vector such that for all $1\leqslant i\leqslant n$, the i-th component of ε is ε_i. The model is then written

$$Y = X\theta_\star + \varepsilon \;.$$

Remark 3.4 *If the matrix X has full rank, i.e. its columns are linearly independent, then $X^\top X$ is positive definite since for all $u\in\mathbb{R}^d$, $u^\top X^\top X u = \|Xu\|_2^2$ and therefore $u^\top X^\top Xu \geqslant 0$ and $u^\top X^\top X u = 0$ if and only if $Xu=0$ i.e. if $u=0$.*

Proposition 3.5 *If the matrix X has full rank, then, $\widehat{\theta}_n = (X^\top X)^{-1}X^\top Y$. If for all $1\leqslant i\leqslant n$, $Y_i = X_i^\top\theta_\star+\varepsilon_i$ for some unknown $\theta_\star\in\mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered random variables in $\mathbb{R}$ with variance $\sigma_\star^2$, $\widehat{\theta}_n$ is an unbiased estimator of $\theta_\star$ and it satisfies $\mathbb{V}[\widehat{\theta}_n] = \sigma_\star^2(X^\top X)^{-1}$.*

PROOF. For all $\theta\in\mathbb{R}^d$,

$$\|Y - X\theta\|_2^2 = \|Y\|_2^2 + \theta^\top X^\top X\theta + 2Y^\top X\theta \;.$$

The function $\ell : \theta\mapsto \|Y\|_2^2 + \theta^\top X^\top X\theta + 2Y^\top X\theta$ is convex and for all $\theta\in\mathbb{R}^d$,

$$\nabla\ell(\theta) = 2X^\top X\theta + 2X^\top Y \;.$$

As the matrix X has full rank, $X^\top X$ is nonsingular and $\nabla\ell(\theta)=0$ has a unique solution given by

$$\widehat{\theta}_n = (X^\top X)^{-1}X^\top Y \;.$$

First, note that $\widehat{\theta}_n$ is unbiased as

$$\mathbb{E}[\widehat{\theta}_n] = (X^\top X)^{-1}X^\top\mathbb{E}[Y] = (X^\top X)^{-1}X^\top X\theta_\star = \theta_\star \;.$$

In addition,

$$\mathbb{V}[\widehat{\theta}_n] = (X^\top X)^{-1}X^\top\mathbb{V}[Y]X(X^\top X)^{-1} = \sigma_\star^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} = \sigma_\star^2(X^\top X)^{-1} \;.$$

■

Remark 3.6 *If we assume that for all $1 \leqslant i \leqslant n$, $Y_i = X_i^\top \theta_\star + \varepsilon_i$ for some unknown $\theta_\star \in \mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i \leqslant n}$ are i.i.d. random variables in $\mathbb{R}$ with distribution $\mathcal{N}(0,\sigma_*^2)$, $\widehat{\theta}_n$ is the maximum likelihood estimator of $\theta_\star$. The loglikelihood of the observations writes, for all $\theta \in \Theta$,*

$$\log p_\theta(Y_{1:n}) = \sum_{i=1}^n \log p_\theta(Y_i) = \sum_{i=1}^n \left\{ -\frac{1}{2}\log(2\pi\sigma_\star^2) - \frac{1}{2\sigma_\star^2}\left(Y_i - \theta^\top X_i\right)^2 \right\} .$$

Therefore, maximizing $\theta \mapsto \log p_\theta(Y_{1:n})$ amounts to minimizing $\theta \mapsto \sum_{i=1}^n (Y_i - \theta^\top X_i)^2 = \|Y - X\theta\|_2^2$.

Remark 3.7 *The matrix $X(X^\top X)^{-1}X^\top \in \mathbb{R}^{n\times n}$ is the matrix of the orthogonal projection onto* $\mathrm{Range}(X)$, *i.e., the vector space generated by the column vectors of X. First, let $v \in \mathrm{Range}(X)$, then, there exists $u \in \mathbb{R}^d$ such that $v = Xu$ and $X(X^\top X)^{-1}X^\top v = X(X^\top X)^{-1}X^\top Xu = Xu = v$. Therefore, for all $v \in \mathrm{Range}(X)$, $X(X^\top X)^{-1}X^\top v = v$. In addition, for all $v \in \mathrm{Range}(X)^\perp$, $X^\top v = 0$ so that $X(X^\top X)^{-1}X^\top v = 0$.*

Remark 3.8 *The projected value of Y is*

$$\widehat{Y} = X\widehat{\theta}_n = X(X^\top X)^{-1}X^\top Y .$$

In the special case where $d = 1$, $X \in \mathbb{R}^n$ and $\widehat{Y} = \{X^\top Y/(X^\top X)\}X = \{\langle X;Y\rangle/\langle X;X\rangle\}X$.

Proposition 3.9 *If $Y = X\theta_\star + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,\sigma_\star^2 \mathrm{I}_n)$, the random variable*

$$\widehat{\sigma}_n^2 = \frac{\|Y - X\widehat{\theta}_n\|_2^2}{n-d}$$

is an unbiased estimator of $\sigma_\star$. In addition, $(n-d)\widehat{\sigma}_n^2/\sigma_\star^2 \sim \chi^2(n-d)$, $\widehat{\theta}_n \sim \mathcal{N}(\theta_\star, \sigma_\star^2(X^\top X)^{-1})$ and $\widehat{\theta}_n$ and $\widehat{\sigma}_n^2$ are independent.

Proof. By definition of $\widehat{\theta}_n$,

$$\widehat{\sigma}_n^2 = \frac{\|Y - X\widehat{\theta}_n\|_2^2}{n-d} = \frac{\|Y - X(X^\top X)^{-1}X^\top Y\|_2^2}{n-d} = \frac{\|(I_n - X(X^\top X)^{-1}X^\top)Y\|_2^2}{n-d}$$

By Remark 3.7, the matrix of the orthogonal projection on $\mathrm{Range}(X)$ is $X(X^\top X)^{-1}X^\top$ and therefore $(I_n - X(X^\top X)^{-1}X^\top$ is the matrix of the orthogonal projection on $\mathrm{Range}(X)^\perp$. Then,

$$(I_n - X(X^\top X)^{-1}X^\top)Y = (I_n - X(X^\top X)^{-1}X^\top)(X\theta_\star + \varepsilon) = (I_n - X(X^\top X)^{-1}X^\top)\varepsilon .$$

By Theorem 3.3, $\|\sigma_\star^{-1}(I_n - X(X^\top X)^{-1}X^\top)\varepsilon\|_2^2$ has a χ^2 distribution with $n-d$ degrees of freedom which yields

$$\mathbb{E}[\|(I_n - X(X^\top X)^{-1}X^\top)Y\|_2^2] = \sigma_\star^2(n-d)$$

and $\mathbb{E}[\widehat{\sigma}_n^2] = \sigma_\star^2$. By Proposition 3.5, $\mathbb{E}[\widehat{\theta}_n] = \theta_\star$ and $\mathbb{V}[\widehat{\theta}_n] = \sigma_\star^2(X^\top X)^{-1}$ and $\widehat{\theta}_n$ is a Gaussian vector as an affine transformation of a Gaussian vector. Note that $(n-d)\widehat{\sigma}_n^2 = \|(I_n - X(X^\top X)^{-1}X^\top)\varepsilon\|_2^2$ and $\widehat{\theta}_n = (X^\top X)^{-1}X^\top X\theta_\star + (X^\top X)^{-1}X^\top\varepsilon$ and that $(X^\top X)^{-1}X^\top\varepsilon$ and $(I_n - X(X^\top X)^{-1}X^\top)\varepsilon$ are not correlated as

$$\mathbb{E}[(I_n - X(X^\top X)^{-1}X^\top)\varepsilon\varepsilon^\top X(X^\top X)^{-1}] = \sigma_\star^2\mathbb{E}[(I_n - X(X^\top X)^{-1}X^\top)X(X^\top X)^{-1}] = 0 .$$

The independence follows from Proposition 3.2. ■

3.2.3 Computational issues

Eventhough it is possible to compute the inverse of $X^\top X$ in a full rank setting, this matrix can be ill conditioned which may lead to numerical instability.

- In Scikit-learn, the fit function of `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html` uses a SVD-based solver. By Proposition 7.11, if X has rank $r \geqslant 1$, there exist $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ such that
$$X = \sum_{k=1}^{r} \sigma_k u_k v_k^\top \;,$$
where $\{u_1,\ldots,u_r\} \in (\mathbb{R}^n)^r$ and $\{v_1,\ldots,v_r\} \in (\mathbb{R}^d)^r$ are two orthonormal families. The vectors $\{\sigma_1,\ldots,\sigma_r\}$ are called singular values of A and $\{u_1,\ldots,u_r\}$ (resp. $\{v_1,\ldots,v_r\}$) are the left-singular (resp. right-singular) vectors of X. If U denotes the $\mathbb{R}^{n\times r}$ matrix with columns given by $\{u_1,\ldots,u_r\}$ and V denotes the $\mathbb{R}^{d\times r}$ matrix with columns given by $\{v_1,\ldots,v_r\}$, then the singular value decomposition of A may also be written as
$$X = UD_rV^\top \;,$$
where $D_r = \mathrm{diag}(\sigma_1,\ldots,\sigma_r)$. Therefore
$$X^\top X = VD_r^2V^\top \quad \text{and} \quad (X^\top X)^{-1} = VD_r^{-2}V^\top \;.$$
In this case, it is enough to compute V and D to compute $(X^\top X)^{-1}$.
- Using QR decomposition, we know that there exist an orthogonal matrix $Q \in \mathbb{R}^{n\times d}$, $Q^\top Q = I_d$, and an upper triangular matrix $R \in \mathbb{R}^{d\times d}$ such that $X = QR$. Then,
$$X^\top X\widehat{\theta}_n = X^\top Y \Leftrightarrow R\widehat{\theta}_n = Q^\top Y \;.$$
As the matrix R is upper triangular, the last equation can be solved using backsubstitution using for instance. The estimator $\widehat{\theta}_n$ can then de computed by i) computing the QR factorization of X, ii) computing $Q^\top Y$ and iii) solving $R\widehat{\theta}_n = Q^\top Y$.

3.3 Risk analysis of the full-rank multivariate regression

In our fixed-design setting, where the matrix X is deterministic, our aim is to minimize the fixed design risk:
$$\mathsf{R}(\theta) = \frac{1}{n}\mathbb{E}\left[\|Y - X\theta\|_2^2\right] \;.$$

If for all $1 \leqslant i \leqslant n$, $Y_i = X_i^\top \theta_\star + \varepsilon_i$ for some unknown $\theta_\star \in \mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered random variables in $\mathbb{R}$ with variance $\sigma_\star^2$, note that
$$\mathsf{R}(\theta_\star) = \frac{1}{n}\mathbb{E}\left[\|Y - X\theta_\star\|_2^2\right] = \frac{1}{n}\mathbb{E}\left[\|\varepsilon\|_2^2\right] = \sigma_\star^2 \;.$$

Therefore, for all $\theta \in \Theta$,
$$\begin{aligned}
\mathsf{R}(\theta) - \mathsf{R}(\theta_\star) &= \frac{1}{n}\mathbb{E}\left[\|X\theta_\star + \varepsilon - X\theta\|_2^2\right] - \sigma_\star^2 \;,\\
&= \frac{1}{n}\mathbb{E}\left[\|X(\theta_\star - \theta)\|_2^2 + \|\varepsilon\|_2^2 + 2(\theta_\star - \theta)^\top X^\top \varepsilon\right] - \sigma_\star^2 \;,\\
&= (\theta_\star - \theta)^\top \left(\frac{1}{n}X^\top X\right)(\theta_\star - \theta) \;,
\end{aligned}$$

since $\mathbb{E}[\varepsilon]=0$. On the other hand, a standard bias-variance decomposition yields

$$\begin{aligned}\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right] &= \mathbb{E}\left[\left(\theta_\star-\widehat{\theta}_n\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\theta_\star-\widehat{\theta}_n\right)\right]\,,\\ &= \mathbb{E}\left[\left(\theta_\star-\mathbb{E}\left[\widehat{\theta}_n\right]+\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\theta_\star-\mathbb{E}\left[\widehat{\theta}_n\right]+\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)\right]\,,\\ &= \left(\theta_\star-\mathbb{E}\left[\widehat{\theta}_n\right]\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\theta_\star-\mathbb{E}\left[\widehat{\theta}_n\right]\right)+\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)\right]\,.\end{aligned}$$

Proposition 3.10 *If for all $1\leqslant i\leqslant n$, $Y_i=X_i^\top\theta_\star+\varepsilon_i$ for some unknown $\theta_\star\in\mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered random variables in $\mathbb{R}$ with variance $\sigma_\star^2$,*

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right]=\sigma_\star^2\frac{d}{n}\,.$$

PROOF. By Proposition 3.5, $\mathbb{E}[\widehat{\theta}_n]=\theta_\star$ so that

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right]=\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)^\top\left(\frac{1}{n}X^\top X\right)\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)\right]\,.$$

In addition, by Proposition 3.5, $\mathbb{V}[\widehat{\theta}_n]=\sigma_\star^2(X^\top X)^{-1}$, hence

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right]=\frac{\sigma_\star^2}{n}\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)^\top\mathbb{V}[\widehat{\theta}_n]^{-1}\left(\mathbb{E}\left[\widehat{\theta}_n\right]-\widehat{\theta}_n\right)\right]\,.$$

By Lemma 7.7,

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n)-\mathsf{R}(\theta_\star)\right]=\frac{\sigma_\star^2}{n}\mathrm{Trace}(\mathbb{V}[\widehat{\theta}_n]^{-1}\mathbb{V}[\widehat{\theta}_n])=\sigma_\star^2\frac{d}{n}\,.$$

■

3.4 Confidence intervals and tests

Student's t-statistics

Proposition 3.11 *Assume that for all $1\leqslant i\leqslant n$, $Y_i=X_i^\top\theta_\star+\varepsilon_i$ for some unknown $\theta_\star\in\mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. with distribution $\mathscr{N}(0,\sigma_\star^2)$. For all $1\leqslant j\leqslant d$,*

$$\frac{\widehat{\theta}_{n,j}-\theta_{\star,j}}{\widehat{\sigma}_n\sqrt{(X^TX)^{-1}_{j,j}}}\sim\mathscr{S}(n-d)\,,$$

where $\mathscr{S}(n-d)$ is the Student's t-distribution with $n-p$ degrees of freedom, i.e. the law of $X/\sqrt{Y/(n-d)}$ where $X\sim\mathscr{N}(0,1)$ is independent of $Y\sim\chi^2(n-d)$.

PROOF. By definition, for all $1\leqslant j\leqslant d$,

$$\frac{\widehat{\theta}_{n,j}-\theta_{\star,j}}{\widehat{\sigma}_n\sqrt{(X^\top X)^{-1}_{j,j}}}=\frac{\sigma_\star^{-1}(\widehat{\theta}_{n,j}-\theta_{\star,j})}{\sigma_\star^{-1}\widehat{\sigma}_n\sqrt{(X^\top X)^{-1}_{j,j}}}=\frac{e_j^\top(\sigma_\star^{-1}(\widehat{\theta}_n-\theta_\star))}{\sigma_\star^{-1}\widehat{\sigma}_n\sqrt{(X^\top X)^{-1}_{j,j}}}\,.$$

Note that $\sigma_\star^{-1}(\widehat{\theta}_n-\theta_\star)\sim\mathscr{N}(0,(X^\top X)^{-1})$ so that $e_j^\top(\sigma_\star^{-1}(\widehat{\theta}_n-\theta_\star))\sim\mathscr{N}(0,e_j^\top(X^\top X)^{-1}e_j)$ and

$$\frac{e_j^\top(\sigma_\star^{-1}(\widehat{\theta}_n-\theta_\star))}{\sqrt{(X^\top X)^{-1}_{j,j}}}\sim\mathscr{N}(0,1)\,.$$

In addition,

$$\sigma_\star^{-1}\widehat{\sigma}_n=\sqrt{\sigma_\star^{-2}\widehat{\sigma}_n^2}=\sqrt{\|\sigma_\star^{-1}(I_n-X(X^\top X)^{-1}X^\top)\varepsilon\|_2^2/(n-d)}\,,$$

where $\sigma_\star^{-2}\widehat{\sigma}_n^2=\|\sigma_\star^{-1}(I_n-X(X^\top X)^{-1}X^\top)\varepsilon\|_2^2\sim\chi^2(n-d)$. The proof is concluded by noting that $\widehat{\theta}_n$ and $\widehat{\sigma}_n^2$ are independent. ■

By Proposition 3.11, for $\alpha\in(0,1)$, if $s^{n-d}_{1-\alpha/2}$ denotes the quantile of order $1-\alpha/2$ of the law $\mathscr{S}(n-d)$, then

$$\mathbb{P}\left(\left|\frac{\widehat{\theta}_{n,j}-\theta_{\star,j}}{\widehat{\sigma}_n\sqrt{(X^TX)^{-1}_{j,j}}}\right|\leqslant s^{n-d}_{1-\alpha/2}\right)=1-\alpha\,.$$

Therefore,

$$I^{n-p}_{n,j}(\theta_\star)=\left[\widehat{\theta}_{n,j}-\widehat{\sigma}_n s^{n-d}_{1-\alpha/2}\sqrt{(X^\top X)^{-1}_{j,j}}\;;\;\widehat{\theta}_{n,j}+\widehat{\sigma}_n s^{n-d}_{1-\alpha/2}\sqrt{(X^\top X)^{-1}_{j,j}}\right]$$

is a confidence interval for $\theta_{\star,j}$ with confidence level $1-\alpha$. The result of Proposition 3.11 may also be used to perform the test

$$\mathrm{H}_0:\ \theta_{\star,j}=0\quad\text{vs}\quad\mathrm{H}_1:\ \theta_{\star,j}\neq0\,.$$

Under H_0, the random variable $T_{n,j}$ defined by

$$T_{n,j}=\frac{\widehat{\theta}_{n,j}}{\widehat{\sigma}_n\sqrt{(X^\top X)^{-1}_{j,j}}}$$

does not depend on $\theta_\star$ neither on $\sigma_\star$ and is distributed as a Student $\mathscr{S}(n-d)$ random variable. A statistical test with statistical signifiance $1-\alpha$ to decide wether $\theta_\star\neq0$ is $T_{n,j}<s^{n-d}_{1-\alpha/2}$.

Fisher statistics

Proposition 3.12 *Assume that for all $1\leqslant i\leqslant n$, $Y_i=X_i^\top\theta_\star+\varepsilon_i$ for some unknown $\theta_\star\in\mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. with distribution $\mathscr{N}(0,\sigma_\star^2)$. Let L be a $\mathbb{R}^{q\times d}$ matrix with rank $q\leqslant d$. Then,*

$$\frac{(\widehat{\theta}_n-\theta_\star)^\top L^\top(L(X^\top X)^{-1}L^\top)^{-1}L(\widehat{\theta}_n-\theta_\star)}{q\widehat{\sigma}_n^2}\sim\mathscr{F}(q,n-d)\,,$$

where $\mathscr{F}(q,n-d)$ is the Fisher distribution with q and $n-d$ degrees of freedom, i.e. the law of $(X/q)/(Y/(n-d))$ where $X\sim\chi^2(q)$ is independent of $Y\sim\chi^2(n-d)$.

Proof. Note that $\mathrm{rank}(L(X^\top X)^{-1}L^\top)=\mathrm{rank}(LL^\top)=q$. The matrix $L(X^\top X)^{-1}L^\top$ is therefore positive definite. There exists a diagonal matrix $D\in\mathbb{R}^{q\times q}$ with positive diagonal terms and an orthogonal matrix $Q\in\mathbb{R}^{q\times q}$ such that $L(X^\top X)^{-1}L^\top=QDQ^{-1}$. The matrix $(L(X^\top X)^{-1}L^\top)^{-1/2}$ may be defined as $(L(X^\top X)^{-1}L^\top)^{-1/2}=QD^{-1/2}Q^{-1}$. It is then enough to note that $(L(X^\top X)^{-1}L^\top)^{-1/2}L(\widehat{\theta}_n-\theta_\star)/\sigma_\star\sim\mathscr{N}(0,I_q)$. Therefore,

$$\sigma_\star^{-2}\|(L(X^\top X)^{-1}L^\top)^{-1/2}L(\widehat{\theta}_n-\theta_\star)\|^2=(\widehat{\theta}_n-\theta_\star)^\top L^\top(L(X^\top X)^{-1}L^\top)^{-1}L(\widehat{\theta}_n-\theta_\star)/\sigma_\star^2\sim\chi^2(q)\,.$$

On the other hand, by Proposition 3.5,

$$(n-d)\sigma_\star^{-2}\widehat{\sigma}_n^2\sim\chi^2(n-d)\,.$$

The proof is concluded by noting that $\widehat{\theta}_n$ and $\widehat{\sigma}_n^2$ are independent. ■

By Proposition 3.12, for $\alpha\in(0,1)$, if $f^{q,n-d}_{1-\alpha}$ denotes the quantile of order $1-\alpha$ of the law $\mathscr{F}(q,n-p)$, then

$$\mathbb{P}\left(\theta_\star \in \left\{\theta \in \mathbb{R}^d \,;\, (\widehat{\theta}_n - \theta)^\top L^\top (L(X^\top X)^{-1}L^\top)^{-1} L(\widehat{\theta}_n - \theta) \leqslant q\widehat{\sigma}_n^2 f_{1-\alpha}^{q,n-d}\right\}\right) = 1-\alpha\,.$$

Therefore,

$$I_n^{q,n-d}(\theta_\star) = \left\{\theta \in \mathbb{R}^d \,;\, (\widehat{\theta}_n - \theta)^\top L^\top (L(X^\top X)^{-1}L^\top)^{-1} L(\widehat{\theta}_n - \theta) \leqslant q\widehat{\sigma}_n^2 f_{1-\alpha}^{q,n-d}\right\}$$

is a confidence region for $\theta_\star$ with confidence level $1-\alpha$. The result of Proposition 3.12 may also be used to perform the test

$$\mathrm{H}_0 : \, L\theta_\star = \bar{\theta} \quad \text{vs} \quad \mathrm{H}_1 : \, L\theta_\star \neq \bar{\theta}\,,$$

for a given $\bar{\theta} \in \mathbb{R}^d$.

Chapter 4

Penalized and sparse multivariate regression

Contents

4.1 Ridge regression

4.1.1 Ridge estimator

In the case where $X^\top X$ is singular (resp. has eigenvalues close to zero), the least squares estimate cannot be computed (resp. is not robust). A common approach to control the estimator variance is to solve the surrogate Ridge regression problem:

$$\widehat{\theta}_{n,\lambda}^{\text{ridge}} \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|Y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\} ,$$

where $\lambda > 0$.

Remark 4.1 *The matrix $n^{-1}X^\top X + \lambda \mathrm{I}_d$ is definite positive for all $\lambda > 0$ as for all $u \in \mathbb{R}^d$,*

$$u^\top (n^{-1}X^\top X + \lambda \mathrm{I}_d)u = n^{-1}\|Xu\|_2^2 + \lambda \|u\|_2^2 ,$$

which is positive for all $u \neq 0$. This remark allows to obtain the following result.

Proposition 4.2 *The unique solution to the Ridge regression problem is given by*

$$\widehat{\theta}_{n,\lambda}^{\text{ridge}} = \frac{1}{n} \left(\frac{1}{n} X^\top X + \lambda \mathrm{I}_d \right)^{-1} X^\top Y .$$

If for all $1 \leqslant i \leqslant n$, $Y_i = X_i^\top \theta_\star + \varepsilon_i$ *for some unknown* $\theta_\star \in \mathbb{R}^d$ *where the* $(\varepsilon_i)_{1\leqslant i \leqslant n}$ *are i.i.d. centered random variables in* $\mathbb{R}$ *with variance* $\sigma_\star^2$, *this estimator is biased and satisfies*

$$\mathbb{E}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] - \theta_* = -\lambda \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} \theta_\star \,,$$

$$\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{\sigma_\star^2}{n} \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-2} \frac{1}{n}X^\top X \,.$$

PROOF. The unique expression of $\widehat{\theta}^{\text{ridge}}_{n,\lambda}$ is obtained similarly as in the proof of Proposition 3.5. Then,

$$\mathbb{E}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{1}{n}\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} X^\top \mathbb{E}[Y] = \frac{1}{n}\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} X^\top X \theta_\star \,.$$

As the matrix $n^{-1}X^\top X$ is symmetric and real, $n^{-1}X^\top X$ is diagonalizable and $n^{-1}X^\top X$, $n^{-1}X^\top X + \lambda I_d$ and $(n^{-1}X^\top X + \lambda I_d)^{-1}$, are diagonalizable in the same orthonormal basis. Then, there exists nonnegative eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_d$ and orthonormal eigenvectors $u_1, \ldots u_d$ in $\mathbb{R}^d$ such that $n^{-1}X^\top X = \sum_{i=1}^d \lambda_i u_i u_i^\top$ and $(n^{-1}X^\top X + \lambda I_d)^{-1} = \sum_{i=1}^d (\lambda_i + \lambda)^{-1} u_i u_i^\top$. Therefore,

$$\mathbb{E}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] - \theta_\star = \sum_{i=1}^d \lambda_i (\lambda_i + \lambda)^{-1} u_i u_i^\top \theta_\star - \sum_{i=1}^d u_i u_i^\top \theta_\star = -\lambda \sum_{i=1}^d (\lambda_i + \lambda)^{-1} u_i u_i^\top \theta_\star = -\lambda \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} \theta_\star \,.$$

Similarly,

$$\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{1}{n^2}\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} X^\top \mathbb{V}[Y] X \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} = \frac{\sigma_\star^2}{n^2}\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} X^\top X \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} \,.$$

Therefore,

$$\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{\sigma_\star^2}{n} \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-2} \frac{1}{n}X^\top X \,.$$

■

Remark 4.3 *By Proposition 4.2,*

$$\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] = \frac{\sigma_\star^2}{n^2}\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} X^\top X \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} \,.$$

On the other hand, the variance of the full rank estimator is

$$\mathbb{V}[\widehat{\theta}_n] = \sigma_\star^2 \left(X^\top X\right)^{-1} \,.$$

Therefore, writing $M_\lambda = (X^\top X + \lambda n \mathrm{I}_d)^{-1} X^\top X$,

$$\begin{aligned}
\mathbb{V}[\widehat{\theta}^{\text{ridge}}_{n,\lambda}] - \mathbb{V}[\widehat{\theta}_n] &= \frac{\sigma_\star^2}{n^2}\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} X^\top X \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} - \sigma_\star^2 \left(X^\top X\right)^{-1} \\
&= \sigma_\star^2 \left(X^\top X + \lambda n \mathrm{I}_d\right)^{-1} X^\top X \left(X^\top X + \lambda n \mathrm{I}_d\right)^{-1} - \sigma_\star^2 \left(X^\top X\right)^{-1} \\
&= \sigma_\star^2 M_\lambda \left(X^\top X\right)^{-1} M_\lambda^\top - \sigma_\star^2 \left(X^\top X\right)^{-1} \\
&= \sigma_\star^2 M_\lambda \left((X^\top X)^{-1} - M_\lambda^{-1} (X^\top X)^{-1} (M_\lambda^\top)^{-1}\right) M_\lambda^\top \\
&= -\sigma_\star^2 M_\lambda \left(2n\lambda (X^\top X)^{-2} + (n\lambda)^2 (X^\top X)^{-3}\right) M_\lambda^\top \\
&= -\sigma_\star^2 (X^\top X + \lambda n \mathrm{I}_d)^{-1} \left(2n\lambda \mathrm{I}_d + (n\lambda)^2 (X^\top X)^{-1}\right) (X^\top X + \lambda n \mathrm{I}_d)^{-1} \,.
\end{aligned}$$

Therefore, the matrix $\mathbb{V}[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}] - \mathbb{V}[\widehat{\theta}_n]$ *is negative semi-definite i.e.* $\mathbb{V}[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}] \leqslant \mathbb{V}[\widehat{\theta}_n]$.

Remark 4.4 *If* X *is an orthonormal matrix,* $X^\top X = \mathrm{I}_d$ *and*

$$\mathbb{V}[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}] = \frac{\sigma_\star^2}{n^2}\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} X^\top X \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} = \frac{\sigma_\star^2}{(1+n\lambda)^2}\mathrm{I}_d\,.$$

4.1.2 Risk

Proposition 4.5 *Assume that for all* $1 \leqslant i \leqslant n$, $Y_i = X_i^\top \theta_\star + \varepsilon_i$ *for some unknown* $\theta_\star \in \mathbb{R}^d$ *where the* $(\varepsilon_i)_{1\leqslant i \leqslant n}$ *are i.i.d. centered random variables in* $\mathbb{R}$ *with variance* $\sigma_\star^2$*. For all* $\lambda > 0$,

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}) - \mathsf{R}(\theta_\star)\right] = \lambda^2 \theta_\star^\top \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-2} \frac{1}{n}X^\top X \theta_\star + \frac{\sigma_\star^2}{n}\mathrm{Trace}\left((n^{-1}X^\top X)^2 (n^{-1}X^\top X + \lambda \mathrm{I}_d)^{-2}\right)\,.$$

PROOF. Following the full-rank risk analysis given in Section 3.3, we have

$$\begin{aligned}&\mathbb{E}\left[\mathsf{R}(\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}) - \mathsf{R}(\theta_\star)\right]\\ &\quad= \left(\theta_\star - \mathbb{E}\left[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right]\right)^\top \left(\frac{1}{n}X^\top X\right)\left(\theta_\star - \mathbb{E}\left[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right]\right) + \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right] - \widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right)^\top \left(\frac{1}{n}X^\top X\right)\left(\mathbb{E}\left[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right] - \widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right)\right]\,.\end{aligned}$$

By Proposition 4.2, the bias term is given by

$$\begin{aligned}\left(\theta_\star - \mathbb{E}\left[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right]\right)^\top \left(\frac{1}{n}X^\top X\right)\left(\theta_\star - \mathbb{E}\left[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right]\right) &= \frac{\lambda^2}{n}\left(\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1}\theta_\star\right)^\top X^\top X \left(\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1}\theta_\star\right)\,,\\ &= \frac{\lambda^2}{n}\theta_\star^\top \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1} X^\top X \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-1}\theta_\star\,,\\ &= \lambda^2 \theta_\star^\top \left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-2}\frac{1}{n}X^\top X\theta_\star\,.\end{aligned}$$

By Lemma 7.7 and Proposition 4.2, the variance term is given by

$$\begin{aligned}\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right] - \widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right)^\top \left(\frac{1}{n}X^\top X\right)\left(\mathbb{E}\left[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right] - \widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}\right)\right] &= \mathrm{Trace}\left(\frac{1}{n}X^\top X \mathbb{V}[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}]\right)\,,\\ &= \frac{\sigma_\star^2}{n}\mathrm{Trace}\left(\frac{1}{n}X^\top X\left(\frac{1}{n}X^\top X + \lambda \mathrm{I}_d\right)^{-2}\frac{1}{n}X^\top X\right)\,,\\ &= \frac{\sigma_\star^2}{n}\mathrm{Trace}\left((n^{-1}X^\top X)^2(n^{-1}X^\top X + \lambda \mathrm{I}_d)^{-2}\right)\,,\end{aligned}$$

which concludes the proof. ■

Remark 4.6
- *The bias term increases with* λ*. It is 0 when* $\lambda = 0$ *and it converges to* $\theta_\star^\top X^\top X \theta_\star / n$ *when* $\lambda \to \infty$.
- *The variance term decreases with* λ*. It is* $\sigma_\star^2 d/n$ *when* $\lambda = 0$ *and it converges to* 0 *when* $\lambda \to \infty$.
- *The mean square error of the estimator is then given by*

$$\mathbb{E}\left[\left\|\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda} - \theta_*\right\|_2^2\right] = \mathrm{Trace}\left(\mathbb{V}[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}]\right) + \left\|\mathbb{E}[\widehat{\theta}^{\mathrm{ridge}}_{n,\lambda}] - \theta_*\right\|_2^2\,.$$

Let $(\vartheta_1, \ldots, \vartheta_d)$ *be an orthonormal basis of* $\mathbb{R}^d$ *of eigenvectors of* $X^\top X$ *associated with the eigenvalues* $(\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$*. Then,*

$$\mathbb{E}\left[\left\|\widehat{\theta}_{n,\lambda}^{\mathrm{ridge}}-\theta_*\right\|_2^2\right]=\sigma_\star^2\sum_{j=1}^d\frac{\lambda_j}{(\lambda_j+\lambda)^2}+\lambda^2\sum_{j=1}^d\frac{\langle\theta_*\,;\vartheta_j\,\rangle^2}{(\lambda_j+\lambda)^2}\,.$$

The mean square error is therefore a sum of two contributions, a bias related term which increases with λ and a variance related term which decreases with λ. In practice, the value of λ is chosen using cross-validation.

Using the risk analysis for the Ridge-based estimator, we can tune the regularization parameter λ to obtain a better bound than the $\sigma_\star^2 d/n$ bound of the case $\lambda = 0$.

Proposition 4.7 *Choosing $\lambda = \lambda_\star$ where*

$$\lambda_\star=\frac{\sigma_\star\mathrm{Trace}\left(X^\top X\right)^{1/2}}{\sqrt{n}\|\theta_\star\|_2}\,.$$

yields

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_{n,\lambda_\star}^{\mathrm{ridge}})-\mathsf{R}(\theta_\star)\right]\leqslant\frac{\sigma_\star\|\theta_\star\|_2\mathrm{Trace}\left(X^\top X\right)^{1/2}}{\sqrt{n}}\,.$$

PROOF. Let $(\vartheta_1,\ldots,\vartheta_d)$ be an orthonormal basis of $\mathbb{R}^d$ of eigenvectors of $n^{-1}X^\top X$ associated with the eigenvalues $(\lambda_1,\ldots,\lambda_d)\in\mathbb{R}^d$. Therefore,

$$\lambda^2\theta_\star^\top\left(\frac{1}{n}X^\top X+\lambda I_d\right)^{-2}\frac{1}{n}X^\top X\theta_\star=\lambda\sum_{i=1}^d\theta_\star^\top\frac{\lambda\lambda_i}{(\lambda_i+\lambda)^2}u_iu_i^\top\theta_\star\leqslant\frac{\lambda}{2}\|\theta_\star\|_2^2\,,$$

since for all $1\leqslant i\leqslant d$, $2\lambda\lambda_i\leqslant(\lambda+\lambda_i)^2$ implies $\lambda\lambda_i/(\lambda+\lambda_i)^2\leqslant 1/2$. On the other hand,

$$\frac{\sigma_\star^2}{n}\mathrm{Trace}\left((n^{-1}X^\top X)^2(n^{-1}X^\top X+\lambda I_d)^{-2}\right)=\frac{\sigma_\star^2}{n}\mathrm{Trace}\left(n^{-1}X^\top X\sum_{i=1}^d\frac{\lambda_i}{(\lambda+\lambda_i)^2}u_iu_i^\top\right)\leqslant\frac{\sigma_\star^2}{2n\lambda}\mathrm{Trace}\left(n^{-1}X^\top X\right)\,.$$

Therefore, by Proposition 4.5,

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_n^{\mathrm{ridge}})-\mathsf{R}(\theta_\star)\right]\leqslant\frac{\lambda}{2}\|\theta_\star\|_2^2+\frac{\sigma_\star^2}{2n\lambda}\mathrm{Trace}\left(n^{-1}X^\top X\right)\,.$$

The upper-bound is then minimized by choosing

$$\lambda_\star=\frac{\sigma_\star\mathrm{Trace}\left(X^\top X\right)^{1/2}}{\sqrt{n}\|\theta_\star\|_2}\,.$$

■

4.1.3 Choice of λ

Information-based metrics such as Akaike Information Criterion (AIC, [Akaike, 1974]) or Bayesian Information Criterion (BIC, [Schwarz, 1978]) are popular approaches to select λ in practical applications to balance between accuracy and complexity. On the other hand cross-validation techniques are widely used in machine learning and foster models with good prediction performance.

A standard approach is to use a K-fold cross-validation to split data between train and test subsets. The dataset $\mathscr{D}=\{(X_i,Y_i)\}_{1\leq i\leq n}$ is randomly divided into a partition made of K subsets approximately equally sized $\mathscr{D}_1,\ldots\mathscr{D}_K$. For all $1\leqslant k\leqslant K$, let n_k be the number of couples (X_i,Y_i) in $\mathscr{D}_k$. For all $1\leqslant k\leqslant K$, write $\mathbf{X_k}$ (resp. $\mathbf{X_{-k}}$) the matrix containing all input features X_i, $1\leqslant i\leqslant n$, such that $(X_i,Y_i)\in\mathscr{D}_k$ (resp. containing all input features X_i, $1\leqslant i\leqslant n$, such that $(X_i,Y_i)\notin\mathscr{D}_k$). Similarly, write $\mathbf{Y_k}$ (resp. $\mathbf{Y_{-k}}$) the vector containing all observations Y_i, $1\leqslant i\leqslant n$, such that $(X_i,Y_i)\in\mathscr{D}_k$ (resp. containing all observations Y_i, $1\leqslant i\leqslant n$, such that $(X_i,Y_i)\notin\mathscr{D}_k$). To assess the predictive performance of a model, we choose a loss function, and for each model, the model is trained K times using successively $\mathbf{X_{-k}}$ and $\mathbf{Y_{-k}}$, $1 \leqslant k \leqslant K$, to estimate the parameters of the model and $\mathbf{X_k}$ and $\mathbf{Y_k}$ to compute the loss function, i.e each model is associated with K predictive scores computed on a dataset which was not used to train the model. Usually, the model with best average score on the test sets is chosen (but considering the empirical variance of the test scores is also interesting). An example of the procedure is displayed in Algorithm 1.

There exist many ways to build the different groups of samples. In Leave One Out (LOO) strategies, training sets are created using all samples except one $\{(X_i, Y_i)\}_{1 \leqslant i \neq i_* \leqslant n}$, the test score being being evaluated only with (X_{i_*}, Y_{i_*}).

Cross-validation can be performed using Scikit-learn and its ridge regression with built-in cross-validation, see for instance `https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.RidgeCV.html`. Figure 4.1.3 illustrate a way data can be split to perform a cross-validation[1].

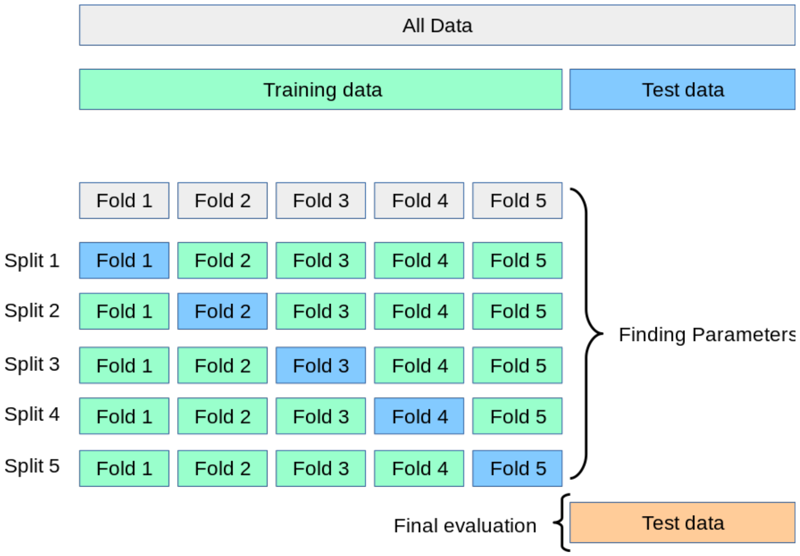

Fig. 4.1 An illustration on how to split data to select hyperparameters using K-fold cross-validation, see `https://scikit-learn.org/stable/modules/cross_validation.html`.

4.2 Lasso regression

The Least Absolute Shrinkage and Selection Operator (Lasso) regression, introduced in [Tibshirani, 1996], is a L_1 based regularized regression which aims at fostering sparsity. The objective is to solve the following minimization problem,

$$\widehat{\theta}^{\text{lasso}}_{\lambda,n} \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n} \|Y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\} , \tag{4.1}$$

where $\lambda > 0$ and

$$\|\theta\|_1 = \sum_{j=1}^{d} |\theta_j| .$$

The function $\theta \mapsto n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1$ is convex but not differentiable and the solution to this problem may not be unique. This approach shrinks some coefficients of θ and sets others to 0. The objective is to

[1] `https://scikit-learn.org/stable/modules/cross_validation.html`

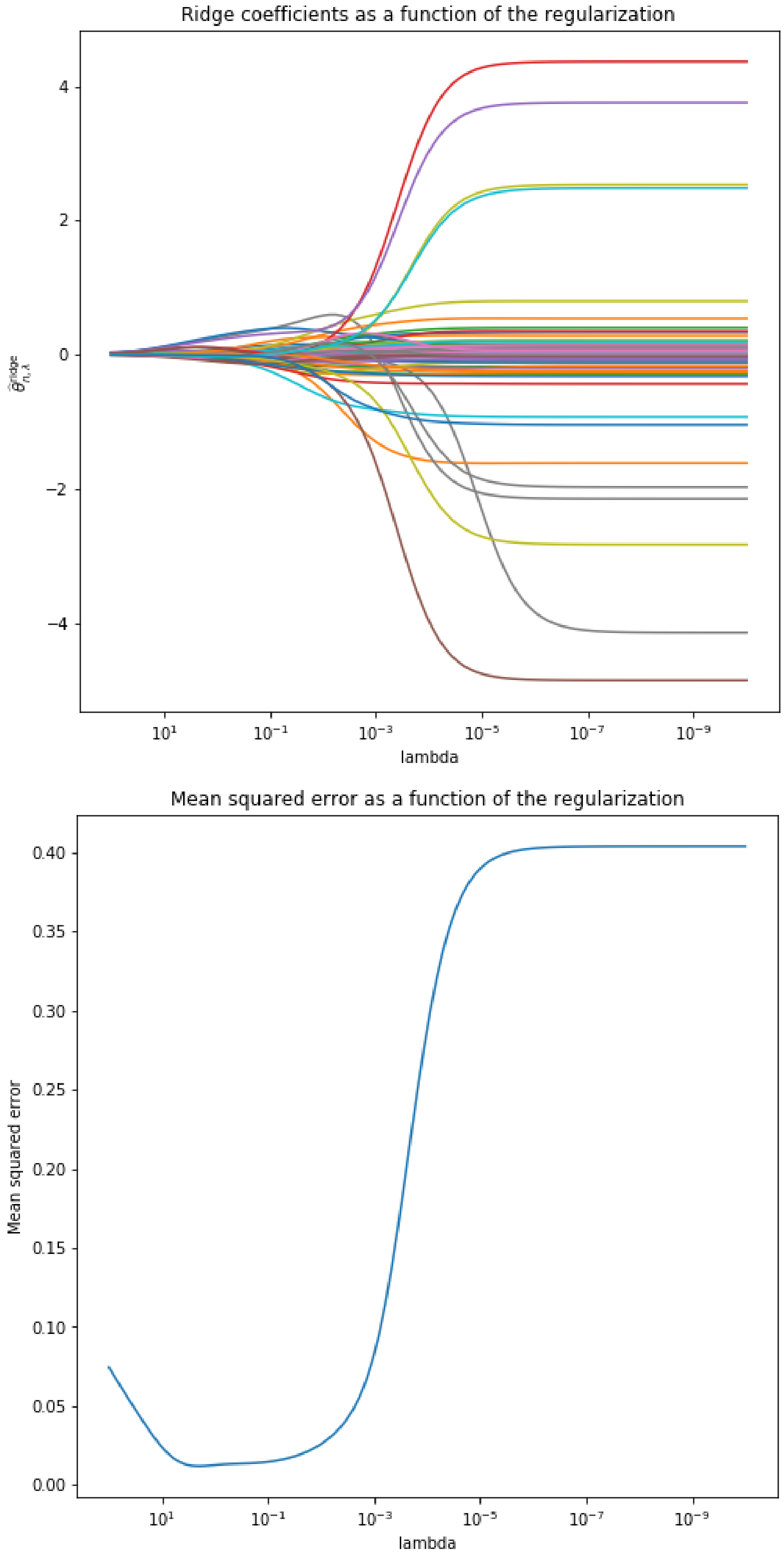

Fig. 4.2 Ridge regression is used to predict the Brazilian inflation based on many observed variables, see `https://github.com/gabrielrvsc/HDeconometrics/`. The model is trained using $n = 140$ data with for each $1 \leqslant i \leqslant n$, $X_i \in \mathbb{R}^{93}$, i.e. $d = 93$. The features are econometric data available each month. (Top) Estimated coefficient $\widehat{\theta}^{\text{ridge}}_{n,\lambda}$ as a function of λ. (Bottom) Mean squared error between the true observations and the predictions over the test set with 15 new data points.

Algorithm 1 K-fold cross-validation to select λ

```
Require: $K$, $\mathscr{D}_1, \dots \mathscr{D}_K$, candidate values $\{\lambda_1, \dots, \lambda_p\}$, $p \geqslant 1$.
  for $j \in \{1, \dots, p\}$ do
      for $k \in \{1, \dots, K\}$ do
          Compute
              $\widehat{\theta}^{\mathrm{ridge}}_{n-n_k,k,\lambda_j} = \frac{1}{n-n_k}\left(\frac{1}{n-n_k}\mathbf{X}_{-\mathbf{k}}^{\top}\mathbf{X}_{-\mathbf{k}} + \lambda_{\mathbf{j}}\mathbf{I}_{\mathbf{d}}\right)^{-1}\mathbf{X}_{-\mathbf{k}}^{\top}\mathbf{Y}_{-\mathbf{k}}$ .
          Compute
              $\mathscr{L}_k(\lambda_j) = \left\|\mathbf{Y}_{\mathbf{k}} - \mathbf{X}_{\mathbf{k}}\widehat{\theta}^{\mathrm{ridge}}_{\mathbf{n-n_k,k,\lambda_j}}\right\|_2^2$ .
      end for
      Set
          $\overline{\mathscr{L}}(\lambda_j) = \frac{1}{K}\sum_{k=1}^{K}\mathscr{L}_k(\lambda_j)$ .
  end for
  Set
      $\widehat{\lambda} = \mathrm{Argmin}_{\lambda \in \{\lambda_1,\dots\lambda_p\}}\overline{\mathscr{L}}(\lambda)$ .
```

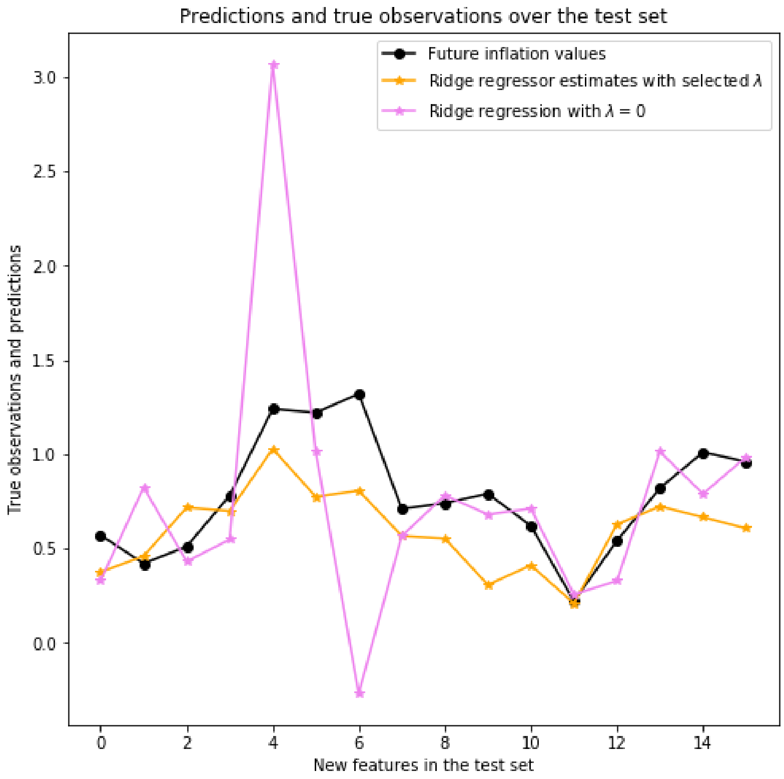

Fig. 4.3 Ridge regression is used to predict the Brazilian inflation based on many observed variables, see `https://github.com/gabrielrvsc/HDeconometrics/`. The model is trained using $n = 140$ data with for each $1 \leqslant i \leqslant n$, $X_i \in \mathbb{R}^{93}$, i.e. $d = 93$. The features are econometric data available each month. In this experiment, 15 new data points in a test set are used to evaluate the Ridge estimator. We present an ordinary least squares estimate (i.e with $\lambda = 0$) and the estimate obtained by selecting λ with a leave-one-out Cross-Validation. The MSE on the test set are 0.016 for the cross-validated λ and 0.398 for $\lambda = 0$.

balance the impact of feature selection and regularization. By Lemma 4.8, even when the Lasso solution is not unique, the solutions to (4.1) share some properties.

Lemma 4.8 ([Tibshirani, 2013]) *For all $\lambda > 0$, $n \geqslant 1$, all solutions $\widehat{\theta}_{\lambda,n}^{\mathrm{lasso}}$ to* (4.1) *have the same fitted value $X\widehat{\theta}_{\lambda,n}^{\mathrm{lasso}}$ and the same* L_1*-norm $\|\widehat{\theta}_{\lambda,n}^{\mathrm{lasso}}\|_1$.*

PROOF.

- Let $\widehat{\theta}_{1,\lambda,n}^{\mathrm{lasso}}$ and $\widehat{\theta}_{2,\lambda,n}^{\mathrm{lasso}}$ be two solutions of (4.1). Assume that $X\theta_{1,\lambda,n}^{\mathrm{lasso}} \neq X\theta_{2,\lambda,n}^{\mathrm{lasso}}$. Then, for all $\gamma \in (0,1)$, writing $\theta_{\gamma,\lambda,n}^{\mathrm{lasso}} = \gamma\theta_{1,\lambda,n}^{\mathrm{lasso}} + (1-\gamma)\theta_{2,\lambda,n}^{\mathrm{lasso}}$ and since $\|\cdot\|_2^2$ is stricly convex,
$$\|Y - X\theta_{\gamma,\lambda,n}^{\mathrm{lasso}}\|_2^2 = \|\gamma(Y - X\theta_{1,\lambda,n}^{\mathrm{lasso}}) + (1-\gamma)(Y - X\theta_{2,\lambda,n}^{\mathrm{lasso}})\|_2^2 < \gamma\|Y - X\theta_{1,\lambda,n}^{\mathrm{lasso}}\|_2 + (1-\gamma)\|Y - X\theta_{2,\lambda,n}^{\mathrm{lasso}}\|_2^2 \,.$$
Writing
$$\ell_{\min} = \frac{1}{n}\|Y - X\theta_{1,\lambda,n}^{\mathrm{lasso}}\|_2^2 + \lambda\|\theta_{1,\lambda,n}^{\mathrm{lasso}}\|_1 = \frac{1}{n}\|Y - X\theta_{2,\lambda,n}^{\mathrm{lasso}}\|_2^2 + \lambda\|\theta_{2,\lambda,n}^{\mathrm{lasso}}\|_1$$
yields
$$\frac{1}{n}\|Y - X\theta_{\gamma,\lambda,n}^{\mathrm{lasso}}\|_2^2 + \lambda\|\theta_{\gamma,\lambda,n}^{\mathrm{lasso}}\|_1 < \gamma\ell_{\min} + (1-\gamma)\ell_{\min} < \ell_{\min} \,,$$
which is not possible. Therefore, $X\theta_{1,\lambda,n}^{\mathrm{lasso}} = X\theta_{2,\lambda,n}^{\mathrm{lasso}}$.
- Let $\widehat{\theta}_{1,\lambda,n}^{\mathrm{lasso}}$ and $\widehat{\theta}_{2,\lambda,n}^{\mathrm{lasso}}$ be two solutions of (4.1). As $X\theta_{1,\lambda,n}^{\mathrm{lasso}} = X\theta_{2,\lambda,n}^{\mathrm{lasso}}$, $\|Y - X\theta_{1,\lambda,n}^{\mathrm{lasso}}\|_2^2 = \|Y - X\theta_{2,\lambda,n}^{\mathrm{lasso}}\|_2^2$. Since $\widehat{\theta}_{1,\lambda,n}^{\mathrm{lasso}}$ and $\widehat{\theta}_{2,\lambda,n}^{\mathrm{lasso}}$ are two solutions of (4.1), $\|\theta_{1,\lambda,n}^{\mathrm{lasso}}\|_1 = \|\theta_{2,\lambda,n}^{\mathrm{lasso}}\|_1$.

■

Remark 4.9 *By [Tibshirani, 2013, Lemma 4], when the entries of X have a joint distribution that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{n\times d}$, the Lasso solution is unique with probability 1.*

4.2.1 Computational issues

A coordinate descent can be applied to solve the Lasso optimization problem. In this case, solving (4.1) amounts to producing iterative estimators, where at each iteration, a coordinate is selected to be updated.

Othonormal design

Assume first that X is orthonormal, i.e. that $X^\top X = \mathrm{I}_d$. The loss of the Lasso optimization problem writes for all $\theta \in \mathbb{R}^d$,

$$\begin{aligned}\mathscr{L}_\lambda(\theta) &= \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1 \,,\\ &= \frac{1}{n}(\|Y\|_2^2 + \theta^\top\theta - 2\theta^\top X^\top Y) + \lambda\|\theta\|_1 \,,\\ &= \frac{1}{n}(\|Y\|_2^2 + \theta^\top\theta - 2\theta^\top X^\top Y) + \lambda\|\theta\|_1 \,,\\ &= \frac{1}{n}\|Y\|_2^2 + \frac{1}{n}\sum_{j=1}^d(\theta_j^2 - 2(X^\top Y)_j\theta_j) + \lambda\sum_{j=1}^d|\theta_j| \,.\end{aligned}$$

Therefore,
$$\widehat{\theta}_{\lambda,n}^{\mathrm{lasso}} \in \underset{\theta\in\mathbb{R}^d}{\mathrm{argmin}}\left\{\frac{1}{n}\sum_{j=1}^d(\theta_j^2 - 2(X^\top Y)_j\theta_j) + \lambda\sum_{j=1}^d|\theta_j|\right\} \,,$$
and $\widehat{\theta}_{\lambda,n}^{\mathrm{lasso}}$ can be computed coordinate per coordinate. Then, the objective function is optimized explicitly with respect to the selected coordinate. For all $\theta \in \mathbb{R}^d$, $j_0 \in \{1,\ldots,d\}$,

$$\partial_{\theta_{j_0}}\left\{\frac{1}{n}\sum_{j=1}^{d}(\theta_j^2-2(X^\top Y)_j\theta_j)\right\}=-\frac{2}{n}((X^\top Y)_{j_0}-\theta_{j_0})=-\frac{2}{n}(X_{\cdot,j_0}^\top Y-\theta_{j_0})\,.$$

Define

$$\upsilon_{j_0}=X_{\cdot,j_0}^\top Y\,.$$

Then,

$$\partial_{\theta_{j_0}}\left\{\frac{1}{n}\sum_{j=1}^{d}(\theta_j^2-2(X^\top Y)_j\theta_j)\right\}=-2(\upsilon_{j_0}-\theta_{j_0})\,.$$

Consequently, for all $\theta_j\neq 0$,

$$(\nabla_\theta(n^{-1}\|Y-X\theta\|_2^2+\lambda\|\theta\|_1))_j=\frac{2}{n}(\theta_j-\upsilon_j+\lambda n\cdot\mathrm{sign}(\theta_j)/2)\,.$$

For all $1\leqslant j\leqslant d$, $\theta_j\mapsto n^{-1}\|Y-X\theta\|_2^2+\lambda\|\theta\|_1$ is convex and grows to infinity when $|\theta_j|\to\infty$ and admits thus a minimum at some $\theta_j^\star\in\mathbb{R}$.

- If $\theta_j^\star\neq 0$, then

  $$\theta_j^\star=\upsilon_j\left(1-\frac{\lambda n\cdot\mathrm{sign}(\theta_j^\star)}{2\upsilon_j}\right)\,,$$

  which yields, as $\mathrm{sign}(\theta_j^\star)=\mathrm{sign}(\upsilon_j)$,

  $$\theta_j^\star=\upsilon_j\left(1-\frac{\lambda n}{2|\upsilon_j|}\right)$$

  and

  $$1-\frac{\lambda n}{2|\upsilon_j|}\geqslant 0\,.$$

- If $1-\lambda n/(2|\upsilon_j|)<0$, there is no solution to $(\nabla_\theta(n^{-1}\|Y-X\theta\|_2^2+\lambda\|\theta\|_1))_j=0$ for $\theta_j\neq 0$. Since $\theta_j\mapsto n^{-1}\|Y-X\theta\|_2^2+\lambda\|\theta\|_1$ admits a minimum, $\theta_j^\star=0$.

Therefore,

$$\theta_j^\star=\upsilon_j\left(1-\frac{\lambda n}{2|\upsilon_j|}\right)_+=\max\left(0;\upsilon_j\left(1-\frac{\lambda n}{2|\upsilon_j|}\right)\right)\,.$$

Normalized design

Assume first that for all $j\in\{1,\ldots,d\}$, $X_{\cdot,j}^\top X_{\cdot,j}=1$. For all $\theta\in\mathbb{R}^d$,

$$\nabla_\theta\|Y-X\theta\|_2^2=-2X^\top(Y-X\theta)\,.$$

Then, for all $1\leqslant j\leqslant d$, $(\nabla_\theta\|Y-X\theta\|_2^2)_j=-2X_{\cdot,j}^\top(Y-X\theta)$. Define, for all $1\leqslant j\leqslant d$,

$$\upsilon_j=X_{\cdot,j}^\top\left(Y-\sum_{\substack{i=1\\i\neq j}}^{d}\theta_iX_{\cdot,i}\right)\,.$$

Then,

$$(\nabla_\theta\|Y-X\theta\|_2^2)_j=-2(\upsilon_j-\theta_j)\,.$$

Consequently, for all $\theta_j\neq 0$,

$$(\nabla_\theta(n^{-1}\|Y-X\theta\|_2^2+\lambda\|\theta\|_1))_j=\frac{2}{n}(\theta_j-\upsilon_j+\lambda n\cdot\mathrm{sign}(\theta_j)/2)\,.$$

For all $1 \leqslant j \leqslant d$, $\theta_j \mapsto n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1$ is convex and grows to infinity when $|\theta_j| \to \infty$ and admits thus a minimum at some $\theta_j^\star \in \mathbb{R}$.

- If $\theta_j^\star \neq 0$, then
$$\theta_j^\star = \upsilon_j\left(1 - \frac{\lambda n \cdot \mathrm{sign}(\theta_j^\star)}{2\upsilon_j}\right) ,$$
which yields, as $\mathrm{sign}(\theta_j^\star) = \mathrm{sign}(\upsilon_j)$,
$$\theta_j^\star = \upsilon_j\left(1 - \frac{\lambda n}{2|\upsilon_j|}\right)$$
and
$$1 - \frac{\lambda n}{2|\upsilon_j|} \geqslant 0 .$$
- If $1 - \lambda n/(2|\upsilon_j|) < 0$, there is no solution to $(\nabla_\theta(n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1))_j = 0$ for $\theta_j \neq 0$. Since $\theta_j \mapsto n^{-1}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1$ admits a minimum, $\theta_j^\star = 0$.

Therefore,
$$\theta_j^\star = \upsilon_j\left(1 - \frac{\lambda n}{2|\upsilon_j|}\right)_+ = \max\left(0; \upsilon_j\left(1 - \frac{\lambda n}{2|\upsilon_j|}\right)\right) .$$

An algorithm to approximatively solve the Lasso regression problem proceeds as described in Algorithm 2.

Algorithm 2 Coordinate descent LASSO solver

```
Choose randomly an initial estimate $\widehat{\theta}_n^{(0)} \in \mathbb{R}^d$.
for $p = 1$ to $p = n_{\mathrm{iter}}$ do
    Choose randomly a coordinate $j \in \{1, \ldots, d\}$.
    Compute
        $$\upsilon_j = \mathbf{X}_j^\top\left(Y - \sum_{\substack{i=1\\ i\neq j}}^{d} \widehat{\theta}_{n,i}^{(p-1)}\mathbf{X}_i\right) .$$
    If $1 - \lambda n/(2|\upsilon_j|) > 0$, set
        $$\widehat{\theta}_{n,j}^{(p)} = \upsilon_j\left(1 - \frac{\lambda n}{2|\upsilon_j|}\right) .$$
    If $1 - \lambda n/(2|\upsilon_j|) < 0$, set $\widehat{\theta}_{n,j}^{(p)} = 0$.
    For all $1 \leqslant k \leqslant d, k \neq j$, set $\widehat{\theta}_{n,k}^{(p)} = \widehat{\theta}_{n,k}^{(p-1)}$.
end for
```

4.2.2 *Risk analysis of LASSO regression problem*

Proposition 4.10 *Assume that for all $1 \leqslant i \leqslant n$, $Y_i = X_i^\top\theta_\star + \varepsilon_i$ for some unknown $\theta_\star \in \mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. with distribution $\mathcal{N}(0, \sigma_\star^2)$. Then, choosing $n\lambda_\star^2/(16\sigma_\star^2\|\Sigma\|_\infty) = \log(dn)$ yields*
$$\frac{1}{n}\mathbb{E}\left[\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2\right] \leqslant 16\sigma_\star\sqrt{\frac{\log(dn)}{n}}\|\Sigma\|_\infty^{1/2}\|\theta_\star\|_1 + 12\sqrt{2}\frac{\sigma_\star^2}{n\sqrt{d}} .$$

PROOF. By definition of $\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n}$, for all $\theta \in \mathbb{R}^d$,
$$\frac{1}{n}\|Y - X\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n}\|_2^2 + \lambda\|\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n}\|_1 \leqslant \frac{1}{n}\|Y - X\theta_\star\|_2^2 + \lambda\|\theta_\star\|_1 .$$

As $Y = X\theta_\star + \varepsilon$, this yields

$$\frac{1}{n}\|\varepsilon - X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2 + \lambda\|\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n}\|_1 \leqslant \frac{1}{n}\|\varepsilon\|_2^2 + \lambda\|\theta_\star\|_1 \,.$$

Therefore,

$$\frac{1}{n}\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2 + \lambda\|\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n}\|_1 \leqslant \frac{2}{n}\varepsilon^\top X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star) + \lambda\|\theta_\star\|_1 \,.$$

and

$$\begin{aligned}\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2 &\leqslant 2\varepsilon^\top X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star) + \lambda n\|\theta_\star\|_1 - \lambda n\|\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n}\|_1 \,, \\ &\leqslant 2\|X^\top\varepsilon\|_\infty\|\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star\|_1 + \lambda n\|\theta_\star\|_1 - \lambda n\|\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n}\|_1 \,.\end{aligned} \tag{4.2}$$

Let $A = \{\|X^\top\varepsilon\|_\infty < n\lambda/2\}$. Writing $\Sigma = n^{-1}X^\top X$, as $X^\top\varepsilon \sim \mathcal{N}(0, n\sigma_\star^2\Sigma)$,

$$\mathbb{P}(A^c) = \mathbb{P}\left(\|X^\top\varepsilon\|_\infty \geqslant \frac{n\lambda}{2}\right) \leqslant \sum_{j=1}^d \mathbb{P}\left(|X^\top\varepsilon|_j \geqslant \frac{n\lambda}{2}\right) \leqslant 2\sum_{j=1}^d \exp\left\{-n\lambda^2/(8\sigma_\star^2\Sigma_{jj})\right\} \leqslant 2d\exp\left\{-n\lambda^2/(8\sigma_\star^2\|\Sigma\|_\infty\right\} \,.$$

Therefore, with probability at least $1 - 2d\exp\left\{-n\lambda^2/(8\sigma_\star^2\|\Sigma\|_\infty\right\}$,

$$\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2 \leqslant \lambda n\|\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star\|_1 + \lambda n\|\theta_\star\|_1 - \lambda n\|\theta^{\mathrm{lasso}}_{\lambda_\star,n}\|_1$$

and

$$\frac{1}{n}\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2 \leqslant 2\lambda\|\theta_\star\|_1 \,.$$

Then,

$$\mathbb{E}\left[\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2\right] \leqslant 2n\lambda\|\theta_\star\|_1 + \mathbb{E}\left[\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2\mathbb{1}_{A^c}\right] \,.$$

Using that for all $x, y \geq 0$, $2xy \leq x^2/2 + 2y^2$, with $x = \|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2$ and $y = \|\varepsilon\|_2$, by (4.2),

$$\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2 \leqslant 2\|\varepsilon\|_2\|X(\theta^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2 + \lambda n\|\theta_\star\|_1 \leqslant 2\|\varepsilon\|_2^2 + \|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2/2 + \lambda n\|\theta_\star\|_1 \,.$$

Then,

$$\mathbb{E}\left[\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2\right] \leqslant 2n\lambda\|\theta_\star\|_1 + \mathbb{E}\left[(4\|\varepsilon\|_2^2 + 2\lambda n\|\theta_\star\|_1)\mathbb{1}_{A^c}\right]$$

and by Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2\right] \leqslant 4n\lambda\|\theta_\star\|_1 + 4\mathbb{E}\left[\|\varepsilon\|_2^4\right]^{1/2}\mathbb{P}(A^c)^{1/2} \,.$$

Using that $\mathbb{E}[\|\varepsilon\|_2^4]^{1/2} \leqslant 3n\sigma_\star^2$ yields

$$\frac{1}{n}\mathbb{E}\left[\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2\right] \leqslant 4\lambda\|\theta_\star\|_1 + 12\sigma_\star^2\cdot\sqrt{2d}\exp\left\{-n\lambda^2/(16\sigma_\star^2\|\Sigma\|_\infty\right\} \,.$$

By choosing λ so that $n\lambda^2/(16\sigma_\star^2\|\Sigma\|_\infty) = \log(dn)$, we obtain

$$\frac{1}{n}\mathbb{E}\left[\|X(\widehat{\theta}^{\mathrm{lasso}}_{\lambda_\star,n} - \theta_\star)\|_2^2\right] \leqslant 16\sigma_\star\sqrt{\frac{\log(dn)}{n}}\|\Sigma\|_\infty^{1/2}\|\theta_\star\|_1 + 12\sqrt{2}\frac{\sigma_\star^2}{n\sqrt{d}} \,.$$

■

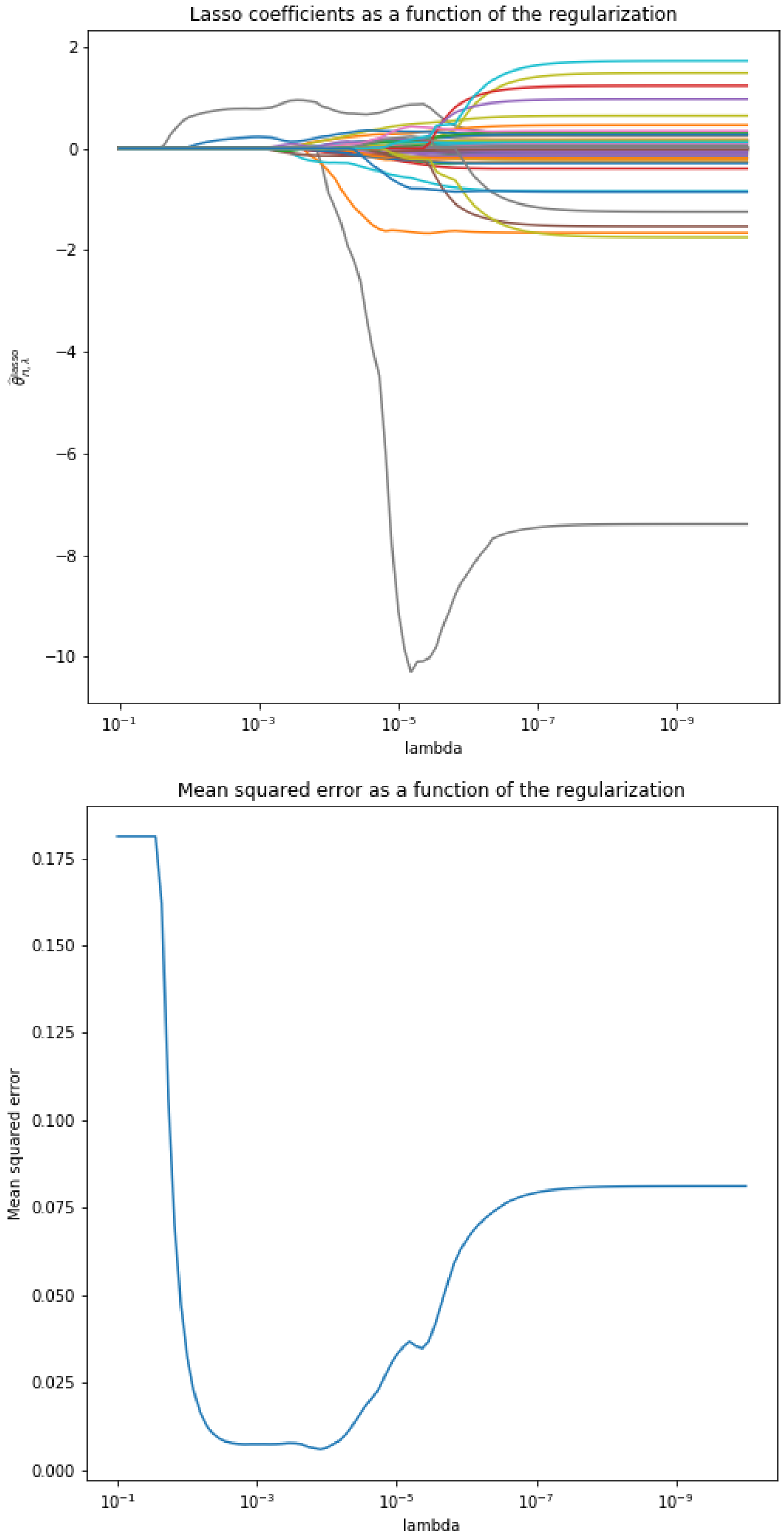

Fig. 4.4 Lasso regression is used to predict the Brazilian inflation based on many observed variables, see `https://github.com/gabrielrvsc/HDeconometrics/`. The model is trained using $n = 140$ data with for each $1 \leqslant i \leqslant n$, $X_i \in \mathbb{R}^{93}$, i.e. $d = 93$. The features are econometric data available each month. (Top) Estimated coefficient $\widehat{\theta}_{n,\lambda}^{\text{lasso}}$ as a function of λ. (Bottom) Mean squared error between the true observations and the predictions over the test set with 15 new data points.

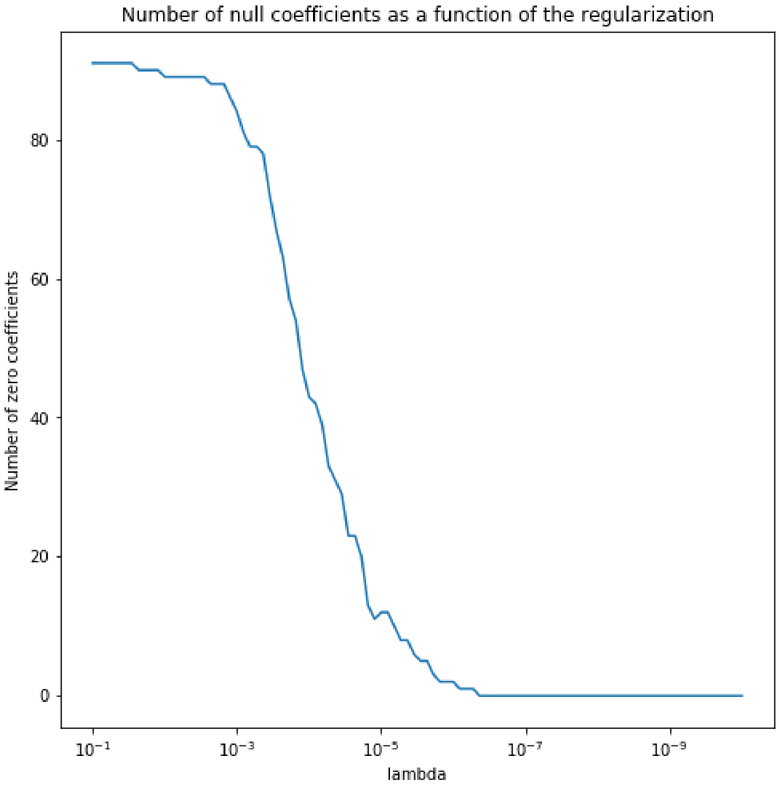

Fig. 4.5 Lasso regression is used to predict the Brazilian inflation based on many observed variables, see `https://github.com/gabrielrvsc/HDeconometrics/`. The model is trained using $n = 140$ data with for each $1 \leqslant i \leqslant n$, $X_i \in \mathbb{R}^{93}$, i.e. $d = 93$. The features are econometric data available each month. Number of null coefficient in $\widehat{\theta}^{\text{lasso}}_{n,\lambda}$ as a function of λ.

Chapter 5

Kernel-based regression

Contents

5.1 Nonparametric regression

In a nonparametric regression framework, it is not assumed that the observations depend linearly on the covariates and a more general model is introduced. For all $1 \leqslant i \leqslant n$, the observation model is given by

$$Y_i = f^*(X_i) + \xi_i \, ,$$

where for all $1 \leqslant i \leqslant n$, $X_i \in \mathscr{X}$, and the $(\xi_i)_{1 \leqslant i \leqslant n}$ are i.i.d. centered Gaussian random variables with variance σ^2. The function f^* is unknown and has to be estimated using the observations $(X_i, Y_i)_{1 \leqslant i \leqslant n}$. A simple approach consists in defining an estimator of f^* as a linear combination of $M \geqslant 1$ known functions $(\varphi_1, \dots, \varphi_M)$ defined on $\mathscr{X}$. Define $\mathscr{F}_\varphi$ as

$$\mathscr{F}_\varphi = \left\{ \sum_{j=1}^{M} \alpha_j \varphi_j \, ; \, (\alpha_1, \dots, \alpha_M) \in \mathbb{R}^M \right\} .$$

Then, the least squares estimator of f^* on $\mathscr{F}_\varphi$ is defined as

$$\widehat{f}_n^\varphi \in \underset{f \in \mathscr{F}_\varphi}{\arg\min} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \, .$$

Let Ψ be the $\mathbb{R}^{n \times M}$ matrix such as, for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant M$, $\Psi_{i,j} = \varphi_j(X_i)$. Then, for all $f \in \mathscr{F}_\varphi$, there exists $\alpha = (\alpha_1, \dots, \alpha_M) \in \mathbb{R}^M$ such that,

$$\sum_{i=1}^{n} (Y_i - f(X_i))^2 = \|Y - \Psi\alpha\|_2^2 \, .$$

Then, following the same steps as in Section 3.2, in the case where $\Psi^\top \Psi$ is nonsingular, the least squares estimate is

$$\widehat{f}_n^{\varphi} : x \mapsto \sum_{j=1}^{M} \widehat{\alpha}_{n,j} \varphi_j \;, \tag{5.1}$$

where

$$\widehat{\alpha}_n = (\Psi^\top \Psi)^{-1} \Psi^\top Y \;.$$

Introducing the function $\varphi : x \mapsto (\varphi_1(x) \ldots, \varphi_M(x))^\top$ yields the linear estimator

$$\widehat{f}_n^{\varphi} : x \mapsto \sum_{i=1}^{n} w_i(x) Y_i \;,$$

where, for all $1 \leqslant i \leqslant n$,

$$w_i(x) = \left(\varphi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top\right)_i \;.$$

Proposition 5.1 *Let* $W = (w_i(X_j))_{1 \leqslant i,j \leqslant n}$ *and* $\bar{f}^* = (f^*(X_1), \ldots, f^*(X_n))^\top$. *Then,*

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} (\widehat{f}_n^{\varphi}(X_i) - f^*(X_i))^2\right] = \frac{1}{n}\sum_{i=1}^{n} ((W\bar{f}^*)_i - f^*(X_i))^2 + \frac{\sigma^2}{n}\mathrm{Trace}(W^\top W) \;,$$

where $\widehat{f}_n^{\varphi}$ *is defined by* (5.1).

Proof. See the exercises. ■

5.2 Kernels

Let $\mathscr{F}$ be a set of functions from $\mathscr{X} = \mathbb{R}^d$ to $\mathbb{R}$. The multivariate regression problem considered so far is part of the more general framework

$$\widehat{f}_{\mathscr{F}}^{n} \in \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) + \lambda \|f\|_{\mathscr{F}}^2 \;, \tag{5.2}$$

where $\lambda > 0$ and $\|\cdot\|_{\mathscr{F}}$ is a norm on the space $\mathscr{F}$. In the case of the Ridge regression $\ell : (y,y') :\mapsto \|y - y'\|_2^2$, $\mathscr{F} = \{f : \mathbb{R}^d \to \mathbb{R} \;;\; \exists \theta \in \mathbb{R}^d \; \forall x \in \mathbb{R}^d \;,\; f(x) = f_\theta(x) = \theta^\top x\}$ and, for $\theta \in \mathbb{R}^d$ and $f_\theta \in \mathscr{F}$, $\|f_\theta\|^2 = \|\theta\|_2^2$. Before considering kernel-based regression it is useful to understand how we can use our knowledge on linear regression in a more general setting. Consider $\mathscr{F}$ a Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathscr{F}}$. We will focus on settings where for all $f \in \mathscr{F}$ there exist a function $\phi : \mathbb{R}^d \to \mathscr{F}$ and $\theta \in \mathscr{F}$, such that for all $x \in \mathbb{R}^d$,

$$f(x) = \langle \theta, \phi(x) \rangle_{\mathscr{F}} \;.$$

Example 5.2 *Let f be defined on $\mathbb{R}^2$ by $f : (x_1, x_2) \to 2x_1^2 + x_2^2/2 - \sqrt{2}x_1x_2$. Then, choosing $\mathscr{F} = \mathbb{R}^3$ with its canonical inner product, $\theta = (2, 1/2, -1)^\top$ and $\phi : (x_1, x_2) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2)^\top$ yields $f(x) = \langle \theta, \phi(x) \rangle$.*

If the function ϕ is known, we can use this representation to apply the results on linear regression.

Definition 5.3. A function $k : \mathscr{X} \times \mathscr{X} :\to \mathbb{R}$ is said to be a positive semi-definite (PSD) kernel if and only if it is symmetric and if for all $n \geqslant 1$, $(x_1, \ldots, x_n) \in \mathscr{X}^n$ and all $(a_1, \ldots, a_n) \in \mathbb{R}^n$,

$$\sum_{1\leqslant i,j\leqslant n} a_i a_j k(x_i,x_j) \geqslant 0\,.$$

Remark 5.4 *The following functions, defined on $\mathbb{R}^d\times\mathbb{R}^d$, are positive semi-definite kernels:*

$$k:(x,y)\mapsto x^\top y \quad\text{and}\quad k:(x,y)\mapsto \exp\left(-\|x-y\|^2/(2\sigma^2)\right)\,,\ \sigma>0\,.$$

Proposition 5.5 *Assume that $\mathscr{F}$ is a Hilbert space endowed with the inner product $\langle\cdot,\cdot\rangle_{\mathscr{F}}$. If $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ is defined by $k:(x,y)\to\langle\phi(x),\phi(y)\rangle_{\mathscr{F}}$ where $\phi:\mathbb{R}^d\to\mathscr{F}$, then k is a PSD kernel.*

PROOF. By definition of an inner product, k is symmetric. For all $n\geqslant 1$, $(x_1,\ldots,x_n)\in\mathscr{X}^n$ and all $(a_1,\ldots,a_n)\in\mathbb{R}^n$,

$$\begin{aligned}\sum_{1\leqslant i,j\leqslant n} a_i a_j k(x_i,x_j) &= \sum_{1\leqslant i,j\leqslant n} a_i a_j \left\langle\phi(x_i),\phi(x_j)\right\rangle_{\mathscr{F}}\\ &= \left\langle\sum_{i=1}^n a_i\phi(x_i),\sum_{j=1}^n a_j\phi(x_j)\right\rangle_{\mathscr{F}}\\ &= \left\|\sum_{i=1}^n a_i\phi(x_i)\right\|_{\mathscr{F}}^2\,,\end{aligned}$$

which completes the proof. ■

Proposition 5.6 *Let $k_1:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ and $k_2:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ be two PSD kernels. Then the following functions are also PSD kernels.*

1. $\alpha k_1+\beta k_2$ *for* $\alpha,\beta>0$.
2. $k_1\cdot k_2$.
3. $\exp(k_1)$.

5.3 Reproducing kernel Hilbert spaces

A useful case in practice consists in choosing $\mathscr{F}$ as a Reproducing Kernel Hilbert Space with positive definite reproducing kernel k on $\mathscr{X}\times\mathscr{X}$. A first important result that we do not prove in these notes is the following theorem.

Theorem 5.7. *Let $k:\mathscr{X}\times\mathscr{X}:\to\mathbb{R}$ be a PSD kernel. Then, there exist Hilbert space $(\mathscr{F},\langle\cdot,\cdot\rangle_{\mathscr{F}})$ and a function $\phi:\mathbb{R}^d\to\mathscr{F}$ such that $k:(x,y)\to\langle\phi(x),\phi(y)\rangle_{\mathscr{F}}$.*

We can now introduce a canonical construction of $\mathscr{F}$ and ϕ associated with a PSD kernel $k:\mathscr{X}\times\mathscr{X}:\to\mathbb{R}$.

Definition 5.8. Let $\mathscr{F}$ be a Hilbert space of functions $f:\mathscr{X}\to\mathbb{R}$. A symmetric function $k:\mathscr{X}\times\mathscr{X}:\to\mathbb{R}$ is said to be a reproducing kernel of $\mathscr{F}$ if and only if for all $x\in\mathscr{X}$, $k(x,\cdot)\in\mathscr{F}$ and for all $x\in$ X and all $f\in\mathscr{F}$, $\langle f;k(x,\cdot)\rangle=f(x)$. The space $\mathscr{F}$ is said to be a reproducing kernel Hilbert space (RKHS) with kernel k.

Remark 5.9 *For all $x \in \mathscr{X}$, the function $k(x,\cdot)$ is called a feature map and often written $\phi(x)$. In this setting, all $x,x' \in \mathscr{X}$, $k(x,x') = \langle \phi(x);\phi(x')\rangle$. An important result states that a function $k : \mathscr{X} \times \mathscr{X} :\to \mathbb{R}$ is a positive semi-definite kernel if and only if there exist a Hilbert space $\mathscr{F}$ and a function $\phi : \mathscr{X} \to \mathscr{F}$ such that $k(x,x') = \langle \phi(x);\phi(x')\rangle$.*

A reproducing kernel associated with a reproducing kernel Hilbert space is positive semi-definite since for all $n \geqslant 1$, $(x_1,\ldots,x_n) \in \mathscr{X}^n$ and all $(a_1,\ldots,a_n) \in \mathbb{R}^n$,

$$\sum_{1\leqslant i,j\leqslant n} a_i a_j k(x_i,x_j) = \sum_{1\leqslant i,j\leqslant n} a_i a_j \langle k(x_i,\cdot);k(x_j,\cdot)\rangle = \left\| \sum_{1\leqslant i\leqslant n} a_i \langle k(x_i,\cdot) \right\|^2 \geqslant 0\,.$$

Remark 5.10 *The positive semi-definite kernel $k : (x,y) \mapsto x^\top y$ defined on $\mathbb{R}^d \times \mathbb{R}^d$ is a reproducing kernel of the space*

$$\mathscr{F} = \left\{ f : \mathbb{R}^d \to \mathbb{R}\,;\, \exists \omega \in \mathbb{R}^d\ \forall x \in \mathbb{R}^d\,,\ f(x) = \omega^\top x \right\}\,,$$

equipped with the inner product defined, for all $(f,g) \in \mathscr{F} \times \mathscr{F}$, by

$$\langle f;g\rangle = \omega_f^\top \omega_g\,,$$

where $\omega_f, \omega_g \in \mathbb{R}^d$ and $f : x \mapsto \omega_f^\top x$, $g : x \mapsto \omega_g^\top x$.

Proposition 5.11 proves that the minimization of the penalized empirical loss amounts to solving a convex optimization problem on $\mathbb{R}^n$ for which many efficient numerical solution exist.

Proposition 5.11 *Let $k : \mathscr{X} \times \mathscr{X} :\to \mathbb{R}$ be a positive definite kernel and $\mathscr{F}$ the RKHS with kernel k. Then,*

$$\widehat{f}^n_{\mathscr{F}} \in \operatorname*{arg\,min}_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \|f\|^2_{\mathscr{F}}\,,$$

where $\|f\|^2_{\mathscr{F}} = \langle f\,;\,f\rangle$, is given by $\widehat{f}^n_{\mathscr{F}} : x \mapsto \sum_{i=1}^n \widehat{\alpha}_i k(X_i,x)$, where

$$\widehat{\alpha} \in \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i, \sum_{j=1}^n \alpha_j k(X_j,X_i) \right) + \lambda \sum_{1\leqslant i,j\leqslant n} \alpha_i\alpha_j k(X_i,X_j) \right\}\,.$$

PROOF. Let V be the linear space spanned by $(k(X_i,\cdot))_{1\leqslant i\leqslant n}$. For all $f \in \mathscr{F}$, f can be written $f = f_V + f_{V^\perp}$ with $f_V \in V$ and $f_{V^\perp} \in V^\perp$. Since $\mathscr{F}$ is a RKHS with kernel k, for all $1 \leqslant i \leqslant n$,

$$f_{V^\perp}(X_i) = 0 \quad \text{and} \quad f(X_i) = \langle f\,;\,k(X_i,\cdot)\rangle = f_V(X_i)\,.$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^n \ell(Y_i,f(X_i)) + \lambda\|f\|^2 = \frac{1}{n}\sum_{i=1}^n \ell(Y_i,f_V(X_i)) + \lambda\|f_V\|^2 + \lambda\|f_{V^\perp}\|^2$$

and any minimizer of the target function is in V. There exist $(\alpha_1,\ldots,\alpha_n) \in \mathbb{R}^n$ such that

$$\widehat{f}^n_{\mathscr{F}} : x \mapsto \sum_{i=1}^n \alpha_i k(X_i,x)\,,$$

which concludes the proof. ■

Therefore, Proposition 5.11 establishes that solving

$$\widehat{f}^n_{\mathscr{F}} \in \operatorname*{arg\,min}_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \|f\|^2_{\mathscr{F}}\,,$$

amounts to compute $\widehat{f}^n_{\mathscr{F}} : x \mapsto \sum_{i=1}^n \widehat{\alpha}_i k(X_i,x)$ where

$$\widehat{\alpha} \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, K\alpha_i) + \lambda \alpha^\top K \alpha \right\} .$$

In a Ridge regression setting thus yields:

$$\widehat{\alpha} \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \|Y - K\alpha\|_2^2 + \lambda \alpha^\top K \alpha \right\} .$$

Chapter 6

Feed Forward neural networks

Contents

Keywords 6.1 *Multi-layer Perceptron, Feed Forward Neural Networks*

6.1 Feed Forward Neural Networks - Multi layer perceptron

Following the logistic regression approach where $(X,Y) \in \mathbb{R}^d \times \{1,\ldots,M\}$ for $M \geqslant 2$, the Multi-layer Perceptron (MLP) provides a parametric function to model $\mathbb{P}(Y=k|X)$ for each possible class k. The first mathematical model for a neuron was the Threshold Logic Unit (McCulloch and Pitts, 1943), with Boolean inputs and outputs. In this setting, the response associated with an input $x \in \{0,1\}^d$ is defined as

$$f : x \mapsto \mathbb{1}_{\omega \sum_{j=1}^d x_j + b \geqslant 0} \,.$$

This construction allows to build any boolean function from elementary units

$$x \vee y = \mathbb{1}_{x+y-1/2\geqslant 0}\,, \quad x \wedge y = \mathbb{1}_{x+y-3/2\geqslant 0} \quad \text{and} \quad 1-x = \mathbb{1}_{-x+1/2\geqslant 0}$$

This elementary model can be extended to real valued inputs (Rosenblatt, 1957) with

$$f : x \mapsto \mathbb{1}_{\sum_{j=1}^d \omega_j x_j + b \geqslant 0} \,.$$

In this case, the nonlinear activation function is $\sigma : x \mapsto \mathbb{1}_{x\geqslant 0}$ and the ouput in $\{0,1\}$ defined as:

$$f : x \mapsto \sigma(\omega^\top x + b)\,.$$

Linear discriminant analysis and logistic regression are other instances with the sigmoid activation function $\sigma : x \mapsto \mathrm{e}^x/(1+\mathrm{e}^x)$ and $\sigma(\omega^\top X + b)$ in $(0,1)$ models $\mathbb{P}(Y=1|X)$. The Multi-layer Perceptron (MLP) or Feed Forward Neural Network (FFNN) weakens the modeling assumptions of LDA or logistic regression and composed in parallel L of these perceptron units to produce the output. Let $x \in \mathbb{R}^d$ be the input and define all layers as follows.

$$\begin{aligned} h_\theta^0(x) &= x\,,\\ z_\theta^k(x) &= b^k + W^k h_\theta^{k-1}(x) \quad \text{for all } 1 \leqslant k \leqslant L\,,\\ h_\theta^k(x) &= \varphi_k(z_\theta^k(x)) \quad \text{for all } 1 \leqslant k \leqslant L\,, \end{aligned}$$

where $b^1 \in \mathbb{R}^{d_1}$, $W^1 \in \mathbb{R}^{d_1 \times d}$ and for all $2 \leqslant k \leqslant L$, $b^k \in \mathbb{R}^{d_k}$, $W^k \in \mathbb{R}^{d_k \times d_{k-1}}$. For all $1 \leqslant k \leqslant L$, $\varphi_k : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ is a nonlinear activation function. Let $\theta = \{b^1, W^1, \ldots, b^L, W^L\}$ be the unknown parameters of the MLP and $f_\theta(x) = h_\theta^L(x)$ be the output layer of the MLP. As there is no modelling assumptions anymore, virtually any activation functions φ^m, $1 \leqslant m \leqslant L-1$ may be used. In this section, it is assumed that these intermediate activation functions apply elementwise and, with a minor abuse of notations, we write for all $1 \leqslant m \leqslant L-1$ and all $z \in \mathbb{R}^{d_m}$,

$$\varphi^m(z) = (\varphi^m(z_1), \ldots, \varphi^m(z_{d_m}))\,,$$

with $\varphi^m : \mathbb{R} \to \mathbb{R}$ the seleced scalar activation function. The rectified linear unit (RELU) activation function $x \mapsto \max(0, x)$ and its extensions are the default recommendation in modern implementations (Jarrettet al., 2009; Nair and Hinton, 2010; Glorot et al., 2011a), (Maas et al.,2013), (He et al., 2015). One of the major motivations arise from the gradient based parameter optimization which is numerically more stable with this choice. The choice of the last activation function φ^L greatly relies on the task the network is assumed to perform.

- **biclass classification**. The output $h_\theta^L(x)$ is the estimate of the probability that the class is 1 given the input x. The common choice in this case is the sigmoid function. Then, $d_L = 1$ and $h_\theta^L(x)$ contains $\mathbb{P}(Y = 1|X)$ and is enough to use as a plug-in Bayes classifier.

$$\varphi^m(z) = \frac{\mathrm{e}^z}{1 + \mathrm{e}^z}\,.$$

- **multiclass classification**. The output $h_\theta^L(x)$ is the estimate of the probability that the class is k for all $1 \leqslant k \leqslant M$, given the input x. The common choice in this case is the softmax function: for all $1 \leqslant i \leqslant M$

$$\varphi^m(z)_i = \mathrm{softmax}(z)_i = \frac{\mathrm{e}^{z_i}}{\sum_{j=1}^M \mathrm{e}^{z_j}}\,.$$

In this case $d_L = M$ and each component k of $h_\theta^L(x)$ contains $\mathbb{P}(Y = k|X)$.

6.2 Gradients and backpropagation

6.2.1 Classification: loss function and gradient

The standard approach to estimate the parameters is by maximizing the logarithm of the likelihood i.e. by minimizing the opposite of the normalized loglikelihood:

$$\ell_n : \theta \mapsto -\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^M \mathbb{1}_{Y_i = k} \log \mathbb{P}_\theta(Y_i = k|X_i) = -\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^M \mathbb{1}_{Y_i = k} \log f_\theta(X_i)_k\,.$$

Assume in this section that the last activation function is the softmax function. The output $h_\theta^L(x) = f_\theta(x)$ is the estimate of the probability that the class is k for all $1 \leqslant k \leqslant M$, given the input x which is obtained with

$$(\varphi_L(z))_i = \mathrm{softmax}(z)_i = \frac{\mathrm{e}^{z_i}}{\sum_{j=1}^M \mathrm{e}^{z_j}}\,.$$

Therefore, for all $1 \leqslant i, j \leqslant M$,

$$\partial_{z_i}(\varphi_L(z))_j = \begin{cases} \mathrm{softmax}(z)_i(1-\mathrm{softmax}(z)_i) & \text{if } i=j\,,\\ -\mathrm{softmax}(z)_i\mathrm{softmax}(z)_j & \text{otherwise.}\end{cases}$$

Proposition 6.1 (Back propagation - classification) *Write $\ell_\theta(X,Y) = -\sum_{k=1}^M \mathbb{1}_{Y=k}\log f_\theta(X)_k$ so that*

$$\ell_n : \theta \mapsto \frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i,Y_i)\,.$$

Therefore, the gradient with respect to all parameters can be computed as follows.

$$\begin{aligned}\nabla_{W^L}\ell_\theta(X,Y) &= (f_\theta(X)-\mathbb{1}_Y)(h_\theta^{L-1}(X))^\top\,,\\ \nabla_{b^L}\ell_\theta(X,Y) &= f_\theta(X)-\mathbb{1}_Y\,,\end{aligned}$$

where $\mathbb{1}_Y$ is the vector where all entries equal to 0 except the entry with index Y which equals 1. Then, for all $1\leqslant m\leqslant L-1$,

$$\begin{aligned}\nabla_{W^m}\ell_\theta(X,Y) &= \nabla_{z_\theta^m(X)}\ell_\theta(X,Y)(h_\theta^{m-1}(X))^\top\,,\\ \nabla_{b^m}\ell_\theta(X,Y) &= \nabla_{z_\theta^m(X)}\ell_\theta(X,Y)\,,\end{aligned}$$

where $\nabla_{z_\theta^m(X)}$ is computed recursively as follows.

$$\begin{aligned}\nabla_{z^L(X)}\ell_\theta(X,Y) &= f_\theta(X)-\mathbb{1}_Y\,,\\ \nabla_{h_\theta^m(X)}\ell_\theta(X,Y) &= (W^{m+1})^\top\nabla_{z_\theta^{m+1}(X)}\ell_\theta(X,Y)\,,\\ \nabla_{z_\theta^m(X)}\ell_\theta(X,Y) &= \nabla_{h_\theta^m(X)}\ell_\theta(X,Y)\odot\varphi_m'(z_\theta^m(X))\,,\end{aligned}$$

where $\odot$ is the elementwise multiplication.

PROOF. For all $1\leqslant j\leqslant M$,

$$\begin{aligned}\partial_{(z^L(X))_j}\ell_\theta(X,Y) &= -\sum_{k=1}^M \mathbb{1}_{Y=k}\partial_{(z^L(X))_j}\log f_\theta(X)_k\,,\\ &= -\sum_{k=1}^M \mathbb{1}_{Y=k}\partial_{(z^L(X))_j}\log\mathrm{softmax}(z^L(X))_k\,,\\ &= -\sum_{k=1}^M \mathbb{1}_{Y=k}\frac{\mathrm{softmax}(z^L(X))_j(1-\mathrm{softmax}(z^L(X))_j)\mathbb{1}_{j=k}-\mathrm{softmax}(z^L(X))_j\mathrm{softmax}(z^L(X))_k\mathbb{1}_{j\neq k}}{\mathrm{softmax}(z^L(X))_k}\,,\\ &= -\sum_{k=1}^M \mathbb{1}_{Y=k}\left\{(1-\mathrm{softmax}(z^L(X))_k)\mathbb{1}_{j=k}-\mathrm{softmax}(z^L(X))_k\mathbb{1}_{j\neq k}\right\}\,.\end{aligned}$$

Therefore,

$$\nabla_{z^L(X)}\ell_\theta(X,Y) = f_\theta(X)-\mathbb{1}_Y\,.$$

Then, for all $1\leqslant i\leqslant M$ and all $1\leqslant j\leqslant d_{L-1}$, by the chain rule, and using that $z_\theta^L(X) = b^L + W^L h_\theta^{L-1}(X)$,

$$\begin{aligned}\partial_{W_{i,j}^L}\ell_\theta(X,Y) &= \sum_{k=1}^M \partial_{(z_\theta^L(X))_k}\ell_\theta(X,Y)\partial_{W_{i,j}^L}(z_\theta^L(X))_k\,,\\ &= \sum_{k=1}^M (f_\theta(X)-\mathbb{1}_Y)_k\mathbb{1}_{i=k}(h_\theta^{L-1}(X))_j\,,\\ &= (f_\theta(X)-\mathbb{1}_Y)_i(h_\theta^{L-1}(X))_j\,.\end{aligned}$$

Therefore,

$$\nabla_{W^L}\ell_\theta(X,Y) = (f_\theta(X)-\mathbb{1}_Y)(h_\theta^{L-1}(X))^\top\,.$$

Similarly, for all $1 \leqslant i \leqslant M$, using that $z^L_\theta(X) = b^L + W^L h^{L-1}_\theta(X)$,

$$\begin{aligned}\partial_{b^L_i}\ell_\theta(X,Y) &= \sum_{k=1}^{M} \partial_{(z^L_\theta(X))_k}\ell_\theta(X,Y)\partial_{b^L_i}(z^L_\theta(X))_k\,,\\ &= \sum_{k=1}^{M} (\ell_\theta(X,Y) - \mathbb{1}_Y)_k \mathbb{1}_{i=k}\,,\\ &= (f_\theta(X) - \mathbb{1}_Y)_i\,.\end{aligned}$$

Therefore,

$$\nabla_{b^L}\ell_\theta(X,Y) = f_\theta(X) - \mathbb{1}_Y\,.$$

To obtain the recursive formulation of the gradient computations, known as the *back propagation* of the gradient, write, for all $1 \leqslant m \leqslant L-1$ and all $1 \leqslant j \leqslant d_m$, , using that $z^{m+1}_\theta(X) = b^{m+1} + W^{m+1} h^m_\theta(X)$,

$$\begin{aligned}\partial_{(h^m_\theta(X))_j}\ell_\theta(X,Y) &= \sum_{i=1}^{d_{m+1}} \partial_{(z^{m+1}_\theta(X))_i}\ell_\theta(X,Y)\partial_{(h^m_\theta(X))_j}(z^{m+1}_\theta(X))_i\,,\\ &= \sum_{i=1}^{d_{m+1}} \partial_{(z^{m+1}_\theta(X))_i}\ell_\theta(X,Y)W^{m+1}_{i,j}\,.\end{aligned}$$

Therefore,

$$\nabla_{h^m_\theta(X)}\ell_\theta(X,Y) = (W^{m+1})^\top \nabla_{z^{m+1}_\theta(X)}\ell_\theta(X,Y)\,.$$

Then, for all $1 \leqslant m \leqslant L-1$ and all $1 \leqslant j \leqslant d_m$, , using that $h^m_\theta(X)_j = \varphi_m(z^m_\theta(X)_j)$,

$$\begin{aligned}\partial_{(z^m_\theta(X))_j}\ell_\theta(X,Y) &= \sum_{i=1}^{d_{m+1}} \partial_{(h^m_\theta(X))_i}\ell_\theta(X,Y)\partial_{(z^m_\theta(X))_j}(h^m_\theta(X))_i\,,\\ &= \sum_{i=1}^{d_{m+1}} \partial_{(h^m_\theta(X))_i}\ell_\theta(X,Y)\mathbb{1}_{i=j}\varphi'_m(z^m_\theta(X)_i)\,,\\ &= \partial_{(h^m_\theta(X))_j}\ell_\theta(X,Y)\varphi'_m(z^m_\theta(X)_j)\,.\end{aligned}$$

Therefore,

$$\nabla_{z^m_\theta(X)}\ell_\theta(X,Y) = \nabla_{h^m_\theta(X)}\ell_\theta(X,Y) \odot \varphi'_m(z^m_\theta(X))\,.$$

Then, for all $1 \leqslant i \leqslant d_m$ and all $1 \leqslant j \leqslant d_{m-1}$, and using that $z^m_\theta(X) = b^m + W^m h^{m-1}_\theta(X)$,

$$\begin{aligned}\partial_{W^m_{i,j}}\ell_\theta(X,Y) &= \sum_{k=1}^{d_m} \partial_{(z^m_\theta(X))_k}\ell_\theta(X,Y)\partial_{W^m_{i,j}}(z^m_\theta(X))_k\,,\\ &= \sum_{k=1}^{d_m} \partial_{(z^m_\theta(X))_k}\ell_\theta(X,Y)\mathbb{1}_{i=k}(h^{m-1}_\theta(X))_j\,,\\ &= \partial_{(z^m_\theta(X))_i}\ell_\theta(X,Y)(h^{m-1}_\theta(X))_j\,.\end{aligned}$$

Therefore,

$$\nabla_{W^m}\ell_\theta(X,Y) = \nabla_{z^m_\theta(X)}\ell_\theta(X,Y)(h^{m-1}_\theta(X))^\top\,.$$

Similarly, for all $1 \leqslant i \leqslant d_m$, using that $z^m_\theta(X) = b^m + W^m h^{m-1}_\theta(X)$,

$$\begin{aligned}\partial_{b^m_i}\ell_\theta(X,Y) &= \sum_{k=1}^{d_m} \partial_{(z^m_\theta(X))_k}\ell_\theta(X,Y)\partial_{b^m_i}(z^m_\theta(X))_k\,,\\ &= \sum_{k=1}^{d_m} \partial_{(z^m_\theta(X))_k}\ell_\theta(X,Y)\mathbb{1}_{i=k}\,,\\ &= \partial_{(z^m_\theta(X))_k}\ell_\theta(X,Y)_i\,.\end{aligned}$$

Therefore,

$$\nabla_{b^m}\ell_\theta(X,Y) = \nabla_{z^m_\theta(X)}\ell_\theta(X,Y)\,.$$

■

6.2.2 Regression: loss function and gradient

In a regression setting, it is assumed that the observations satisfy for all $1 \leqslant i \leqslant n$, $Y_i = f_\star(X_i) + \varepsilon_i$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered random variables in $\mathbb{R}^M$, $X_i \in \mathbb{R}^d$ and $f_\star$ is an unknown function. A Feed Forward Neural Network may be used to estimate $f_\star(X_i)$ (i.e. Y_i) as follows.

$$\begin{aligned}
&h_\theta^0(X_i) = X_i \; , \\
&z_\theta^k(X_i) = b^k + W^k h_\theta^{k-1}(X_i) \quad \text{for all } 1 \leqslant k \leqslant L \; , \\
&h_\theta^k(X_i) = \varphi_k(z_\theta^k(X_i)) \quad \text{for all } 1 \leqslant k \leqslant L \; ,
\end{aligned}$$

where $b^1 \in \mathbb{R}^{d_1}$, $W^1 \in \mathbb{R}^{d_1\times d}$ and for all $2 \leqslant k \leqslant L$, $b^k \in \mathbb{R}^{d_k}$, $W^k \in \mathbb{R}^{d_k\times d_{k-1}}$. Let $\theta = \{b^1, W^1, \ldots, b^L, W^L\}$ be the unknown parameters of the MLP and $f_\theta(x) = h_\theta^L(x)$ be the output layer of the MLP. The standard approach to estimate the parameters is by minimizing the mean square error:

$$\ell_n : \theta \mapsto \frac{1}{n}\sum_{i=1}^n \|f_\theta(X_i) - Y_i\|^2 \; .$$

In a regression setting, the last activation function is usually the identity function (i.e. the output is the linear transform $z_\theta^L(X_i)$). The output $h_\theta^L(X_i) = f_\theta(X_i)$ is the estimate of Y_i.

Proposition 6.2 (Back propagation - regression) *Write $\ell_\theta(X,Y) = \|f_\theta(X) - Y\|_2^2$ so that*

$$\ell_n : \theta \mapsto \frac{1}{n}\sum_{i=1}^n \ell_\theta(X_i, Y_i) \; .$$

Therefore, the gradient with respect to all parameters can be computed as follows.

$$\begin{aligned}
\nabla_{W^L}\ell_\theta(X,Y) &= -2(Y - f_\theta(X))(h_\theta^{L-1}(X))^\top \; , \\
\nabla_{b^L}\ell_\theta(X,Y) &= -2(Y - f_\theta(X)) \; .
\end{aligned}$$

Then, for all $1 \leqslant m \leqslant L-1$,

$$\begin{aligned}
\nabla_{W^m}\ell_\theta(X,Y) &= \nabla_{z_\theta^m(X)}\ell_\theta(X,Y)(h_\theta^{m-1}(X))^\top \; , \\
\nabla_{b^m}\ell_\theta(X,Y) &= \nabla_{z_\theta^m(X)}\ell_\theta(X,Y) \; ,
\end{aligned}$$

where $\nabla_{z_\theta^m(X)}$ is computed recursively as follows.

$$\begin{aligned}
\nabla_{h_\theta^m(X)}\ell_\theta(X,Y) &= (W^{m+1})^\top \nabla_{z_\theta^{m+1}(X)}\ell_\theta(X,Y) \; , \\
\nabla_{z_\theta^m(X)}\ell_\theta(X,Y) &= \nabla_{h_\theta^m(X)}\ell_\theta(X,Y) \odot \varphi_m'(z_\theta^m(X)) \; ,
\end{aligned}$$

where $\odot$ is the elementwise multiplication.

PROOF. By definition, $\ell_\theta(X,Y) = \|f_\theta(X) - Y\|_2^2 = \left\|b^L + W^L h_\theta^{L-1}(X) - Y\right\|_2^2$. For all $1 \leqslant i \leqslant M$ and $1 \leqslant j \leqslant d_{L-1}$,

$$\partial_{W_{i,j}^L}\ell_\theta(X,Y) = -2(Y_i - (f_\theta(X))_i)h_\theta^{L-1}(X)_j \; ,$$

where Y_i is the i-th coordinate of Y. Therefore,

$$\nabla_{W^L}\ell_\theta(X,Y) = -2(Y - f_\theta(X))(h_\theta^{L-1}(X))^\top \; .$$

Similarly,

$$\nabla_{b^L}\ell_\theta(X,Y) = -2(Y - f_\theta(X)) \; .$$

The other identities are the same as in the classification setting. ■

Chapter 7

Technical results

Contents

7.1 Refresher

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.

Definition 7.1 (conditional expectation). Let X be a nonnegative random variable and $\mathscr{G}$ be a sub-σ-field of $\mathscr{F}$. There exists a nonnegative and $\mathscr{G}$-measurable random variable Y such that, for all nonnegative and $\mathscr{G}$-measurable random variables Z,

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] \, . \tag{7.1}$$

Remark 7.2 *If $\mathbb{E}[X] < \infty$, then $\mathbb{E}[Y] < \infty$. If $\widetilde{Y}$ is another nonnegative and $\mathscr{G}$-measurable random variable satisfying* (7.1)*, then $\widetilde{Y} = Y$ $\mathbb{P}$-a.s.*

A random variable Y satisfying the assumptions of Definition 7.1 is called (a verion of) the conditional espectation of X given $\mathscr{G}$ and written $\mathbb{E}[X|\mathscr{G}]$. For all measurable sets A, we also write $\mathbb{E}[\mathbb{1}_A|\mathscr{G}] = \mathbb{P}(A|\mathscr{G})$. For all random variables X define

$$\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X_+|\mathscr{G}] - \mathbb{E}[X_-|\mathscr{G}] \, ,$$

where $X_+ = \max(0, X)$ and $X_- = \max(-X, 0)$ are nonnegative random variables.

Proposition 7.3 *Let X be a random variable such that $\mathbb{E}[X_-|\mathscr{G}] < \infty$ and $\mathscr{G}$ be a sub-σ-field of $\mathscr{F}$. Then, $\mathbb{P}$-a.s., the following properties hold.*

- *If $\mathscr{H}$ is a sub-σ-field of $\mathscr{G}$, then $\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[\mathbb{E}[X|\mathscr{H}]|\mathscr{G}]$.*
- *If $\mathscr{G} = \{\emptyset, \Omega\}$, $\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X]$.*
- *If X is independent of $\mathscr{G}$, then $\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X]$.*

- *If X is $\mathscr{G}$-measurable and $\mathbb{E}[|XY|] < \infty$ and $\mathbb{E}[|Y|] < \infty$, then $\mathbb{E}[XY|\mathscr{G}] = X\mathbb{E}[Y|\mathscr{G}]$.*

7.2 Probabilistic inequalities

Theorem 7.4 (Hoeffding's inequality). *Let $(X_i)_{1\leqslant i\leqslant n}$ be n independent random variables such that for all $1 \leqslant i \leqslant n$, $\mathbb{P}(a_i \leqslant X_i \leqslant b_i) = 1$ where a_i, b_i are real numbers such that $a_i < b_i$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]\right| > t\right) \leqslant 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) .$$

PROOF. Without loss of generality, assume that $\mathbb{E}\left[X_i\right] = 0$ for all $1 \leqslant i \leqslant n$. It is enough to prove that, for all $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) \leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) . \tag{7.2}$$

Equation (7.2) implies Hoeffding's inequality by noting that $\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > t\right) \leq \mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) + \mathbb{P}\left(-\sum_{i=1}^{n} X_i > t\right)$ and by applying (7.2) to $(X_i)_{1\leqslant i\leqslant n}$ and $(-X_i)_{1\leqslant i\leqslant n}$. Write, for any $s,t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) = \mathbb{P}\left(\mathrm{e}^{s\sum_{i=1}^{n} X_i} > \mathrm{e}^{st}\right) < \mathrm{e}^{-st}\mathbb{E}\left[\mathrm{e}^{s\sum_{i=1}^{n} X_i}\right] = \mathrm{e}^{-st}\prod_{i=1}^{n}\mathbb{E}\left[\mathrm{e}^{sX_i}\right]$$

To bound the right hand side of this inequality, set, for all $1 \leqslant i \leqslant n$, $\phi_i : s \mapsto \log\left(\mathbb{E}\left[\mathrm{e}^{sX_i}\right]\right)$. Since X_i is almost surely bounded, ϕ_i is differentiable and for all $s > 0$, $\phi_i'(s) = \mathbb{E}\left[X_i\mathrm{e}^{sX_i}\right]/\mathbb{E}\left[\mathrm{e}^{sX_i}\right]$. Then, differentiating again,

$$\phi_i''(s) = \log''\left(\mathbb{E}\left[\mathrm{e}^{sX_i}\right]\right) = \frac{\mathbb{E}\left[X_i^2\mathrm{e}^{sX_i}\right]}{\mathbb{E}\left[\mathrm{e}^{sX_i}\right]} - \left(\frac{\mathbb{E}\left[X_i\mathrm{e}^{sX_i}\right]}{\mathbb{E}\left[\mathrm{e}^{sX_i}\right]}\right)^2 = \widetilde{\mathbb{E}}_i[X^2] - (\widetilde{\mathbb{E}}_i[X])^2 = \widetilde{\mathbb{E}}_i[(X - \widetilde{\mathbb{E}}_i[X])^2] ,$$

where

$$\widetilde{\mathbb{E}}_i[Z] = \frac{\mathbb{E}\left[Z\mathrm{e}^{sX_i}\right]}{\mathbb{E}\left[\mathrm{e}^{sX_i}\right]} .$$

Then,

$$\phi_i''(s) = \inf_{x\in[a_i,b_i]} \widetilde{\mathbb{E}}_i[(X - x)^2] \leqslant \widetilde{\mathbb{E}}_i\left[\left(X - \frac{a_i + b_i}{2}\right)^2\right] \leqslant \left(\frac{b_i - a_i}{2}\right)^2 .$$

Finally, using Taylor's expansion,

$$\phi_i(s) \leq \phi_i(0) + \phi_i'(0) + \frac{s^2}{2}\sup_{\alpha\in[0,1]} \phi_i''(\alpha s) \leq \frac{s^2(b_i - a_i)^2}{8} . \tag{7.3}$$

This implies

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) \leqslant \mathrm{e}^{-st}\mathrm{e}^{s^2\sum_{i=1}^{n}\frac{(b_i - a_i)^2}{8}} .$$

Choosing $s = 4t/(\sum_{i=1}^{n}(b_i - a_i)^2)$ minimizes the right hand side and yields (7.2).

■

Lemma 7.5 *Let X be a Bernoulli random variable. Then, for all $t > 0$,*

$$\Psi(t) = \mathbb{E}\left[\mathrm{e}^{t(X-\mathbb{E}[X])}\right] \leqslant \mathrm{e}^{t^2/8} \,.$$

PROOF. Let $p \in (0,1)$ be such that $p = \mathbb{P}(X=1)$ (cases $p=0$ and $p=1$ are straightforward). For all $t>0$,

$$\varphi(t) = \log \Psi(t) = \log\left(1-p+p\mathrm{e}^t\right) - pt \,.$$

The proof then follows from proof of the Hoeffding inequality, i.e. (7.3) with $b_i = 1-p$ and $a_i = -p$. ■

7.3 Matrix calculus

Lemma 7.6 *Let X be a random vector in $\mathbb{R}^d$ with covariance matrix Σ and A be a real matrix in $\mathbb{R}^{m\times d}$. Then, $\mathbb{V}[AX] = A\Sigma A^\top$.*

PROOF. By definition, using the linearity of the expectation,

$$\mathbb{V}[AX] = \mathbb{E}\left[(AX-\mathbb{E}[AX])(AX-\mathbb{E}[AX])^\top\right] = \mathbb{E}\left[A(X-\mathbb{E}[X])(X-\mathbb{E}[X])^\top A^\top\right] = A\mathbb{E}\left[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^\top\right]A^\top = A\Sigma A^\top \,.$$

■

Lemma 7.7 *Let X be a random vector in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and A a symmetric matrix in $\mathbb{R}^{d\times d}$. Then,*

$$\mathbb{E}[X^\top AX] = \mu^\top A\mu + \mathrm{Trace}(A\Sigma) \,.$$

PROOF. As $X^\top AX$ is a real number, $\mathbb{E}[X^\top AX] = \mathbb{E}[\mathrm{Trace}(X^\top AX)] = \mathbb{E}[\mathrm{Trace}(AXX^\top)]$. By linearity, $\mathbb{E}[X^\top AX] = \mathrm{Trace}(A\mathbb{E}[XX^\top])$ which yields,

$$\mathbb{E}[X^\top AX] = \mathrm{Trace}(A\{\mathbb{V}[X] + \mathbb{E}[X]\mathbb{E}[X]^\top\}) = \mu^\top A\mu + \mathrm{Trace}(A\Sigma) \,.$$

■

Lemma 7.8 *Let A be a $n\times d$ matrix with real entries. Then,* $\mathrm{range}(A) = \mathrm{range}(AA^\top)$.

PROOF. First note that for all $x \in \mathbb{R}^n$, $AA^\top x = 0$ implies $\langle A^\top x; A^\top x\rangle = 0$ so that $A^\top x = 0$. The converse is obvious. Therefore, $\mathrm{Ker}(AA^\top) = \mathrm{Ker}(A^\top)$. Using that for any matrix B, $\mathrm{Ker}(B^\top) = (\mathrm{range}(B))^\perp$, yields $\mathrm{range}(AA^\top)^\perp = \mathrm{range}(A)^\perp$, which concludes the proof. ■

Lemma 7.9 *Let $\{U_k\}_{1\leqslant k\leqslant r}$ be a family of r orthonormal vectors of $\mathbb{R}^d$. Then, $\sum_{k=1}^r U_k U_k^\top$ is the matrix of the orthogonal projection onto*

$$\mathbf{H} = \left\{\sum_{k=1}^r \alpha_k U_k \,;\, \alpha_1,\ldots,\alpha_r \in \mathbb{R}\right\} \,.$$

Remark 7.10 *If A is a $n\times d$ matrix with real entries such that each column of A is in* H*, then,*

$$\left(\sum_{k=1}^{r} U_k U_k^\top\right) A = A\ .$$

PROOF. For all $X \in \mathbb{R}^d$, let $\pi_{\mathbf{H}}(X)$ be the orthogonal projection of X onto $\mathbf{H}$. Since $\{U_k\}_{1\leqslant k\leqslant r}$ is an orthonormal basis of $\mathbf{H}$,

$$\pi_{\mathbf{H}}(X) = \sum_{k=1}^{r} \langle X; U_k\rangle U_k = \left(\sum_{k=1}^{r} U_k U_k^\top\right) X\ .$$

This implies in particular that for each $X \in \mathbf{H}$, $X = \left(\sum_{k=1}^{r} U_k U_k^\top\right) X$. ■

Proposition 7.11 (Singular value decomposition) *For all $\mathbb{R}^{n\times d}$ matrix A with rank r, there exist $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ such that*

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^\top\ ,$$

where $\{u_1,\ldots,u_r\} \in (\mathbb{R}^n)^r$ and $\{v_1,\ldots,v_r\} \in (\mathbb{R}^d)^r$ are two orthonormal families. The vectors $\{\sigma_1,\ldots,\sigma_r\}$ are called singular values of A and $\{u_1,\ldots,u_r\}$ (resp. $\{v_1,\ldots,v_r\}$) are the left-singular (resp. right-singular) vectors of A.

Remark 7.12 *If U denotes the $\mathbb{R}^{n\times r}$ matrix with columns given by $\{u_1,\ldots,u_r\}$ and V denotes the $\mathbb{R}^{p\times r}$ matrix with columns given by $\{v_1,\ldots,v_r\}$, then the singular value decomposition of A may also be written as*

$$A = UD_rV^\top\ ,$$

where $D_r = \mathrm{diag}(\sigma_1,\ldots,\sigma_r)$.

Remark 7.13 *The singular value decomposition is closely related to the spectral theorem for symmetric semipositive definite matrices. In the framework of Proposition 7.11, $A^\top A$ and $AA^\top$ are positive semidefinite such that*

$$A^\top A = VD_r^2V^\top \quad \text{and} \quad AA^\top = UD_r^2U^\top\ .$$

PROOF. Since the matrix $AA^\top$ is positive semidefinite, its spectral decomposition is given by

$$AA^\top = \sum_{k=1}^{r} \lambda_k u_k u_k^\top\ ,$$

where $\lambda_1 \geqslant \ldots \geqslant \lambda_r > 0$ are the nonzero eigenvalues of $AA^\top$ and $\{u_1,\ldots,u_r\}$ is an orthonormal family of $\mathbb{R}^n$. For all $1 \leqslant k \leqslant r$, define $v_k = \lambda_k^{-1/2} A^\top u_k$ so that

$$\|v_k\|^2 = \lambda_k^{-1}\langle A^\top u_k; A^\top u_k\rangle = \lambda_k^{-1} u_k^\top AA^\top u_k = 1\ ,$$
$$A^\top A v_k = \lambda_k^{-1/2} A^\top AA^\top u_k = \lambda_k v_k\ .$$

On the other hand, for all $1 \leqslant k \neq j \leqslant r$, $\langle v_k; v_j\rangle = \lambda_k^{-1/2}\lambda_j^{-1/2} u_k^\top AA^\top u_j = \lambda_k^{-1/2}\lambda_j^{1/2} u_k^\top u_j = 0$. Therefore, $\{v_1,\ldots,v_r\}$ is an orthonormal family of eigenvectors of $A^\top A$ associated with the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_r > 0$. Define, for all $1 \leqslant k \leqslant r$, $\sigma_k = \lambda_k^{1/2}$ which yields

$$\sum_{k=1}^{r} \sigma_k u_k v_k^\top = \sum_{k=1}^{r} u_k u_k^\top A = \left(\sum_{k=1}^{r} u_k u_k^\top\right) A\ .$$

As $\{u_1,\ldots,u_r\}$ is an orthonormal family, by Lemma 7.9$UU^\top = \sum_{k=1}^{r} u_k u_k^\top$ is the orthogonal projection onto the range$(AA^\top)$. And, by Lemma 7.8, range$(AA^\top) =$ range(A), which implies

$$\sum_{k=1}^{r} \sigma_k u_k v_k^\top = \left(\sum_{k=1}^{r} u_k u_k^\top\right) A = A\ .$$

■

Let M_d^+ the space of real-valued $d \times d$ symmetric positive matrices.

Lemma 7.14 *The function $\Sigma \mapsto \log\det\Sigma$ is concave on M_d^+.*

Proof. Let $\Sigma, \Gamma \in M_d^+$ and $\lambda \in [0,1]$. Since $\Sigma^{-1/2}\Gamma\Sigma^{-1/2} \in M_d^+$, it is diagonalisable in some orthonormal basis and write $\mu_1, \ldots, \mu_d$ the (possibly repeated) entries of the diagonal. Note in particular that $\det\left(\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\right) = \prod_{i=1}^d \mu_i$. Then,

$$\begin{aligned}\log\det((1-\lambda)\Sigma + \lambda\Gamma) &= \log\det\left[\Sigma^{1/2}\left((1-\lambda)I + \lambda\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\right)\Sigma^{1/2}\right] \\ &= \log\det\Sigma + \log\det\left((1-\lambda)I + \lambda\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\right) \\ &= \log\det\Sigma + \sum_{i=1}^d \log(1-\lambda+\lambda\mu_i) \\ &\geq \log\det\Sigma + \sum_{i=1}^d (1-\lambda)\log(1) + \lambda\log(\mu_i) := D\end{aligned}$$

where the last inequality follows from the concavity of the log. Now, rewrite the rhs D as:

$$\begin{aligned}D &= (1-\lambda)\log\det\Sigma + \lambda\left(\log\det\Sigma^{1/2} + \log\det\Sigma^{-1/2}\Gamma\Sigma^{-1/2} + \log\det\Sigma^{1/2}\right) \\ &= (1-\lambda)\log\det\Sigma + \lambda\log\det\Gamma\end{aligned}$$

This finishes the proof. ■

Lemma 7.15 *Let Σ be a symmetric and invertible matrix in $\mathbb{R}^{d\times d}$.*

(i) The derivative of the real valued function $\Sigma \mapsto \log\det(\Sigma)$ defined on $\mathbb{R}^{d\times d}$ is given by:

$$\partial_\Sigma\{\log\det(\Sigma)\} = \Sigma^{-1} ,$$

where, for all real valued function f defined on $\mathbb{R}^{d\times d}$, $\partial_\Sigma f(\Sigma)$ denotes the $\mathbb{R}^{d\times d}$ matrix such that for all $1 \leqslant i, j \leqslant d$, $\{\partial_\Sigma f(\Sigma)\}_{i,j}$ is the partial derivative of f with respect to $\Sigma_{i,j}$.

(ii) The derivative of the real valued function $x \mapsto x^\top \Sigma x$ defined on $\mathbb{R}^d$ is given by:

$$\partial_x\{x^\top\Sigma x\} = 2\Sigma x .$$

Proof.

(i) Recall that for all $i \in \{1, \ldots, d\}$ we have $\det(\Sigma) = \sum_{k=1}^d \Sigma_{i,k}\Delta_{i,k}$ where $\Delta_{i,j}$ is the (i,j)-cofactor associated to Σ. For any fixed i, j, the component $\Sigma_{i,j}$ does not appear in anywhere in the decomposition $\sum_{k=1}^d \Sigma_{i,k}\Delta_{i,k}$, except for the term $k = j$. This implies

$$\frac{\partial\log\det(\Sigma)}{\partial\Sigma_{i,j}} = \frac{1}{\det\Sigma}\frac{\partial\det(\Sigma)}{\partial\Sigma_{i,j}} = \frac{\Delta_{i,j}}{\det\Sigma}$$

Recalling the identity $\Sigma\ [\Delta_{j,i}]_{1\leq i,j\leq d} = (\det\Sigma)\ I_d$ so that $\Sigma^{-1} = \frac{[\Delta_{j,i}]^\top_{1\leq i,j\leq d}}{\det\Sigma}$, we finally get

$$\left[\frac{\partial\log\det(\Sigma)}{\partial\Sigma_{i,j}}\right]_{1\leq i,j\leq d} = (\Sigma^{-1})^\top = \Sigma^{-1}$$

where the last equality follows from the fact that Σ is symmetric.

(ii) Define $\varphi(x) = x^\top\Sigma x$. Then, by straightforward algebra, $\varphi(x+h) = \varphi(x) + 2h^\top\Sigma x + \varphi(h) = \varphi(x) + 2h^\top\Sigma x + o(\|h\|)$, which concludes the proof.

■

Chapter 8

Exercices

Contents

8.1 Order of a kernel

Let $(\phi_k)_{k\geqslant 0}$ be a family of polynomials such that

- for all $k \geqslant 0$, ϕ_k has degree k ;
- for all $k,k' \geqslant 0$, $\int_{-1}^{1} \phi_k(u)\phi_{k'}(u)\mathrm{d}u = \delta_{k,k'}$.

1. Write ϕ_0, ϕ_1 and ϕ_2.

   It is straightforward to show that $\phi_0 : x \mapsto 1/\sqrt{2}$, $\phi_1 : x \mapsto \sqrt{3/2}x$ *and* $\phi_2 : x \mapsto \sqrt{5/8}(3x^2-1)$.
2. Show that for all $\ell \geqslant 0$, $K : u \mapsto \sum_{m=0}^{\ell} \phi_m(0)\phi_m(u)\mathbb{1}_{[-1,1]}(u)$ is a kernel of order ℓ.

   The $(\phi_k)_{k\geqslant 0}$ *provide an orthonormal basis of* $\mathrm{L}^2([-1,1])$. *For all* $0 \leqslant j \leqslant \ell$, *there exist* $(a_{j,k})_{0\leqslant k\leqslant j}$ *such that* $X^j = \sum_{k=0}^{j} a_{j,k}\phi_k(X)$. *Then,*

$$\int u^j K(u)\mathrm{d}u = \sum_{k=0}^{j}\sum_{m=0}^{\ell}\int \phi_m(0)\phi_m(u)\mathbb{1}_{[-1,1]}(u)a_{j,k}\phi_k(u)\mathrm{d}u = \sum_{k=0}^{j} a_{j,k}\phi_k(0) = 0^j\ .$$

8.2 Estimate of the integratred mean square error for bandwith selection

Let $(X_1 \dots, X_n)$ be i.i.d. random variables with probability density function $p_\star$ with respect to the Lebesgue measure. For all $h > 0$, an unbiased estimator J_n^h of $\overline{\mathscr{E}} - \int_{\mathbb{R}} (p_\star(x))^2 \mathrm{d}x$ is given by:

$$J_n^h = \int_{\mathbb{R}} p_\star^2(x)\mathrm{d}x + \int_{\mathbb{R}} (\widehat{p}_n^h)^2(x)\mathrm{d}x - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j\neq i}^n K\left(\frac{X_i - X_j}{h}\right) ,$$

where

$$\widehat{p}_n^h : x \mapsto \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) .$$

Evaluate J_n^h with the Gaussian kernel $K : u \mapsto (2\pi)^{-1/2} \mathrm{e}^{-u^2/2}$.

It is enough to compute $\int_{\mathbb{R}} (\widehat{p}_n^h)^2(x)\mathrm{d}x$. *In the case of the Gaussian kernel,*

$$\int_{\mathbb{R}} (\widehat{p}_n^h)^2(x)\mathrm{d}x = \frac{1}{n^2h^2} \sum_{i=1}^n \int_{\mathbb{R}} K\left(\frac{X_i - x}{h}\right)^2 \mathrm{d}x + \frac{1}{n^2h^2} \sum_{i=1}^n \sum_{i=1\,;j\neq i}^n \int_{\mathbb{R}} K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right) \mathrm{d}x .$$

For all $1 \leqslant i \leqslant n$,

$$\int_{\mathbb{R}} K\left(\frac{X_i - x}{h}\right)^2 \mathrm{d}x = \frac{1}{2\pi} \int \mathrm{e}^{-(X_i - x)^2/h^2} \mathrm{d}x = \frac{h}{2\sqrt{\pi}} \frac{1}{\sqrt{2\pi(h/\sqrt{2})^2}} \int \mathrm{e}^{-(X_i - x)^2/2(h/\sqrt{2})^2} \mathrm{d}x = \frac{h}{2\sqrt{\pi}} .$$

For all $1 \leqslant i \neq j \leqslant n$,

$$\begin{aligned}\int_{\mathbb{R}} K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right) \mathrm{d}x &= \frac{1}{2\pi} \int \mathrm{e}^{-(X_i - x)^2/2h^2 - (X_j - x)^2/2h^2} \mathrm{d}x ,\\ &= \frac{1}{2\pi} \mathrm{e}^{-X_i^2/2h^2} \mathrm{e}^{-X_j^2/2h^2} \mathrm{e}^{(X_i + X_j)^2/4h^2} \int \mathrm{e}^{-(x - (X_i + X_j)/2)^2/h^2} \mathrm{d}x ,\\ &= \frac{h}{2\sqrt{\pi}} \mathrm{e}^{-(X_i - X_j)^2/4h^2} ,\end{aligned}$$

which concludes the proof.

8.3 Moving window based kernel density estimate

Let $(X_1 \dots, X_n)$ be i.i.d. random variables with probability density function $p_\star$ with respect to the Lebesgue measure. The unknown density $p_\star$ is estimated by:

$$\widehat{p}_n^{h_n} : x \mapsto \frac{1}{2nh_n} \sum_{i=1}^n \mathbb{1}_{(x - h_n, x + h_n)}(X_i) .$$

1. Write the bias-variance decompostion for this estimator.

   For all $x \in \mathbb{R}$,

$$\mathbb{E}\left[\left|\widehat{p}_n^{h_n}(x) - p_\star(x)\right|^2\right] = \left(\mathbb{E}\left[\widehat{p}_n^{h_n}(x)\right] - p_\star(x)\right)^2 + \mathbb{V}\left[\widehat{p}_n^{h_n}(x)\right] .$$

2. Show that when $h_n \to 0$ and $nh_n \to +\infty$ then for almost all x, $\widehat{p}_n^{h_n}(x)$ converges in L^2 (and therefore in probability) to $p_\star(x)$.

   For all $x \in \mathbb{R}$,

$$\widehat{p}_n^{h_n}(x) = \frac{1}{2nh_n} Z_{j,n}(x) \ ,$$

where $Z_{j,n}(x)$ has a binomial distribution with parameters n and $\tilde{q}_n(x) = F(x+h_n) - F(x-h_n)$, F being the distribution function of X_1. Therefore,

$$\mathbb{V}\left[\widehat{p}_n^{h_n}(x)\right] = \frac{1}{4n^2h_n^2} n\tilde{q}_n(x)(1-\tilde{q}_n(x)) \leqslant \frac{1}{2nh_n}\frac{\tilde{q}_n(x)}{2h_n}$$

and the variance of the estimate converges to 0 for almost all x as $\tilde{q}_n(x)/(2h_n)$ converges to $p_\star(x)$. In addition,

$$\mathbb{E}\left[\widehat{p}_n^{h_n}(x)\right] = \frac{1}{2nh_n} n\tilde{q}_n(x)$$

and the bias converges to 0 for almost all x. The bias-variance decomposition completes the proof.

3. Let $(S_n)_{n\geqslant 1}$ be a sequence of i.i.d. random variables with binomial distribution with parameters n and p_n with $np_n \to +\infty$ et $\limsup p_n < 1$. For all $t \in \mathbb{R}$, noting that

$$p_n \frac{t^2}{np_n(1-p_n)} = O\left(\frac{1}{n}\right) \quad \text{et} \quad p_n^2 \frac{t^2}{np_n(1-p_n)} = O\left(\frac{1}{n}\right) \ ,$$

provide the asymptotic behavior of $\mathbb{E}[\mathrm{e}^{itZ_n}]$ where

$$Z_n = \frac{S_n - np_n}{\sqrt{np_n(1-p_n)}} \ .$$

For all $t \in \mathbb{R}$,

$$\mathbb{E}[\mathrm{e}^{itZ_n}] = \mathrm{e}^{-itnp_n/\sqrt{np_n(1-p_n)}} \mathbb{E}\left[\mathrm{e}^{-itS_n/\sqrt{np_n(1-p_n)}}\right]$$

and, using that S_n is the sum of independent Bernoulli random variables,

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{itZ_n}] &= \mathrm{e}^{-itnp_n/\sqrt{np_n(1-p_n)}} \left(1 - p_n + p_n \mathrm{e}^{-it/\sqrt{np_n(1-p_n)}}\right)^n \ ,\\
&= \left(1 - \frac{itp_n}{\sqrt{np_n(1-p_n)}} - \frac{t^2p_n^2}{2np_n(1-p_n)} + O\left(\frac{1}{n}\right)\right)^n \\
&\qquad\qquad \times \left(1 - \frac{itp_n}{\sqrt{np_n(1-p_n)}} - \frac{t^2p_n}{2np_n(1-p_n)} + O\left(\frac{1}{n}\right)\right)^n \ ,\\
&= \left(1 - \frac{t^2p_n^2}{np_n(1-p_n)} - \frac{t^2p_n^2}{2np_n(1-p_n)} - \frac{t^2p_n}{2np_n(1-p_n)} + O\left(\frac{1}{n}\right)\right)^n \ ,\\
&= \left(1 - \frac{t^2}{2n} + O\left(\frac{1}{n}\right)\right)^n \ .
\end{aligned}$$

4. Show that $(Z_n)_{n\geqslant 1}$ converges in distribution to $\mathcal{N}(0,1)$.

By the previous question, for all $t \in \mathbb{R}$,

$$\mathbb{E}[\mathrm{e}^{itZ_n}] \underset{n\to+\infty}{\longrightarrow} \mathrm{e}^{-t^2/2} \ ,$$

which completes the proof.

5. Let $x \in \mathbb{R}$ be such that $p_\star(x) > 0$ and such that $p_\star$ is differentiable on a neighborhood $V(x)$ of x, with a bounded derivative on $V(x)$. Show that if $nh_n \to +\infty$ and $nh_n^3 \to 0$, then

$$\sqrt{\frac{2nh_n}{\widehat{p}_n^{h_n}(x)}}\left(\widehat{p}_n^{h_n}(x)-p_\star(x)\right)\Rightarrow\mathscr{N}(0,1)\,.$$

By the previous lemma,

$$\frac{2nh_n}{\sqrt{n\tilde{q}_n(x)(1-\tilde{q}_n(x))}}\left(\widehat{p}_n^{h_n}(x)-\mathbb{E}\left[\widehat{p}_n^{h_n}(x)\right]\right)\Rightarrow\mathscr{N}(0,1)\,.$$

By Slutsky's lemma,

$$\sqrt{\frac{2nh_n}{p_\star(x)}}\left(\widehat{p}_n^{h_n}(x)-\mathbb{E}\left[\widehat{p}_n^{h_n}(x)\right]\right)\Rightarrow\mathscr{N}(0,1)\,.$$

Then,

$$\left|\mathbb{E}\left[\widehat{p}_n^{h_n}(x)\right]-p_\star(x)\right|=\left|\frac{1}{2h_n}\int_{x-h_n}^{x+h_n}p_\star(u)\mathrm{d}u-p_\star(x)\right|\leqslant\frac{1}{2h_n}\int_{x-h_n}^{x+h_n}|p_\star(u)-p_\star(x)|\,\mathrm{d}u$$

For all sufficiently large n, $(x-h_n,x+h_n)$ is included in the neighborhood of x on which $p_\star$ has a bounded derivative. Write M_x the bound of $p'_\star$ on this neighborhood. By the meanvalue theorem,

$$\left|\mathbb{E}\left[\widehat{p}_n^{h_n}(x)\right]-p_\star(x)\right|\leqslant\frac{M_xh_n}{2}\,.$$

Puisque $nh_n^3\to+\infty$,

$$\sqrt{\frac{2nh_n}{p_\star(x)}}\left(\widehat{p}_n^{h_n}(x)-p_\star(x)\right)\Rightarrow\mathscr{N}(0,1)\,.$$

The proof is completed with Sutsky's lemma.

8.4 Linear discriminant analysis

Linear discriminant analysis assumes that the random variables $(X,Y)\in\mathbb{R}^p\times\{0,1\}$ has the following distribution. For all $A\in\mathscr{B}(\mathbb{R}^p)$ and all $y\in\{0,1\}$,

$$\mathbb{P}(X\in A;Y=y)=\pi_y\int_A g_y(x)\mathrm{d}x\,,$$

where π_0 and π_1 are positive real numbers such that $\pi_0+\pi_1=1$ and g_0 (resp. g_1) is the probability density of a Gaussian random variable with mean $\mu_0\in\mathbb{R}^d$ (resp. μ_1) and positive definite covariance matrix $\Sigma_0\in\mathbb{R}^{d\times d}$ (resp. Σ_1). The Bayes classifier $h_*:\mathbb{R}^p\to\{0,1\}$ is defined by

$$h_*:x\mapsto\mathbb{1}_{\{\pi_1g_1(x)>\pi_0g_0(x)\}}\,.$$

1. Give the distribution of the random variable X and prove that

$$\mathbb{P}(h_*(X)\neq Y)=\min_{h:\mathbb{R}^p\to\{0,1\}}\{\mathbb{P}(h(X)\neq Y)\}\,.$$

For all $A\in\mathscr{B}(\mathbb{R}^p)$,

$$\begin{aligned}\mathbb{P}(X\in A)&=\mathbb{P}(Y=0)\mathbb{P}(X\in A|Y=0)+\mathbb{P}(Y=1)\mathbb{P}(X\in A|Y=1)\,,\\&=\pi_0\int_A g_0(x)\mathrm{d}x+\pi_1\int_A g_1(x)\mathrm{d}x\,.\end{aligned}$$

The probability density of the random variable X is given, for all $x \in \mathbb{R}^d$, by

$$g(x) = \pi_0 g_0(x) + \pi_1 g_1(x) \,.$$

Then, note that for all $x \in \mathbb{R}^d$,

$$\eta(x) = \mathbb{P}(Y=1|X)|_{X=x} = \frac{\mathbb{P}(X|Y=1)|_{X=x}\,\mathbb{P}(Y=1)}{g(x)} = \frac{\pi_1 g_1(x)}{\pi_0 g_0(x) + \pi_1 g_1(x)} \,,$$

and the condition $\eta(x) \leqslant 1/2$ can be rewritten as

$$\frac{\pi_1 g_1(x)}{\pi_0 g_0(x) + \pi_1 g_1(x)} \leqslant 1/2 \,,$$

that is $\pi_1 g_1(x) \leqslant \pi_0 g_0(x)$.

2. Assume that $\mu_0 \neq \mu_1$. Prove that when $\Sigma_0 = \Sigma_1 = \Sigma$, for all $x \in \mathbb{R}^p$,

$$h_*(x) = 1 \Leftrightarrow (\mu_1 - \mu_0)^T \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_0}{2} \right) > \log(\pi_0/\pi_1) \,.$$

Provide a geometrical interpretation.

For all $x \in \mathbb{R}^d$,

$$\begin{aligned}
\pi_1 g_1(x) > \pi_0 g_0(x) &\Leftrightarrow \log(\pi_1 g_1(x)) > \log(\pi_0 g_0(x)) \,,\\
&\Leftrightarrow -\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0) > \log(\pi_0/\pi_1) \,,\\
&\Leftrightarrow -\frac{1}{2}\Big(-\mu_1^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1}\mu_1 - x^T \Sigma^{-1}\mu_1 + \mu_0^T \Sigma^{-1} x - \mu_0^T \Sigma^{-1}\mu_0 + x^T \Sigma^{-1}\mu_0 \Big) > \log(\pi_0/\pi_1) \,,\\
&\Leftrightarrow x^T \Sigma^{-1}\mu_1 - x^T \Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T \Sigma^{-1}\mu_0 > \log(\pi_0/\pi_1) \,,\\
&\Leftrightarrow (\mu_1 - \mu_0)^T \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_0}{2} \right) > \log(\pi_0/\pi_1) \,.
\end{aligned}$$

Therefore, all $x \in \mathbb{R}^d$ is classified according to its position with respect to an affine hyperplane orthogonal to $\Sigma^{-1}(\mu_1 - \mu_{-1})$.

3. Prove that when $\pi_1 = \pi_0$,

$$\mathbb{P}(h_*(X) = 1|Y=0) = \Phi(-d(\mu_1,\mu_0)/2) \,,$$

where Φ is the cumulative distribution function of a standard Gaussian random variable and

$$d(\mu_1,\mu_0)^2 = (\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0) \,.$$

Let Z_0 be a Gaussian random variable with mean μ_0 and variance Σ. Note that

$$\mathbb{P}(h_*(X) = 1|Y=0) = \mathbb{P}\Big(\underbrace{(\mu_1 - \mu_0)^T \Sigma^{-1}(Z_0 - \frac{\mu_1 + \mu_0}{2})}_{Z} > 0 \Big) \,,$$

where, using $\delta = d(\mu_1,\mu_0)$,

$$\mathbb{E}[Z] = (\mu_1 - \mu_0)^T \Sigma^{-1}(\frac{\mu_0 - \mu_1}{2}) = -\frac{\delta^2}{2}$$

and

$$\mathbb{V}[Z] = \mathbb{V}\left[(\mu_1 - \mu_0)^T \Sigma^{-1} X\right] = \left((\mu_1 - \mu_0)^T \Sigma^{-1}\right) \Sigma \left(\Sigma^{-1}(\mu_1 - \mu_0)\right) = \delta^2.$$

Hence,

$$\mathbb{P}(h_*(X) = 1 | Y = 0) = \mathbb{P}\Big(-\frac{\delta^2}{2} + \delta\varepsilon > 0 \Big) = \mathbb{P}\Big(\varepsilon > \frac{\delta}{2} \Big) = \Phi\Big(-\frac{\delta}{2} \Big).$$

4. When $\Sigma_1 \neq \Sigma_0$, what is the nature of the frontier between $\{x\,;\, h_*(x) = 1\}$ and $\{x\,;\, h_*(x) = 0\}$?

   In this case, for all $x \in \mathbb{R}^d$,

$$\begin{aligned}
\pi_1 g_1(x) > \pi_0 g_0(x) &\Leftrightarrow \log(\pi_1 g_1(x)) > \log(\pi_0 g_0(x))\,, \\
&\Leftrightarrow -\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) > \log(\pi_0/\pi_1)\,, \\
&\Leftrightarrow \frac{1}{2}x'\Sigma_0^{-1}x - \frac{1}{2}x'\Sigma_1^{-1}x + x^T\Sigma_1^{-1}\mu_1 - x^T\Sigma_0^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 > \log(\pi_0/\pi_1)\,.
\end{aligned}$$

   As the quadratic term does not vanish anymore, the frontier between $\{x\,;\, h_*(x) = 1\}$ and $\{x\,;\, h_*(x) = 0\}$ is a quadric.

8.5 Plug-in classifier

Let (X,Y) be random variables defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. For any classifier $h : \mathscr{X} \to \{-1,1\}$, define its classification error by

$$R(h) = \mathbb{P}(Y \neq h(X))\,.$$

The best classifier in terms of the classification error R is the Bayes classifier

$$h_*(x) = \mathrm{sign}(\eta(x) - 1/2)\,,$$

where

$$\eta : x \mapsto \mathbb{P}(Y = 1|X)_{|X=x}\,.\,.$$

Given n independent couples $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ with the same distribution as (X,Y), an empirical surrogate for h_* is obtained from a possibly nonparametric estimator $\widehat{\eta}_n$ of η:

$$\widehat{h}_n : x \mapsto \mathrm{sign}(\widehat{\eta}_n(x) - 1/2)\,.$$

1. Prove that for any classifier $h : \mathscr{X} \to \{-1,1\}$,

$$\mathbb{P}(Y \neq h(X)|X) = (2\eta(X) - 1)\mathbb{1}_{h(X)=-1} + 1 - \eta(X)$$

   and

$$R(h) - R(h_*) = 2\mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}_{h(X)\neq h_*(X)}\right]\,.$$

   For all $x \in \mathscr{X}$,

$$\begin{aligned}
\mathbb{P}(Y \neq h(X)|X) &= \mathbb{P}(Y = -1, h(X) = 1|X) + \mathbb{P}(Y = 1, h(X) = -1|X)\,, \\
&= \mathbb{1}_{h(X)=1}\mathbb{P}(Y = -1|X) + \mathbb{1}_{h(X)=-1}\mathbb{P}(Y = 1|X)\,, \\
&= \mathbb{1}_{h(X)=-1}(2\eta(X) - 1) + 1 - \eta(X)\,.
\end{aligned}$$

   On the other hand,

$$
\begin{aligned}
R(h)-R(h_*) &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{P}(Y\neq h(X)|X)-\mathbb{P}(Y\neq h_*(X)|X)|X\right]\right] ,\\
&= \mathbb{E}\left[(2\eta(X)-1)\left(\mathbb{1}_{h(X)=-1}-\mathbb{1}_{h_*(X)=-1}\right)\middle|X\right] ,\\
&= \mathbb{E}\left[\mathbb{1}_{h_*(X)\neq h(X)}\left((2\eta(X)-1)\mathbb{1}_{h_*(X)=1}-(2\eta(X)-1)\mathbb{1}_{h_*(X)=-1}\right)\middle|X\right] ,\\
&= 2\mathbb{E}\left[|\eta(X)-1/2|\,\mathbb{1}_{h_*(X)\neq h(X)}\right] .
\end{aligned}
$$

2. Prove that
$$|\eta(x)-1/2|\mathbb{1}_{\widehat{h}_n(x)\neq h_*(x)} \leqslant |\eta(x)-\widehat{\eta}_n(x)|\mathbb{1}_{\widehat{h}_n(x)\neq h_*(x)} ,$$
where
$$\widehat{h}_n : x \mapsto \mathrm{sign}(\widehat{\eta}_n(x)-1/2) .$$
Deduce that
$$R(\widehat{h}_n)-R(h_*) \leqslant 2\|\widehat{\eta}_n-\eta\|_{\mathrm{L}^2(\mathbb{P}_X)} ,$$
where $\mathbb{P}_X$ is the distribution of X.

Note that
$$\{x\in\mathscr{X}\,;\,\widehat{h}_n(x)\neq h_*(x)\} = \{x\in\mathscr{X}\,;\,\eta(x)\geqslant 1/2\,,\,\widehat{\eta}_n(x)\leqslant 1/2\}\cup\{x\in\mathscr{X}\,;\,\eta(x)\leqslant 1/2\,,\,\widehat{\eta}_n(x)\geqslant 1/2\} .$$
For all $x\in\{x\in\mathscr{X}\;;\;\eta(x)\geqslant 1/2\;,\;\widehat{\eta}_n(x)\leqslant 1/2\}$,
$$|\eta(x)-\widehat{\eta}_n(x)| = \eta(x)-\widehat{\eta}_n(x) \geqslant \eta(x)-1/2$$
On the other hand, for all $x\in\{x\in\mathscr{X}\;;\;\eta(x)\leqslant 1/2\;,\;\widehat{\eta}_n(x)\geqslant 1/2\}$,
$$|\eta(x)-\widehat{\eta}_n(x)| = \widehat{\eta}_n(x)-\eta(x) \geqslant 1/2-\eta(x) .$$
Therefore, for all $x\in\mathscr{X}$,
$$|\eta(x)-1/2|\mathbb{1}_{\widehat{h}_n(x)\neq h_*(x)} \leq |\eta(x)-\widehat{\eta}_n(x)|\mathbb{1}_{\widehat{h}_n(x)\neq h_*(x)}.$$
By the first question and Cauchy-Schwarz inequality,
$$R(\widehat{h}_n)-R(h_*) = 2\mathbb{E}\left[|\eta(X)-1/2|\,\mathbb{1}_{h_*(X)=\widehat{h}_n(X)}\right] \leqslant 2\mathbb{E}\left[|\eta(X)-\widehat{\eta}_n(X)|\,\mathbb{1}_{\widehat{h}_n(X)\neq h_*(X)}\right] \leqslant 2\|\eta-\widehat{\eta}_n\|_{\mathrm{L}^2(\mathbb{P}_X)} .$$

8.6 Logistic Regression

The *logistic model* assumes that the random variables $(X,Y)\in\mathbb{R}^p\times\{0,1\}$ are such that

$$\mathbb{P}(Y=1|X) = \frac{\exp\left(\langle\beta^*,X\rangle\right)}{1+\exp\left(\langle\beta^*,X\rangle\right)} , \tag{8.1}$$

with $\beta^*\in\mathbb{R}^d$. In this case, for all $x\in\mathbb{R}^d$, $\mathbb{P}(Y=1|X)|_{X=x}>1/2$ if and only if $\langle\beta^*,x\rangle>0$, so the frontier between $\{x\,;\,h_*(x)=1\}$ and $\{x\,;\,h_*(x)=0\}$ is an hyperplane, with orthogonal direction β^*. The unknown parameter β^* may be estimated by maximizing the conditional likelihood of Y given X

$$\widehat{\beta}_n \in \mathrm{argmax}_{\beta\in\mathbb{R}^d}\prod_{i=1}^n\left[\left(\frac{\exp\left(\langle\beta,x_i\rangle\right)}{1+\exp\left(\langle\beta,x_i\rangle\right)}\right)^{Y_i}\left(\frac{1}{1+\exp\left(\langle\beta,x_i\rangle\right)}\right)^{1-Y_i}\right],$$

to define the empirical classifier
$$\widehat{h}_n : x \mapsto \mathbb{1}_{\langle\widehat{\beta}_n,x\rangle>0} .$$

1. Compute the gradient and the Hessian H_n of

$$\ell_n : \beta \mapsto -\sum_{i=1}^{n} \left[Y_i \langle x_i, \beta \rangle - \log(1+\exp(\langle x_i, \beta \rangle)) \right] \ .$$

What can be said about the function ℓ_n when for all $\beta \in \mathbb{R}^d$, $H_n(\beta)$ is nonsingular? This assumption is supposed to hold in the following questions.

Since for all $u \in \mathbb{R}^d$, $\nabla_\beta \langle u, \beta \rangle = u$,

$$\nabla \ell_n(\beta) = -\sum_{i=1}^{n} Y_i x_i + \sum_{i=1}^{n} \frac{\exp(\langle x_i, \beta \rangle)}{1+\exp(\langle x_i, \beta \rangle)} x_i \ .$$

On the other hand, for all $1 \leqslant i \leqslant n$ and all $1 \leqslant j \leqslant d$,

$$\partial_j \left(\frac{\exp(\langle x_i, \beta \rangle)}{1+\exp(\langle x_i, \beta \rangle)} x_i \right) = \frac{\exp(\langle x_i, \beta \rangle)}{(1+\exp(\langle x_i, \beta \rangle))^2} x_{ij} x_i \ ,$$

where x_{ij} is the jth component of x_i. Then

$$\left(H_n(\beta)\right)_{\ell j} = \sum_{i=1}^{n} \frac{\exp(\langle x_i, \beta \rangle)}{(1+\exp(\langle x_i, \beta \rangle))^2} x_{ij} x_{i\ell} \ ,$$

that is,

$$H_n(\beta) = \sum_{i=1}^{n} \frac{\exp(\langle x_i, \beta \rangle)}{(1+\exp(\langle x_i, \beta \rangle))^2} x_i x_i' \ .$$

$H_n(\beta)$ is a semi positive definite matrix, which implies that $\ell_n(\beta)$ is convex. If we assume that H_n is nonsingular, ℓ_n is strictly convex.

2. Prove that there exists $\widetilde{\beta}_n \in \mathbb{R}^d$ such that $\|\widetilde{\beta}_n - \beta^*\| \leq \|\widehat{\beta}_n - \beta^*\|$ and

$$\widehat{\beta}_n - \beta^* = -H_n(\widetilde{\beta}_n)^{-1} \nabla \ell_n(\beta^*) \ .$$

Using a Taylor expansion between $\beta^\star$ and $\widehat{\beta}_n$, there exists $\tilde{\beta}_n \in B(\beta^\star, \|\widehat{\beta}_n - \beta^\star\|)$ such that

$$\nabla \ell_n(\widehat{\beta}_n) = \nabla \ell_n(\beta^\star) + H_n(\tilde{\beta}_n)(\hat{\beta}_n - \beta^\star) \ .$$

By definition, $\ell_n(\widehat{\beta}_n) = 0$. Therefore,

$$\widehat{\beta}_n - \beta^\star = -H_n(\tilde{\beta}_n)^{-1} \nabla \ell_n(\beta^\star) \ ,$$

where $H_n(\tilde{\beta}_n)^{-1}$ exists since $H_n(\tilde{\beta})$ is assumed to be non-singular for all β.

In the following it is assumed that the $(x_i)_{1 \leqslant i \leqslant n}$ are uniformly bounded, $\widehat{\beta}_n \to \beta^*$ a.s. and that there exists a continuous and nonsingular function H such that $n^{-1} H_n(\beta)$ converges to $H(\beta)$, uniformly in a ball around β^*.

3. Define for all $1 \leqslant i \leqslant n$, $p_i(\beta) = e^{\langle x_i, \beta \rangle} / \left(1 + e^{\langle x_i, \beta \rangle}\right)$. Check that

$$\begin{aligned} \mathbb{E}\left[e^{-n^{-1/2} \langle t, \nabla \ell_n(\beta^*) \rangle} \right] &= \prod_{i=1}^{n} \left(1 - p_i(\beta^*) + p_i(\beta^*) e^{\langle t, x_i \rangle / \sqrt{n}} \right) e^{-p_i(\beta^*) \langle t, x_i \rangle / \sqrt{n}} \ , \\ &= \exp\left(\frac{1}{2} t^T \left(n^{-1} H_n(\beta^*) \right) t + O(n^{-1/2}) \right) \ . \end{aligned}$$

For all $t \in \mathbb{R}^d$,

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}}\langle t,\nabla\ell_n(\beta^\star)\rangle\right)\right] &= \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{1}{\sqrt{n}}(Y_i-p_i(\beta^\star))\langle x_i,t\rangle\right)\right]\,,\\
&= \prod_{i=1}^{n}\left[\left(1-p_i(\beta^\star)+p_i(\beta^\star)\exp\left(\frac{1}{\sqrt{n}}\langle x_i,t\rangle\right)\right)\exp\left(-\frac{p_i(\beta^\star)}{\sqrt{n}}\langle x_i,t\rangle\right)\right]\,.
\end{aligned}
$$

Note that

$$
\log\left(1-p_i+p_i\exp(u/\sqrt{n})\right)=\log\left(1+p_i\frac{u}{\sqrt{n}}+p_i\frac{u^2}{2n}+O\left(n^{-3/2}\right)\right)=p_i\frac{u}{\sqrt{n}}+\frac{p_iu^2}{2n}-\frac{p_i^2u^2}{2n}+O\left(n^{-3/2}\right)\,.
$$

Finally,

$$
\mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}}\langle t,\nabla\ell_n(\beta^\star)\rangle\right)\right]=\exp\left(\frac{1}{2n}\underbrace{\sum_{i=1}^{n}p_i(\beta^\star)(1-p_i(\beta^\star))\langle t,x_i\rangle^2}_{t^TH_n(\beta^\star)t}+O(n^{-1/2})\right)\,.
$$

4. What is the asymptotic distribution of $-n^{-1/2}\nabla\ell_n(\beta^*)$ and of $\sqrt{n}(\widehat{\beta}_n-\beta^*)$?

   For all $t\in\mathbb{R}^d$, since $n^{-1}H_n(\beta^\star)\to_{n\to\infty}H(\beta^\star)$,

   $$
   \mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}}\langle t,\nabla\ell_n(\beta^\star)\rangle\right)\right]\to_{n\to\infty}\exp\left(\frac{1}{2}t^TH(\beta^\star)t\right)\,.
   $$

   Therefore, $-\nabla\ell_n(\beta^\star)/\sqrt{n}$ converges in distribution to $Z\sim\mathscr{N}(0,H(\beta^\star))$. On the other hand,

   $$
   \sqrt{n}(\widehat{\beta}_n-\beta^\star)=-\left(\frac{1}{n}H_n(\tilde{\beta}_n)\right)^{-1}\frac{1}{\sqrt{n}}\nabla\ell_n(\beta^\star)\,.
   $$

   As for all $n\geqslant 1$, $\tilde{\beta}_n\in B(\beta^\star,\|\widehat{\beta}_n-\beta^\star\|)$, $\tilde{\beta}_n$ converges to $\beta^\star$ almost surely as n grows to infinity. Hence, almost surely

   $$
   \left(\frac{1}{n}H_n(\tilde{\beta}_n)\right)^{-1}\to H(\beta^\star)^{-1}
   $$

   and, by Slutsky lemma, $\sqrt{n}(\widehat{\beta}_n-\beta^\star)$ converges in distribution to $Z\sim\mathscr{N}(0,H(\beta^\star)^{-1})$.

5. For all $1\leqslant j\leqslant d$ and all $\alpha\in(0,1)$, propose a confidence interval $\mathscr{I}_{n,\alpha}$ such that $\beta_j^*\in\mathscr{I}_{n,\alpha}$ with asymptotic probability $1-\alpha$.

   According to the last question, $\sqrt{n}(\widehat{\beta}_j-\beta_j^\star)$ converges in distribution to a centered Gaussian random variable with variance $(H(\beta^\star)^{-1})_{jj}$. On the other hand, almost surely,

   $$
   \widehat{\sigma}^2_{n,j}=(nH_n(\widehat{\beta}_n)^{-1})_{jj}\to_{n\to\infty}(H(\beta^\star)^{-1})_{jj}\,.
   $$

   Then,

   $$
   \sqrt{\frac{n}{\widehat{\sigma}^2_{n,j}}}(\widehat{\beta}_{n,j}-\beta_j^\star)\to_{n\to\infty}\mathscr{N}(0,1).
   $$

   An asymptotic confidence interval $\mathscr{I}_{n,\alpha}$ of level $1-\alpha$ is then given by

   $$
   \mathscr{I}_{n,\alpha}=\left[\widehat{\beta}_{n,j}-z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{n,j}}{n}}\;,\;\widehat{\beta}_{n,j}+z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{n,j}}{n}}\right]\,,
   $$

   where $z_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of $\mathscr{N}(0,1)$.

6. Propose a confidence ellipsoid $\mathscr{E}_{n,\alpha}$ such that the probability that $\beta^* \in \mathscr{E}_{n,\alpha}$ is asymptotically $1-\alpha$.

8.7 K-means algorithm

The K-means algorithm is a procedure which aims at partitioning a data set into K distinct, non-overlapping clusters. Consider $n \geqslant 1$ observations $(X_1,\ldots,X_n)$ taking values in $\mathbb{R}^p$. The K-means algorithm seeks to minimize over all partitions $C = (C_1,\ldots,C_K)$ of $\{1,\ldots,n\}$ the following criterion

$$\mathrm{crit}(C) = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b\in C_k} \|X_a - X_b\|^2\,,$$

where for all $1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant K$, $i \in C_k$ if and only if X_i is in the k-th cluster.

1. Define the distance between two clusters $1 \leqslant i,j \leqslant K$ as

$$d(C_i,C_j) = \sum_{a\in C_i\cup C_j} \|X_a - \bar{X}_{C_i\cup C_j}\|^2 - \sum_{a\in C_i} \|X_a - \bar{X}_{C_i}\|^2 - \sum_{a\in C_j} \|X_a - \bar{X}_{C_j}\|^2\,.$$

Prove that for all $1 \leqslant i,j \leqslant K$,

$$d(C_i,C_j) = \frac{|C_i||C_j|}{|C_i|+|C_j|}\|\bar{X}_{C_i} - \bar{X}_{C_j}\|^2\,.$$

For all $1 \leqslant i,j \leqslant K$*, note tthat*

$$\bar{X}_{C_i\cup C_j} = \frac{|C_i|}{|C_i|+|C_j|}\bar{X}_{C_i} + \frac{|C_j|}{|C_i|+|C_j|}\bar{X}_{C_j}\,,$$

so that

$$\begin{aligned}\sum_{a\in C_i} \|X_a - \bar{X}_{C_i\cup C_j}\|^2 &= \sum_{a\in C_i} \left\|X_a - \bar{X}_{C_i} + \frac{|C_j|}{|C_i|+|C_j|}(\bar{X}_{C_i} - \bar{X}_{C_j})\right\|^2\,,\\ &= \sum_{a\in C_i} \|X_a - \bar{X}_{C_i}\|^2 + 2\sum_{a\in C_i}\left\langle X_a - \bar{X}_{C_i}; \frac{|C_j|}{|C_i|+|C_j|}(\bar{X}_{C_i} - \bar{X}_{C_j})\right\rangle\\ &\qquad + |C_i|\left\|\frac{|C_j|}{|C_i|+|C_j|}(\bar{X}_{C_i} - \bar{X}_{C_j})\right\|^2\,,\\ &= \sum_{a\in C_i} \|X_a - \bar{X}_{C_i}\|^2 + \frac{|C_i||C_j|^2}{(|C_i|+|C_j|)^2}\left\|\bar{X}_{C_i} - \bar{X}_{C_j}\right\|^2\,.\end{aligned}$$

Similarly,

$$\sum_{a\in C_j} \|X_a - \bar{X}_{C_i\cup C_j}\|^2 = \sum_{a\in C_j} \left\|X_a - \bar{X}_{C_j}\right\|^2 + \frac{|C_j||C_i|^2}{(|C_i|+|C_j|)^2}\left\|\bar{X}_{C_i} - \bar{X}_{C_j}\right\|^2\,.$$

Therefore,

$$\sum_{a\in C_i\cup C_j} \|X_a - \bar{X}_{C_i\cup C_j}\|^2 = \sum_{a\in C_i} \|X_a - \bar{X}_{C_i}\|^2 + \sum_{a\in C_j} \left\|X_a - \bar{X}_{C_j}\right\|^2 + \frac{|C_i||C_j|}{|C_i|+|C_j|}\left\|\bar{X}_{C_i} - \bar{X}_{C_j}\right\|^2\,,$$

which concludes the proof.

2. Establish that

$$\mathrm{crit}(C) = 2\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}\langle X_a, X_a - X_b\rangle = 2\sum_{k=1}^{K}\sum_{a\in C_k}\|X_a - \bar{X}_{C_k}\|^2\,,$$

where

$$\bar{X}_{C_k} = \frac{1}{|C_k|}\sum_{b\in C_k} X_b\,.$$

Note that

$$\begin{aligned}
crit(C) &= \sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in G_k}\|X_a - X_b\|^2\,,\\
&= \sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}\langle X_a - X_b, X_a - X_b\rangle\,,\\
&= \sum_{k=1}^{K}\frac{1}{|C_k|}\left\{\sum_{a,b\in C_k}\langle X_a - X_b, X_a\rangle + \langle X_b - X_a, X_b\rangle\right\}\,,\\
&= 2\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}\langle X_a - X_b, X_a\rangle\,.
\end{aligned}$$

which concludes the proof of the first inequality. For the second inequality, write

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{a\in C_k}\|X_a - \bar{X}_{C_k}\|^2 &= \sum_{k=1}^{K}\sum_{a\in C_k}\langle X_a - \frac{1}{|C_k|}\sum_{b\in C_k}X_b, X_a - \frac{1}{|C_k|}\sum_{c\in C_k}X_c\rangle\,,\\
&= \sum_{k=1}^{K}\frac{1}{|C_k|^2}\sum_{a,b,c\in C_k}\langle X_a - X_b, X_a - X_c\rangle\,,\\
&= \sum_{k=1}^{K}\frac{1}{|C_k|^2}\sum_{a,b,c\in C_k}\langle X_a - X_b, X_a\rangle - \sum_{k=1}^{K}\frac{1}{|C_k|^2}\sum_{a,b,c\in C_k}\langle X_a - X_b, X_c\rangle\,,
\end{aligned}$$

where

$$\sum_{a,b,c\in C_k}\langle X_a - X_b, X_c\rangle = |C_k|\sum_{a,c\in C_k}\langle X_a, X_c\rangle - |C_k|\sum_{b,c\in C_k}\langle X_b, X_c\rangle = 0\,.$$

Thus,

$$crit(C) = 2\sum_{k=1}^{K}\sum_{a\in C_k}\|X_a - \bar{X}_{C_k}\|^2\,.$$

3. Prove that the criterion monotonically decreases with the iterations of the K-means algorithm.

   For any cluster C in and any $z\in\mathbb{R}^p$,

   $$\sum_{a\in C}\|X_a - z\|^2 = \sum_{a\in C}\|X_a - \bar{X}_C\|^2 + \sum_{a\in C}\|\bar{X}_C - z\|^2 + 2\sum_{a\in C}\langle\bar{X}_C - z; X_a - \bar{X}_C\rangle = \sum_{a\in C}\|X_a - \bar{X}_C\|^2 + |C|\|\bar{X}_C - z\|^2\,,$$

   so that

   $$\sum_{a\in C}\|X_a - z\|^2 \geqslant \sum_{a\in C}\|X_a - \bar{X}_C\|^2\,,$$

   which is enough to conclude the proof.

4. Assume that the observations are independent random variables. Define $\mu_a\in\mathbb{R}^p$ as the expectation of X_a so that $X_a = \mu_a + \varepsilon_a$ with $(\varepsilon_1,\ldots,\varepsilon_n)$ centered and independent. Define also $v_a = \mathrm{trace}(cov(X_a))$. Prove that

   $$\mathbb{E}[\mathrm{crit}(C)] = \sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{a,b\in C_k}\left(\|\mu_a - \mu_b\|^2 + v_a + v_b\right)\mathbf{1}_{a\neq b}\,.$$

What is the value of $\mathbb{E}[\mathrm{crit}(C)]$ when all the within-group variables have the same mean?

The expectation of $crit(C)$ is given by

$$\mathbb{E}\left[crit(C)\right] = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b\in C_k} \mathbb{E}\left[\|X_a - X_b\|^2\right] \;.$$

Let $1 \leqslant k \leqslant K$ and $a,b \in C_k, a \neq b$,

$$\begin{aligned}
\mathbb{E}\left[\|X_a - X_b\|^2\right] &= \mathbb{E}\left[\|\mu_a - \mu_b + \varepsilon_a - \varepsilon_b\|^2\right] \;, \\
&= \mathbb{E}\left[\|\mu_a - \mu_b\|^2\right] + \mathbb{E}\left[\|\varepsilon_a - \varepsilon_b\|^2\right] + 2\underbrace{\mathbb{E}\left[\langle \mu_a - \mu_b, \varepsilon_a - \varepsilon_b\rangle\right]}_{=0} \;, \\
&= \|\mu_a - \mu_b\|^2 + \mathbb{E}\left[\|\varepsilon_a\|^2\right] + \mathbb{E}\left[\|\varepsilon_b\|^2\right] + 2\underbrace{\mathbb{E}\left[\langle \varepsilon_a, \varepsilon_b\rangle\right]}_{=0} \;,
\end{aligned}$$

since ε_a and ε_b are independent and centred. Finally, since for all $a \in C_k, \mathbb{E}\left[\|\varepsilon_a\|^2\right] = v_a$,

$$\mathbb{E}\left[crit(C)\right] = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b\in C_k} \left(\|\mu_a - \mu_b\|^2 + v_a + v_b\right) \mathbf{1}_{a\neq b} \;.$$

Assume now that for all $1 \leqslant k \leqslant K$, there exists $m_k \in \mathbb{R}^p$ such that for all $a \in C_k$, $\mu_a = m_k$. In this setting,

$$\mathbb{E}\left[crit(C)\right] = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b\in C_k} (v_a + v_b) \mathbf{1}_{a\neq b} \;,$$

where

$$\begin{aligned}
\frac{1}{|C_k|} \sum_{a,b\in C_k} (v_a + v_b) \mathbf{1}_{a\neq b} &= \frac{1}{|C_k|} \left(\sum_{a,b\in C_k} (v_a + v_b) - \sum_{a,b\in C_k} (v_a + v_b) \mathbf{1}_{a=b} \right) \;, \\
&= \frac{1}{|C_k|} \left(2|C_k| \sum_{a\in C_k} v_a - 2 \sum_{a\in C_k} v_a \right) \;, \\
&= \frac{2(|C_k| - 1)}{|C_k|} \sum_{a\in C_k} v_a \;.
\end{aligned}$$

Consequently, if, for all $a \in C_k$, $\mu_a = m_k$,

$$\mathbb{E}\left[crit(C)\right] = 2\sum_{k=1}^{K} \frac{|C_k| - 1}{|C_k|} \sum_{a\in C_k} v_a \;.$$

8.8 Gaussian vectors

1. Let X be a Gaussian vector with mean $\mu \in \mathbb{R}^n$ and definite positive covariance matrix Σ. Pove that the characteristic function of X is given, for all $t \in \mathbb{R}^n$, by

$$\mathbb{E}[\mathrm{e}^{i\langle t\,;\,X\rangle}] = \mathrm{e}^{i\langle t\,;\,\mu\rangle - t'\Sigma t/2} \;.$$

2. Let Σ be a positive definite matrix of $\mathbb{R}^{n\times n}$. Provide a solution to sample a Gaussian vector with covariance matrix Σ based on i.i.d. standard Gaussian variables.

3. Let ε be a random variable in $\{-1,1\}$ such that $\mathbb{P}(\varepsilon=1)=1/2$. If $(X,Y)'\sim\mathscr{N}(0,I_2)$ explain why the following vectors are or are not Gaussian vectors.
   - $(X,\varepsilon X)$.
     Not Gaussian since the probability that $X+\varepsilon X=0$ is $1/2$.
   - $(X,\varepsilon Y)$.
     Gaussian since coordinates are independent Gaussian random variables.
   - $(X,\varepsilon X+Y)$.
     Not Gaussian since the characteristic function of $(1+\varepsilon)X+Y$ is not the Gaussian characteristic function.
   - $(X,X+\varepsilon Y)$.
     Gaussian as a linear transform of (b). Indeed,

$$\begin{pmatrix} X \\ X+\varepsilon Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X \\ \varepsilon Y \end{pmatrix} .$$

4. Let X be a Gaussian vector in $\mathbb{R}^n$ with mean $\mu\in\mathbb{R}^n$ and covariance matrix $\sigma^2 I_n$. Prove that the random variables $\bar{X}_n$ and $\widehat{\sigma}_n^2$ defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad \widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar{X}_n)^2$$

are independent.

8.9 Regression: prediction of a new observation

Consider the regression model given, for all $1\leqslant i\leqslant n$, by

$$Y_i = X\beta_\star + \xi_i ,$$

where $X\in\mathbb{R}^{n\times d}$ the $(\xi_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered Gaussian random variables with variance $\sigma_\star^2$. Assume that $X'X$ has full rank and that $\beta_\star$ and $\sigma_\star^2$ are estimated by

$$\widehat{\beta}_n = (X'X)^{-1}X'Y \quad \text{and} \quad \widehat{\sigma}_n^2 = \frac{\|Y-X\widehat{\beta}_n\|^2}{n-d} .$$

Let $x_\star\in\mathbb{R}^d$ and assume that its associated observation $Y_\star = x_\star'\beta_\star+\varepsilon_\star$ is predicted by $\widehat{Y}_\star = x_\star'\widehat{\beta}_n$.

1. Provide the expression of $\mathbb{E}[(\widehat{Y}_\star - x_\star'\beta_\star)^2]$?

   By definition of $\widehat{\beta}_n$,

$$\widehat{Y}_\star - x_\star^T\beta_\star = x_\star^T(\widehat{\beta}_n-\beta_\star) ,$$

   so that $\mathbb{E}[\widehat{Y}_\star]=x_\star^T\beta_\star$ and

$$\mathbb{E}[(\widehat{Y}_\star - x_\star^T\beta_\star)^2] = \mathbb{V}[\widehat{Y}_\star] = x_\star^T\mathbb{V}[\widehat{\beta}_n]x_\star .$$

   On the other hand,

$$\mathbb{V}[\widehat{\beta}_n] = (X^TX)^{-1}X^T\mathbb{V}[Y]X(X^TX)^{-1} = \sigma^2(X^TX)^{-1} .$$

   Therefore,

$$\mathbb{E}[(\widehat{Y}_\star - x_\star^T\beta_\star)^2] = \sigma^2 x_\star^T(X^TX)^{-1}x_\star .$$

2. Provide a confidence interval for $x_\star'\beta_\star$ with statistical significiance $1-\alpha$ for $\alpha\in(0,1)$?

By the first question, $\widehat{Y}_\star$ is a Gaussian random variable with mean $x_\star^T\beta_\star$ and variance $\sigma_\star^2 x_\star^T(X^TX)^{-1}x_\star$. If $z_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of the standard Gaussian variable,

$$\mathbb{P}\left(\frac{\left|\widehat{Y}_\star - x_\star^T\beta_\star\right|}{\sigma_\star(x_\star^T(X^TX)^{-1}x_\star)^{1/2}} \leqslant z_{1-\alpha/2}\right) \geqslant 1-\alpha\,.$$

Therefore, with probability larger than $1-\alpha$,

$$x_\star^T\beta_\star \in \left(\widehat{Y}_\star - \sigma_\star(x_\star^T(X^TX)^{-1}x_\star)^{1/2}z_{1-\alpha/2}\,;\,\widehat{Y}_\star + \sigma_\star(x_\star^T(X^TX)^{-1}x_\star)^{1/2}z_{1-\alpha/2}\right)\,.$$

8.10 Regression: linear estimators

Consider the regression model given, for all $1 \leqslant i \leqslant n$, by

$$Y_i = f^*(X_i) + \xi_i\,,$$

where for all $1 \leqslant i \leqslant n$, $X_i \in \mathscr{X}$, and the $(\xi_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered Gaussian random variables with variance σ^2. In this exercise, f^* is estimated by a linear estimator of the form

$$\widehat{f}_n : x \mapsto \sum_{i=1}^n w_i(x)Y_i\,.$$

Prove that

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n(\widehat{f}_n(X_i) - f^*(X_i))^2\right] = \|Wf^*(X) - f^*(X)\|^2 + \frac{\sigma^2}{n}\mathrm{Trace}(W'W)\,,$$

where $W = (w_i(X_j))_{1\leqslant i,j\leqslant n}$ and $f^*(X) = (f^*(X_1),\dots,f^*(X_n))'$.

8.11 Kernels

Let $\mathscr{H}$ be a RKHS associated with a positive definite kernel $k : \mathscr{X}\times\mathscr{X} \to \mathbb{R}$.

1. Prove that for all $(x,y) \in \mathscr{X}\times\mathscr{X}$,

$$|f(x) - f(y)| \leqslant \|f\|_{\mathscr{H}}|k(x,\cdot) - k(y,\cdot)|_{\mathscr{H}}\,.$$

The proof follows from Cauchy-Schwarz inequality as, for all $(x,y) \in \mathscr{X}^2$,

$$|f(x) - f(y)| = |\langle f, k(x,\cdot)\rangle_{\mathscr{H}} - \langle f, k(x,\cdot)\rangle_{\mathscr{H}}| = |\langle f, k(x,\cdot) - k(y,\cdot)\rangle_{\mathscr{H}}|\,.$$

2. Prove that the kernel k associated with $\mathscr{H}$ is unique, i.e. if $\widetilde{k}$ is another potitive definite kernel satisfying the RKHS properties for $\mathscr{H}$, then $k = \widetilde{k}$.

Write, for all $x \in \mathscr{X}$,

$$\|k(x,\cdot) - \widetilde{k}(x,\cdot)\|^2_{\mathscr{H}} = \langle k(x,\cdot) - \widetilde{k}(x,\cdot), k(x,\cdot) - \widetilde{k}(x,\cdot)\rangle = k(x,x) - \widetilde{k}(x,x) + \widetilde{k}(x,x) - k(x,x) = 0\,.$$

3. Prove that for all $x \in \mathscr{X}$, the function defined on $\mathscr{H}$ by $\delta_x : f \mapsto f(x)$ is continuous.

8.12 Kernel Principal Component Analysis

Let $(X_i)_{1\leqslant i\leqslant n}$ be n observations in a general space X and $k : X\times X \to \mathbb{R}$ a positive kernel. W denotes the Reproducing Kernel Hilbert Space associated with k and for all $x\in X$, $\phi(x)$ denotes the function $\phi(x): y \to k(x,y)$. The aim is now to perform a PCA on $(\phi(X_1),\dots,\phi(X_n))$. It is assumed that

$$\sum_{i=1}^{n} \phi(X_i) = 0\ .$$

Define $K = (k(X_i,X_j))_{1\leqslant i,j\leqslant n}$.

1. Prove that
$$f_1 = \underset{f\in \mathsf{W}\,;\,\|f\|_{\mathsf{W}}=1}{\operatorname{argmax}} \sum_{i=1}^{n} \langle \phi(X_i), f\rangle_{\mathsf{W}}^2$$
may be written
$$f_1 = \sum_{i=1}^{n} \alpha_1(i)\phi(X_i)\ , \quad \text{where} \quad \alpha_1 = \underset{\alpha\in\mathbb{R}^n\,;\,\alpha^T K\alpha=1}{\operatorname{argmax}} \alpha^T K^2 \alpha\ .$$
Any solution to the optimization problem lies in the vectorial subspace $V = \mathrm{span}\{\phi(X_i),\dots,\phi(X_n)\}$. Let $f = \sum_{i=1}^{n} \alpha(i)\phi(X_i)$ be such that $\|f\|_{\mathsf{W}} = 1$. Then,
$$\|f\|_{\mathsf{W}}^2 = \sum_{i,j=1}^{n} \alpha_i\alpha_j \langle \phi(X_i), \phi(X_j)\rangle_{\mathsf{W}} = \alpha^T K\alpha\ .$$
On the other hand, $\langle \phi(X_i), f\rangle_{\mathsf{W}} = f(X_i) = [K\alpha](i)$ so that,
$$\sum_{i=1}^{n} \langle \phi(X_i), f\rangle_{\mathsf{W}}^2 = \sum_{i=1}^{n} f^2(X_i) = \sum_{i=1}^{n} ([K\alpha](i))^2 = (K\alpha_1)^T K\alpha_1 = \alpha^T K^2\alpha\ .$$

2. Prove that $\alpha_1 = \lambda_1^{-1/2} b_1$ where b_1 is the unit eigenvector associated with the largest eigenvalue λ_1 of K.
Let $\lambda_1 \geqslant \ldots \geqslant \lambda_n \geq 0$ be the eigenvalues of K associated with the orthonormal basis of eigenvectors $(b_1,\dots,b_n)$. For any $\alpha\in\mathbb{R}^n$ such that $\alpha^T K\alpha = 1$,
$$\alpha^T K^2\alpha = \alpha^T \left(\sum_{i=1}^{n} \lambda_i b_i b_i^T\right)^2 \alpha = \sum_{i=1}^{n} \lambda_i^2 \langle \alpha, b_i\rangle^2 \leqslant \lambda_1 \underbrace{\sum_{i=1}^{n} \lambda_i \langle \alpha, b_i\rangle^2}_{=1} = \lambda_1\ ,$$
as $\alpha^T K\alpha = \sum_{i=1}^{n} \lambda_i\langle\alpha,b_i\rangle^2 = 1$. On the other hand,
$$\left(\lambda_1^{-1/2} b_1\right)^T K^2 \left(\lambda_1^{-1/2} b_1\right) = \lambda_1^{-1} \sum_{i=1}^{n} \lambda_i^2 \langle b_1, b_i\rangle^2 = \lambda_1\ .$$
Following the same steps, f_j may be written $f_j = \sum_{i=1}^{n} \alpha_j(i)\phi(x_i)$ with $\alpha_j = \lambda_j^{-1/2} b_j$.

3. Write $H_d = \mathrm{span}\{f_1,\dots,f_d\}$. Prove that, for all $1\leqslant i\leqslant n$,
$$\pi_{H_d}(\phi(X_i)) = \sum_{j=1}^{d} \lambda_j \alpha_j(i) f_j\ .$$
Note first that the $(f_1,\dots,f_d)$ is an orthonormal family. Therefore,

$$\pi_{H_d}(\phi(X_i)) = \sum_{j=1}^{d} \langle \phi(X_i), f_j \rangle_{\mathsf{W}} f_j = \sum_{j=1}^{d} \langle \phi(X_i), \sum_{\ell=1}^{n} \alpha_j(\ell)\phi(X_\ell) \rangle_{\mathsf{W}} f_j = \sum_{j=1}^{d} [K\alpha_j](i) f_j \,.$$

Therefore,

$$\pi_{H_d}(\phi(x_i)) = \sum_{j=1}^{d} \lambda_j^{-1/2} [Kb_j](i) f_j = \sum_{j=1}^{d} \lambda_j^{1/2} b_j(i) f_j = \sum_{j=1}^{d} \lambda_j \alpha_j(i) f_j \,.$$

8.13 Penalized kernel regression

Consider the regression model given, for all $1 \leqslant i \leqslant n$, by

$$Y_i = f^*(X_i) + \xi_i \,,$$

where for all $1 \leqslant i \leqslant n$, $X_i \in \mathscr{X}$, and the $(\xi_i)_{1 \leqslant i \leqslant n}$ are i.i.d. centered Gaussian random variables with variance σ^2. In this exercise, f^* is estimated by

$$\widehat{f}_n = \underset{f \in \mathscr{H}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(X_i))^2 + \frac{\lambda}{n} \|f\|_{\mathscr{H}}^2 \right\} \,,$$

with $\lambda > 0$ and $\mathscr{H}$ a RKHS on $\mathscr{X}$ with kernel k.

1. Check that $\widehat{f}(x) = \sum_{j=1}^{n} \widehat{\beta}_{n,j} k(X_j, x)$ where $\widehat{\beta}_n$ is solution to

$$\widehat{\beta}_n = \underset{\beta \in \mathsf{R}^n}{\arg\min} \left\{ \|y - K\beta\|^2 + \lambda \beta' K \beta \right\} \,,$$

with K defined, for all $1 \leqslant i, j \leqslant n$, by $K_{i,j} = k(X_i, X_j)$. Provide the explicit expression of $\widehat{\beta}_n$ when K is nonsingular.

Let

$$V = \left\{ \sum_{i=1}^{n} \alpha_i k(X_i, \cdot) \,;\, (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \right\} \,.$$

For all $f \in \mathscr{H}$, write $f = f_V + f_{V^\perp}$ where $f_V \in V$ and $f_{V^\perp} \in V^\perp$. Therefore,

$$\frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - f(X_i) \Big)^2 + \frac{\lambda}{n} \|f\|_{\mathscr{H}}^2 = \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - f_V(X_i) \Big)^2 + \frac{\lambda}{n} \Big(\|f_V\|_{\mathscr{H}}^2 + \|f_{V^\perp}\|_{\mathscr{H}}^2 \Big) \,,$$

since, by definition of $V^\perp$, for all $1 \leqslant i \leqslant n$,

$$f_{V^\perp}(X_i) = \langle f_{V^\perp}, k(X_i, \cdot) \rangle_{\mathscr{H}} = 0 \,.$$

Thus, the initial optimization problem can be written as

$$\widehat{f}_n = \underset{f \in V}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \frac{\lambda}{n} \|f\|_{\mathscr{H}}^2 \right\} .$$

Therefore, there exists $\beta \in \mathbb{R}^n$ such that, for all $x \in \mathscr{X}$,

$$\widehat{f}_n(x) = \sum_{j=1}^{n} \widehat{\beta}_j k(X_j, x) \,.$$

This yields,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \frac{\lambda}{n}\|f\|_{\mathscr{H}}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{n}\beta_j k(X_j,X_i))^2 + \frac{\lambda}{n}\langle\sum_{j=1}^{n}\beta_j k(X_j,\cdot),\sum_{i=1}^{n}\beta_i k(X_i,\cdot)\rangle_{\mathscr{H}}\,.$$

The proof is completed by noting that,

$$\langle\sum_{j=1}^{n}\beta_j k(X_j,\cdot),\sum_{i=1}^{n}\beta_i k(X_i,\cdot)\rangle_{\mathscr{H}} = \sum_{i,j=1}^{n}\beta_i\beta_j k(X_i,X_j)\,.$$

Let

$$L(\beta) = \|Y - K\beta\|_2^2 + \lambda\beta^T K\beta.$$

The gradient of L is then given by

$$\nabla L(\beta) = -2K^T(Y-K\beta) + \lambda(K\beta + K^T\beta) = -2K(Y-K\beta) + 2\lambda K\beta\,.$$

The minimum $\widehat{\beta}_n$ of L satisfies

$$\widehat{\beta}_n = (K+\lambda I_n)^{-1}Y\,.$$

2. Assume that $f^* \in \mathscr{H}$ and write

$$f_V^* : x \mapsto \sum_{i=1}^{n}\beta_i^* k(X_i,x)$$

the projection of f^* onto the vector subspace generated by $(k(X_i,\cdot))_{1\leqslant i\leqslant n}$, with respect to the scalar product $\langle\ ;\ \rangle_{\mathscr{H}}$. Let $K = \sum_{i=1}^{n}\lambda_i u_i u_i^T$ be an eigenvalue decomposition of K. Check that

$$K\widehat{\beta} = \sum_{i=1}^{n}\frac{\lambda_i}{\lambda_i+\lambda}\langle Y,u_i\rangle u_i$$

and

$$\|\mathbb{E}[K\widehat{\beta}_n] - K\beta^*\|^2 = \sum_{i=1}^{n}\left(\frac{\lambda\lambda_i}{\lambda_i+\lambda}\right)^2\langle\beta^*,u_i\rangle^2\,.$$

Since $(u_i)_{1\leq i\leq n}$ is an orthonormal basis of $\mathsf{R}^{\,n}$*, one can write*

$$\begin{aligned}K\widehat{\beta}_n &= \sum_{i=1}^{n}\langle K\widehat{\beta}_n,u_i\rangle u_i\,,\\ &= \sum_{i=1}^{n}\langle K(K+\lambda I_n)^{-1}Y,u_i\rangle u_i\,,\\ &= \sum_{i=1}^{n}\langle Y,(K+\lambda I_n)^{-1}Ku_i\rangle u_i\,,\\ &= \sum_{i=1}^{n}\frac{\lambda_i}{\lambda+\lambda_i}\langle Y,u_i\rangle u_i\,.\end{aligned}$$

3. Prove that

$$\mathbb{V}[K\widehat{\beta}_n] = \sum_{i=1}^{n}\left(\frac{\lambda_i\sigma}{\lambda_i+\lambda}\right)^2 u_i u_i^T\ .$$

Since $\widehat{\beta}_n = (K+\lambda I_n)^{-1}Y$,

$$
\begin{aligned}
\mathbb{V}[K\widehat{\beta}_n] &= K\mathbb{V}\left[(K+\lambda I_n)^{-1}Y\right]K^T\,,\\
&= K(K+\lambda I)^{-1}\mathbb{V}[Y](K+\lambda I_n)^{-1}K\,,\\
&= \sigma^2 K^2(K+\lambda I_n)^{-2}\,,\\
&= \sum_{i=1}^{n}\left(\frac{\lambda_i\sigma}{\lambda_i+\lambda}\right)^2 u_i u_i^T\,,
\end{aligned}
$$

using the eigenvector decomposition of K.

8.14 Expectation Maximization algorithm

In the case where we are interested in estimating unknown parameters $\theta \in \mathbb{R}^m$ characterizing a model with missing data, the Expectation Maximization (EM) algorithm (Dempster et al. 1977) can be used when the joint distribution of the missing data X and the observed data Y is explicit. For all $\theta \in \mathbb{R}^m$, let p_θ be the probability density function of (X,Y) when the model is parameterized by θ with respect to a given reference measure μ. The EM algorithm aims at computing iteratively an approximation of the maximum likelihood estimator which maximizes the observed data loglikelihood:

$$
\ell(\theta;Y) = \log p_\theta(Y) = \log\int f_\theta(x,Y)\mu(\mathrm{d}x)\,.
$$

As this quantity cannot be computed explicitly in general cases, the EM algorithm finds the maximum likelihood estimator by iteratively maximizing the expected complete data loglikelihood. Start with an inital value $\theta^{(0)}$ and let $\theta^{(t)}$ be the estimate at the t-th iteration for $t \geqslant 0$, then the next iteration of EM is decomposed into two steps.

1. **E step**. Compute the expectation of the complete data loglikelihood, with respect to the conditional distribution of the missing data given the observed data parameterized by $\theta^{(t)}$:

$$
Q(\theta,\theta^{(t)}) = \mathbb{E}_{\theta^{(t)}}\left[\log p_\theta(X,Y)|Y\right]\,.
$$

2. **M step**. Determine $\theta^{(t+1)}$ by maximizing the function Q:

$$
\theta^{(t+1)} \in \mathrm{argmax}_\theta Q(\theta,\theta^{(t)})\,.
$$

1. Prove the following crucial property motivates the EM algorithm. For all $\theta,\theta^{(t)}$,

$$
\ell(Y;\theta) - \ell(Y;\theta^{(t)}) \geqslant Q(\theta,\theta^{(t)}) - Q(\theta^{(t)},\theta^{(t)})\,.
$$

This may be proved by noting that

$$
\ell(Y;\theta) = \log\left(\frac{p_\theta(X,Y)}{p_\theta(X|Y)}\right)\,.
$$

Considering the conditional expectation of both terms given Y when the parameter value is $\theta^{(t)}$ yields

$$
\ell(Y;\theta) = Q(\theta,\theta^{(t)}) - \mathbb{E}_{\theta^{(t)}}[\log p_\theta(X|Y)|Y]\,.
$$

Then,

$$
\ell(Y;\theta) - \ell(Y;\theta^{(t)}) = Q(\theta,\theta^{(t)}) - Q(\theta^{(t)},\theta^{(t)}) + H(\theta,\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)})\,,
$$

where

$$
H(\theta,\theta^{(t)}) = -\mathbb{E}_{\theta^{(t)}}[\log p_\theta(X|Y)|Y]\,.
$$

The proof is completed by noting that

$$H(\theta,\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \geqslant 0\,,$$

as this difference if a Kullback-Leibler divergence.

In the following, $X = (X_1,\dots,X_n)$ and $Y = (Y_1,\dots,Y_n)$ where $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ are i.i.d. in $\{-1,1\}\times\mathbb{R}^d$. For $k\in\{-1,1\}$, write $\pi_k = \mathbb{P}(X_1 = k)$. Assume that, conditionally on the event $\{X_1 = k\}$, Y_1 has a Gaussian distribution with mean $\mu_k\in\mathbb{R}^d$ and covariance matrix $\Sigma\in\mathbb{R}^{d\times d}$. In this case, the parameter $\theta = (\pi_1,\mu_1,\mu_{-1},\Sigma)$ belongs to the set $\Theta = [0,1]\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^{d\times d}$.

1. Write the complete data loglikelihood.

   The complete data loglikelihood is given by

$$\begin{aligned}\log p_\theta(X,Y) &= -\frac{nd}{2}\log(2\pi) + \sum_{i=1}^n\sum_{k\in\{-1,1\}}\mathbb{1}_{X_i=k}\left(\log\pi_k - \frac{\log\det\Sigma}{2} - \frac{1}{2}(Y_i-\mu_k)^T\Sigma^{-1}(Y_i-\mu_k)\right)\,,\\ &= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^n\mathbb{1}_{X_i=1}\right)\log\pi_1 + \left(\sum_{i=1}^n\mathbb{1}_{X_i=-1}\right)\log(1-\pi_1)\\ &\quad -\frac{1}{2}\sum_{i=1}^n\mathbb{1}_{X_i=1}(Y_i-\mu_1)^T\Sigma^{-1}(Y_i-\mu_1) - \frac{1}{2}\sum_{i=1}^n\mathbb{1}_{X_i=-1}(Y_i-\mu_{-1})^T\Sigma^{-1}(Y_i-\mu_{-1})\,.\end{aligned}$$

2. Let $\theta^{(t)}$ be the current parameter estimate. Compute $\theta\mapsto Q(\theta,\theta^{(t)})$.

   Write $\omega_t^i = \mathbb{P}_{\theta^{(t)}}(X_i = 1|Y_i)$. The intermediate quantity of the EM algorithm is given by

$$\begin{aligned}Q(\theta,\theta^{(t)}) &= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^n\omega_t^i\right)\log\pi_1 + \sum_{i=1}^n\left(1-\omega_t^i\right)\log(1-\pi_1)\\ &\quad -\frac{1}{2}\sum_{i=1}^n\omega_t^i(Y_i-\mu_1)^T\Sigma^{-1}(Y_i-\mu_1) - \frac{1}{2}\sum_{i=1}^n(1-\omega_t^i)(Y_i-\mu_{-1})^T\Sigma^{-1}(Y_i-\mu_{-1})\,.\end{aligned}$$

3. Compute $\theta^{(t+1)}$.

   The gradient of $Q(\theta,\theta^{(t)})$ with respect to θ is therefore given by

$$\begin{aligned}\frac{\partial Q(\theta,\theta^{(t)})}{\partial\pi_1} &= \frac{\sum_{i=1}^n\omega_t^i}{\pi_1} - \frac{n-\sum_{i=1}^n\omega_t^i}{1-\pi_1}\,,\\ \frac{\partial Q(\theta,\theta^{(t)})}{\partial\mu_1} &= \sum_{i=1}^n\omega_t^i\left(2\Sigma^{-1}Y_i - 2\Sigma^{-1}\mu_1\right)\,,\\ \frac{\partial Q(\theta,\theta^{(t)})}{\partial\mu_{-1}} &= \sum_{i=1}^n(1-\omega_t^i)\left(2\Sigma^{-1}Y_i - 2\Sigma^{-1}\mu_{-1}\right)\,,\\ \frac{\partial Q(\theta,\theta^{(t)})}{\partial\Sigma^{-1}} &= \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^n\omega_t^i(Y_i-\mu_1)(Y_i-\mu_1)^T - \frac{1}{2}\sum_{i=1}^n(1-\omega_t^i)(Y_i-\mu_{-1})(Y_i-\mu_{-1})^T\,.\end{aligned}$$

   Then, $\theta^{(t+1)}$ is defined as the only parameter such that all these equations are set to 0. It is given by

$$\begin{aligned}\widehat{\pi}_1^{(t+1)} &= \frac{1}{n}\sum_{i=1}^n\omega_t^i\,,\\ \widehat{\mu}_1^{(t+1)} &= \frac{1}{\sum_{i=1}^n\omega_t^i}\sum_{i=1}^n\omega_t^iY_i\,,\\ \widehat{\Sigma}^{(t+1)} &= \frac{1}{n}\sum_{i=1}^n\omega_t^i(Y_i-\mu_1)(Y_i-\mu_1)^T + \frac{1}{n}\sum_{i=1}^n(1-\omega_t^i)(Y_i-\mu_{-1})(Y_i-\mu_{-1})^T\,.\end{aligned}$$

References

Akaike, 1974. Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Arlot and Celisse, 2010. Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Geisser, 1975. Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328.

Hotelling, 1933. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Schwarz, 1978. Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Tibshirani, 1996. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tibshirani, 2013. Tibshirani, R. J. (2013). The lasso problem and uniqueness.

Index