Randal Douc & Sylvain Le Corff

Introduction to machine learning

Telecom SudParis

Contents

Chapter 1

Dimension reduction

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 [**?**] who first introduced the principal components as a way to reduce the dimensionality of the data. The objective is to find a low-dimensional representation that captures the statistical properties of high-dimensional data. Reducing the dimensionality of the data is motivated by several practical reasons such as the following (in addition to better understand the observations which are difficult to plot and interpret).

- Compression, denoising, data completion, anomaly detection.
- Preprocessing before supervised learning (improve performances / regularization to reduce overfitting).
- Simplifying the description of massive datasets.
- Providing tools to analyze both observations and variables.

Principal component analysis relies on the implicit assumption that the data lies in a low-dimensional manifold of the orginial space. 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^T. In this chapter, it is assumed that the columns of X are centered. This means that for all $1 \leqslant k \leqslant d$, $\sum_{i=1}^n X_{i,k} = 0$. Let Σ_n be the empirical covariance matrix:

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

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 $p \leqslant d$ so that for each $1 \leqslant i \leqslant n$, WX_i ia 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:

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

The source codes in R and/or Python of this chapter may be found at:

https : //sylvainlc.github.io/project/teaching/

1.1 Principal Component Analysis as a singular value decomposition problem

1.1.1 Refresher on matrices

Lemma 1.1 *Let A be a $n \times d$ matrix with real entries. Then,* $\operatorname{range}(A) = \operatorname{range}(AA^T)$.

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

Lemma 1.2 *Let $\{U_k\}_{1\leqslant k\leqslant r}$ be a family of r orthonormal vectors of $\mathbb{R}^d$. Then, $\sum_{k=1}^r U_k U_k^T$ 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 1.3 *If A is a $n \times d$ matrix with real entries such that each column of A is in* $\mathbf{H}$*, then,*

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

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 $\mathbf{H}$,

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

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

1.1.2 Singular value decomposition

Proposition 1.4 *For all $\mathbb{R}^{n\times d}$ matrix A with rank r, there exist $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ such that*

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

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

$$A = UD_rV^T \ ,$$

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

Remark 1.6 *The singular value decomposition is closely related to the spectral theorem for symmetric semipositive definite matrices. In the framework of Proposition 1.8, A^TA and AA^T are positive semidefinite such that*

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

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

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

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

$$\begin{aligned} \|v_k\|^2 &= \lambda_k^{-1}\langle A^Tu_k; A^Tu_k\rangle = \lambda_k^{-1}u_k^TAA^Tu_k = 1 \ , \\ A^TAv_k &= \lambda_k^{-1/2}A^TAA^Tu_k = \lambda_k v_k \ . \end{aligned}$$

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

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

As $\{u_1,\dots,u_r\}$ is an orthonormal family, by Lemma 1.2 $UU^T = \sum_{k=1}^r u_k u_k^T$ is the orthogonal projection onto the range(AA^T). And, by Lemma 1.1, range(AA^T) = range(A), which implies

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

■

1.1.3 Application to Principal Component Analysis

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

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

Lemma 1.7 *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 $W_\star = U_\star^T$.*

PROOF. Let $(U,W) \in \mathbb{R}^{d\times p}\times\mathbb{R}^{p\times d}$ and V be a matrix whose columns are given by an orthonormal basis of the vector space E generated by the UW. Therefore, for all $x \in \mathbb{R}^d$,

$$VV^Tx = \underset{z\in\mathsf{E}}{\mathrm{argmin}} \ \|x - z\|^2 \ .$$

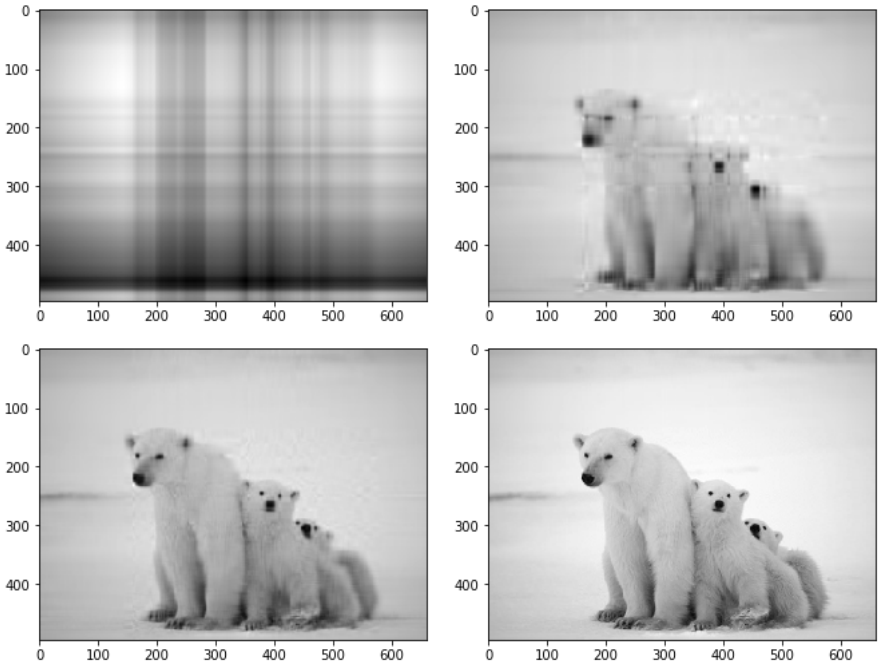


Fig. 1.1 Application of SVD to image reconstruction. The original image is a $\mathbb{R}^{495\times 660}$ matrix of pixels (grey scale). Its SVD is obtained using the np.linalg.svd function in Python and the image is then reconstructed using only the largest (top left), the first ten (top right), the first 25 (bottom left) and the first 100 (bottom right) singular values.

As for all $1 \leqslant i \leqslant n$, $UWX_i \in \mathsf{E}$,

$$\sum_{i=1}^{n} \|X_i - UWX_i\|^2 \geqslant \sum_{i=1}^{n} \|X_i - VV^T X_i\|^2 ,$$

which concludes the proof. ■

Let $U \in \mathbb{R}^{d\times p}$ be such that $U^T U = I_p$. Then,

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

Therefore, by Lemma 1.7, solving (1.1) boils down to computing

$$U_\star \in \underset{U\in\mathbb{R}^{d\times p},\, U^TU=I_p}{\mathrm{argmax}} \{\mathrm{trace}(U^T \Sigma_n U)\} . \tag{1.2}$$

Proposition 1.8 *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 $W_\star = U_\star^T$.*

```python
# write a function with input the path of an image "path_image" and an integer "k"
# and return in gray scale the reconstructed picture with the first k singular values
def svd_decomposition(path_image,k):
    img           = Image.open(path_image)
    img_mat       = np.array(list(img.getdata(band=0)), float)
    img_mat.shape = (img.size[1], img.size[0])
    img_mat       = np.matrix(img_mat)

    # Perform Singular Value Decomposition
    U, sigma, V = np.linalg.svd(img_mat)
    print('Size left singular eigenvectors  ' + str(np.shape(U)))
    print('Size right singular eigenvectors  ' + str(np.shape(V)))
    print('Size eigenvalues matrix ' + str(np.shape(sigma)))

    # Image reconstruction
    reconstimg = np.matrix(U[:, :k]) * np.diag(sigma[:k]) * np.matrix(V[:k, :])

    fig = plt.figure(1)
    plt.plot(sigma[0:12]*100/np.sum(sigma))
    plt.title("Normalized values of the singular values (in %)")
    fig = plt.figure(2)
    plt.title("Image reconstruction with %g singular values"%k)
    plt.imshow(reconstimg, cmap='gray')
    plt.axis('off')
```

```
Size left singular eigenvectors  (1415, 1415)
Size right singular eigenvectors  (2122, 2122)
Size eigenvalues matrix (1415,)
```

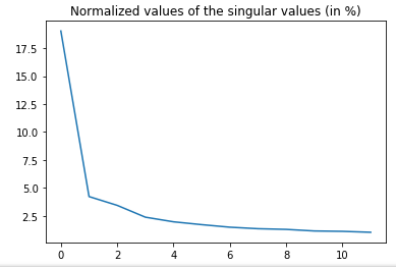


Fig. 1.2 Application of a singular value decomposition in Python

Remark 1.9 *Note that* $\sum_{i=1}^n X_i X_i^T = X^T X$ *so that* $\Sigma_n = (X/\sqrt{n})^T (X/\sqrt{n})$ *and the eigenvalue decomposition of* Σ_n *allows to recover the singular values of* $X/\sqrt{n}$ *and its right-singular vectors.*

PROOF. By Lemma 1.7 and (1.2), the proof is equivalent to prove that $U_\star$ is a solution to (1.2). Let $\Sigma_n = V D_n V^T$ 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 $U \in \mathbb{R}^{d\times p}$ matrix with orthonormal colums define $B = V^T U$ so that, as $V \in \mathbb{R}^{d\times d}$ is an orthogonal matrix,

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

Therefore,

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

On the other hand,

$$B^T B = U^T V V^T U^T = U^T U = I_p \,,$$

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),

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

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

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

As the columns of $U_\star$ are$\{\vartheta_1,\dots,\vartheta_p\}$, 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),

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

which completes the proof. ■

1.2 Principal Component Analysis as an optimal projection

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

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

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

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

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

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

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

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

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

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

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,

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

and, as $(\vartheta_i)_{1 \leqslant i \leqslant d}$ is an orthonormal basis of $\mathbb{R}^d$, $\sum_{i=1}^{d} \langle v, \vartheta_i \rangle^2 = \|v\|^2 = 1$. 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 $\|v\| = 1$ 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 $\mathbb{R}^d$, $\sum_{i=1}^{d}\langle v, \vartheta_i\rangle^2 = \sum_{i=k}^{d}\langle v, \vartheta_i\rangle^2 = \|v\|^2 = 1$. 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, $V_p = \text{span}\{\vartheta_1, \dots \vartheta_p\}$ 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 $U_\star U_\star^T$ where $U_\star$ is a $\mathbb{R}^{d\times p}$ matrix with columns $\{\vartheta_1, \dots \vartheta_p\}$.

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 follows. Then, as $V_d = \text{span}\{\vartheta_1, \dots, \vartheta_d\}$, for all $1 \leqslant i \leqslant n$,

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

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 coordinates of each X_i, $1 \leqslant i \leqslant n$, relative to the basis $\{\vartheta_1, \dots, \vartheta_d\}$ of V_d.

1.3.2 Projection of the variables

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

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

as $\{\vartheta_1, \dots, \vartheta_d\}$ is an orthonormal family. Let W_d be the vector subspace of $\mathbb{R}^n$ generated by $\{c_1, \dots, c_d\}$. For all $1 \leqslant j \leqslant d$, write $\mathbf{X}_j$ the j-th colunm of X, i.e. the values of the j-th variable along all individuals. Since $(c_j)_{1\leqslant j\leqslant d}$ form a orthogonal basis of W_d, for all $1 \leqslant j \leqslant d$,

$$\pi_{W_p}(\mathbf{X}_j) = \sum_{\ell=1}^{p}\frac{\langle c_\ell, \mathbf{X}_j\rangle}{\|c_\ell\|^2}c_\ell \ .$$

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

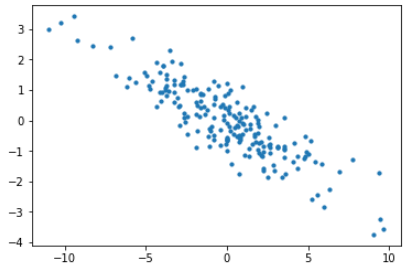
$$\langle c_\ell, \mathbf{X}_j\rangle = \langle X\vartheta_\ell, \mathbf{X}_j\rangle = \mathbf{X}_j^T X\vartheta_\ell = (X^T X\vartheta_\ell)_j = (n\Sigma_n\vartheta_\ell)_j = n\lambda_\ell\vartheta_\ell(j) \ .$$

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

$$\pi_{W_p}(\mathbf{X}_j) = \sum_{\ell=1}^{p}\vartheta_\ell(j)c_\ell \ .$$

```
# Build a data set with n=100 bi-dimensional i.i.d. data with centered Gaussian distribution
X = np.dot(np.random.normal(0,1,[2,2]), np.random.normal(0,1,[2,200])).T
plt.scatter(X[:, 0], X[:, 1], s = 10)
```

```
<matplotlib.collections.PathCollection at 0x2b7a92b99e8>
```




```
# perform a PCA with one component using PCA(n_components=1) and the function fit.
from sklearn.decomposition import PCA
pca = PCA(n_components=1)
pca.fit(X)
print('The principal component is:')
print(pca.components_)

print('The explained variance is %g'%pca.explained_variance_)
print('The associated singular value is %g'%pca.singular_values_)

# Apply the dimensionality reduction on X
# X_pca contains the coordinates of each data in the space generated by the principal components
X_pca = pca.transform(X)

# in this case pca.components_[k] contains the coordinates of the k-th principal component in
# the original space (here the usual Euclidian plane). In a general case pca.components_[k] is a
# d-dimensional vector.
# X_pca[i] contains the coordinates of the i-th data in the vector space generated by the principal
# components.
# Therefore, X_pca[i] is a vector with n_components entries.
```

```
The principal component is:
[[-0.95642601  0.29197481]]
The explained variance is 13.5057
The associated singular value is 51.8423
```

```
# transform the reduced data set in the original space
# X_inverse[i,:] contains the coordinates of the projection of Xi in the original space
X_inverse = pca.inverse_transform(X_pca)
plt.scatter(X[:, 0], X[:, 1], alpha=0.2, s=10)
plt.scatter(X_inverse[:, 0], X_inverse[:, 1], alpha=0.8, s=10)
```

```
<matplotlib.collections.PathCollection at 0x2b7ab9442e8>
```

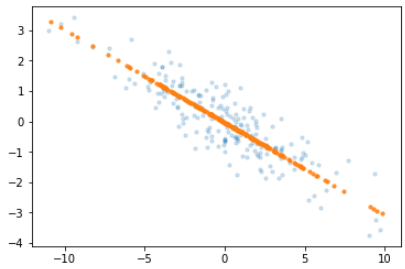


Fig. 1.3 Principal component analysis with one component in Python.

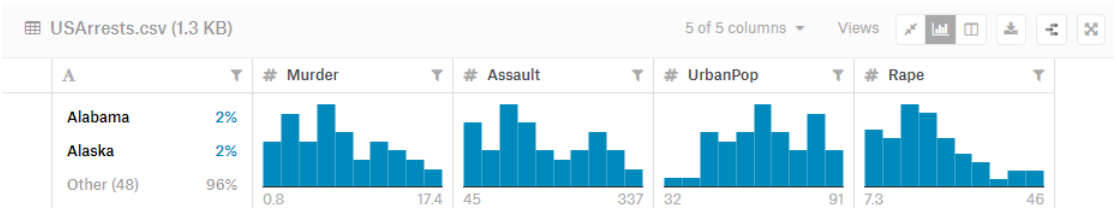
1.3.3 Explained variance and projection quality

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

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

1.3.4 Application to the USArrests dataset [?]

Dataset containing arrests per 100.000 residents for assault, murder and rape for the 50 US states in 1973. Percent of the population living in urban areas are also given.



	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

In this setting $n = 50$ (number states) and $d = 4$ (number of variables - Murder, Assault, Rape and UrbanPop).

$$\vartheta_1 = (0.53, 0.58, 0.28, 0.54) ,$$
$$\vartheta_2 = (-0.42, -0.19, 0.87, 0.17) .$$

The coordinates of the projected observations (the states) in the plane generated by the two first eigenvectors are:

$$\pi_{V_2}(X_i) = (X_i^T \vartheta_1)\vartheta_1 + (X_i^T \vartheta_2)\vartheta_2 . \tag{1.7}$$

The coordinates of the projected variables (Murder, Assault, Rape and UrbanPop) in the plane generated by the two first components are:

$$\pi_{W_2}(\mathbf{X}_j) = \sum_{\ell=1}^{2} \vartheta_\ell(j) c_2 . \tag{1.8}$$

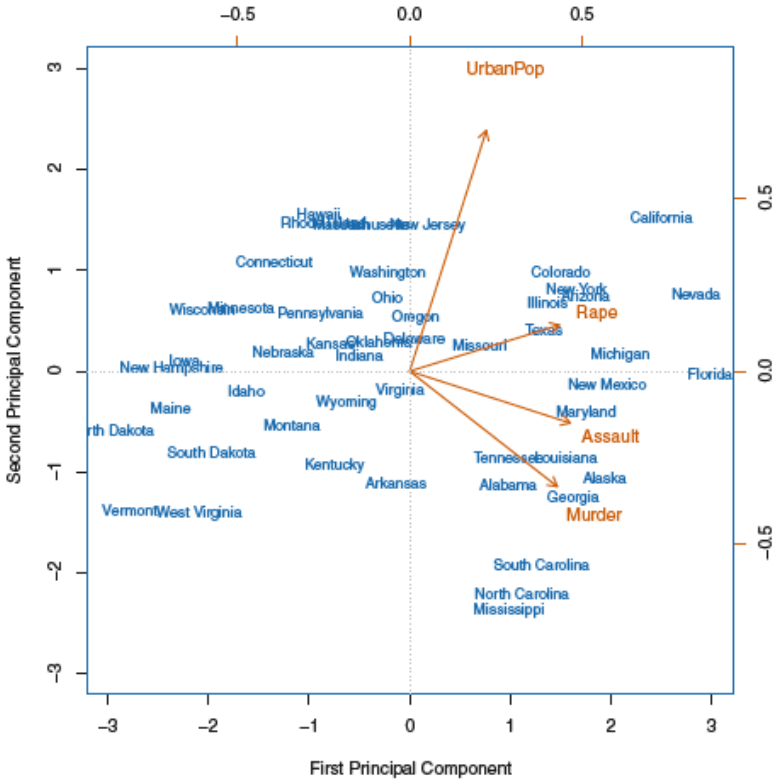


Fig. 1.4 This biplot simultaneously displays the observations and the variables. States names in blue are the the scores for the two first principal components (coordinates of the projected observations) with left and bottom axis, see (1.7). Orange arrows are the loadings for the two first principal components (coordinates of the projected variables) with right and top axis, see (1.8).

Chapter 2

Probability density estimation

Contents

Keywords 2.1 *Probability density, histogram, kernel, bias-variance trade-off*

Probability density estimation is an all-pervasive problem in statistical analysis across a wide range of applied science and engineering domains. In a classification setting for instance it may be useful to obtain a nonparametric estimation of the likelihood (i.e. the probability density function) of the data to (a) analyze the behavior of individuals within each group and the differences between groups and (b) to make plausible assumptions on the distribution of the data before designing a classification rule. In this chapter, the aim is to estimate a probability density function $p_\star$ on $\mathsf{X} \subset \mathbb{R}^d$ based on independent and identically distributed random variables $(X_i)_{1\leqslant i\leqslant n}$ such that the law of X_1 has $p_\star$ as probability density function with respect to a reference measure on X.

In a parametric approach, strict limitations are introduced on $p_\star$ which is assumed to belong to a family of distributions which depend on an unknown parameter $\theta \in \Theta \subset \mathbb{R}^q$. For instance, a parametric approach may assume $p_\star$ to be a Gaussian distribution with unknown mean and variance to be estimated. Such assumptions significantly simplify the problem, since only the parameters of the chosen family need to be determined. Nonparametric approaches are more appealing when such assumptions cannot be introduced easily, one of the strengths of such methods is that the available data drive the estimation process directly without any parametric assumption. Let $\widehat{p}_n(X_1,\dots,X_n)$ be an estimator $p_\star$ (in the following, $(X_1,\dots,X_n)$ are dropped from the notations of the estimator). In this chapter, the performance of this estimator is measured by the two following quantities.

i) *The mean square error* defined, for all $x \in \mathsf{X}$ by

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

ii) *The integrated mean square error* defined by

$$\overline{\mathscr{E}} = \mathbb{E}\left[\int_0^1 (\widehat{p}_n(x) - p_\star(x))^2 \,\mathrm{d}x\right].$$

2.1 Histograms

In this section, it is assumed that X is bounded and $\mathsf{X} = [0,1)$ without loss of generality. For all $m \geqslant 1$, let $(A_1,\dots,A_m)$ be the uniform partition of $[0,1)$ so that for all $1 \leqslant i \leqslant m$, $A_i = [(i-1)/m, i/m)$. Let $(X_1 \dots, X_n)$ i.i.d. with probability density $p_\star$ with respect to the Lebesgue measure on $[0,1)$. The histogram estimator $p_\star$ is defined as follows:

$$\widehat{p}_n^h : x \mapsto \frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^m 1_{A_j}(X_i) 1_{A_j}(x),$$

where $h = 1/m$.

Proposition 2.1 *For all $x \in [0,1)$,*

$$\mathscr{E}(x) = \sum_{j=1}^m 1_{A_j}(x) \frac{p_j(1-p_j)}{nh^2} + \left(\sum_{j=1}^m 1_{A_j}(x) \frac{p_j}{h} - p_\star(x) \right)^2$$

and

$$\overline{\mathscr{E}} = \int_0^1 p_\star^2(x)dx + \sum_{j=1}^m \frac{p_j(1-p_j)}{mn} - \sum_{j=1}^m \frac{p_j^2}{h}.$$

PROOF. For all $x \in [0,1)$, let j_x be the unique $1 \leqslant j \leqslant m$ such that $x \in A_j$, which yields

$$\widehat{p}_n^h(x) = \frac{1}{nh} \sum_{i=1}^n 1_{A_{j_x}}(X_i) = \frac{Z_{j_x}}{nh},$$

where for all $1 \leqslant j \leqslant m$, Z_j has a binomial distribution with parameters n and $p_j = \mathbb{P}(X_1 \in A_j)$. Then,

$$\mathbb{E}\left[\widehat{p}_n^h(x)\right] = \frac{p_{j_x}}{h}$$

and

$$\mathbb{V}\left[\widehat{p}_n^h(x)\right] = \frac{p_{j_x}(1-p_{j_x})}{nh^2}.$$

Therefore,

$$\mathbb{E}\left[\left(\widehat{p}_n^h(x) - p_\star(x)\right)^2\right] = \sum_{j=1}^m 1_{A_j}(x) \frac{p_j(1-p_j)}{nh^2} + \left(\sum_{j=1}^m 1_{A_j}(x) \frac{p_j}{h} - p_\star(x) \right)^2.$$

The integrated mean square error is then given by

$$\begin{aligned}
\overline{\mathscr{E}} &= \sum_{j=1}^m \frac{p_j(1-p_j)}{mnh^2} + \sum_{j=1}^m \int_0^1 1_{A_j}(x) \left(\frac{p_j}{h} - p_\star(x) \right)^2 dx, \\
&= \sum_{j=1}^m \frac{p_j(1-p_j)}{mn} + \int_0^1 p_\star^2(x)dx + \sum_{j=1}^m \left\{ \frac{p_j^2}{h} - 2\frac{p_j^2}{h} \right\}, \\
&= \int_0^1 p_\star^2(x)dx + \sum_{j=1}^m \frac{p_j(1-p_j)}{mn} - \sum_{j=1}^m \frac{p_j^2}{h}.
\end{aligned}$$

■

2.2 Kernel based estimators

The empirical distribution function associated with $(X_i)_{1 \leqslant i \leqslant n}$ is given by

$$F_n : x \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leqslant x} .$$

It is straightforward to show that F_n is an unbiased and consistent estimator of the true distribution function $F_\star : x \mapsto \int_{-\infty}^{x} p_\star(u)\mathrm{d}u$. Numerical differentiation may the be used to obtain an intuitive estimate of $p_\star$. Therefore, choosing $h_n > 0$ allows to introduce

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

where $K : x \mapsto \mathbb{1}_{-1<x\leqslant 1}/2$. This intuitive estimate motivates the analysis of kernel based density estimators with several *kernel* functions K satisfying a few assumptions defined more precisely in this section.

2.2.1 Mean square error when $p_\star$ has bounded second derivative

In this section, $p_\star$ is estimated by:

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

where $K : \mathbb{R} \to \mathbb{R}$ is such that $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$ and where $h_n > 0$.

Proposition 2.2 *Assume that K is such that $\int_{\mathbb{R}} K^2(u)\,\mathrm{d}u < \infty$ and $\int_{\mathbb{R}} u^2 |K(u)|\,\mathrm{d}u < \infty$ and that $p_\star$ is twice continuously differentiable with a bounded second derivative. Then, there exist c_1 and c_2 such that for all x,*

$$\mathscr{E}(x) = \left|\mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x)\right|^2 + \mathbb{V}[\widehat{p}_n^{h_n}(x)] \leqslant \frac{c_1}{nh_n} + c_2 h_n^4 .$$

PROOF. As the $(X_i)_{1\leqslant i\leqslant n}$ are i.i.d.,

$$\begin{aligned}
\mathbb{V}[\widehat{p}_n^{h_n}(x)] = \frac{1}{nh_n^2}\mathbb{V}\left[K\left(\frac{X_1 - x}{h_n}\right)\right] &\leqslant \frac{1}{nh_n^2}\mathbb{E}\left[K^2\left(\frac{X_1 - x}{h_n}\right)\right] \\
&\leqslant \frac{1}{nh_n^2}\int_{\mathbb{R}} K^2\left(\frac{u-x}{h_n}\right) p_\star(u)\mathrm{d}u , \\
&\leqslant \frac{1}{nh_n}\int_{\mathbb{R}} K^2(u)\, p_\star(x+uh_n)\mathrm{d}u , \\
&\leqslant \frac{1}{nh_n}\|p_\star\|_\infty \int_{\mathbb{R}} K^2(u)\,\mathrm{d}u .
\end{aligned}$$

In addition,

$$\begin{aligned}
\mathbb{E}[\widehat{p}_n^{h_n}(x)] &= \frac{1}{h_n}\mathbb{E}\left[K\left(\frac{X_1 - x}{h_n}\right)\right] , \\
&= \frac{1}{h_n}\int_{\mathbb{R}} K\left(\frac{u-x}{h_n}\right) p_\star(u)\mathrm{d}u , \\
&= \int_{\mathbb{R}} K(u)\, p_\star(x+uh_n)\mathrm{d}u .
\end{aligned}$$

Therefore, the bias may be written:

$$\mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x) = \int_{\mathbb{R}} K(u)\,(p_\star(x+uh_n) - p_\star(x))\mathrm{d}u .$$

For all $u \in \mathbb{R}$, there exists $\xi_{u,x}$ between x and $x+uh_n$ such that

$$p_\star(x+uh_n) - p_\star(x) = uh_n p'_\star(x) + \frac{u^2 h_n^2}{2} p''_\star(\xi_{u,x})\,.$$

Using that $\int_{\mathbb{R}} uK(u)\,\mathrm{d}u = 0$ yields

$$\mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x) = \int_{\mathbb{R}} K(u)\,\frac{u^2 h_n^2}{2} p''_\star(\xi_{u,x})\mathrm{d}u$$

and

$$\left|\mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x)\right| = \frac{h_n^2}{2}\int_{\mathbb{R}} u^2\,|K(u)|\,\frac{u^2 h_n^2}{2} p''_\star(\xi_{u,x})\mathrm{d}u \leqslant \frac{h_n^2}{2}\|p''_\star\|_\infty \int_{\mathbb{R}} u^2\,|K(u)|\,\mathrm{d}u\,.$$

■

Therefore, there exist c_1 and c_2 such that for all x,

$$\mathscr{E}(x) = \left|\mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x)\right|^2 + \mathbb{V}[\widehat{p}_n^{h_n}(x)] \leqslant \frac{c_1}{nh_n} + c_2 h_n^4\,.$$

Minimizing this upper bound with respect to h_n, yields a solution h_n proportional to $n^{-1/5}$ and a uniform upper bound for $\mathscr{E}$ or order $n^{-4/5}$.

2.2.2 Mean square error when $p_\star$ belongs to a Hölder class of functions

The bias of the estimator may be controlled with more general assumptions in particular on the function $p_\star$.

Definition 2.3. Let $\beta > 0$ and $\eta > 0$. A function $f : \mathbb{R} \to \mathbb{R}$ is said to be Hölder with parameters β and η if $f^{\lfloor\beta\rfloor}$ exists and for all $(x,y) \in \mathbb{R}^2$,

$$\left|f^{\lfloor\beta\rfloor}(y) - f^{\lfloor\beta\rfloor}(x)\right| \leqslant \eta\,|y-x|^{\beta-\lfloor\beta\rfloor}\,.$$

Definition 2.4. A function $K : \mathbb{R} \to \mathbb{R}$ is said to be a kernel of order $m \geqslant 1$ if for all $0 \leqslant j \leqslant m$, $\int |u|^j K(u)\mathrm{d}u < \infty$ and if $\int K(u)\mathrm{d}u = 1$ and for $1 \leqslant j \leqslant m$, $\int u^j K(u)\mathrm{d}u = 0$.

Proposition 2.5 *Let $\beta > 0$ and $\eta > 0$. Assume that K is a kernel of order $\lfloor\beta\rfloor$ such that $\int_{\mathbb{R}} K^2(u)\,du < \infty$ and $\int_{\mathbb{R}} |u|^\beta |K(u)|\,\mathrm{d}u < \infty$ and that $p_\star$ is Hölder with parameters β and η. Then, there exist c_1 and c_2 such that for all x,*

$$\mathscr{E}(x) = \left|\mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x)\right|^2 + \mathbb{V}[\widehat{p}_n^{h_n}(x)] \leqslant \frac{c_1}{nh_n} + c_2 h_n^{2\beta}\,.$$

PROOF. The control of the variance follows the same lines as the proof of Proposition 2.2. In addition,

$$\mathbb{E}[\widehat{p}_n^{h_n}(x)] = \frac{1}{h_n}\mathbb{E}\left[K\left(\frac{X_1 - x}{h_n}\right)\right] = \frac{1}{h_n}\int_{\mathbb{R}} K\left(\frac{u-x}{h_n}\right) p_\star(u)\mathrm{d}u = \int_{\mathbb{R}} K(u)\,p_\star(x+uh_n)\mathrm{d}u\,.$$

Therefore, the bias may be written:

$$\mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x) = \int_{\mathbb{R}} K(u)\,(p_\star(x+uh_n) - p_\star(x))\mathrm{d}u\,.$$

For all $u \in \mathbb{R}$, there exists $\xi_{u,x}$ between x and $x+uh_n$ such that

$$p_\star(x+uh_n) - p_\star(x) = \sum_{j=1}^{\lfloor\beta\rfloor-1} \frac{(uh_n)^j}{j!} p_\star^{(j)}(x) + \frac{(uh_n)^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!} p_\star^{(\lfloor\beta\rfloor)}(\xi_{u,x}) .$$

Using that $\int_{\mathbb{R}} u^j K(u)\,\mathrm{d}u = 0$ for $1 \leqslant j \leqslant \lfloor\beta\rfloor$ yields

$$\mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x) = \int_{\mathbb{R}} K(u) \frac{(uh_n)^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!} p_\star^{(\lfloor\beta\rfloor)}(\xi_{u,x})\mathrm{d}u = \int_{\mathbb{R}} K(u) \frac{(uh_n)^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!} \left(p_\star^{(\lfloor\beta\rfloor)}(\xi_{u,x}) - p_\star^{(\lfloor\beta\rfloor)}(x) \right) \mathrm{d}u$$

and

$$\left| \mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x) \right| \leqslant \eta \int_{\mathbb{R}} |K(u)| \frac{|uh_n|^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!} |x - \xi_{u,x}|^{\beta-\lfloor\beta\rfloor}\,\mathrm{d}u \leqslant \eta \int_{\mathbb{R}} |K(u)| \frac{|uh_n|^{\beta}}{\lfloor\beta\rfloor!}\mathrm{d}u .$$

■

Therefore, there exist c_1 and c_2 such that for all x,

$$\mathscr{E}(x) = \left| \mathbb{E}[\widehat{p}_n^{h_n}(x)] - p_\star(x) \right|^2 + \mathbb{V}[\widehat{p}_n^{h_n}(x)] \leqslant \frac{c_1}{nh_n} + c_2 h_n^{2\beta} .$$

Minimizing this upper bound with respect to h_n, yields a uniform upper bound for $\mathscr{E}$ or order $n^{-1/(2\beta+1)}$.

2.2.3 Choice of the bandwith h_n

The estimator highly depends on the choice of h_n whose optimal value depends on unknown quantities. A common way to choose h_n automatically is to minimize an estimator of $\overline{\mathscr{E}}$. Note that

$$\overline{\mathscr{E}} = \mathbb{E}\left[\int_{\mathbb{R}} \left(\widehat{p}_n^h(x) - p_\star(x) \right)^2 \mathrm{d}x \right] = \int_{\mathbb{R}} p_\star^2(x)\mathrm{d}x + \mathbb{E}\left[\int_{\mathbb{R}} (\widehat{p}_n^h(x))^2 \mathrm{d}x \right] - 2\int_0^1 p_\star(x)\mathbb{E}\left[\widehat{p}_n^h(x) \right] \mathrm{d}x .$$

Minimizing $\overline{\mathscr{E}}$ as a function of h is equivalent to minimizing

$$h \mapsto \mathbb{E}\left[\int_{\mathbb{R}} (\widehat{p}_n^h(x))^2 \mathrm{d}x \right] - 2\int_{\mathbb{R}} p_\star(x)\mathbb{E}\left[\widehat{p}_n^h(x) \right] \mathrm{d}x .$$

This quantity is not available explicitly but may be estimated unbiasedly.

Proposition 2.6 *For all $h < 0$, an unbiased estimator of $\overline{\mathscr{E}} - \int_{\mathbb{R}} p_\star^2(x)\mathrm{d}x$ is given by*

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

PROOF. It is straightforward that $\int_{\mathbb{R}} (\widehat{p}_n^h(x))^2 \mathrm{d}x$ is an unbiased estimate of $\mathbb{E}\left[\int_{\mathbb{R}} (\widehat{p}_n^h(x))^2 \mathrm{d}x \right]$. Then, as for all $1 \leqslant i \neq j \leqslant n$, the law of (X_i, X_j) has probability distribution function $(x,y) \mapsto p_\star(x)p_\star(y)$,

$$\frac{2}{n(n-1)h} \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1, j\neq i}^n K\left(\frac{X_i - X_j}{h} \right) \right] = \frac{2}{h} \int_{\mathbb{R}^2} K\left(\frac{x-y}{h} \right) p_\star(x)p_\star(y)\mathrm{d}x\mathrm{d}y = 2\int_{\mathbb{R}} p_\star(x) \left(\frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x-y}{h} \right) p_\star(y) dy \right) \mathrm{d}x .$$

The proof is completed by noting that

$$\frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x-y}{h} \right) p_\star(y)\mathrm{d}y = \mathbb{E}\left[\widehat{p}_n^h(x) \right] .$$

■

Chapter 3

Supervised classification

Contents

Keywords 3.1 *Bayes classifier, empirical risk, oracle inequality, linear discriminant analysis, logistic regression, support vector machines.*

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 this introduction, the focus is set on supervised learning with M classes so that an individual is associated with a label in $\{1,\dots,M\}$. The state space $\mathscr{X}$ is usually a subset of $\mathbb{R}^d$ an element of $\mathscr{X}$ contains all the features the label prediction has to be based on. The prediciton rule relies on a training data set $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ where for all $1\leqslant i\leqslant n$, $X_i\in\mathscr{X}$ is an individual which has been associated with the class $Y_i\in\{1,\dots,M\}$. Using these training examples, supervised learning aims at designing a *classifier* $h:\mathscr{X}\to\{1,\dots,M\}$ employed to predict the label of new individuals. The source codes in R and/or Python of this chapter may be found at:

$$\text{https}://\text{/sylvainlc.github.io/project/teaching/}$$

3.1 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 probability of misclassification of h is

$$L_{\text{miss}}(h)=\mathbb{P}\left(Y\neq h(X)\right)\ .$$

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 3.1 *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

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

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

$$L_{\text{miss}}(h) = \mathbb{P}(Y \neq h(X)) = (1 - \mathbb{E}[Yh(X)])/2\,. \tag{3.1}$$

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

$$\mathbb{E}[Yh(X)] = \mathbb{E}[\mathbb{E}[Y|X]h(X)] \leq \mathbb{E}[|\mathbb{E}[Y|X]|\underbrace{|h(X)|}_{=1}] = \mathbb{E}[\mathbb{E}[Y|X]\underbrace{\operatorname{sgn}(\mathbb{E}[Y|X])}_{h_\star(X)}] = \mathbb{E}[Yh_\star(X)]$$

Plugging this into (3.1) yields $L_{\text{miss}}(h) \geq L_{\text{miss}}(h_\star)$, 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 3.2. 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 L_{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 $L_{\text{miss}}(\widehat{h}_n) - L_{\text{miss}}(h_\star)$ may be controlled.

3.2 Parametric and semiparametric classifiers

3.2.1 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,

$$\mathbb{P}(Y=1|X)=\sigma\left(X^T\omega+b\right)\ , \tag{3.2}$$

where

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

Since

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

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

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

In this case, the Bayes classifier is given by

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

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

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

By Lemma 7.4, the gradient of $\log\mathbb{P}_\theta\left(X_{1:n},Y_{1:n}\right)$ with respect to θ is therefore given by

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

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

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

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.

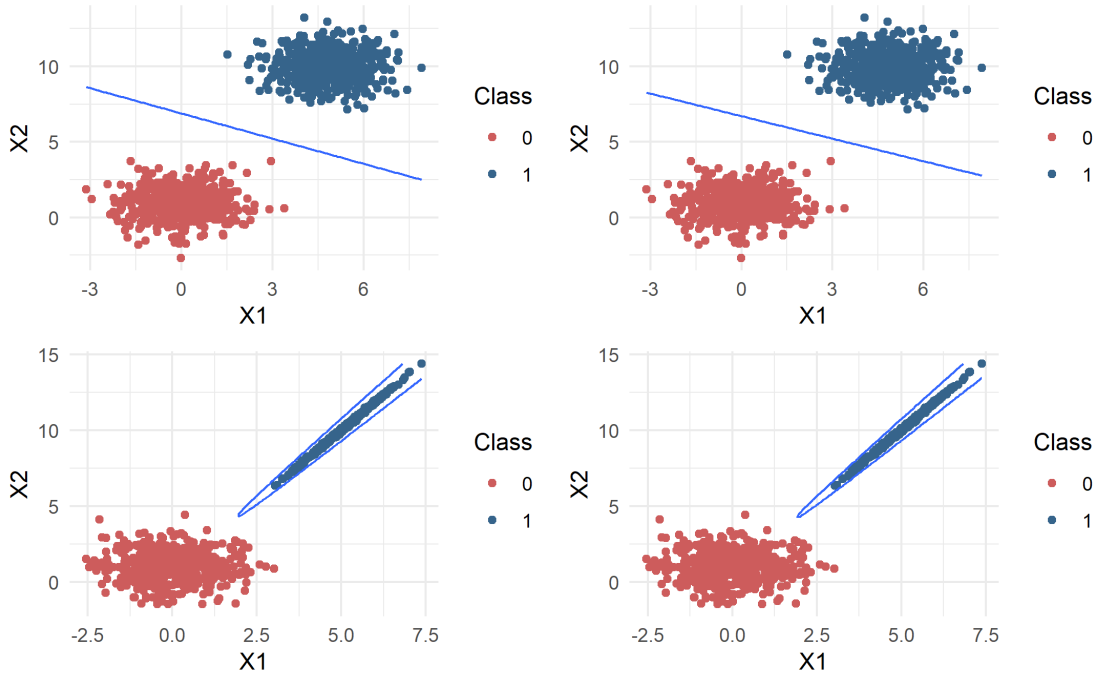


Fig. 3.1 (Top) linear discriminant analysis when both classes are parameterized by the same covariance matrix with the true parameters (left) and maximum likelihood estimates (right). (Bottom) quadratic classification frontier when classes are parameterized by difference covariance matrices with the true parameters (left) and maximum likelihood estimates (right).

3.2.2 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 3.1 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

$$\mathbb{P}(Y=1|X)=\sigma(\alpha+\beta^T X)\,, \tag{3.3}$$

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:

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

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:

$$(\widehat{\alpha}_n,\widehat{\beta}_n)\in\underset{\theta=(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^d}{\arg\max}\ \log\mathbb{P}_\theta(Y_{1:n}|X_{1:n})\,, \tag{3.4}$$

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.

Chapter 4

Feed Forward and convolutional neural networks

Contents

Keywords 4.1 *Multi-layer Perceptron, Feed Forward Neural Networks, Convolution*

4.1 Feed Forward Neural Networks - Multi layer perceptron

4.1.1 Classification: loss function and gradient

Following the logistic regression approach, 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

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

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

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

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

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

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 the output $\sigma(\omega^T X + b)$ in $(0,1)$ is $\mathbb{P}(Y = 1|X)$. 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 $\mathbb{P}(Y = k|X)$.

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:

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

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

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

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

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

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* $1 \leqslant m \leqslant L-1$*,*

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

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

$$\begin{aligned}
\nabla_{z^L(\mathbf{X})} \ell_\theta(\mathbf{X},Y) &= \ell_\theta(\mathbf{X},Y) - \mathbb{1}_Y \ , \\
\nabla_{h_\theta^m(\mathbf{X})} \ell_\theta(\mathbf{X},Y) &= (W^{m+1})^T \nabla_{z_\theta^{m+1}(\mathbf{X})} \ell_\theta(\mathbf{X},Y) \ , \\
\nabla_{z_\theta^m(\mathbf{X})} \ell_\theta(\mathbf{X},Y) &= \nabla_{h_\theta^m(\mathbf{X})} \ell_\theta(X,Y) \odot \varphi_m'(z_\theta^m(\mathbf{X})) \ ,
\end{aligned}$$

where $\odot$ *is the elementwise multiplication.*

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

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

Therefore,

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

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

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

Therefore,

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

Similarly, for all $1 \leqslant i \leqslant M$, using that $z_\theta^L(\mathbf{X}) = b^L + W^L h_\theta^{L-1}(\mathbf{X})$,

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

Therefore,

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

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 $z_\theta^{m+1}(\mathbf{X}) = b^{m+1}+W^{m+1}h_\theta^m(\mathbf{X})$,

$$\begin{aligned}\partial_{(h_\theta^m(\mathbf{X}))_j}\ell_\theta(X,Y) &= \sum_{i=1}^{d_{m+1}}\partial_{(z_\theta^{m+1}(\mathbf{X}))_i}\ell_\theta(X,Y)\partial_{(h_\theta^m(\mathbf{X}))_j}(z_\theta^{m+1}(\mathbf{X}))_i\ ,\\ &= \sum_{i=1}^{d_{m+1}}\partial_{(z_\theta^{m+1}(\mathbf{X}))_i}\ell_\theta(X,Y)W_{i,j}^{m+1}\ .\end{aligned}$$

Therefore,

$$\nabla_{h_\theta^m(\mathbf{X})}\ell_\theta(\mathbf{X},Y) = (W^{m+1})^T\nabla_{z_\theta^{m+1}(\mathbf{X})}\ell_\theta(\mathbf{X},Y)\ .$$

Then, for all $1\leqslant m\leqslant L-1$ and all $1\leqslant j\leqslant d_m$, , using that $h_\theta^m(\mathbf{X})_j = \varphi_m(z_\theta^m(\mathbf{X})_j)$,

$$\begin{aligned}\partial_{(z_\theta^m(\mathbf{X}))_j}\ell_\theta(X,Y) &= \sum_{i=1}^{d_{m+1}}\partial_{(h_\theta^m(\mathbf{X}))_i}\ell_\theta(X,Y)\partial_{(z_\theta^m(\mathbf{X}))_j}(h_\theta^m(\mathbf{X}))_i\ ,\\ &= \sum_{i=1}^{d_{m+1}}\partial_{(h_\theta^m(\mathbf{X}))_i}\ell_\theta(X,Y)\mathbb{1}_{i=j}\varphi_m'(z_\theta^m(\mathbf{X})_i)\ ,\\ &= \partial_{(h_\theta^m(\mathbf{X}))_j}\ell_\theta(X,Y)\varphi_m'(z_\theta^m(\mathbf{X})_j)\ .\end{aligned}$$

Therefore,

$$\nabla_{z_\theta^m(\mathbf{X})}\ell_\theta(\mathbf{X},Y) = \nabla_{h_\theta^m(\mathbf{X})}\ell_\theta(X,Y)\odot\varphi_m'(z_\theta^m(\mathbf{X}))\ .$$

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

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

Therefore,

$$\nabla_{W^m}\ell_\theta(\mathbf{X},Y) = \nabla_{z_\theta^m(\mathbf{X})}\ell_\theta(X,Y)(h_\theta^{m-1}(\mathbf{X}))^T\ .$$

Similarly, for all $1\leqslant i\leqslant d_m$, using that $z_\theta^m(\mathbf{X}) = b^m+W^m h_\theta^{m-1}(\mathbf{X})$,

$$\begin{aligned}\partial_{b_i^m}\ell_\theta(X,Y) &= \sum_{k=1}^{d_m}\partial_{(z_\theta^m(\mathbf{X}))_k}\ell_\theta(X,Y)\partial_{b_i^m}(z_\theta^m(\mathbf{X}))_k\ ,\\ &= \sum_{k=1}^{d_m}\partial_{(z_\theta^m(\mathbf{X}))_k}\ell_\theta(X,Y)\mathbb{1}_{i=k}\ ,\\ &= \partial_{(z_\theta^m(\mathbf{X}))_k}\ell_\theta(X,Y)_i\ .\end{aligned}$$

Therefore,

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

■

These Feed Forward Neural Networks may be used to perform classification on the MNIST dataset. This dataset contains images representing handwritten digits. Each image is made of 28 x 28 pixels, and each pixel is represented by an integer (gray level). These arrays can be flattened into vec-

tors in $\mathbb{R}^{784}$. Visualisations of this vector space are given here: `http://colah.github.io/posts/2014-10-Visualizing-MNIST/`. The labels in $\{0,\dots,9\}$ are represented using one-hot-encoding and grayscale of each pixel in $\{0,\dots,255\}$ are normalized to be in $(0,1)$.

```
from keras.datasets import mnist
# Number of classes
num_classes = 10
# input image dimensions
img_rows, img_cols = 28, 28

# the data, shuffled and split between train and test sets
(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_train     = x_train.reshape(x_train.shape[0], img_rows, img_cols, 1)
x_test      = x_test.reshape(x_test.shape[0], img_rows, img_cols, 1)
input_shape = (img_rows, img_cols, 1)

x_train = x_train.astype('float32')
x_test  = x_test.astype('float32')

print('x_train shape:', x_train.shape)
print('x_test shape:', x_test.shape)
print('y_train shape:', y_train.shape)
print('y_test shape:', y_test.shape)
print(x_train.shape[0], 'train samples')
print(x_test.shape[0], 'test samples')
```

```
x_train shape: (60000, 28, 28, 1)
x_test shape: (10000, 28, 28, 1)
y_train shape: (60000,)
y_test shape: (10000,)
60000 train samples
10000 test samples
```

```
plt.figure(figsize=(8, 2))
for i in range(4):
    plt.subplot(1, 4, i+1)
    plt.imshow(x_train[i].reshape(28, 28),
               interpolation="none", cmap="gray_r")
    plt.title('Label = %d' % y_train[i], fontsize=14)
    plt.axis("off")
plt.tight_layout()
```

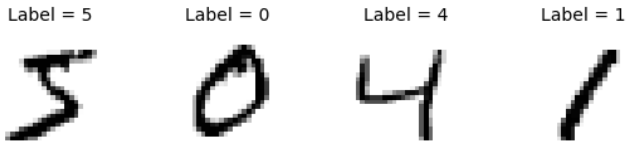


Fig. 4.1 MNIST dataset.

This models relies on more than 100.000 unknown parameters which should be estimated. As for the logistic regression and the discriminant analysis, a common choice is to minimize the negative loglikelihood of the data:

$$\theta \mapsto -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{10} \mathbb{1}_{Y_i=k} \log \mathbb{P}_\theta(Y_i = k|\mathbf{X}_i)\,.$$

The negative loglikelihood is computed using $n = 60.000$ training samples and minimized using gradient descent algorithms. Then, the performance of the model is assessed using 10.000 new (test) samples: the accuracy is the frequency of labels which are well predicted by the model with the estimated parameters.

```
model_ffnn = Sequential()

model_ffnn.add(Flatten(input_shape=input_shape))

model_ffnn.add(Dense(128, activation='relu'))

model_ffnn.add(Dense(num_classes, activation='softmax'))
```

```
model_ffnn.compile(
    loss=keras.losses.categorical_crossentropy,
    optimizer=keras.optimizers.Adagrad(),
    metrics=['accuracy']
)

model_ffnn.summary()
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
flatten_1 (Flatten)          (None, 784)               0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               100480
_________________________________________________________________
dense_2 (Dense)              (None, 10)                1290
=================================================================
Total params: 101,770
Trainable params: 101,770
Non-trainable params: 0
_________________________________________________________________
```

Fig. 4.2 Feed Forward Neural network. h_1 is obtained with the RELU activation function and is in $\mathbb{R}^{128}$. The last layer is $h_2 \in \mathbb{R}^{10}$ and is obtained with the softmax activation function so that each component k models $\mathbb{P}(Y = k|X)$. This neural network with one hidden layer relies on 101.770 parameters.

4.1.2 Regression: loss function and gradient

In a regression setting, it is assumed that the observations satisfy for all $1 \leqslant i \leqslant n$, $Y_i = f_\star(\mathbf{X}_i) + \varepsilon_i$ 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 $f_\star(\mathbf{X}_i)$ (i.e. Y_i) as follows.

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

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:

$$\ell_n : \theta \mapsto \frac{1}{n}\sum_{i=1}^{n} \|f_\theta(\mathbf{X}_i) - Y_i\|^2 \,.$$

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

Proposition 4.2 (Back propagation - regression) *Write* $\ell_\theta(\mathbf{X}, Y) = \|f_\theta(\mathbf{X}) - Y\|^2$ *so that*

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

```
batch_size = 32
epochs = 8

# Run the train
history = model_ffnn.fit(x_train, y_train,
                         batch_size=batch_size,
                         epochs=epochs,
                         verbose=1,
                         validation_data=(x_test, y_test))
score = model_ffnn.evaluate(x_test, y_test, verbose=0)
print('Test loss:', score[0])
print('Test accuracy:', score[1])
```

```
plt.figure(figsize=(5, 4))
plt.plot(history.epoch, history.history['acc'], lw = 1, label='Training')
plt.plot(history.epoch, history.history['val_acc'], lw = 1, label='Testing')
plt.legend()
plt.title('Accuracy of softmax regression', fontsize=16)
plt.xlabel('Epoch', fontsize=14)
plt.ylabel('Accuracy', fontsize=14)
plt.tick_params(labelright=True)
plt.grid('True')
plt.tight_layout()
```

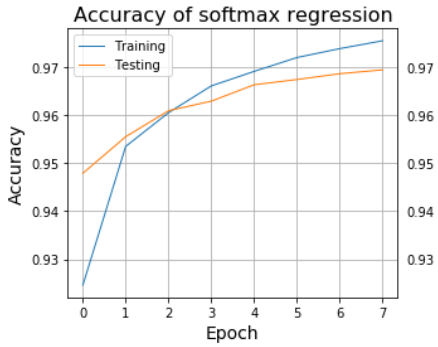


Fig. 4.3 Minimization of the negative liglikelihood using a gradient descent algorithm (here AdaGrad). The gradient is computed using batches of 32 observations and the whole data set is used 8 times.

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

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

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

$$\nabla_{W^m}\ell_\theta(\mathbf{X},Y) = \nabla_{z_\theta^m(\mathbf{X})}\ell_\theta(X,Y)(h_\theta^{m-1}(\mathbf{X}))^T \,,$$
$$\nabla_{b^m}\ell_\theta(\mathbf{X},Y) = \nabla_{z_\theta^m(\mathbf{X})}\ell_\theta(X,Y) \,,$$

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

$$\nabla_{h_\theta^m(\mathbf{X})}\ell_\theta(\mathbf{X},Y) = (W^{m+1})^T\nabla_{z_\theta^{m+1}(\mathbf{X})}\ell_\theta(\mathbf{X},Y) \,,$$
$$\nabla_{z_\theta^m(\mathbf{X})}\ell_\theta(\mathbf{X},Y) = \nabla_{h_\theta^m(\mathbf{X})}\ell_\theta(X,Y) \odot \varphi_m'(z_\theta^m(\mathbf{X})) \,,$$

where $\odot$ *is the elementwise multiplication.*

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

$$\partial_{W^L_{i,j}} \ell_\theta(X,Y) = -2(Y_i - (f_\theta(\mathbf{X}))_i) h^{L-1}_\theta(\mathbf{X})_j \ ,$$

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

$$\nabla_{W^L} \ell_\theta(\mathbf{X},Y) = -2(Y - f_\theta(\mathbf{X}))(h^{L-1}_\theta(\mathbf{X}))^T \ .$$

Similarly,

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

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

Chapter 5

Stochastic gradient descent

Contents

Keywords 5.1

Supervised learning applications and nonparametric regression settings are usually based on the minimization of an objective function on $\mathbb{R}^d$, see for instance Proposition **??** for kernel based SVM algorithms, Exercise 8.13 for penalized kernel based regression or (3.4) for the maximum likelihood estimation of the parameters of a logistic regression application. In these cases, the optimization problem to be solved may be written

$$\widehat{\theta}_n \in \underset{\theta\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \theta'\Phi(X_i)) + \lambda\,\omega(\theta) \right\}, \tag{5.1}$$

where $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ are i.i.d., ℓ is a loss function, Φ is a known function, $\lambda > 0$ and ω is a penalization function. In this framework, the first term if the empirical loss function associated with the expected risk

$$R(\theta) = \mathbb{E}[\ell(Y, \theta'\Phi(X))]\,,$$

which is usually not available explicitly. This chapter introduces widely used gradient descent algorithms to minize (5.1). In the following sections, gradient descent algorithms are used to obtain a sequence $(x_k)_{k\geqslant 0}$ to approximate $x_\star = \arg\min_{x\in\mathbb{R}} f(x)$ (resp. $x_\star = \arg\min_{x\in\mathbb{R}^d} f(x)$) for a convex funxtion $f : \mathbb{R} \to \mathbb{R}$ (resp. $f : \mathbb{R}^d \to \mathbb{R}$).

5.1 Gentle introduction, minimization of a convex Lipschitz function on $\mathbb{R}$

Definition 5.1. Let I be an interval of $\mathbb{R}$ and $f : I \to \mathbb{R}$.The function f is said to be convex if and only if, for all $(x,y) \in I^2$ and all $\lambda \in [0,1]$,

$$f(\lambda x+(1-\lambda)y) \leqslant \lambda f(x)+(1-\lambda)f(y)\,.$$

If $-f$ is convex, f is concave. f is said to be **strictly** convex if the inequality if stict when $x \neq y$ and $\lambda \in (0,1)$.

Proposition 5.2 *Let I be an interval of $\mathbb{R}$ and $f : I \to \mathbb{R}$. Then, f is convex if and only if for all $x_0 \in I$, the function φ_{x_0} defined on $I \setminus \{x_0\}$ by*

$$\varphi_{x_0} : x \mapsto \frac{f(x)-f(x_0)}{x-x_0}$$

is nondecreasing.

PROOF. Assume that f is convex. Let $x_0 \in I$ and choose $a < b$ in I. Assume for instance that $a < x_0 < b$ (other cases are dealt with similarly). Write

$$x_0 = \frac{b-x_0}{b-a}a+\frac{x_0-a}{b-a}b$$

so that, by convexity of f,

$$f(x_0) \leqslant \frac{b-x_0}{b-a}f(a)+\frac{x_0-a}{b-a}f(b)\,.$$

Then, using

$$f(x_0) = \frac{b-x_0}{b-a}f(x_0)+\frac{x_0-a}{b-a}f(x_0)\,,$$

yields

$$\frac{b-x_0}{b-a}(f(a)-f(x_0))+\frac{x_0-a}{b-a}(f(b)-f(x_0)) \geqslant 0$$

and

$$\frac{f(a)-f(x_0)}{x_0-a}+\frac{f(b)-f(x_0)}{b-x_0} \geqslant 0\,.$$

Therefore, $\varphi_{x_0}(a) \leqslant \varphi_{x_0}(b)$, which concludes the proof. Assume now that for all $x_0 \in I$, φ_{x_0} is nondecreasing. Let a and b in I be such that $a < b$, choose $\lambda \in (0,1)$ and write $x_0 = \lambda a+(1-\lambda)b$. Following the same steps as above yields $f(x_0) \leqslant \lambda f(a)+(1-\lambda)f(b)$ and f is convex. ■

Theorem 5.3. *Let I be an interval of $\mathbb{R}$ and $f : I \to \mathbb{R}$ a differentiable function on I. Then, f is convex if and only if f' is nondecreasing. Equivalently, f is convex if and only if, for all $(x,y) \in \mathbb{R}^2$,*

$$f(y) \geqslant f(x)+f'(x)(y-x)\,.$$

PROOF. Assume that f is convex. Then, for all $a < x < b$ in I, by Proposition 5.2,

$$\frac{f(x)-f(a)}{x-a} \leqslant \frac{f(b)-f(a)}{b-a} \leqslant \frac{f(b)-f(x)}{b-x}\,.$$

Taking the limit of the left hand term when x tends to a and of the right hand term when x tends tends b yields $f'(a) \leqslant f'(b)$. f' is therefore nondecreasing. Assume now that f' is nondecreasing and let $a < x_0 < b$ be in I. By the mean value theorem, tere exist $\alpha_1 \in (a,x_0)$ and $\alpha_2 \in (x_0,b)$ such that

$$f(x_0)-f(a) = f'(\alpha_1)(x_0-a) \quad \text{et} \quad f(b)-f(x_0) = f'(\alpha_2)(b-x_0)\,.$$

As f' is nondecreasing,

$$\frac{f(x_0)-f(a)}{x_0-a} \leqslant \frac{f(b)-f(x_0)}{b-x_0}$$

and f is convex by Proposition 5.2. ■

Corollary 5.4 *Let I be an interval of $\mathbb{R}$ and $f : I \to \mathbb{R}$ be a twice differentiable function on I. Then, f is convex if and only if for all $x \in I$, $f''(x) \geqslant 0$.*

Consider a function $f : \mathbb{R} \to \mathbb{R}$ convex, differentiable and such that the derivative of f is Lipschitz: there exists $L \in \mathbb{R}$ such that for all $x \in \mathbb{R}$,

$$\left|f'(x) - f'(y)\right| \leqslant L|x-y| \ . \tag{5.2}$$

Note first that for al $(x,y) \in \mathbb{R}^2$,

$$f(x) + f'(x)(y-x) \leqslant f(y) \leqslant f(x) + f'(x)(y-x) + \frac{L}{2}(x-y)^2 \ .$$

The first inequality is a direct consequence of Theorem 5.3.To prove the second inequality, define, for all $(x,y) \in I^2$, the function $\varphi_{x,y}$ on $[0,1]$ by

$$\varphi_{x,y} : h \mapsto f((1-h)x + hy) \ .$$

This function is differentiable and for all $h \in [0,1]$,

$$\varphi'_{x,y}(h) = (y-x)f'((1-h)x + hy) \ .$$

Then,

$$\begin{aligned} f(y) - f(x) = \varphi_{x,y}(1) - \varphi_{x,y}(0) &= \int_0^1 (y-x)f'((1-h)x+hy)\mathrm{d}h \ , \\ &= \int_0^1 (y-x)\left\{f'((1-h)x+hy) - f'(x)\right\}\mathrm{d}h + (y-x)f'(x) \ , \\ &\leqslant (y-x)f'(x) + (y-x)^2 L \int_0^1 h\mathrm{d}h \ , \\ &\leqslant (y-x)f'(x) + \frac{L}{2}(y-x)^2 \ . \end{aligned}$$

Define the sequence $(x_n)_{n\geqslant 0}$ by choosing $x_0 \in \mathbb{R}$ and setting, for all $n \in \mathbb{N}$,

$$x_{n+1} = x_n - \frac{1}{L}f'(x_n) \ .$$

As for all $y \in \mathbb{R}$,

$$\begin{aligned} f(y) &\leqslant f(x_n) + (y-x_n)f'(x_n) + \frac{L}{2}(y-x_n)^2 \ , \\ &\leqslant f(x_n) - \frac{1}{2L}|f'(x_n)|^2 + \frac{L}{2}\left|x_n - y - \frac{1}{L}f'(x_n)\right|^2 \ , \end{aligned}$$

the following upper bound holds

$$f(x_{n+1}) \leqslant f(x_n) - \frac{1}{2L}|f'(x_n)|^2 \ . \tag{5.3}$$

On the other hand, for all $n \in \mathbb{N}$, $f(x_n) + f'(x_n)(x_\star - x_n) \leqslant f(x_\star)$, then $f(x_n) - f(x_\star) \leqslant f'(x_n)(x_n - x_\star)$ and, for all $n \in \mathbb{N}$ such that $x_n \neq x_\star$,

$$\left|f'(x_n)\right| \geqslant \frac{f(x_n)-f(x_\star)}{|x_n-x_\star|} .$$

Plugging this inequality in (5.3) yields, for all $n \in \mathbb{N}$ such that $x_n \neq x_\star$,

$$f(x_{n+1})-f(x_\star) \leqslant f(x_n)-f(x_\star)-\frac{1}{2L}\frac{(f(x_n)-f(x_\star))^2}{|x_n-x_\star|^2} .$$

Proposition 5.5 *For all $n \in \mathbb{N}^\star$,*

$$f(x_n)-f(x_\star) \leqslant \frac{2L}{n}|x_0-x_\star|^2 .$$

PROOF. Note that for all $n \in \mathbb{N}$, by definition of x_{n+1},

$$\begin{aligned} |x_{n+1}-x_\star|^2 &= |x_n-x_\star|^2-\frac{2}{L}f'(x_n)(x_n-x_\star)+\frac{1}{L^2}(f'(x_n))^2 , \\ &\leqslant |x_n-x_\star|^2-\frac{2}{L}(f(x_n)-f(x_\star))+\frac{2}{L}(f(x_n)-f(x_{n+1})) , \\ &\leqslant |x_n-x_\star|^2 , \end{aligned}$$

by (5.3) and by (5.2), $f'(x_n)(x_n-x_\star) \geqslant f(x_n)-f(x_\star)$. The sequence $(|x_n-x_\star|)_{n\in\mathbb{N}}$ is then nonincreasing and for all $n \in \mathbb{N}$ such that $x_n \neq x_\star$,

$$f(x_{n+1})-f(x_\star) \leqslant f(x_n)-f(x_\star)-\frac{1}{2L}\frac{(f(x_n)-f(x_\star))^2}{|x_0-x_\star|^2} .$$

Using that for all x, $1-x \leqslant (1+x)^{-1}$,

$$\begin{aligned} f(x_{n+1})-f(x_\star) &\leqslant \{f(x_n)-f(x_\star)\}\left(1-\frac{f(x_n)-f(x_\star)}{2L|x_0-x_\star|^2}\right) , \\ &\leqslant \{f(x_n)-f(x_\star)\}\left(1+\frac{f(x_n)-f(x_\star)}{2L|x_0-x_\star|^2}\right)^{-1} . \end{aligned}$$

Writing, for all $n \in \mathbb{N}$, $\alpha_n = (f(x_{n+1})-f(x_\star))^{-1}$,

$$\alpha_n \geqslant \alpha_{n-1}+\frac{1}{2L|x_0-x_\star|^2} \geqslant \ldots \geqslant \alpha_0+\frac{n}{2L|x_0-x_\star|^2} \geqslant \frac{n}{2L|x_0-x_\star|^2} .$$

Then, for all $n \in \mathbb{N}$,

$$f(x_{n+1})-f(x_\star) \leqslant \frac{2L|x_0-x_\star|^2}{n} .$$

■

5.2 Gradient descent on $\mathbb{R}^d$

In the following, f is assumed to be a convex differentiable function on $\mathbb{R}^d$.

Proposition 5.6 (L-smooth function) *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is a convex differentiable function such that ∇f is L-Lipschitz: for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$,*

$$\|\nabla f(x)-\nabla f(y)\| \leqslant L\|x-y\| .$$

Then, for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$,

i)
$$f(x)+\langle\nabla f(x),y-x\rangle\leqslant f(y)\leqslant f(x)+\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|y-x\|^2\,.$$

ii)
$$f(y)-f(x)\leqslant\langle\nabla f(y),y-x\rangle-\frac{1}{2L}\|\nabla f(y)-\nabla f(x)\|^2\,.$$

iii)
$$\langle\nabla f(y)-\nabla f(x),y-x\rangle\geqslant\frac{1}{L}\|\nabla f(y)-\nabla f(x)\|^2\,.$$

PROOF.

i) For all $(x,h)\in\mathbb{R}^d\times\mathbb{R}^d$ the function $\varphi_{x,h}$ defined on $[0,1]$ by
$$\varphi_{x,h}:t\mapsto f(x+th)$$
is differentiable and convex. Then, $\varphi'_{x,h}(0)\leqslant\varphi_{x,h}(1)-\varphi_{x,h}(0)$. As for all $t\in[0,1]$, $\varphi'_{x,h}(t)=\langle\nabla f(x+th),h\rangle$,
$$f(y)-f(x)\geqslant\langle\nabla f(x),y-x\rangle\,. \tag{5.4}$$
The other inequality follows exactly the same steps as the proof for convex functions defined on $\mathbb{R}$.

ii) For all x,y,z in $\mathbb{R}^d$, by convexity and L-smoothness,
$$f(y)-f(x)\leqslant\langle\nabla f(y),y-z\rangle+\langle\nabla f(x),z-x\rangle+\frac{L}{2}\|z-x\|^2\,.$$
Then, choosing $z=x-(\nabla f(x)-\nabla f(y))/L$ yields
$$\begin{aligned}f(y)-f(x)&\leqslant\langle\nabla f(y),y-z\rangle+\langle\nabla f(x),z-x\rangle+\frac{L}{2}\|z-x\|^2\,,\\&\leqslant\langle\nabla f(y),y-x\rangle+L^{-1}\langle\nabla f(y),\nabla f(x)-\nabla f(y)\rangle-L^{-1}\langle\nabla f(x),\nabla f(x)-\nabla f(y)\rangle+\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2\,,\\&\leqslant\langle\nabla f(y),y-x\rangle-\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2\,,\end{aligned}$$
which concludes the proof.

iii) By ii), for all x,y in $\mathbb{R}^d$,
$$f(y)-f(x)\leqslant\langle\nabla f(y),y-x\rangle-\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2$$
and
$$f(x)-f(y)\leqslant\langle\nabla f(x),x-y\rangle-\frac{1}{2L}\|\nabla f(x)-\nabla f(y)\|^2\,.$$
The proof is completed by summing these two inequalities.

■

Gradient descent algorithm on $\mathbb{R}^d$ follows the same steps as in the scalar case.

Choosing $x_0\in\mathbb{R}^d$ and $\eta>0$, define, for all $n\in\mathbb{N}$,
$$x_{n+1}=x_n-\eta\nabla f(x_n)\,.$$

Proposition 5.7 *Let f be a convex differentiable function on $\mathbb{R}^d$ such that ∇f is L-Lipschitz. Assume that $\eta=1/L$, then for all $n\in\mathbb{N}^\star$,*
$$f(x_n)-f(x_\star)\leqslant\frac{2L\|x_0-x_*\|^2}{n+4}\,.$$

PROOF. First, for all $n \geqslant 1$, by Proposition 5.6, if $\eta \leqslant 2/L$,

$$\begin{aligned}\|x_n - x_*\|^2 = \|x_{n-1} - x_* - \eta \nabla f(x_{n-1})\|^2 &\leqslant \|x_{n-1} - x_*\|^2 + \eta^2 \|\nabla f(x_{n-1})\|^2 - 2\eta \langle x_{n-1} - x_*, \nabla f(x_{n-1}) \rangle \ , \\ &\leqslant \|x_{n-1} - x_*\|^2 + \eta^2 \|\nabla f(x_{n-1})\|^2 - 2\frac{\eta}{L} \|\nabla f(x_{n-1})\|^2 \ , \\ &\leqslant \|x_{n-1} - x_*\|^2 \ , \\ &\leqslant \|x_0 - x_*\|^2 \ .\end{aligned}$$

On the other hand, by L-smoothness,

$$f(x_n) - f(x_{n-1}) \leqslant \langle \nabla f(x_{n-1}), -\eta \nabla f(x_{n-1}) \rangle + \frac{L\eta^2}{2} \|\nabla f(x_{n-1})\|^2 \leqslant -\frac{1}{2L} \|\nabla f(x_{n-1})\|^2 \ .$$

Then, by convexity and the Cauchy-Schwarz inequality,

$$f(x_{n-1}) - f(x_*) \leqslant \langle \nabla f(x_{n-1}), x_{n-1} - x_* \rangle \leqslant \|\nabla f(x_{n-1})\| \cdot \|x_{n-1} - x_*\| \ .$$

This yields,

$$\begin{aligned}f(x_n) - f(x_*) \leqslant f(x_{n-1}) - f(x_*) - \frac{1}{2L} \|\nabla f(x_{n-1})\|^2 &\leqslant f(x_{n-1}) - f(x_*) - \frac{(f(x_{n-1}) - f(x_*))^2}{2L\|x_{n-1} - x_*\|^2} \ , \\ &\leqslant f(x_{n-1}) - f(x_*) - \frac{(f(x_{n-1}) - f(x_*))^2}{2L\|x_0 - x_*\|^2} \ .\end{aligned}$$

Hence,

$$\frac{1}{f(x_{n-1}) - f(x_*)} \leqslant \frac{1}{f(x_n) - f(x_*)} - \frac{f(x_{n-1}) - f(x_*)}{2L\|x_0 - x_*\|^2 \, (f(x_n) - f(x_*))} \leqslant \frac{1}{f(x_n) - f(x_*)} - \frac{1}{2L\|x_0 - x_*\|^2}$$

so that

$$\frac{1}{f(x_{n-1}) - f(x_*)} \geqslant \frac{n}{2L\|x_0 - x_*\|^2} + \frac{1}{f(x_0) - f(x_*)}$$

and

$$f(x_{n-1}) - f(x_*) \leqslant \frac{f(x_0) - f(x_*)}{1 + \frac{n(f(x_0) - f(x_*))}{2L\|x_0 - x_*\|^2}} \ .$$

The proof is completed by $f(x_0) - f(x_*) \leqslant L\|x_0 - x_*\|^2/2$. ■

5.3 Gradient descent projected on a bounded convex subset of $\mathbb{R}^d$

Let $\mathscr{C}$ be a convex set of $\mathbb{R}^d$ and $\Pi_{\mathscr{C}}$ the orthogonal projection on $\mathscr{C}$: for all $x \in \mathbb{R}^d$,

$$\Pi_{\mathscr{C}}(x) = \operatorname*{arg\,min}_{u \in \mathscr{C}} \|u - x\|^2 \ .$$

$\mathscr{C}$ is assumed to be bounded, so that there exists $r > 0$ such that $\mathscr{C} \subset \mathscr{B}(0, r)$. In the following, f is assumed to be a convex differentiable function on $\mathbb{R}^d$.

Choosing $x_0 \in \mathscr{C}$ and $\eta > 0$, define, for all $n \in \mathbb{N}$,

$$y_{n+1} = x_n - \eta \nabla f(x_n) \quad \text{and} \quad x_{n+1} = \Pi_{\mathscr{C}}(y_{n+1}) \ .$$

Note that for all $c \in \mathscr{C}$ and all $x \in \mathbb{R}^d$, $\|x - c\| \geqslant \|x - \Pi_{\mathscr{C}}(x)\|$ and for all $\lambda \in [0,1]$, since $\lambda c + (1 - \lambda)\Pi_{\mathscr{C}}(x) \in \mathscr{C}$,

$$\|x - (\lambda c + (1 - \lambda)\Pi_{\mathscr{C}}(x))\|^2 \geqslant \|x - \Pi_{\mathscr{C}}(x)\|^2 \tag{5.5}$$

and

$$\lambda \left\| c - \Pi_{\mathscr{C}}(x) \right\|^2 \geqslant 2 \left\langle x - \Pi_{\mathscr{C}}(x), c - \Pi_{\mathscr{C}}(x) \right\rangle . \tag{5.6}$$

Note that

$$\left\| x - \Pi_{\mathscr{C}}(x) - \lambda \left(c - \Pi_{\mathscr{C}}(x) \right) \right\|^2 = \left\| x - \Pi_{\mathscr{C}}(x) \right\|^2 + \lambda^2 \left\| c - \Pi_{\mathscr{C}}(x) \right\|^2 - 2\lambda \left\langle x - \Pi_{\mathscr{C}}(x), \left(c - \Pi_{\mathscr{C}}(x) \right) \right\rangle .$$

By (5.6), $\left\langle x - \Pi_{\mathscr{C}}(x), c - \Pi_{\mathscr{C}}(x) \right\rangle \leqslant 0$ and

$$\begin{aligned} \left\| x - \Pi_{\mathscr{C}}(x) \right\|^2 + \left\| c - \Pi_{\mathscr{C}}(x) \right\|^2 &= \left\| x - \Pi_{\mathscr{C}}(x) + \Pi_{\mathscr{C}}(x) - c \right\|^2 - 2 \left\langle \Pi_{\mathscr{C}}(x) - x, c - \Pi_{\mathscr{C}}(x) \right\rangle , \\ &= \left\| x - c \right\|^2 + 2 \left\langle c - \Pi_{\mathscr{C}}(x), x - \Pi_{\mathscr{C}}(x) \right\rangle , \\ &\leqslant \left\| x - c \right\|^2 . \end{aligned} \tag{5.7}$$

Proposition 5.8 *Let f be a convex differentiable function on $\mathbb{R}^d$ such that ∇f is L-Lipschitz. Then, for all $n \in \mathbb{N}^\star$, choosing $\eta = r/(L\sqrt{n})$ yields*

$$f\left(\frac{1}{n} \sum_{k=1}^{n} x_k \right) - f(x_\star) \leqslant \frac{rL}{\sqrt{n}} .$$

PROOF. By Proposition 5.6, for all $k \in \mathbb{N}$,

$$f(x_\star) - f(x_k) \geqslant \left\langle \nabla f(x_k), x_\star - x_k \right\rangle .$$

As $\nabla f(x_k) = \eta^{-1}(x_k - y_{k+1})$ and

$$\begin{aligned} f(x_k) - f(x_\star) \leqslant \eta^{-1} \left\langle x_k - y_{k+1}, x_k - x_\star \right\rangle &\leqslant \eta^{-1} \left\langle x_k - y_{k+1}, x_n - y_{k+1} \right\rangle + \eta^{-1} \left\langle x_k - y_{k+1}, y_{k+1} - x_\star \right\rangle , \\ &\leqslant \eta \|\nabla f(x_k)\|^2 + \frac{1}{2\eta} \left(\|x_k - x_\star\|^2 - \|x_k - y_{k+1}\|^2 - \|y_{k+1} - x_\star\|^2 \right) , \\ &\leqslant \frac{\eta}{2} \|\nabla f(x_k)\|^2 + \frac{1}{2\eta} \left(\|x_k - x_\star\|^2 - \|y_{k+1} - x_\star\|^2 \right) . \end{aligned}$$

This yields

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^{n} f(x_k) - f(x_\star) &\leqslant \frac{\eta L^2}{2} + \frac{1}{2\eta n} \sum_{k=1}^{n} \left(\|x_k - x_\star\|^2 - \|y_{k+1} - x_\star\|^2 \right) , \\ &\leqslant \frac{\eta L^2}{2} + \frac{1}{2\eta n} \sum_{k=2}^{n} \left(\|x_k - x_\star\|^2 - \|y_k - x_\star\|^2 \right) + \frac{1}{2\eta n} \left(\|x_1 - x_\star\|^2 - \|y_{n+1} - x_\star\|^2 \right) . \end{aligned}$$

Since $\pi_{\mathscr{C}}(y_{k+1}) = x_{k+1}$, by Proposition 5.6,

$$\|\pi_{\mathscr{C}}(y_{k+1}) - y_{k+1}\|^2 + \|\pi_{\mathscr{C}}(y_{k+1}) - x_\star\|^2 \leqslant \|y_{k+1} - x_\star\|^2$$

and

$$\|x_{k+1} - x_\star\|^2 - \|y_{k+1} - x_\star\|^2 \leqslant -\|\pi_{\mathscr{C}}(y_{k+1}) - y_{k+1}\|^2 \leqslant 0 .$$

Then,

$$\frac{1}{n} \sum_{k=1}^{n} f(x_k) - f(x_\star) \leqslant \frac{\eta L^2}{2} + \frac{\|x_1 - x_\star\|^2}{2\eta n} ,$$

which concludes the proof by convexity of f choosing $\eta = r/(L\sqrt{n})$. ■

Definition 5.9. Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is a convex differentiable function. f is said to be α-strongly convex if and only if for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$f(y) \leqslant f(x) + \langle \nabla f(y), y - x \rangle - \frac{\alpha}{2} \|y - x\|^2 .$$

Proposition 5.10 *Let f be a convex differentiable function on $\mathbb{R}^d$ such that ∇f is L-Lipschitz. Assume that f is α-strongly convex. Then, for all $n \in \mathbb{N}^\star$, choosing $\eta = 1/L$ yields*

$$\|x_{n+1} - x_*\|^2 \leqslant \mathrm{e}^{-\alpha n/L} \|x_1 - x_*\|^2 .$$

PROOF. For all $n \geqslant 0$,

$$\|x_{n+1} - x_*\|^2 = \|x_n - \eta \nabla f(x_n) - x_*\|^2 = \|x_n - x_*\|^2 + \eta^2 \|\nabla f(x_n)\|^2 - 2\eta \langle x_n - x_*, \nabla f(x_n) \rangle$$

By strong convexity,

$$f(x_{n+1}) - f(x_*) \leqslant \langle \nabla f(x_n), x_n - x_* \rangle - \frac{1}{2L} \|\nabla f(x_n)\|^2 - \frac{\alpha}{2} \|x_n - x_*\|^2 ,$$

which yields

$$-2\eta \langle x_n - x_*, \nabla f(x_n) \rangle \leqslant -\eta^2 \|\nabla f(x_n)\|^2 - \alpha\eta \|x_n - x_*\|^2 .$$

Therefore,

$$\|x_{n+1} - x_*\|^2 \leqslant (1 - \alpha\eta) \|x_n - x_*\|^2 \leqslant (1 - \alpha\eta)^n \|x_1 - x_*\|^2 ,$$

which concludes the proof. ■

Proposition 5.11 *Let f be a convex differentiable function on $\mathbb{R}^d$ such that ∇f is L-Lipschitz. Assume that f is α-strongly convex. Then, for all $n \in \mathbb{N}^\star$, choosing $\eta = 2/(\alpha + L)$ yields*

$$f(x_{n+1}) - f(x_*) \leqslant \frac{L}{2} \mathrm{e}^{-4n/(L/\alpha + 1)} \|x_1 - x_*\|^2 .$$

PROOF. Note first that the function $x \mapsto f(x) - \alpha\|x\|^2/2$ is convex as f is α-strongly convex. On the other hand,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geqslant \frac{1}{L - \alpha} \|\nabla f(x) - \nabla f(y)\|^2 .$$

As $\nabla f(x_*) = 0$,

$$f(x_n) - f(x_*) \leqslant \frac{L}{2} \|x_n - x_*\|^2 .$$

For all $n \geqslant 0$,

$$\|x_{n+1} - x_*\|^2 = \|x_n - \eta \nabla f(x_n) - x_*\|^2 = \|x_n - x_*\|^2 + \eta^2 \|\nabla f(x_n)\|^2 - 2\eta \langle x_n - x_*, \nabla f(x_n) \rangle$$

Therefore,

$$\|x_{n+1} - x_*\|^2 \leqslant \|x_n - \eta \nabla f(x_n) - x_*\|^2 = \|x_n - x_*\|^2 + \eta^2 \|\nabla f(x_n)\|^2 - 2\eta \langle x_n - x_*, \nabla f(x_n) \rangle$$

and

$$\|x_{n+1} - x_*\|^2 \leqslant (1 - \alpha\eta) \|x_n - x_*\|^2 \leqslant (1 - \alpha\eta)^n \|x_1 - x_*\|^2 ,$$

which concludes the proof. ■

5.4 Stochastic gradient descent algorithm

The previous algorithm allows to obtain a sequence $(x_k)_{k \geqslant 0}$ to approximate $x_\star = \arg\min_{x \in \mathbb{R}^d} f(x)$. At each time step $k \geqslant 0$, this gradient descent algorithm requires to compute $\nabla f(x_k)$ which may be computationally prohibitive when f is a function based on all the available data (as in Proposition **??**, Exercise 8.13 or

(3.4)). Stochastic gradient descent algorithms offer an appealing alternative by computing an expectedly less costly uunbiased estimate of this gradient.

Let $x_0 \in \mathbb{R}^d$ and $\eta > 0$, define, for all $k \in \mathbb{N}$,

$$x_{k+1} = x_k - \eta\, v_k \ ,$$

where v_k is an unbiased estimate of $\nabla f(x_k)$. The $(v_k)_{k\geqslant 0}$ are assumed to be independent and such that

$$\mathrm{Trace}\left(\mathbb{V}[v_k]\right) = \sigma_k^2 \ .$$

Proposition 5.12 *Let f be a convex differentiable function defined on $\mathbb{R}^d$ such that $\|\nabla f\|$ is L-Lipschitz. Assume that there exists σ^2 such that for all $k \geqslant 0$, $\sigma_k^2 \leqslant \sigma^2$ and that $\eta L \leqslant 1$. Then, for all $n \geqslant 1$,*

$$\mathbb{E}\left[f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)\right] - f(x_\star) \leqslant \frac{1}{2\eta n}\|x_0 - x_\star\|^2 + \eta\sigma^2 \ .$$

Therefore, choosing

$$\eta_n = \frac{\|x_0 - x_\star\|}{\sigma\sqrt{2n}}$$

yields, for n sufficiently large,

$$\mathbb{E}\left[f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)\right] - f(x_\star) \leqslant \frac{\sqrt{2}\sigma\|x_0 - x_\star\|}{\sqrt{n}} \ .$$

PROOF. By Proposition 5.6,

$$f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle \leqslant f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

which yields

$$f(x_{k+1}) \leqslant f(x_k) - \eta\,\langle \nabla f(x_k), v_k\rangle + \frac{L\eta^2}{2}\|v_k\|^2 \ .$$

Therefore, if $\mathscr{F}_k$ denotes the σ-algebra generated by $(v_0, \ldots, v_{k-1})$,

$$\mathbb{E}[f(x_{k+1})|\mathscr{F}_k] \leqslant f(x_k) - \eta\|\nabla f(x_k)\|^2 + \frac{L\eta^2}{2}\left(\|\nabla f(x_k)\|^2 + \sigma_k^2\right)$$

and

$$\mathbb{E}[f(x_{k+1})|\mathscr{F}_k] \leqslant f(x_k) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(x_k)\|^2 + \frac{L\eta^2}{2}\sigma_k^2 \ .$$

Using the fact that $L\eta \leqslant 1$,

$$\mathbb{E}[f(x_{k+1})|\mathscr{F}_k] \leqslant f(x_k) - \frac{\eta}{2}\|\nabla f(x_k)\|^2 + \frac{\eta}{2}\sigma_k^2 \ .$$

and, combined with Proposition 5.6,

$$\mathbb{E}[f(x_{k+1})|\mathscr{F}_k] \leqslant f(x_\star) + \langle \nabla f(x_k), x_k - x_\star\rangle - \frac{\eta}{2}\|\nabla f(x_k)\|^2 + \frac{\eta}{2}\sigma_k^2 \ .$$

Then,

$$\mathbb{E}[f(x_{k+1})|\mathscr{F}_k] \leqslant f(x_\star) + \mathbb{E}\left[\langle v_k, x_k - x_\star\rangle - \frac{\eta}{2}\|v_k\|^2|\mathscr{F}_k\right] + \eta\sigma_k^2$$

which yields

$$\mathbb{E}[f(x_{k+1})|\mathscr{F}_k] \leqslant f(x_\star) + \frac{1}{2\eta}\mathbb{E}\left[\|x_k - x_\star\|^2 - \|x_k - x_\star - \eta v_k\|^2|\mathscr{F}_k\right] + \eta\sigma_k^2 ,$$
$$\leqslant f(x_\star) + \frac{1}{2\eta}\mathbb{E}\left[\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2|\mathscr{F}_k\right] + \eta\sigma_k^2$$

and

$$\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}[f(x_{k+1})] - f(x_\star) \leqslant \frac{1}{2\eta n}\mathbb{E}\left[\|x_0 - x_\star\|^2 - \|x_n - x_\star - \eta v_k\|^2\right] + \eta\sigma^2 .$$

By convexity,

$$\mathbb{E}\left[f\left(\frac{1}{n}\sum_{k=0}^{n-1}x_{k+1}\right)\right] - f(x_\star) \leqslant \frac{1}{2\eta n}\|x_0 - x_\star\|^2 + \eta\sigma^2 ,$$

which concludes the proof. ■

5.5 Optimization methods for neural networks

5.5.1 AdaGrad

The ADAptive GRADient algorithm introduced by (Duchi et al. 2011) starts from $w^{(0)}$ and uses a learning rate $\eta > 0$ and a momentum α and defines, for all $k \geqslant 0$ and all $j \in \{1, \ldots, d\}$,

$$w_j^{(k+1)} \leftarrow w_j^{(k)} - \frac{\eta}{\sqrt{\sum_{\tau=1}^{k}(\nabla f(w^{(\tau)}))_j^2}}(\nabla f(w^{(k)}))_j .$$

The rationale of this method is that different rates are used for all coordinates which is crucial for neural networks in which gradient at different layers can be of different order of magnitude. It is proved in (Ward et al., 2018) that AdaGrad achieves the same convergence rate as gradient descent with optimal fixed stepsize up to a log factor. The adaptive step size grows with the inverse of the gradient magnitudes, so that large gradients have small learning rates and small gradients have large learning rates.

5.5.2 AdaDelta

Introduced in (Zeiler, 2012), was introduced to reduce the sensitivity to initial conditions of AdaGrad. Indeed, if the initial gradients are large, the learning rates of AdaGrad will be low for all updates which can be overcome by increasing η, but making the AdaGrad method highly sensitive to the choice of η.

5.5.3 RMSprop optimizer

Unpublished method, from the course of Geoff Hinton.

5.5.4 ADAM: Adaptive moment estimation

Introduced in (Kingma et al., 2014) and considered as the state of the art to otpimize neural networks, the ADAM procedure update the parameter estimate as follows. Starting from $m_0 = 0$ and $v_0 = 0$ and choosing $\beta_1, \beta_2, \eta, \varepsilon \in (0,1)$, compute first and second moment estimate

$$m_k = \beta_1 m_{k-1} + (1-\beta_1)\nabla f(w^{(k)}) \quad \text{and} \quad v_k = \beta_2 v_{k-1} + (1-\beta_2)(\nabla f(w^{(k)}))^2\,,$$

then, compute the correction terms

$$\hat{m}_k = \frac{m_k}{1-\beta_1^k} \quad \hat{v}_k = \frac{v_k}{1-\beta_2^k}\,,$$

and update the parameter estimate with

$$w^{(k+1)} = w^{(k)} - \frac{\eta}{\sqrt{\hat{v}_k}+\varepsilon}\hat{m}_k\,.$$

First convergence results can be found in (Kingma et al., 2014) and examples where ADAM algorithm does not converge to the optimum are given in (Reddi et al., 2018). Recent analysis by (Barakat et al., 2018).

Chapter 6

Multivariate regression

Contents

Keywords 6.1

6.1 Gaussian vectors

Definition 6.1. A random variable $X \in \mathbb{R}^n$ is a Gaussian vector if and only if, for all $a \in \mathbb{R}^n$, the random variable $\langle a\,;\,X\rangle$ is a Gaussian random variable.

For all random variable $X \in \mathbb{R}^n$, $X \sim \mathcal{N}(\mu, \Sigma)$ 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

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

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 :

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

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

Proposition 6.2 *Let $X \in \mathbb{R}^n$ be a Gaussian vector. Let $\{i_1,\dots,i_d\}$ be a subset of $\{1,\dots,n\}$, $d \geqslant 1$. If for all $1 \leqslant k \neq j \leqslant d$, $\mathrm{Cov}(X_{i_k}, X_{i_j}) = 0$, then $(X_{i_1},\dots,X_{i_d})$ are independent.*

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

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

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

Theorem 6.3 (Cochran). *Let $X \sim \mathcal{N}(0,I_n)$ be a Gaussian vector in $\mathbb{R}^n$, F be a vector subspace of $\mathbb{R}^n$ 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, $\|\pi_F(X)\|^2 \sim \chi^2(p)$ and $\|\pi_{F^\perp}(X)\|^2 \sim \chi^2(n-p)$, where p is the dimension of F.*

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

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

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

$$\begin{pmatrix} \pi_F(X) \\ \pi_{F^\perp}(X) \end{pmatrix} = \begin{pmatrix} U_{(p)}U_{(p)}^T \\ U_{(n-p)}^\perp (U_{(n-p)}^\perp)^T \end{pmatrix} X$$

is a centered Gaussian vector with covariance matrix given by

$$\begin{pmatrix} U_{(p)}U_{(p)}^T & 0 \\ 0 & U_{(n-p)}^\perp (U_{(n-p)}^\perp)^T \end{pmatrix} .$$

By Proposition 6.2, $\pi_F(X)$ and $\pi_{F^\perp}(X)$ are independent. On the other hand,

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

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

6.2 Full rank multivariate regression

6.2.1 Least squares estimator

It is assumed that for all $1 \leqslant i \leqslant n$, $Y_i = X_i^T\beta_\star + \varepsilon_i$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. random variables in $\mathbb{R}^d$, $X_i \in \mathbb{R}^d$ and $\beta_\star$ is an unknown vector in $\mathbb{R}^d$. Let $Y \in \mathbb{R}^d$ (resp. $\varepsilon \in \mathbb{R}^d$) be the random vector such that for all $1 \leqslant i \leqslant n$, the i-th component of Y (resp. ε) is Y_i (resp. ε_i) and $X \in \mathbb{R}^{n\times d}$ the matrix with line i equal to X_i^T. The model is then written

$$Y = X\beta_\star + \varepsilon .$$

In this section, it is assumed that $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon\varepsilon^T] = \sigma_\star^2 I_n$ and that the matrix X has full rank, i.e. the columns of X are linearly independent. The least squares estimate of $\beta_\star$ is defined as a solution to

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

Proposition 6.4 *If the matrix X has full rank, then, $\widehat{\beta}_n = (X^TX)^{-1}X^TY$. This estimator is unbiased and satisfies $\mathbb{V}[\widehat{\beta}_n] = \sigma_\star^2(X^TX)^{-1}$.*

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

$$\|Y - X\beta\|^2 = \|Y\|^2 + \beta^TX^TX\beta + 2Y^TX\beta .$$

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

$$\nabla\ell(\beta) = 2X^TX\beta + 2X^TY .$$

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

$$\widehat{\beta}_n = (X^TX)^{-1}X^TY .$$

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

$$\mathbb{E}[\widehat{\beta}_n] = (X^TX)^{-1}X^T\mathbb{E}[Y] = (X^TX)^{-1}X^TX\beta_\star = \beta_\star .$$

In addition,

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

■

Proposition 6.5 *In the case where $\varepsilon \sim \mathcal{N}(0, \sigma_\star^2 I_n)$, the random variable*

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

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

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

$$\widehat{\sigma}_n^2 = \frac{\|Y - X\widehat{\beta}_n\|^2}{n-d} = \frac{\|Y - X(X^TX)^{-1}X^TY\|^2}{n-d} = \frac{\|(I_n - X(X^TX)^{-1}X^T)Y\|^2}{n-d}$$

The columns of X are linearly independent so that the matrix of the orthogonal projection on Range(X) is $X(X^TX)^{-1}X^T$ and therefore $(I_n - X(X^TX)^{-1}X^T$ is the matrix of the orthogonal projection on Range$(X)^\perp$. Then,

$$(I_n - X(X^TX)^{-1}X^T)Y = (I_n - X(X^TX)^{-1}X^T)(X\beta_\star + \varepsilon) = (I_n - X(X^TX)^{-1}X^T)\varepsilon .$$

By Theorem 6.3, $\|\sigma_\star^{-1}(I_n - X(X^TX)^{-1}X^T)\varepsilon\|^2$ has a χ^2 distribution with $n-d$ degrees of freedom which yields

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

and $\mathbb{E}[\widehat{\sigma}_n^2] = \sigma_\star^2$. By Proposition 6.5, $\mathbb{E}[\widehat{\beta}_n] = \beta_\star$ and $\mathbb{V}[\widehat{\beta}_n] = \sigma_\star^2(X^TX)^{-1}$ and $\widehat{\beta}_n$ is a Gaussian vector as an affine transformation of a Gaussian vector. Note that $(n-d)\widehat{\sigma}_n^2 = \|(I_n - X(X^TX)^{-1}X^T)\varepsilon\|^2$ and $\widehat{\beta}_n = (X^TX)^{-1}X^TX\beta_\star + (X^TX)^{-1}X^T\varepsilon$ and that $(X^TX)^{-1}X^T\varepsilon$ and $(I_n - X(X^TX)^{-1}X^T)\varepsilon$ are not correlated as

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

The independence follows from Proposition 6.2. ■

6.2.2 Confidence intervals and tests

Student's t-statistics

Proposition 6.6 *For all* $1 \leqslant j \leqslant n$,

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

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 $1 \leqslant j \leqslant d$,

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

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

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

In addition,

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

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

By Proposition 6.6, 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

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

Therefore,

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

is a confidence interval for $\beta_{\star,j}$ with confidence level $1-\alpha$. The result of Proposition 6.6 may also be used to perform the test

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

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

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

does not depend on $\beta_\star$ 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 $\beta_\star \neq 0$ is $T_{n,j} < s^{n-d}_{1-\alpha/2}$.

Fisher statistics

Proposition 6.7 *Let L be a $\mathbb{R}^{q\times d}$ matrix with rank $q \leqslant d$. Then,*

$$\frac{(\widehat{\beta}_n-\beta_\star)^T L^T (L(X^TX)^{-1}L^T)^{-1} L(\widehat{\beta}_n-\beta_\star)}{q\widehat{\sigma}_n^2} \sim \mathscr{F}(q,n-d)\ ,$$

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

PROOF. Note that $\operatorname{rank}(L(X^TX)^{-1}L^T)=\operatorname{rank}(LL^T)=q$. The matrix $L(X^TX)^{-1}L^T$ 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 $L(X^TX)^{-1}L^T = QDQ^{-1}$. The matrix $(L(X^TX)^{-1}L^T)^{-1/2}$ may be defined as $(L(X^TX)^{-1}L^T)^{-1/2}=QD^{-1/2}Q^{-1}$. It is then enough to note that $(L(X^TX)^{-1}L^T)^{-1/2}L(\widehat{\beta}_n-\beta_\star)/\sigma_\star\sim\mathscr{N}(0,I_q)$. Therefore,

$$\sigma_\star^{-2}\|(L(X^TX)^{-1}L^T)^{-1/2}L(\widehat{\beta}_n-\beta_\star)\|^2 = (\widehat{\beta}_n-\beta_\star)^TL^T(L(X^TX)^{-1}L^T)^{-1}L(\widehat{\beta}_n-\beta_\star)/\sigma_\star^2\sim\chi^2(q)\ .$$

On the other hand, by Proposition 6.5,

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

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

By Proposition 6.7, for $\alpha\in(0,1)$, if $f_{1-\alpha}^{q,n-d}$ denotes the quantile of order $1-\alpha$ of the law $\mathscr{F}(q,n-p)$, then

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

Therefore,

$$I_n^{q,n-d}(\beta_\star)=\left\{\beta\in\mathbb{R}^d\ ;\ (\widehat{\beta}_n-\beta)^TL^T(L(X^TX)^{-1}L^T)^{-1}L(\widehat{\beta}_n-\beta)\leqslant q\widehat{\sigma}_n^2 f_{1-\alpha}^{q,n-d}\right\}$$

is a confidence region for $\beta_\star$ with confidence level $1-\alpha$. The result of Proposition 6.7 may also be used to perform the test

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

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

6.3 Introduction to regularized multivariate regression

6.3.1 Ridge regression

In the case where $X'X$ 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

$$\widehat{\beta}_n^{\text{ridge}}\in\operatorname*{arg\,min}_{\beta\in\mathbb{R}^d}\|Y-X\beta\|^2+\lambda\|\beta\|^2\ ,$$

where $\lambda>0$. The matrix $X'X+\lambda I_n$ is definite positive for all $\lambda>0$ as for all $u\in\mathbb{R}^d$,

$$u'(X'X+\lambda I_n)u=\|Xu\|^2+\lambda\|u\|^2\ ,$$

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

Proposition 6.8 *The unique solution to the Ridge regression problem is given by*

$$\widehat{\beta}_n^{\text{ridge}} = (X'X + \lambda I_n)^{-1} X'Y \,.$$

This estimator is biased and satisfies

$$\mathbb{E}[\widehat{\beta}_n] - \beta_* = -\lambda (X'X + \lambda I_n)^{-1} \beta_* \,,$$
$$\mathbb{V}[\widehat{\beta}_n] = \sigma_\star^2 (X'X + \lambda I_n)^{-2} X'X \,.$$

PROOF. ■

The mean square error of the estimator is then given by

$$\mathbb{E}\left[\left\|\widehat{\beta}_n - \beta_*\right\|^2\right] = \text{Trace}\left(\mathbb{V}[\widehat{\beta}_n]\right) + \left\|\mathbb{E}[\widehat{\beta}_n] - \beta_*\right\|^2 \,.$$

Let $(\vartheta_1, \ldots, \vartheta_d)$ be an orthonormal basis of $\mathbb{R}^d$ of eigenvectors of $X'X$ associated with the eigenvalues $(\gamma_1, \ldots, \gamma_d) \in \mathbb{R}^d$. Then,

$$\mathbb{E}\left[\left\|\widehat{\beta}_n - \beta_*\right\|^2\right] = \sigma_\star^2 \sum_{j=1}^{d} \frac{\gamma_j}{(\gamma_j + \lambda)^2} + \lambda^2 \sum_{j=1}^{d} \frac{\langle \beta_* \,; \vartheta_j \,\rangle^2}{(\gamma_j + \lambda)^2} \,.$$

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.

6.3.2 Lasso regression

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

$$\widehat{\beta}_n^{\text{lasso}} \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \, \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \,,$$

where $\lambda > 0$ and

$$\|\beta\|_1 = \sum_{j=1}^{d} |\beta_j| \,.$$

The function $\beta \mapsto \|Y - X\beta\|^2 + \lambda \|\beta\|_1$ is convex but not differentiable and the solution to this problem may not be unique. For all $\beta \in \mathbb{R}^d$,

$$\partial_\beta \|Y - X\beta\|^2 = -2X'(Y - X\beta) \,.$$

Then, for all $1 \leqslant j \leqslant d$, $(\partial_\beta \|Y - X\beta\|^2)_j = -2\mathbf{X}_j'(Y - X\beta)$, where $\mathbf{X}_j$ is the j-th column of the matrix X. Define, for all $1 \leqslant j \leqslant d$,

$$\upsilon_j = \mathbf{X}_j' \left(Y - \sum_{\substack{i=1 \\ i \neq j}}^{n} \beta_i \mathbf{X}_i \right) \,,$$

then,

$$(\partial_\beta \|Y - X\beta\|^2)_j = -2(\upsilon_j - \beta_j) .$$

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

$$\partial_j(\|Y - X\beta\|^2 + \lambda\|\beta\|_1) = 2(\beta_j - \upsilon_j + \lambda \operatorname{sign}(\beta_j)/2) .$$

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

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

$$\beta_j^\star = \upsilon_j \left(1 - \frac{\lambda \operatorname{sign}(\beta_j^\star)}{2\upsilon_j}\right) ,$$

  which yields, as $\operatorname{sign}(\beta_j^\star) = \operatorname{sign}(\upsilon_j)$,

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

  and

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

- If $1 - \lambda/(2|\upsilon_j|) < 0$, there is no solution to $\partial_j(\|Y - X\beta\|^2 + \lambda\|\beta\|_1) = 0$ for $\beta_j \neq 0$. Since $\beta_j \mapsto \|Y - X\beta\|^2 + \lambda\|\beta\|_1$ admits a minimum, $\beta_j^\star = 0$.

Therefore,

$$\beta_j^\star = \upsilon_j \left(1 - \frac{\lambda}{2|\upsilon_j|}\right)_+ .$$

An algorithm to approximatively solve the Lasso regression problem proceeds as follows.

6.3.3 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 $(\varphi_1, \ldots, \varphi_M)$ defined on $\mathscr{X}$. Define $\mathscr{F}_\varphi$ as

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

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

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

Let Ψ be the $\mathbb{R}^{M\times n}$ 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 $\alpha = (\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M$ such that,

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

Then, following the same steps as in Section 6.2, in the case where $\Psi'\Psi$ is nonsingular, the least squares estimate is

$$\widehat{f}_n^{\varphi} : x \mapsto \sum_{j=1}^{M} \widehat{\alpha}_{n,j}\varphi_j , \tag{6.1}$$

where

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

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

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

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

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

Proposition 6.9 *Let $W = (w_i(X_j))_{1\leqslant i,j\leqslant n}$ and $\bar{f}^* = (f^*(X_1),\ldots,f^*(X_n))'$. Then,*

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

where $\widehat{f}_n^{\varphi}$ is defined by (6.1)*.*

PROOF. See the exercises. ■

Chapter 7

Technical results

Contents

7.1 Probabilistic inequalities

Theorem 7.1 (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 $t > 0$,*

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

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

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) \leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) . \tag{7.1}$$

Equation (7.1) implies Hoeffding's inequality by noting that $\mathbb{P}(|\sum_{i=1}^n X_i| > t) \leq \mathbb{P}(\sum_{i=1}^n X_i > t) + \mathbb{P}(-\sum_{i=1}^n X_i > t)$ and by applying (7.1) 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,

$$\phi_i(s) \leq \phi_i(0) + \phi_i'(0) + \frac{s^2}{2}\sup_{\alpha\in[0,1]} \phi_i''(\alpha s) \leq \frac{s^2(b_i-a_i)^2}{8} . \tag{7.2}$$

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 (7.1).

■

Lemma 7.2 *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. (7.2) with $b_i = 1-p$ and $a_i = -p$. ■

7.2 Matrix calculus

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

Lemma 7.3 *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,

$$\begin{aligned}
\log\det((1-\lambda)\Sigma+\lambda\Gamma) &= \log\det\left[\Sigma^{1/2}\left((1-\lambda)I+\lambda\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\right)\Sigma^{1/2}\right] \\
&= \log\det\Sigma + \log\det\left((1-\lambda)I+\lambda\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\right) \\
&= \log\det\Sigma + \sum_{i=1}^d \log(1-\lambda+\lambda\mu_i) \\
&\geq \log\det\Sigma + \sum_{i=1}^d (1-\lambda)\underbrace{\log(1)}_{=0} + \lambda\log(\mu_i) := D
\end{aligned}$$

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 7.4 *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 $x \mapsto x'\Sigma x$ defined on $\mathbb{R}^d$ is given by:

$$\partial_x\{x'\Sigma x\} = 2\Sigma x .$$

PROOF.

(i) Recall that for all $i \in \{1,\dots,d\}$ 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 $\Sigma^{-1} = \frac{[\Delta_{j,i}]^T_{1\leq i,j\leq d}}{\det\Sigma}$, we finally get

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

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

(ii) Define $\varphi(x) = x'\Sigma x$. Then, by straightforward algebra, $\varphi(x+h) = \varphi(x) + 2h'\Sigma x + \varphi(h) = \varphi(x) + 2h'\Sigma x + o(||h||)$, 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,

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

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

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

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:

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

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 $x\in\mathscr{X}$,

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

   On the other hand,

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

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

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

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

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

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

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

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

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

Note that

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

Finally,

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

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 $n^{-1}H_n(\beta^\star) \to_{n\to\infty} H(\beta^\star)$,

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

   Therefore, $-\nabla\ell_n(\beta^\star)/\sqrt{n}$ converges in distribution to $Z \sim \mathcal{N}(0, H(\beta^\star))$. 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 $Z \sim \mathcal{N}(0, H(\beta^\star)^{-1})$.

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,

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

   Then,

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

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

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

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

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 $(X_1,\dots,X_n)$ taking values in $\mathbb{R}^p$. The K-means algorithm seeks to minimize over all partitions $C=(C_1,\dots,C_K)$ of $\{1,\dots,n\}$ 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

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

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 $v_a = \text{trace}(cov(X_a))$. Prove that

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

What is the value of $\mathbb{E}[\text{crit}(C)]$ 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 $(X,Y)' \sim \mathcal{N}(0,I_2)$ 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 $(\phi(X_1),\ldots,\phi(X_n))$. 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 $V = \operatorname{span}\{\phi(X_i),\ldots,\phi(X_n)\}$. *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* $(b_1,\ldots,b_n)$. *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 $H_d = \operatorname{span}\{f_1,\ldots,f_d\}$. 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 $(f_1,\ldots,f_d)$ *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

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathscr{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(X_i))^2 + \frac{\lambda}{n} \|f\|_{\mathscr{H}}^2 \right\} \, ,$$

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

$$\widehat{\beta}_n = \operatorname*{arg\,min}_{\beta \in \mathsf{R}^n} \left\{ \|y - K\beta\|^2 + \lambda \beta' K \beta \right\} \, ,$$

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

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in V} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \frac{\lambda}{n} \|f\|_{\mathscr{H}}^2 \right\} .$$

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,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \frac{\lambda}{n}\|f\|^2_{\mathscr{H}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{n}\beta_j k(X_j,X_i))^2 + \frac{\lambda}{n}\langle\sum_{j=1}^{n}\beta_j k(X_j,\cdot),\sum_{i=1}^{n}\beta_i k(X_i,\cdot)\rangle_{\mathscr{H}}\,.$$

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 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 $\widehat{\beta}_n = (K+\lambda I_n)^{-1}Y$,

$$\begin{aligned}\mathbb{V}[K\widehat{\beta}_n] &= K\mathbb{V}\left[(K+\lambda I_n)^{-1}Y\right]K^T\,,\\ &= K(K+\lambda I)^{-1}\mathbb{V}[Y](K+\lambda I_n)^{-1}K\,,\\ &= \sigma^2K^2(K+\lambda I_n)^{-2}\,,\\ &= \sum_{i=1}^{n}\left(\frac{\lambda_i\sigma}{\lambda_i+\lambda}\right)^2 u_iu_i^T\,,\end{aligned}$$

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

$$\begin{aligned}\widehat{\pi}_1^{(t+1)}&=\frac{1}{n}\sum_{i=1}^n\omega_t^i\,,\\\widehat{\mu}_1^{(t+1)}&=\frac{1}{\sum_{i=1}^n\omega_t^i}\sum_{i=1}^n\omega_t^iY_i\,,\\\widehat{\Sigma}^{(t+1)}&=\frac{1}{n}\sum_{i=1}^n\omega_t^i(Y_i-\mu_1)(Y_i-\mu_1)^T+\frac{1}{n}\sum_{i=1}^n(1-\omega_t^i)(Y_i-\mu_{-1})(Y_i-\mu_{-1})^T\,.\end{aligned}$$