Correction du TD : Erreur de prédiction du Lasso – Analyse Discriminante Linéaire (LDA) et Quadratique (QDA)

Agathe Guilloux, Geneviève Robin

17 Décembre 2020

L'objectif de ce TD est de traiter deux sujets relativement indépendants. Dans un premier problème, nous traitons les capacités prédictives du Lasso (Least absolute shrinkage and selection operator); en particulier, nous calculons des bornes supérieures sur l'erreur de prédiction du Lasso, et montrons que celui-ci surpasse dans certains cas la régression linéaire classique. Dans un deuxième problème, nous introduisons une famille de méthodes de classification supervisée : l'Analyse Discriminante Linéaire (LDA) ou Quadratique (QDA).

Problème 1 (Bornes sur l'erreur de prédiction du Lasso)**.** Soit $(X_i, Y_i)_{1\leq i\leq n}$ un n-échantillon i.i.d avec $X_i \in \mathbb{R}^p$ un vecteur de prédicteurs et $Y_i \in \mathbb{R}$ une réponse, pour tout $1 \leq i \leq n$. On note $X \in \mathbb{R}^{n\times p}$ la matrice de design, $Y \in \mathbb{R}^n$ le vecteur de réponses et $\varepsilon \in \mathbb{R}^n$ le vecteur de bruit. On considère le modèle linéaire suivant, avec $\beta^* \in \mathbb{R}^p$ inconnu et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ pour tout $1 \leq i \leq n$:

$$Y_i = X_i^\top \beta^* + \varepsilon_i. \tag{1}$$

1) Dans cette question on suppose $n \geq p$ et la matrice de covariance empirique $X^\top X$ inversible.

   a) Rappeler la formule de l'estimateur des moindres carrés ordinaires $\hat{\beta}^{LS}$.

   Correction : L'estimateur des moindres carrés vérifie

   $$\hat{\beta}^{LS} \in \operatorname{argmin}_{\beta\in\mathbb{R}^p} \|Y - X\beta\|_2^2.$$

   En utilisant la condition d'optimalité du premier ordre, on obtient

   $$\hat{\beta}^{LS} = (X^\top X)^{-1} X^\top Y.$$

   b) Calculer l'erreur moyenne de prédiction $n^{-1}\mathbb{E}[\|X(\hat{\beta}^{LS} - \beta^*)\|_2^2]$.

   Correction : En utilisant la formule $\hat{\beta}^{LS} = (X^\top X)^{-1} X^\top$ ainsi que le modèle $Y = X\beta^* + \varepsilon$, on obtient

   $$X(\hat{\beta}^{LS} - \beta^*) = X(X^\top X)^{-1} X^\top \varepsilon.$$

Il reste à calculer l'espérance $\mathbb{E}[\|X(X^\top X)^{-1}X^\top\varepsilon\|_2^2]$. On définit la matrice $A = X(X^\top X)^{-1}X^\top$; notons que $A^\top A = A$. La matrice A est symétrique et donc diagonalisable dans une base orthonormale. Il existe $P \in \mathbb{R}^{n\times n}$ orthonormale et $\Lambda \in \mathbb{R}^{n\times n}$ diagonale, telles que $A = P^\top \Lambda P$. On pose $\gamma = P\varepsilon$; on note que $\gamma \sim \mathcal{N}(0, \sigma^2 I_p)$. On obtient :

$$\begin{aligned}\mathbb{E}[\|X(X^\top X)^{-1}X^\top\varepsilon\|_2^2] &= \mathbb{E}[\varepsilon^\top A^\top A\varepsilon] = \mathbb{E}[\varepsilon^\top A\varepsilon] \\ &= \mathbb{E}[\gamma^\top\Lambda\gamma] = \sum_{i=1}^{n}\mathbb{E}[\Lambda_{ii}\gamma_i] \\ &= \sigma^2\,\mathrm{Trace}(\Lambda) \\ &= \sigma^2\,\mathrm{Trace}(X(X^\top X)^{-1}X^\top).\end{aligned}$$

On obtient finalement :

$$n^{-1}\mathbb{E}[\|X(\hat\beta^{LS} - \beta^*)\|_2^2 = n^{-1}\sigma^2\,\mathrm{Trace}(X(X^\top X)^{-1}X^\top).$$

c) Préciser la valeur de $n^{-1}\mathbb{E}[\|X(\hat\beta^{LS} - \beta^*)\|_2^2]$ dans le cas d'un design orthogonal, où $X^\top X = I_p$.

Correction : Dans le cas d'un design orthogonal, $X(X^\top X)^{-1}X^\top = I_p$, de sorte que

$$n^{-1}\mathbb{E}[\|X(\hat\beta^{LS} - \beta^*)\|_2^2 = \frac{\sigma^2 p}{n}.$$

d) Que se passe-t-il si $p > n$?

Correction : Lorsque $p > n$, la matrice de covariance empirique $X^\top X$ n'est plus inversible. On ne peut donc pas calculer l'estimateur des moindres carrés de la manière utilisée dans les questions précédentes. Lorsque $p \le n$ avec $p \simeq n$, on peut calculer $\hat\beta^{LS}$, mais l'erreur de prédiction $\frac{\sigma^2 p}{n}$ peut devenir grande.

Dans la suite du problème, on étudie un estimateur parcimonieux de β^*, reposant sur une pénalisation ℓ_1 des moindres carrés. L'estimateur du Lasso est défini comme suit :

$$\hat\beta \in \mathrm{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\}, \tag{2}$$

où $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. Le paramètre $\lambda > 0$ est le paramètre de régularisation, contrôlant la parcimonie de l'estimateur $\hat\beta$. L'objecif du problème est de calculer une borne supérieure sur l'erreur de prédiction du Lasso $n^{-1}\|X(\hat\beta - \beta^*)\|_2^2$, et de la comparer à l'erreur de prédiction des moindres carrés ordinaires. Pour $\beta \in \mathbb{R}^p$, soit

$$\ell_n(\beta, \beta^*) = \frac{1}{n}\|X(\beta - \beta^*)\|_2^2.$$

On va prouver un résultat de la forme suivante :

$$\ell_n(\hat\beta, \beta^*) \le R(\beta^*, \sigma^2, n, p, \delta) \text{ avec probabilité au moins } 1-\delta,\ \delta \in (0,1). \tag{3}$$

Dans l'équation (3), $R(\beta^*, \sigma^2, n, p, \delta)$ est une borne supérieure valide avec grande probabilité (par rapport à la distribution du bruit ε) qui dépend de la dimension du problème p, du nombre d'observations n, de la variance du bruit σ^2, et de la probabilité $1-\delta$ avec laquelle on veut contrôler l'erreur de prédiction $\ell_n(\hat{\beta}, \beta^*)$.

2) En utilisant la définition de $\hat{\beta}$ comme un minimiseur de $\mathcal{F}(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$, montrer que, pour tout $\beta \in \mathbb{R}^p$:

$$\ell_n(\hat{\beta}, \beta^*) \leq \ell_n(\beta, \beta^*) + 4\lambda\|\beta\|_1 + \frac{2}{n}\varepsilon^\top X(\hat{\beta} - \beta) - 2\lambda\left(\|\beta\|_1 + \|\hat{\beta}\|_1\right). \quad (4)$$

Correction : Par définition, $\hat{\beta} \in \operatorname{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$. Donc, pour tout $\beta \in \mathbb{R}^p$:

$$\frac{1}{2n}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

En utilisant $Y = X\beta^* + \varepsilon$, on obtient :

$$\begin{aligned}\frac{1}{2n}\|X(\beta^* - \hat{\beta})\|_2^2 + \frac{1}{2n}\|\varepsilon\|_2^2 + \frac{1}{n}\epsilon^\top(\beta^* - \hat{\beta}) + \lambda\|\hat{\beta}\|_1 \\ \leq \frac{1}{2n}\|X(\beta^* - \beta)\|_2^2 + \frac{1}{2n}\|\varepsilon\|_2^2 + \frac{1}{n}\epsilon^\top(\beta^* - \beta) + \lambda\|\beta\|_1.\end{aligned}$$

De manière équivalente :

$$\ell_n(\hat{\beta}, \beta^*) \leq \ell_n(\beta, \beta^*) + 4\lambda\|\beta\|_1 + \frac{2}{n}\varepsilon^\top X(\hat{\beta} - \beta) - 2\lambda\left(\|\beta\|_1 + \|\hat{\beta}\|_1\right).$$

3) Pour un vecteur $x \in \mathbb{R}^p$, on définit la norme ℓ_∞ de x : $\|x\|_\infty = \max_{1\leq j\leq p} |x_j|$. Montrer que, pour tout $\beta \in \mathbb{R}^p$:

$$\frac{1}{n}\varepsilon^\top X(\hat{\beta} - \beta) - \lambda\left(\|\beta\|_1 + \|\hat{\beta}\|_1\right) \leq \left(\frac{1}{n}\|\varepsilon^\top X\|_\infty - \lambda\right)\left(\|\hat{\beta}\|_1 + \|\beta\|_1\right). \quad (5)$$

Indice : utiliser l'inégalité suivante pour $x, y \in \mathbb{R}^p$, découlant de la dualité des normes ℓ_1 et ℓ_∞ : $x^\top y \leq \|x\|_\infty\|y\|_1$.

Correction : La dualité des normes ℓ_1 et ℓ_∞ implique d'une part :

$$\varepsilon^\top X(\hat{\beta} - \beta) \leq \|\varepsilon^\top X\|_\infty\|\hat{\beta} - \beta\|_1.$$

D'autre part, l'inégalité triangulaire donne $\|\hat{\beta} - \beta\|_1 \leq \|\hat{\beta}\|_1 + \|\beta\|_1$. Finalement,

$$\varepsilon^\top X(\hat{\beta} - \beta) \leq \|\varepsilon^\top X\|_\infty(\|\hat{\beta}\|_1 + \|\beta\|_1),$$

ce qui prouve l'inégalité désirée.

On cherche à présent à montrer que, pour une valeur de λ bien choisie, le membre de droite de l'inégalité (5) est négatif ou nul avec grande probabilité

4) Pour $1 \leq j \leq p$, notons X^j la j-ième colonne de la matrice de design X. Soit $\delta \in (0, 1)$ fixé. On suppose $\|X^j\|_2^2 \leq n$ pour tout $1 \leq j \leq p$, et

$$\lambda = \sigma\sqrt{\frac{2}{n}\ln(p/\delta)}.$$

a) On définit la variable aléatoire $\zeta_j = \frac{\varepsilon^\top X^j}{\sigma\|X^j\|_2}$. Quelle loi suit ζ_j ?

Correction : La variable aléatoire ζ_j est un v.a. Gaussienne. On a $\mathbb{E}[\zeta_j] = \frac{\mathbb{E}[\varepsilon^\top]X^j}{\sigma\|X^j\|_2} = 0$, et

$$Var[\zeta_j] = \mathbb{E}\left[\frac{(\varepsilon^\top X^j)^2}{\sigma^2\|X^j\|_2^2}\right] = \mathbb{E}\left[\frac{(X^j)^\top\varepsilon\varepsilon^\top X^j}{\sigma^2\|X^j\|_2^2}\right] = \frac{(X^j)^\top\sigma^2 I_n X^j}{\sigma^2\|X^j\|_2^2} = 1.$$

Finalement, $\zeta_j \sim \mathcal{N}(0,1)$.

b) Montrer que, pour tout $1 \leq j \leq p$,

$$\mathbb{P}\left(|\varepsilon^\top X^j| > \sigma\sqrt{2n\ln(p/\delta)}\right) \leq \frac{\delta}{p}.$$

Indice : utiliser l'inégalité de concentration Gaussienne $\mathbb{P}(\xi > x) \leq \frac{1}{2}\exp(-x^2/2)$ *pour* $\xi \sim \mathcal{N}(0,1)$.

Correction : Remarquons d'abord que l'hypothèse $\|X^j\|_2^2 \leq n$ implique $\mathbb{P}\left(|\varepsilon^\top X^j| > \sigma\sqrt{2n\ln(p/\delta)}\right) \leq \mathbb{P}\left(|\varepsilon^\top X^j| > \sigma\|X^j\|_2\sqrt{2\ln(p/\delta)}\right)$. Nous avons :

$$\begin{aligned}\mathbb{P}\left(|\varepsilon^\top X^j| > \sigma\|X^j\|_2\sqrt{2\ln(p/\delta)}\right) &= \mathbb{P}\left(\frac{|\varepsilon^\top X^j|}{\sigma\|X^j\|_2} > \sqrt{2\ln(p/\delta)}\right)\\ &\leq \mathbb{P}(|\zeta_j| > \sqrt{2\ln(p/\delta)}) \leq \frac{\delta}{p},\end{aligned}$$

où nous avons utilisé l'inégalité de concentration Gaussienne avec $x = \sqrt{2\ln(p/\delta)}$ pour obtenir la dernière ligne. Finalement,

$$\mathbb{P}\left(|\varepsilon^\top X^j| > \sigma\sqrt{2n\ln(p/\delta)}\right) \leq \frac{\delta}{p}.$$

c) En déduire que $\mathbb{P}\left(n^{-1}\|\varepsilon^\top X\|_\infty > \lambda\right) \leq \delta$.

Correction : D'après la valeur de λ, on a

$$\mathbb{P}\left(n^{-1}\|\varepsilon^\top X\|_\infty > \lambda\right) = \mathbb{P}\left(\|\varepsilon^\top X\|_\infty > \sigma\sqrt{2n\ln(p/\delta)}\right).$$

Par ailleurs,

$$\begin{aligned}\mathbb{P}\left(\|\varepsilon^\top X\|_\infty > \sigma\sqrt{2n\ln(p/\delta)}\right) &= \mathbb{P}\left(\bigcup_{j=1}^{p}\left\{|\varepsilon^\top X_j| > \sigma\sqrt{2n\ln(p/\delta)}\right\}\right)\\ &\leq \sum_{j=1}^{p}\mathbb{P}\left(|\varepsilon^\top X^j| > \sigma\sqrt{2n\ln(p/\delta)}\right) \leq \delta.\end{aligned}$$

5) En utilisant les réponses aux questions précédentes, conclure que, pour $\delta \in (0,1)$ fixé, si $\lambda = \sigma\sqrt{\frac{2}{n}\ln(p/\delta)}$ alors, avec probabilité au moins $1-\delta$,

$$\ell_n(\hat{\beta}, \beta^*) \leq \inf_{\beta\in\mathbb{R}^p}\left\{\ell_n(\beta,\beta^*) + 4\sqrt{2}\sigma\sqrt{\frac{\ln(p/\delta)}{n}}\|\beta\|_1\right\}. \tag{6}$$

Correction : On a montré précédemment que $\|\varepsilon^\top X\|_\infty \leq \sigma\sqrt{2n\ln(p/\delta)}$ avec probabilité au moins $1-\delta$. On en déduit, en utilisant le résultat de la question 3), qu'avec probabilité au moins $1-\delta$:

$$\frac{1}{n}\varepsilon^\top X(\hat{\beta}-\beta) - \lambda\left(\|\beta\|_1 + \|\hat{\beta}\|_1\right) \leq \left(\frac{1}{n}\|\varepsilon^\top X\|_\infty - \lambda\right)\left(\|\hat{\beta}\|_1 + \|\beta\|_1\right) \leq 0.$$

Ainsi, d'après le résultat de la question 2), on obtient que pour tout $\beta \in \mathbb{R}^p$, toujours avec probabilité $1-\delta$, on a :

$$\ell_n(\hat{\beta},\beta^*) \leq \ell_n(\beta,\beta^*) + 4\lambda\|\beta\|_1.$$

Ce résultat étant vrai pour tout $\beta \in \mathbb{R}^p$, par définition de l'infimum, on obtient finalement:

$$\ell_n(\hat{\beta},\beta^*) \leq \inf_{\beta\in\mathbb{R}^p}\left\{\ell_n(\beta,\beta^*) + 4\sqrt{2}\sigma\sqrt{\frac{\ln(p/\delta)}{n}}\|\beta\|_1\right\}.$$

L'inégalité (6) est appelée une "inégalité oracle". En effet, elle compare l'erreur de prédiction de l'estimateur du Lasso, $\ell_n(\hat{\beta},\beta^*)$, à l'erreur de prédiction du meilleur estimateur parcimonieux de β^*. Ce meilleur estimateur, noté $\bar{\beta}$, satisfait :

$$\bar{\beta} \in \mathrm{argmin}_{\beta\in\mathbb{R}^p}\left\{\ell_n(\beta,\beta^*) + 4\sqrt{2}\sigma\sqrt{\frac{\ln(p/\delta)}{n}}\|\beta\|_1\right\}.$$

En pratique, on ne connait pas l'estimateur oracle $\bar{\beta}$, car on ne peut pas calculer la perte $\ell_n(\beta,\beta^*)$ qui dépend de β^*, lui-même inconnu. Cependant, sous certaines conditions sur β^*, on peut obtenir un résultat plus précis.

6) On suppose $\|\beta^*\|_0 = \sum_{j=1}^p 1_{\{|\beta_j|>0\}} \leq s$, et $\|\beta^*\|_\infty \leq a$. Montrer que

$$\ell_n(\hat{\beta},\beta^*) \leq C\sigma\sqrt{\frac{\ln(p/\delta)}{n}}as \tag{7}$$

avec probabilité au moins $1-\delta$, et C une constant numérique que l'on précisera.

Correction : Le résultat de la question précédente implique en particulier que

$$\ell_n(\hat{\beta},\beta^*) \leq \ell_n(\beta^*,\beta^*) + 4\sqrt{2}\sigma\sqrt{\frac{\ln(p/\delta)}{n}}\|\beta^*\|_1.$$

D'une part, $\ell_n(\beta^*,\beta^*) = 0$. D'autre part, d'après les hypothèses, $\|\beta^*\|_1 \leq as$. On obtient donc : $\ell_n(\hat{\beta},\beta^*) \leq 4\lambda as$. En utilisant la valeur donnée pour λ, on a finalement :

$$\ell_n(\hat{\beta},\beta^*) \leq 4\sqrt{2}\sigma\sqrt{\frac{\ln(p/\delta)}{n}}as.$$

7) Comparer la borne supérieure sur l'erreur de prédiction du Lasso obtenue en (7) à l'erreur de prédiction de l'estimateur des moindres carrés calculé à la question 1) et conclure.

Correction : Pour comparer les deux erreurs, on va fixer $\delta = 1/n$. On a d'une part, pour le Lasso, avec probabilité au moins $1 - 1/n$:

$$\ell_n(\hat{\beta}, \beta^*) \leq 4\sqrt{2}\sigma\sqrt{\frac{\ln(p) + \ln(n)}{n}}as.$$

D'autre part, pour les moindres carrés ordinaires,

$$n^{-1}\mathbb{E}[\|X(\hat{\beta}^{LS} - \beta^*)\|_2^2 = \frac{\sigma^2 p}{n}.$$

En fixant la variance du bruit σ et la valeur maximale des coefficients a, on obtient que l'ordre de grandeur de l'erreur de prédiction pour les deux estimateurs est :

$$R^{Lasso}(n, p, s) \leq \frac{s\sqrt{\ln(p) + \ln(n)}}{\sqrt{n}}, \quad R^{LS}(n, p) = \frac{p}{n}.$$

Le ratio entre les deux erreurs est donc de l'ordre :

$$\frac{R^{Lasso}(n, p, s)}{R^{LS}(n, p, s)} \leq \frac{s\sqrt{n}\sqrt{\ln(p) + \ln(n)}}{p}.$$

Lorsque le nombre de paramètres p est grand par rapport $\sqrt{n}$, avec n le nombre d'observations, et lorsque le paramètre β^* est parcimonieux (i.e. contient beaucoup de zéros), l'estimateur du Lasso devient bien meilleur que les moindres carrés en terme d'erreur de prédiction. Par exemple, pour $n = 10,000$, $p = 5,000$, et $s = 5$, on obtient

$$\frac{R^{Lasso}(n, p, s)}{R^{LS}(n, p, s)} \simeq 0.5.$$

Dans ce cas le Lasso améliore la prédiction de 50% par rapport aux moindres carrés.

Remarquons que, sous des hypothèses supplémentaires sur la matrice de design X, on peut montrer que le ratio des deux erreurs est encore plus petit, et de l'ordre de

$$\frac{R^{Lasso}(n, p, s)}{R^{LS}(n, p, s)} \leq \frac{s\sqrt{\ln(p) + \ln(n)}}{p},$$

où l'on a gagné un facteur $\sqrt{n}$. En reprenant l'exemple précédent, on obtient $\frac{R^{Lasso}(n,p,s)}{R^{LS}(n,p,s)} \simeq 0.005$. En d'autres termes, les moindres carrés ont dans ce cas une erreur 200 fois plus grande que le Lasso.

Problème 2 (Analyse discriminante)**.** Dans ce problème, on considère des méthodes de classification supervisée reposant sur une modélisation statistique. Soit $(X_i, Y_i)_{1 \leq i \leq n}$ un n-échantillon i.i.d avec $X_i \in \mathbb{R}^p$ un vecteur de prédicteurs et $Y_i \in \{-1, 1\}$ une réponse binaire, pour tout $1 \leq i \leq n$. On note $X \in \mathbb{R}^{n \times p}$ la matrice de design et $Y \in \mathbb{R}^n$. On suppose le modèle suivant pour $1 \leq i \leq n$:

$$\mathbb{P}(Y_i = k) = \pi_k \text{ et } X_i | Y_i = k \sim \mathcal{N}(\mu_k, \Sigma_k), \ k \in \{-1, 1\}, \tag{8}$$

avec $\mu_k \in \mathbb{R}^p$, $k \in \{-1, 1\}$ deux vecteurs de moyennes et $\Sigma_k \in \mathbb{R}^{p \times p}$, $k \in \{-1, 1\}$, inversibles. On notera $f_{-1}(x)$ et $f_1(x)$ les densités Gaussiennes associées.

1) On suppose dans un premier temps les paramètres du modèles π_k, μ_k, Σ_k, $k \in \{-1, 1\}$ connus. Pour un volume infinitésimal dx autour de $x \in \mathbb{R}^p$, calculer la loi $\mathbb{P}[X \in dx]$.

Correction : Le modèle supposé donne :

$$\mathbb{P}(X_i \in dx | Y_i = 1) = f_1(x), \text{ et } \mathbb{P}(X_i \in dx | Y_i = -1) = f_{-1}(x).$$

D'après la formule

$$\mathbb{P}(X \in dx) = \mathbb{P}(X_i \in dx | Y_i = 1)\mathbb{P}(Y_i = 1) + \mathbb{P}(X_i \in dx | Y_i = -1)\mathbb{P}(Y_i = -1),$$

on obtient :

$$\mathbb{P}(X \in dx) = \pi_1 f_1(x) + \pi_{-1} f_{-1}(x).$$

2) Calculer la probabilité $\mathbb{P}[Y = 1 | X \in dx]$, et proposer un classifieur $h : \mathbb{R}^p \to \{-1, 1\}$. Que peut-on dire de son risque $\mathbb{P}[h(X) \neq Y]$?

Correction : D'après la formule de Bayes, on a :

$$\begin{aligned}\mathbb{P}[Y = 1 | X \in dx] &= \frac{\mathbb{P}[X \in dx | Y = 1]\mathbb{P}[Y = 1]}{\mathbb{P}[X \in dx]} \\ &= \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_{-1} f_{-1}(x)},\end{aligned}$$

et $\mathbb{P}[Y = 1 | X \in dx] \geq 1/2$ si et seulement si $\pi_1 f_1(x) \geq \pi_{-1} f_{-1}(x)$. On en déduit que le classifieur de Bayes, définit par

$$h : x \mapsto \begin{cases} 1 \text{ si } \mathbb{P}[Y = 1 | X \in dx] \geq 1/2 \\ -1 \text{ si } \mathbb{P}[Y = 1 | X \in dx] < 1/2 \end{cases}$$

s'écrit

$$h(x) = \mathbf{1}_{\pi_1 f_1(x) \geq \pi_{-1} f_{-1}(x)} - \mathbf{1}_{\pi_1 f_1(x) < \pi_{-1} f_{-1}(x)}.$$

Ce classifieur minimise le risque $\mathbb{P}[h(X) \neq Y]$ parmi tous les classifieurs.

On va suppose dans un premier temps $\mu_1 \neq \mu_{-1}$ et $\Sigma_1 = \Sigma_{-1} = \Sigma$; il s'agit de l'Analyse Discriminante Linéaire (LDA, cf. Figure 1). On considère le classifieur suivant :

$$h_{LDA}(x) = \mathbf{1}_{\pi_1 f_1(x) > \pi_{-1} f_{-1}(x)} - \mathbf{1}_{\pi_1 f_1(x) \leq \pi_{-1} f_{-1}(x)}.$$

3) Montrer que $h_{LDA}(x) = 1$ si et seulement si x appartient à un demi-espace dont on précisera l'équation.

Correction : $h_{LDA}(x) = 1$ si et seulement si

$$\begin{aligned}&\pi_1 f_1(x) > \pi_{-1} f_{-1}(x) \\ &\Leftrightarrow \log(\pi_1 f_1(x)) > \log(\pi_{-1} f_{-1}(x)) \\ &\Leftrightarrow \log\left(\frac{\pi_1}{\pi_{-1}}\right) \geq \frac{1}{2}\left[(x - \mu_1)^\top \Sigma^{-1}(x - \mu_1) - (x - \mu_{-1})^\top \Sigma^{-1}(x - \mu_{-1})\right] \\ &\Leftrightarrow \log\left(\frac{\pi_1}{\pi_{-1}}\right) \geq (\mu_{-1} - \mu_1)^\top \Sigma^{-1}\left(x - \frac{\mu_1 + \mu_{-1}}{2}\right).\end{aligned}$$

L'équation

$$\log\left(\frac{\pi_1}{\pi_{-1}}\right) \geq (\mu_{-1}-\mu_1)^\top \Sigma^{-1}\left(x - \frac{\mu_1+\mu_{-1}}{2}\right)$$

définit bien un demi-espace dont la frontière est donnée par un hyper-plan.

4) On suppose à présent μ_1, μ_{-1} et Σ inconnus, et la matrice de covariance empirique $X^\top X$ inversible.

   a) Proposer des estimateurs pour les paramètres μ_1, μ_{-1} et Σ.

   Correction : On peut estimer les paramètres μ_1, μ_{-1} et Σ par maximum de vraisemblance (MLE). On obtient les estimateurs suivants :

$$\begin{aligned}
\hat{\mu_1} &= \frac{\sum_{i=1}^n X_i \mathbf{1}_{Y_i=1}}{\sum_{i=1}^n \mathbf{1}_{Y_i=1}},\\
\hat{\mu_{-1}} &= \frac{\sum_{i=1}^n X_i \mathbf{1}_{Y_i=-1}}{\sum_{i=1}^n \mathbf{1}_{Y_i=-1}},\\
\hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^n (X_i-\hat{\mu}_{Y_i})(X_i-\hat{\mu}_{Y_i})^\top.
\end{aligned}$$

   b) Proposer un classifieur $\hat{h} : \mathbb{R}^p \to \{-1,1\}$.

   Correction : À l'aide de ces estimateurs, on peut définir un classifieur

$$\hat{h}(x) = \begin{cases} 1 & \text{si } \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_{-1}}\right) \geq (\hat{\mu}_{-1}-\hat{\mu}_1)^\top \hat{\Sigma}^{-1}\left(x - \frac{\hat{\mu}_1+\hat{\mu}_{-1}}{2}\right) \\ 0 & \text{sinon.} \end{cases}$$

On suppose maintenant $\mu_1 \neq \mu_{-1}$ et $\Sigma_1 \neq \Sigma_{-1}$; il s'agit de l'Analyse Discriminante Quadratique (QDA, cf. Figure 2). On considère le classifieur suivant :

$$h_{QDA}(x) = \mathbf{1}_{\pi_1 f_1(x) > \pi_{-1} f_{-1}(x)} - \mathbf{1}_{\pi_1 f_1(x) \leq \pi_{-1} f_{-1}(x)}.$$

5) On suppose μ_1, μ_{-1}, Σ_1 et Σ_{-1} connus.

   a) Montrer que le classifieur $h_{QDA}(x) = 1$ si et seulement si $Q(x) \geq 0$, où $Q(x)$ est une fonction quadratique que l'on précisera.

   Correction : En faisant un calcul similaire à celui de la question 3), mais dans lequel le terme quadratique en x ne s'annule pas car $\Sigma_1 \neq \Sigma_{-1}$, on obtient que $h_{QDA}(x) = 1$ si et seulement si

$$\begin{aligned}
2\log\left(\frac{\pi_1}{\pi_{-1}}\right) - \log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_{-1})}\right) + \mu_1\Sigma_1^{-1}\mu_1 - \mu_{-1}\Sigma_{-1}^{-1}\mu_{-1} \geq \\
x^\top(\Sigma_1^{-1}-\Sigma_{-1}^{-1})x - 2x^\top(\Sigma_1^{-1}\mu_1 - \Sigma_{-1}^{-1}\mu_{-1})
\end{aligned}$$

   b) Interpréter ce résultat géométriquement.

   Correction : Cela signifie que, dans la QDA, la frontière entre les deux espaces définis par $\{x \in \mathbb{R}^p; h(x) = 1\}$ et $\{x \in \mathbb{R}^p; h(x) = -1\}$ est quadratique – contrairement à la LDA où il s'agit d'une séparation linéaire.

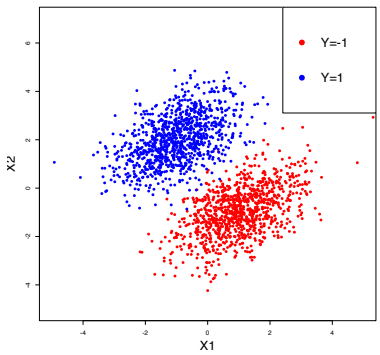

Figure 1: Linear discriminant analysis ($\Sigma_{-1} = \Sigma_1$)

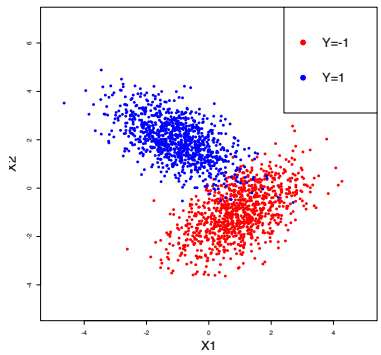

Figure 2: Quadratic discriminant analysis ($\Sigma_{-1} \neq \Sigma_1$)