

1 Résultats

Des explications détaillées des calculs qui mènent aux résultats présentés ici sont données en fin de document.

1.1 Exercice 1 : Modèles graphiques à variables discrètes

Considérons N observations i.i.d et les variables :

$$Z_m^n = \delta(z^n, m) \text{ et } X_k^n = \delta(x^n, k) \\ (n \in \llbracket 1, N \rrbracket, m \in \llbracket 1, M \rrbracket, k \in \llbracket 1, K \rrbracket).$$

La vraisemblance du modèle s'écrit :

$$\begin{aligned} \mathcal{L}(\pi, \theta) &= \prod_{n=1}^N P(x^n = k | z^n = m) P(z^n = m) \\ &= \prod_{n=1}^N \prod_{m=1}^M \prod_{k=1}^K \pi_m^{Z_m^n} \theta_{mk}^{Z_m^n X_k^n} \end{aligned}$$

En utilisant la méthode du Lagrangien, on trouve comme estimateur du maximum de vraisemblance :

$$\boxed{\forall m \in \llbracket 1, M \rrbracket \quad \pi_m^* = \frac{N_m}{N}}$$

$$\boxed{\forall m \in \llbracket 1, M \rrbracket, \forall k \in \llbracket 1, K \rrbracket, \quad \theta_{mk}^* = \frac{N_{mk}}{N_m}}$$

en posant : $\forall m \in \llbracket 1, M \rrbracket, \quad N_m = \sum_{n=1}^N Z_m^n$ et $\forall m \in \llbracket 1, M \rrbracket, \forall k \in \llbracket 1, K \rrbracket, \quad N_{mk} = \sum_{n=1}^N Z_m^n X_k^n$.

1.2 Exercice 2.a : Modèle génératif LDA

On trouve l'estimateur du maximum de vraisemblance $(\pi^*, \mu_1^*, \mu_0^*, \Sigma^*)$:

$$\boxed{\pi^* = \frac{\sum_{i=1}^N y_i}{N}}$$

$$\boxed{\mu_1^* = \frac{\sum_{i=1}^N y_i x_i}{\sum_{i=1}^N y_i}}$$

$$\boxed{\mu_0^* = \frac{\sum_{i=1}^N (1-y_i) x_i}{\sum_{i=1}^N (1-y_i)}}$$

$$\boxed{\Sigma^* = \frac{\sum_{i=1}^N (x_i - \mu_{y_i}^*)(x_i - \mu_{y_i}^*)^T}{N}}$$

où : $\mu_{y_i}^* = y_i \mu_1^* + (1-y_i) \mu_0^*$

Ce en calculant le point d'annulation du gradient de la vraisemblance du modèle donnée par :

$$\begin{aligned} \mathcal{L}(\pi, \mu_1, \mu_0, \Sigma) &= \prod_{i=1}^N \pi^{y_i} (1-\pi)^{(1-y_i)} \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \\ &\times \exp \left[-\frac{1}{2} (x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i}) \right] \end{aligned}$$

où : $\mu_{y_i} = y_i \mu_1 + (1-y_i) \mu_0$

Afin de classifier une nouvelle donnée x , nous montrons également que $P(y=1|x) > P(y=0|x) \iff a + b^T x > 0$, en posant :

$$a = \log \frac{\pi}{1-\pi} - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0$$

$$b = \Sigma^{-1}(\mu_1 - \mu_0)$$

1.3 Exercice 5.a : Modèle génératif QDA

On obtient l'estimateur du maximum de vraisemblance $(\pi^*, \mu_1^*, \mu_0^*, \Sigma_1^*, \Sigma_0^*)$:

$$\boxed{\pi^* = \frac{\sum_{i=1}^N y_i}{N}}$$

$$\boxed{\mu_1^* = \frac{\sum_{i=1}^N y_i x_i}{\sum_{i=1}^N y_i}}$$

$$\boxed{\mu_0^* = \frac{\sum_{i=1}^N (1-y_i) x_i}{\sum_{i=1}^N (1-y_i)}}$$

$$\boxed{\Sigma_1^* = \frac{\sum_{i=1}^N y_i (x_i - \mu_1^*)(x_i - \mu_1^*)^T}{\sum_{i=1}^N y_i}}$$

$$\boxed{\Sigma_0^* = \frac{\sum_{i=1}^N (1-y_i) (x_i - \mu_0^*)(x_i - \mu_0^*)^T}{\sum_{i=1}^N (1-y_i)}}$$

Et $P(y=1|x) > P(y=0|x)$ équivaut à :

$$\log |\Sigma_0| - \log |\Sigma_1| + 2 \log \frac{\pi}{1-\pi}$$

$$-(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + (x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) > 0$$

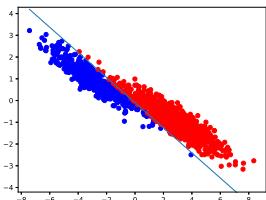
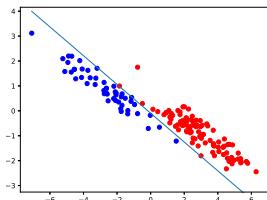
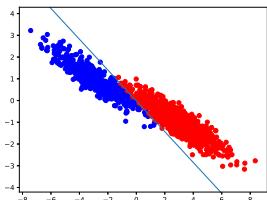
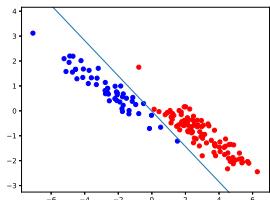


FIGURE 1 – Labels attribués par le modèle LDA aux données de train (gauche) et de test (droite) du jeu A. En bleu la frontière de classification.

FIGURE 2 – Labels attribués par la Régression Logistique aux données de train (gauche) et de test (droite) du jeu A. En bleu la frontière de classification.

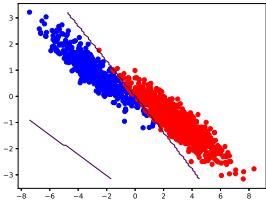
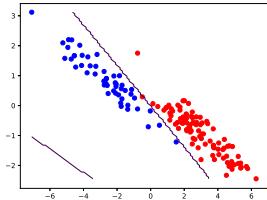
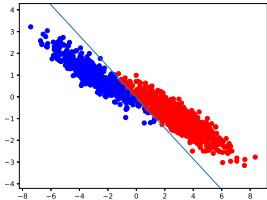
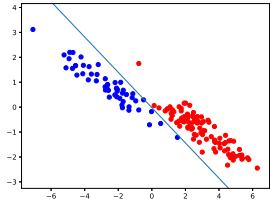


FIGURE 3 – Labels attribués par la Régression Linéaire aux données de train (gauche) et de test (droite) du jeu A. En bleu la frontière de classification.

FIGURE 4 – Labels attribués par le modèle QDA aux données de train (gauche) et de test (droite) du jeu A. En bleu la frontière de classification.

Modèle	Données	Erreur de classification
LDA	Train	0.013333
	Test	0.02
LoR	Train	0.0
	Test	0.034
LiR	Train	0.013333
	Test	0.020667
QDA	Train	0.006667
	Test	0.02

TABLE 1 – Erreur de classification des différents modèles sur le jeu de données A. LDA : Analyse Discriminante Linéaire ; LoR : Régression Logistique ; LiR : Régression Linéaire ; QDA : Analyse Discriminante Quadratique

Sur ce premier jeu de données, on peut tout d'abord noter que les erreurs de classification sur les données de test sont légèrement supérieures à celles sur les données de train sauf pour la Régression Logistique, qui donne une erreur nulle sur le train et les moins bons résultats sur le test. Les trois autres modèles donnent des résultats similaires sur le jeu de train, comme sur celui de test.

- Les données semblent provenir de deux gaussiennes de même matrice de covariance mais de moyennes différentes, ce qui explique les bons résultats du modèle LDA. Le modèle QDA, qui fait des hypothèses plus générales

que LDA (covariances différentes), donc encore cohérentes avec la distribution empirique des données, donne les mêmes résultats que LDA sur l'ensemble de test.

- La Régression Logistique obtient le meilleur résultat sur l'ensemble de train mais le pire sur l'ensemble de test. Contrairement aux autres méthodes, elle ne fait pas d'hypothèse (vérifiée empiriquement) sur la manière dont sont générées les données (et sa solution est obtenue par un algorithme itératif).
- Le modèle génératif LDA et la Régression Linéaire donnent des résultats tout à fait similaires sur le train et le test. Contrairement à QDA, tous deux sont des modèles linéaires, et, contrairement à la Régression Logistique, on peut retrouver l'erreur des moindres carrés de la Régression Linéaire comme la log-vraisemblance d'un modèle génératif dans lequel la variable réelle y (ici restreinte à $\{0,1\}$) suit une distribution gaussienne dont la moyenne est une transformation linéaire de x . Des calculs reposant sur la règle de Bayes seraient probablement nécessaires pour comprendre plus précisément sa relation avec le modèle LDA qui suppose lui que les points x de chaque classe y sont issus d'une distribution gaussienne.

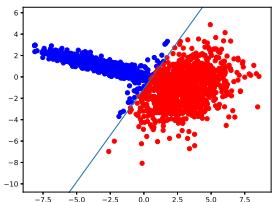
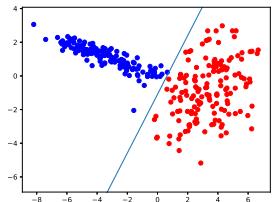
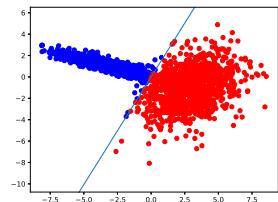
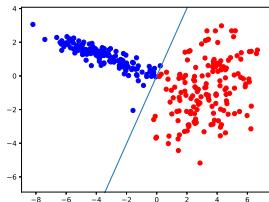


FIGURE 5 – Labels attribués par le modèle LDA aux données de train (gauche) et de test (droite) du jeu B. En bleu la frontière de classification.

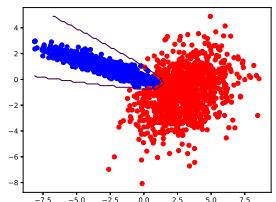
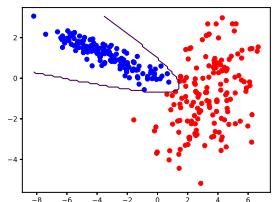
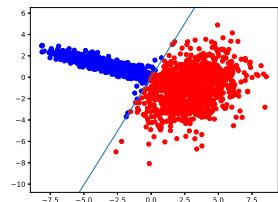
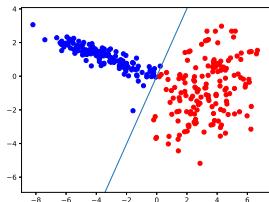


FIGURE 6 – Labels attribués par la Régression Logistique aux données de train (gauche) et de test (droite) du jeu B. En bleu la frontière de classification.

FIGURE 6 – Labels attribués par la Régression Logistique aux données de train (gauche) et de test (droite) du jeu B. En bleu la frontière de classification

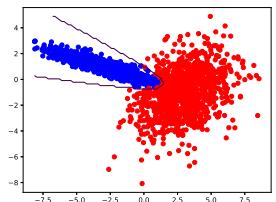
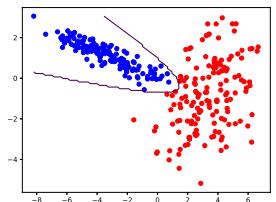
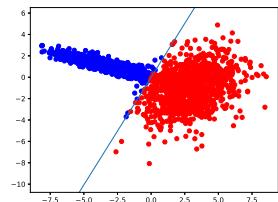
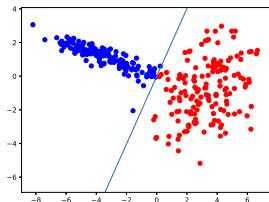


FIGURE 7 – Labels attribués par la Régression Linéaire aux données de train (gauche) et de test (droite) du jeu B. En bleu la frontière de classification.

FIGURE 8 – Labels attribués par le modèle QDA aux données de train (gauche) et de test (droite) du jeu B. En bleu la frontière de classification.

Modèle	Données	Erreur de classification
LDA	Train	0.03
	Test	0.0415
LoR	Train	0.02
	Test	0.0385
LiR	Train	0.03
	Test	0.0415
QDA	Train	0.013333
	Test	0.02

TABLE 2 – Erreur de classification des différents modèles sur le jeu de données B. LDA : Analyse Discriminante Linéaire ; LoR : Régression Logistique ; LiR : Régression Linéaire ; QDA : Analyse Discriminante Quadratique

Sur ce deuxième jeu de données, à l’exception du modèle QDA qui donne les meilleurs résultats à la fois en train et test, les erreurs entre données de train et

de test sont plus marquées qu’avant. QDA est suivi par la Régression Logistique, puis par LDA et la Régression Linéaire qui aboutissent aux mêmes résultats encore une fois.

- Il semble logique que QDA performe le mieux, les données semblant générées à partir de deux gaussiennes de matrices de covariances très différentes.
- LDA suppose au contraire que les gaussiennes ont mêmes matrices de covariance, ce modèle génératif est donc moins adapté ici que la Régression Logistique (laquelle ne fait pas d’hypothèse sur la manière dont sont générées les données).
- Enfin, l’écart plus important entre les erreurs de train et test pour les modèles autres que QDA sont dus au fait que les données de tests ne sont pas linéairement séparables (quand la frontière de QDA est une conique, cf. calculs en page 10).

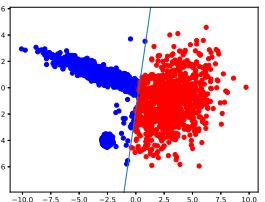
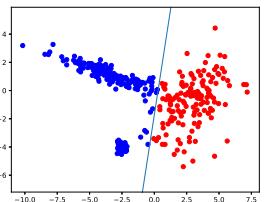
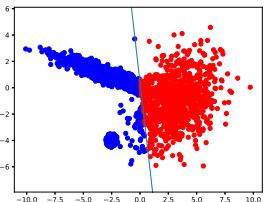
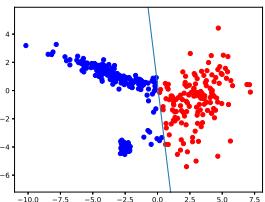


FIGURE 9 – Labels attribués par le modèle LDA aux données de train (gauche) et de test (droite) du jeu C. En bleu la frontière de classification.

FIGURE 10 – Labels attribués par la Régression Logistique aux données de train (gauche) et de test (droite) du jeu C. En bleu la frontière de classification

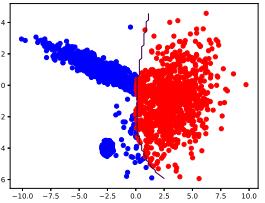
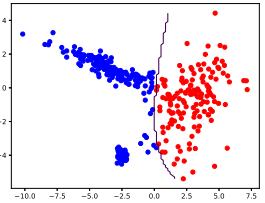
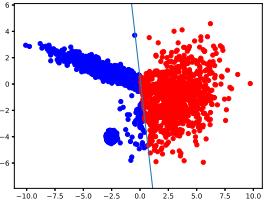
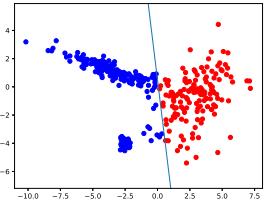


FIGURE 11 – Labels attribués par la Régression Linéaire aux données de train (gauche) et de test (droite) du jeu C. En bleu la frontière de classification.

FIGURE 12 – Labels attribués par le modèle QDA aux données de train (gauche) et de test (droite) du jeu C. En bleu la frontière de classification.

Modèle	Données	Erreur de classification
LDA	Train	0.055
	Test	0.042333
LoR	Train	0.0425
	Test	0.028
LiR	Train	0.055
	Test	0.042333
QDA	Train	0.0525
	Test	0.038333

TABLE 3 – Erreur de classification des différents modèles sur le jeu de données C. LDA : Analyse Discriminante Linéaire ; LoR : Régression Logistique ; LiR : Régression Linéaire ; QDA : Analyse Discriminante Quadratique

Sur ce troisième jeu de données, les erreurs de classification sur les données de test sont cette fois légèrement inférieures à celles sur les données de train. La Régression Logistique offre les meilleurs résultats à la fois en train et en test, le modèle QDA est le deuxième meilleur, le modèle LDA et la Régression Linéaire ferment la marche et conduisent à nouveau aux mêmes erreurs.

- Ici la classe représentée par les points bleus ne semble pas du tout générée depuis une distribution gaussienne (cf. petit agglomérat en bas près de la frontière). Ainsi les modèles génératifs LDA et QDA ne sont pas adaptés pour ce jeu de données. Comme le montrent les résultats, il vaut mieux utiliser une Régression Logistique qui ne fait pas *d'a priori* sur la forme des distributions des données.
- Les erreurs de test sont ici moins grandes que celles sur le train. Cela semble dû à la distribution particulière des données. La classe représentée par les points rouges semble avoir été générée depuis une distribution gaussienne. Le jeu de test étant plus important que celui de train, la proportion de points rouges situés sur la périphérie de la gaussienne est donc plus petite dans le jeu de test que dans le jeu de train. Or les erreurs de classification se situent majoritairement dans cette zone, à cause du petit agglomérat de points bleus qui se situe très près.

2 Preuves plus détaillées

2.1 Exercice 1 : Estimateur du maximum de vraisemblance pour modèles graphiques discrets

Considérons N observations i.i.d et les vecteurs Z^n et X^n ($n \in \llbracket 1, N \rrbracket$) définis comme suit :

$\forall n \in \llbracket 1, N \rrbracket, \forall m \in \llbracket 1, M \rrbracket, \forall k \in \llbracket 1, K \rrbracket$:

$$Z_m^n = \begin{cases} 1 & \text{si pour l'observation } n z = m \\ 0 & \text{sinon.} \end{cases}$$

$$X_k^n = \begin{cases} 1 & \text{si pour l'observation } n x = k \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance du modèle s'écrit :

$$\mathcal{L}(\pi, \theta) = \prod_{n=1}^N P(x^n = k | z^n = m) P(z^n = m) = \prod_{n=1}^N \prod_{m=1}^M \prod_{k=1}^K \pi_m^{Z_m^n} \theta_{mk}^{Z_m^n X_k^n}$$

La log-vraisemblance du modèle est donc :

$$\begin{aligned} l(\pi, \theta) &= \log \mathcal{L}(\pi, \theta) \\ &= \sum_{n=1}^N \sum_{m=1}^M Z_m^n \log \pi_m + \sum_{n=1}^N \sum_{m=1}^M \sum_{k=1}^K Z_m^n X_k^n \log \theta_{mk} \\ &= \sum_{m=1}^M N_m \log \pi_m + \sum_{m=1}^M \sum_{k=1}^K N_{mk} \log \theta_{mk} \end{aligned}$$

en posant : $\forall m \in \llbracket 1, M \rrbracket, N_m = \sum_{n=1}^N Z_m^n$ et $\forall m \in \llbracket 1, M \rrbracket, \forall k \in \llbracket 1, K \rrbracket, N_{mk} = \sum_{n=1}^N Z_m^n X_k^n$.

On cherche donc le maximum de $l(\pi, \theta)$ sous les contraintes :

$$\sum_{m=1}^M \pi_m = 1 \tag{1}$$

$$\forall m \in \llbracket 1, M \rrbracket, \sum_{k=1}^K \theta_{mk} = 1 \tag{2}$$

On note que les contraintes sont affines et que donc les contraintes sont qualifiées. La log-vraisemblance (à maximiser) étant concave, son opposée (à minimiser) est convexe. Donc, s'il existe une solution réalisable, il y a dualité forte par le lemme de Slater.

On introduit λ_π le multiplicateur de Lagrange associé à l'égalité (1) et $\lambda_\theta = (\lambda_{\theta_m})_{m \in \llbracket 1, M \rrbracket}$ les multiplicateurs de Lagrange associés aux égalités (2). On peut alors écrire le Lagrangien $L(\pi, \theta, \lambda_\pi, \lambda_\theta)$:

$$L(\pi, \theta, \lambda_\pi, \lambda_\theta) = \sum_{m=1}^M N_m \log \pi_m + \sum_{m=1}^M \sum_{k=1}^K N_{mk} \log \theta_{mk} + \lambda_\pi \left(\sum_{m=1}^M \pi_m - 1 \right) + \sum_{m=1}^M \lambda_{\theta_m} \left(\sum_{k=1}^K \theta_{mk} - 1 \right)$$

On cherche alors $(\pi^*, \theta^*, \lambda_\pi^*, \lambda_\theta^*)$ tel que $\nabla L(\pi^*, \theta^*, \lambda_\pi^*, \lambda_\theta^*) = 0$

$$\frac{\partial L}{\partial \pi_m}(\pi, \theta, \lambda_\pi, \lambda_\theta) = \frac{N_m}{\pi_m} + \lambda_\pi$$

$$\frac{\partial L}{\partial \theta_{mk}}(\pi, \theta, \lambda_\pi, \lambda_\theta) = \frac{N_{mk}}{\theta_{mk}} + \lambda_{\theta_m}$$

Ainsi on a : $N_m + \lambda_\pi^* \pi_m^* = 0$ et $N_{mk} + \theta_{mk}^* \lambda_{\theta_m}^* = 0$

Les contraintes étant satisfaites, $\sum_{m=1}^M \pi_m^* = 1$ et $\sum_{k=1}^K \theta_{mk}^* = 1$

D'où l'on obtient que l'estimateur du maximum de vraisemblance (MLE) est donné par :

$$\forall m \in \llbracket 1, M \rrbracket \quad \pi_m^* = \frac{N_m}{N}$$
$$\forall m \in \llbracket 1, M \rrbracket, \forall k \in \llbracket 1, K \rrbracket, \quad \theta_{mk}^* = \frac{N_{mk}}{N_m}$$

2.2 Modèle génératif LDA :

Calcul de l'estimateur du maximum de vraisemblance (MLE) :

La vraisemblance du modèle s'écrit :

$$\begin{aligned}\mathcal{L}(\pi, \mu_1, \mu_0, \Sigma) &= \prod_{i=1}^N P(Y = y_i | \pi) P(X = x_i | Y = y_i, \mu_1, \mu_0, \Sigma) \\ &= \prod_{i=1}^N \pi^{y_i} (1 - \pi)^{(1-y_i)} \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i}) \right] \\ \text{où : } \mu_{y_i} &= y_i \mu_1 + (1 - y_i) \mu_0\end{aligned}$$

Ainsi, la log-vraisemblance l est :

$$l(\pi, \mu_1, \mu_0, \Sigma) = \sum_{i=1}^N \left(y_i \log \pi + (1 - y_i) \log (1 - \pi) + \frac{1}{2} \log |\Sigma^{-1}| - \frac{1}{2} [(x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i})] \right) + cste$$

Alors :

$$\begin{aligned}\frac{\partial l}{\partial \pi}(\pi, \mu_1, \mu_0, \Sigma) &= \sum_{i=1}^N y_i \frac{1}{\pi} + \sum_{i=1}^N (1 - y_i) \frac{-1}{1 - \pi} \\ \frac{\partial l}{\partial \mu_1}(\pi, \mu_1, \mu_0, \Sigma) &= \sum_{i=1}^N y_i \Sigma^{-1} (x_i - \mu_1) \\ \frac{\partial l}{\partial \mu_0}(\pi, \mu_1, \mu_0, \Sigma) &= \sum_{i=1}^N (1 - y_i) \Sigma^{-1} (x_i - \mu_0) \\ \frac{\partial l}{\partial \Sigma^{-1}}(\pi, \mu_1, \mu_0, \Sigma) &= \frac{N}{2} \Sigma - \frac{1}{2} \sum_{i=1}^N y_i (x_i - \mu_{y_i}) (x_i - \mu_{y_i})^T\end{aligned}$$

Pour calculer la dérivée par rapport à Σ^{-1} , nous avons en particulier utilisé les deux dérivées suivantes :

- $\nabla_A(A \rightarrow x^T A x) = \nabla_A(A \rightarrow \text{Tr}(x^T A x)) = \nabla_A(A \rightarrow \text{Tr}(A x x^T)) = x x^T$
- $\nabla_X(X \rightarrow \log |X|) = X^{-1}$.

Pour prouver ce deuxième point, posons le taux d'accroissement $\log |X + H| - \log |X|$, pour H matrice de mêmes dimensions que X .

- Comme X est inversible, il existe une matrice Z telle que $Z^2 = X$. Ainsi, on peut écrire :
 $\log |X + H| - \log |X| = \log |Z^2 + H| - \log |X| = \log |Z(I + Z^{-1} H Z^{-1})Z| - \log |X| = \log |I + Z^{-1} H Z^{-1}| = \log |I + H'| - \log |I|$
 posant $H' = Z^{-1} H Z^{-1}$.
- Ainsi on s'est ramené à calculer la différentielle de $X \rightarrow \log |X|$ en le point I dans la direction H' .
- On sait que $|I + H'| = |-(-I - H')| = (-1)^n \chi(-1) = (-1)^n \prod_i (-1 - \lambda_i) = \prod_i (1 + \lambda_i) = 1 + \sum_i \lambda_i + o(||H'||)$
 où $\chi(x) = |xI - H'| = \prod_i |x - \lambda_i|$ est le polynôme caractéristique de H' , avec λ_i les valeurs propres de H' .
 D'où : $d(X \rightarrow |X|)(I).H' = \text{Tr}(H')$
 Ainsi, par composition, $d(X \rightarrow \log |X|)(I).H' = \frac{1}{|I|} \text{Tr}(H') = \text{Tr}(H')$

- Finalement,
 $\log |X + H| - \log |X| = \log |I + H'| - \log |I| = \text{Tr}(H') + o(||H'||)$
 $\text{Tr}(H') = \text{Tr}(Z^{-1} H Z^{-1}) = \text{Tr}(H Z^{-1} Z^{-1}) = \text{Tr}(H(Z^2)^{-1}) = \text{Tr}(H X^{-1})$
 Ce qui donne bien : $\nabla_X(X \rightarrow \log |X|) = X^{-1}$

Notons $(\pi^*, \mu_1^*, \mu_0^*, \Sigma^*)$ un point d'annulation du gradient de l . Alors :

$$\begin{aligned}\frac{\partial l}{\partial \pi}(\pi^*, \mu_1^*, \mu_0^*, \Sigma^*) = 0 &\implies \pi^* = \frac{\sum_{i=1}^N y_i}{N} \\ \frac{\partial l}{\partial \mu_1}(\pi^*, \mu_1^*, \mu_0^*, \Sigma^*) = 0 &\implies \mu_1^* = \frac{\sum_{i=1}^N y_i x_i}{\sum_{i=1}^N y_i} \\ \frac{\partial l}{\partial \mu_0}(\pi^*, \mu_1^*, \mu_0^*, \Sigma^*) = 0 &\implies \mu_0^* = \frac{\sum_{i=1}^N (1-y_i) x_i}{\sum_{i=1}^N (1-y_i)} \\ \frac{\partial l}{\partial \Sigma^{-1}}(\pi^*, \mu_1^*, \mu_0^*, \Sigma^*) = 0 &\implies \Sigma^* = \frac{\sum_{i=1}^N (x_i - \mu_{y_i}^*)(x_i - \mu_{y_i}^*)^T}{N}\end{aligned}$$

où : $\mu_{y_i}^* = y_i \mu_1^* + (1-y_i) \mu_0^*$.

$(\pi^*, \mu_1^*, \mu_0^*, \Sigma^*)$ est donc notre estimateur du maximum de vraisemblance (il faudrait en principe vérifier que la hessienne de l est semi-définie négative pour conclure que le point critique est un maximum).

Note : On peut aussi réécrire Σ^* comme suit : $\Sigma^* = \frac{\sum_{i=1}^N y_i (x_i - \mu_1^*)(x_i - \mu_1^*)^T + \sum_{i=1}^N (1-y_i) (x_i - \mu_0^*)(x_i - \mu_0^*)^T}{N}$

Classification :

Soit x fixé. On cherche ainsi à déterminer laquelle des deux probabilités $P(y=1|x)$ et $P(y=0|x)$ est la plus grande.

Dans le développement suivant, on notera $\theta = (\pi, \mu_1, \mu_0, \Sigma)$ et désignera par C_θ (resp. C_x , resp. $C_{x,\theta}$) toute expression faisant intervenir uniquement θ (resp. x , resp. x et θ). On garde la notation $\mu_y = y\mu_1 + (1-y)\mu_0$.

Soit $y \in \{0, 1\}$

$$P(y|x, \theta) = \frac{P(x|y, \theta)P(y|\theta)}{P(x)} = C_x P(x|y, \theta)P(y|\theta)$$

$$\begin{aligned}P(x|y, \theta) &= C_\theta \exp \left(-\frac{1}{2}(x - \mu_y)^T \Sigma^{-1} (x - \mu_y) \right) \\ &= C_{\theta,x} \exp \left(-\frac{1}{2} y \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} y \mu_0^T \Sigma^{-1} \mu_0 + y x^T \Sigma^{-1} \mu_1 - y x^T \Sigma^{-1} \mu_0 \right) \\ P(y|\theta) &= \exp(y \log \pi + (1-y) \log(1-\pi)) \\ &= C_\theta \exp \left(y \log \frac{\pi}{1-\pi} \right)\end{aligned}$$

d'où :

$$P(y|x, \theta) = C_{x,\theta} \exp[(a + b^T x)y]$$

en posant :

$$\begin{aligned}a &= \log \frac{\pi}{1-\pi} - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 \\ b &= \Sigma^{-1}(\mu_1 - \mu_0)\end{aligned}$$

Ainsi $P(y=1|x) > P(y=0|x) \iff \frac{P(y=1|x)}{P(y=0|x)} > 1 \iff \exp(a + b^T x) > 1 \iff a + b^T x > 0$.

Note : Pour une implémentation plus rapide, on pourra aussi observer que

$$-\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 = -\frac{1}{2} (\mu_1 + \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0) = -\frac{1}{2} (\mu_1 + \mu_0)^T b$$

2.3 Modèle génératif QDA :

Calcul de l'estimateur du maximum de vraisemblance : Soit \mathcal{L} la vraisemblance du modèle et l sa log-vraisemblance.

$$\begin{aligned} \mathcal{L}(\pi, \mu_1, \mu_0, \Sigma_1, \Sigma_0) &= \prod_{i=1}^N \pi^{y_i} (1-\pi)^{(1-y_i)} \frac{1}{(2\pi)^{\frac{d}{2}} (y_i |\Sigma_1| + (1-y_i) |\Sigma_0|)^{\frac{1}{2}}} \\ &\quad \times \exp -\frac{1}{2} [y_i(x_i - \mu_1)^T \Sigma_1^{-1} (x_i - \mu_1) + (1-y_i)(x_i - \mu_0)^T \Sigma_0^{-1} (x_i - \mu_0)] \end{aligned} \quad (3)$$

$$\begin{aligned} l(\pi, \mu_1, \mu_0, \Sigma_1, \Sigma_0) &= \sum_{i=1}^N \left(y_i \log \pi + (1-y_i) \log (1-\pi) - \frac{1}{2} \log (y_i |\Sigma_1| + (1-y_i) |\Sigma_0|) \right. \\ &\quad \left. - \frac{1}{2} [y_i(x_i - \mu_1)^T \Sigma_1^{-1} (x_i - \mu_1) + (1-y_i)(x_i - \mu_0)^T \Sigma_0^{-1} (x_i - \mu_0)] \right) + cste \\ &= \sum_{i=1}^N \left(y_i \log \pi + (1-y_i) \log (1-\pi) + \frac{1}{2} y_i \log |\Sigma_1^{-1}| + \frac{1}{2} (1-y_i) \log |\Sigma_0^{-1}| \right. \\ &\quad \left. - \frac{1}{2} [y_i(x_i - \mu_1)^T \Sigma_1^{-1} (x_i - \mu_1) + (1-y_i)(x_i - \mu_0)^T \Sigma_0^{-1} (x_i - \mu_0)] \right) + cste \end{aligned}$$

On obtient donc :

$$\begin{aligned} \frac{\partial l}{\partial \pi}(\pi, \mu_1, \mu_0, \Sigma_1, \Sigma_0) &= \sum_{i=1}^N y_i \frac{1}{\pi} + \sum_{i=1}^N (1-y_i) \frac{-1}{1-\pi} \\ \frac{\partial l}{\partial \mu_1}(\pi, \mu_1, \mu_0, \Sigma_1, \Sigma_0) &= \sum_{i=1}^N y_i \Sigma_1^{-1} (x_i - \mu_1) \\ \frac{\partial l}{\partial \mu_0}(\pi, \mu_1, \mu_0, \Sigma_1, \Sigma_0) &= \sum_{i=1}^N (1-y_i) \Sigma_0^{-1} (x_i - \mu_0) \\ \frac{\partial l}{\partial \Sigma_1^{-1}}(\pi, \mu_1, \mu_0, \Sigma_1, \Sigma_0) &= \frac{1}{2} \sum_{i=1}^N y_i \Sigma_1 - \frac{1}{2} \sum_{i=1}^N y_i (x_i - \mu_1)(x_i - \mu_1)^T \\ \frac{\partial l}{\partial \Sigma_0^{-1}}(\pi, \mu_1, \mu_0, \Sigma_1, \Sigma_0) &= \frac{1}{2} \sum_{i=1}^N (1-y_i) \Sigma_0 - \frac{1}{2} \sum_{i=1}^N (1-y_i) (x_i - \mu_0)(x_i - \mu_0)^T \end{aligned}$$

Notons $(\pi^*, \mu_1^*, \mu_0^*, \Sigma_0^*, \Sigma_1^*)$ un point d'annulation du gradient de l . Alors :

$$\begin{aligned} \frac{\partial l}{\partial \pi}(\pi^*, \mu_1^*, \mu_0^*, \Sigma_1^*, \Sigma_0^*) = 0 &\implies \pi^* = \frac{\sum_{i=1}^N y_i}{N} \\ \frac{\partial l}{\partial \mu_1}(\pi^*, \mu_1^*, \mu_0^*, \Sigma_1^*, \Sigma_0^*) = 0 &\implies \mu_1^* = \frac{\sum_{i=1}^N y_i x_i}{\sum_{i=1}^N y_i} \\ \frac{\partial l}{\partial \mu_0}(\pi^*, \mu_1^*, \mu_0^*, \Sigma_1^*, \Sigma_0^*) = 0 &\implies \mu_0^* = \frac{\sum_{i=1}^N (1-y_i) x_i}{N - \sum_{i=1}^N y_i} \\ \frac{\partial l}{\partial \Sigma_1^{-1}}(\pi^*, \mu_1^*, \mu_0^*, \Sigma_1^*, \Sigma_0^*) = 0 &\implies \Sigma_1^* = \frac{\sum_{i=1}^N y_i (x_i - \mu_1^*)(x_i - \mu_1^*)^T}{\sum_{i=1}^N y_i} \\ \frac{\partial l}{\partial \Sigma_0^{-1}}(\pi^*, \mu_1^*, \mu_0^*, \Sigma_1^*, \Sigma_0^*) = 0 &\implies \Sigma_0^* = \frac{\sum_{i=1}^N (1-y_i) (x_i - \mu_0^*)(x_i - \mu_0^*)^T}{N - \sum_{i=1}^N y_i} \end{aligned}$$

$(\pi^*, \mu_1^*, \mu_0^*, \Sigma_0^*, \Sigma_1^*)$ est l'estimateur du maximum de vraisemblance du modèle.

Classification : Soit x fixé. On cherche ainsi à déterminer laquelle des deux probabilités $P(y = 1|x)$ et $P(y = 0|x)$ est la plus grande. Notant $\theta = (\pi, \mu_1, \mu_0, \Sigma_0, \Sigma_1)$ et C_x (resp. C_θ) toute expression faisant intervenir uniquement x (resp. θ), on a :

$$\begin{aligned} P(y|x, \theta) &= \frac{P(x|y, \theta)P(y|\theta)}{P(x)} \\ &= C_x P(x|y, \theta)P(y|\theta) \end{aligned}$$

$$\begin{aligned} P(x|y, \theta) &= cste \times \frac{1}{|\Sigma_y|^{\frac{1}{2}}} \exp -\frac{1}{2} \left((x - \mu_y)^T \Sigma_y^{-1} (x - \mu_y) \right) \\ P(y|\theta) &= C_\theta \exp \left(y \log \frac{\pi}{1 - \pi} \right) \end{aligned}$$

Ainsi :

$$\frac{P(y = 1|x, \theta)}{P(y = 0|x, \theta)} = \frac{|\Sigma_0|^{\frac{1}{2}}}{|\Sigma_1|^{\frac{1}{2}}} \exp \left(\log \frac{\pi}{1 - \pi} - \frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{1}{2}(x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) \right)$$

Donc : $P(y = 1|x, \theta) > P(y = 0|x, \theta)$ si et seulement si

$$\log |\Sigma_0| - \log |\Sigma_1| + 2 \log \frac{\pi}{1 - \pi} - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + (x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) > 0$$