**M2 de Statistique, Université Pierre et Marie Curie**
**Bertrand MICHEL**

# Notes de cours

# Analyse Discriminante Linéaire
# Analyse Discriminante Quadratique

On dispose d'une population composée de $k$ sous-populations définies par une variable catégorielle à $k$ modalités que l'on note $\boldsymbol{y}$. Quitte à renommer les classes, on suppose que l'ensemble des modalités est $\{1,\dots,k\}$. Chaque individu de la population est aussi décrit par des variables explicatives $\boldsymbol{x}^1,\dots,\boldsymbol{x}^j,\dots,\boldsymbol{x}^p$.

## 1 Nuages et décomposition de la variance

Dans cette section, on suppose que les variables explicatives $\boldsymbol{x}^1,\dots,\boldsymbol{x}^j,\dots,\boldsymbol{x}^p$ sont toutes de type continu. Soit $\mathbf{M}$ la matrice associée au nuage des individus mesurés sur chacune des $p$ variables explicatives :

$$\mathbf{M}=\begin{bmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \vdots & \vdots \\ x_n^1 & \dots & x_n^p \end{bmatrix}.$$

On appelle nuage de $\mathbb{R}^p$ le n-uplet des observations $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$, où chacun des vecteurs $\boldsymbol{x}_i$ de $\mathbb{R}^p$ est un point du nuage. Par abus on notera aussi $\mathbf{M}$ le nuage. Dans la suite on supposera toujours que le nuage est centré :

$$\left(\frac{1}{n}\sum_{i=1\dots n} x_i^j\right)_{j=1\dots p} = 0.$$

La matrice de variance-covariance des variables du nuage $\mathbf{M}$ est définie par :

$$\boldsymbol{S} := \left[\mathrm{cov}(\boldsymbol{x}^j,\boldsymbol{x}^{j'})\right]_{1\leq j,j'\leq p} = \frac{1}{n}\mathbf{M}'\mathbf{M} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'$$

### 1.1 Géométrie d'un seul nuage

Avant de décrire la géométrie des $k$ nuages, considérons le cas d'un seul nuage en laissant de côté pour le moment le problème de la classification supervisée. L'analyse en composantes principales (voir le cours dédié) nous dit que l'orientation du nuage $\mathbf{M}$ peut être décrite à partir de la décomposition en valeurs propres de la matrice $\boldsymbol{S}$. Plus précisément, la première direction propre (pour la plus grande valeur propre) correspond à la direction de $\mathbb{R}^p$ pour laquelle les données du nuage $\mathbf{M}$ sont le plus dispersée. La seconde direction propre donne la direction orthogonale à la première pour laquelle les données sont le plus dispersée, etc...

### 1.2 Géométrie des $k$ nuages

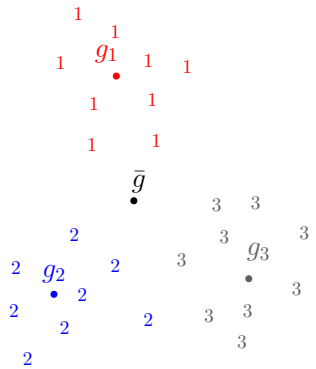
Nous supposons désormais que $k$ nuages sont observés : la variable $y$ présente $k$ modalités distinctes. Pour simplifier les notations, les individus sont numérotés de telle sorte que les premiers individus sont ceux du groupe 1, puis ceux du groupe 2 :

$$\mathbf{M}=\begin{bmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \vdots & \vdots \\ x_n^1 & \dots & x_n^p \end{bmatrix} = \begin{bmatrix} \mathbf{M}(1) \\ \vdots \\ \mathbf{M}(k) \end{bmatrix}$$

où $\mathbf{M}(\ell)$ est la matrice des données du groupe $\ell$. Les $k$ groupes induits par la variable catégorielle $\boldsymbol{y}$ définissent $k$ nuages dans l'espace $\mathbb{R}^p$ des variables explicatives. On supposera de plus que $k \leq p < n$. On note :

— $\mathrm{I}_\ell$ la liste des individus du groupe $\ell$,
— $n_\ell$ l'effectif de la sous-population $\ell$,
— $\boldsymbol{g}_\ell$ le centre de gravité du groupe $\ell$ : $\boldsymbol{g}_\ell = (g_\ell^1, \ldots, g_\ell^p)'$ avec

$$g_\ell^j = \frac{1}{n_\ell} \sum_{i \in \mathrm{I}_\ell} x_i^j.$$




Chacun des $k$ nuages $\mathbf{M}(\ell)$ a une orientation $\mathbb{R}^p$ qui peut être décrite par sa matrice de variance-covariance $\boldsymbol{S}_\ell$ (c.f. paragraphe précédent) où

$$\begin{aligned} \boldsymbol{S}_\ell &:= \frac{1}{n_\ell}(\mathbf{M}(\ell) - \boldsymbol{e}_{n_\ell}\boldsymbol{g}_\ell')'(\mathbf{M}(\ell) - \boldsymbol{e}_{n_\ell}\boldsymbol{g}_\ell') \\ &= \frac{1}{n_\ell} \sum_{i \in \mathrm{I}_l} (\boldsymbol{x}_i - \boldsymbol{g}_\ell)(\boldsymbol{x}_i - \boldsymbol{g}_\ell)'. \end{aligned}$$

où $\boldsymbol{e}_k$ designe le vecteur $(1, \ldots, 1)'$ de $\mathbb{R}^k$. On définit alors la matrice de variance **intra-classe** par

$$\boldsymbol{W} := \sum_{\ell=1\ldots k} \frac{n_\ell}{n} \boldsymbol{S}_\ell \qquad p \times p$$

Cette matrice est la moyenne pondérée des matrices de variance-covariance des $k$ nuages. Elle correspond donc à "l'orientation moyenne" des $k$ nuages. La première direction principale de $\boldsymbol{W}$ correspond à la direction selon laquelle les nuages sont le plus étalés, etc ... En général, la matrice $\boldsymbol{W}$ **est inversible**, ce que l'on supposera dans la suite.

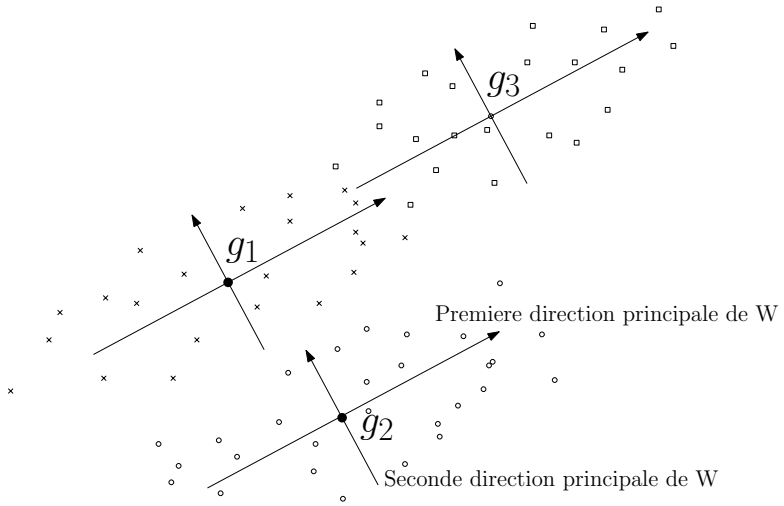



FIGURE 1 – Trois nuages avec des orientations comparables.

Soit $\boldsymbol{G}$ la matrice $k \times p$ du nuage des centres des $k$ nuages. Le centre de gravité $\bar{\boldsymbol{g}}$ de $\boldsymbol{G}$ pondéré des $n_\ell$ vérifie

$$\bar{\boldsymbol{g}} := \frac{1}{n} \sum_{\ell=1\ldots k} n_\ell \boldsymbol{g}_\ell = \bar{\boldsymbol{x}} = 0.$$

La matrice de variance-covariance $\boldsymbol{B}$ des $k$ centres de gravité pondérés par les $n_l$ est appelée matrice

de variance **inter-classes** :

$$\begin{aligned}\boldsymbol{B} &:= \left(\frac{1}{n}\sum_{\ell=1\dots k} n_\ell (g_\ell^j - \bar{g}^j)(g_\ell^u - \bar{g}^u)\right)_{1\le j,u\le k} \\ &= \boldsymbol{G}' \,\mathrm{Diag}(\frac{n_1}{n},\dots,\frac{n_k}{n})\boldsymbol{G} \qquad \text{car} \quad \bar{g}^j = \bar{g}^u = 0 \\ &= \frac{1}{n}\sum_{\ell=1\dots k} n_\ell \boldsymbol{g}_\ell \boldsymbol{g}_\ell' \qquad p\times p\end{aligned}$$

Puisque $\sum_{\ell=1\dots k} n_\ell \boldsymbol{g}_\ell = 0$, le nuage des $\boldsymbol{g}_l$ est contenu dans un sous-espace vectoriel de $\mathbb{R}^p$ de dimension $k-1$ et la matrice $\boldsymbol{B}$ n'est pas inversible. La matrice $\boldsymbol{B}$ décrit la géométrie du nuage des $k$ centres de gravité.

Les variances intra et inter classes permettent de décomposer la variance totale du nuage :

**Proposition 1.** *La variance du nuage* $\mathbf{M}$ *se décompose en variance intra-classe et variance inter-classes :*

$$\boldsymbol{S} = \boldsymbol{W} + \boldsymbol{B}.$$

# 2 Analyses discriminantes linéaire et quadratique

On suppose dans cette section que les variables aléatoires $X^j$ sont toutes de type continu.

## 2.1 Modélisation

On suppose que les variables en jeu sont des variables aléatoires notées $Y$ et $X^j$, dont on observe des réalisations $x_1^j,\dots x_n^j$ et $y_1,\dots,y_n$. Supposons de plus que la distribution de $Y$ admet $k\ge 2$ modalités. Pour chacune de ces modalités on considère la loi conditionnelle de $X=(X^1,\dots,X^p)$ sachant $Y=\ell$ et on suppose que cette loi conditionnelle admet une densité $f_\ell$ pour la mesure de Lebesgue sur $\mathbb{R}^p$ (que l'on note ici $\lambda^p$). On considère

— $\pi_\ell = P(Y=\ell)$ : la probabilité a priori d'appartenance au groupe $\ell$.
— $P(Y=\ell|X=x)$ : la probabilité a posteriori d'appartenance au groupe $\ell$.

**Proposition 2.** *Sous les hypothèses précédentes :*

1. *La distribution du vecteur aléatoire $(X,Y)$ admet la densité*

$$\begin{array}{rccc} f & : & \mathbb{R}^p\times\{1,\dots k\} & \to & \mathbb{R}^+ \\ & & (\boldsymbol{x},y) & \mapsto & \sum_{\ell=1}^k f_\ell(\boldsymbol{x})\pi_\ell \mathbf{1}_{y=\ell} \end{array}$$

   *par rapport à la mesure $\lambda^p\otimes\delta_k$, où $\delta_k$ désigne la mesure ponctuelle d'atomes $\{1,\dots,k\}$.*

2. *La distribution du vecteur $X$ admet la densité suivante par rapport à $\lambda^p$ :*

$$f_X = \sum_{\ell=1}^k \pi_\ell f_\ell.$$

3. *Pour tout $\ell\in\{1\dots,k\}$, la probabilité a posteriori d'appartenance au groupe $\ell$ vérifie :*

$$P(Y=\ell|X=x) = \frac{\pi_\ell f_\ell(x)}{\sum_{s=1}^k \pi_s f_s(x)} \tag{1}$$

Les méthodes d'analyse discriminante linéaire et quadratique sont des méthodes d'analyse discriminantes de type décisionnel. Pour prédire la variable $Y$ à partir des variables $X^1,\dots,X^p$, il est naturel de s'appuyer sur les probabilités a posteriori. Plus précisément, la **règle bayésienne d'attribution** consiste à attribuer une observation au groupe le plus probable pour celle-ci , c'est-à-dire celui pour lequel la probabilité a posteriori est maximale, ce qui équivaut d'après la relation (1) à choisir

$$\hat{Y}(\boldsymbol{x}) = \underset{\ell=1\dots k}{\mathrm{Argmax}}\ f_\ell(\boldsymbol{x})\pi_\ell.$$

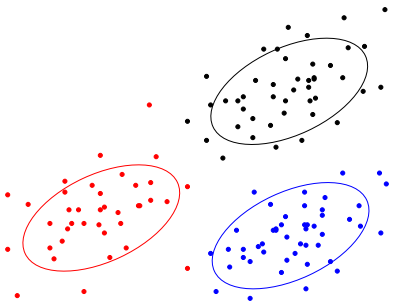
Cependant, en pratique ces quantités sont inconnues et il faut les estimer à partir des observations disponibles. Pour cela on proposera différentes hypothèses de modélisation sur la loi de $X$ sachant $Y$.

**Hypothèse gaussienne.** Pour modéliser le fait que les observations de chaque groupe $\ell$ sont organisés en "clusters", nous allons supposer que la loi du vecteur $X = (X^1, \ldots, X^p)$ peut être modélisée par une loi normale multivariée de densité sur $\mathbb{R}^p$ :

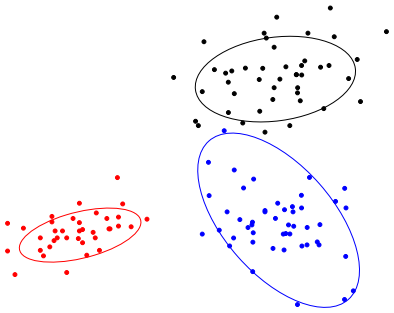
$$\boldsymbol{x} \in \mathbb{R}^p, \quad f_\ell(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \boldsymbol{\Sigma}_\ell}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\ell)\right].$$

Comme dans le cas de la régression linéaire, cette hypothèse n'est jamais rigoureusement vérifiée pour des données réelles. Cependant, cette modélisation est souvent suffisamment souple pour approcher efficacement la véritable loi des données (que l'on ne connaît évidemment pas en pratique).

**ADL et ADQ.** Si toutes les matrices de variance-covariance sont égales : $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, il s'agit de l'analyse discriminante linéaire (ADL) :



Au contraire, si les matrices de variance-covariance ne sont pas supposées égales, il s'agit de l'analyse discriminante quadratique (ADQ) :



## 2.2 Analyse Discriminante Linéaire (ADL)

Dans le but de déterminer les frontières entre les zones d'attribution, on considère le logarithme des rapports entre probabilités a posteriori : $\boldsymbol{x} \in \mathbb{R}^p$

$$\begin{aligned}
\log \frac{P(Y=\ell \mid X=\boldsymbol{x})}{P(Y=h \mid X=\boldsymbol{x})} &= \log \frac{\pi_\ell}{\pi_h} + \log \frac{f_\ell(\boldsymbol{x})}{f_h(\boldsymbol{x})} \\
&= \log \frac{\pi_\ell}{\pi_h} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\ell) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_h)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_h) \qquad (2) \\
&= \log \frac{\pi_\ell}{\pi_h} - \frac{1}{2}\boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell + \frac{1}{2}\boldsymbol{\mu}_h'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_h + \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_h)
\end{aligned}$$

On définit une fonction discriminante $s_\ell$ pour chaque groupe $\ell$ par :

$$\boldsymbol{x} \in \mathbb{R}^p, \quad s_\ell(\boldsymbol{x}) := x'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell + \log \pi_\ell - \frac{1}{2}\boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell.$$

La règle de Bayes consiste ici à affecter une observation $\boldsymbol{x}$ au groupe $y(\boldsymbol{x})$ de score maximum :

$$y(\boldsymbol{x}) = \underset{\ell=1\ldots k}{\operatorname{argmax}}\, s_\ell(\boldsymbol{x}).$$

Ces fonctions discriminantes (ou scores) sont linéaires en $\boldsymbol{x}$, d'où l'appellation d'analyse discriminante linéaire.

**Inférence.** En pratique, les quantités $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ et $\pi_1, \ldots, \pi_k$ sont inconnues. On peut cependant les estimer par la méthode du maximum de vraisemblance :

**Proposition 3.** *En supposant que la loi conditionnelle de $(X|Y = \ell)$ est celle d'une loi normale multivariée p-dimensionnel dont la matrice de variance-covariance est inversible et ne dépend pas de $\ell$ (hypothèse ADL), les estimateurs du maximum de vraisemblance de $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_\ell$ et $\pi_\ell$ vérifient :*

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{W} = \sum_{\ell=1\ldots k} \frac{n_\ell}{n} \boldsymbol{S}_\ell, \quad \hat{\boldsymbol{\mu}}_\ell = \boldsymbol{g}_\ell = \frac{1}{n_\ell} \sum_{i|y_i=\ell} \boldsymbol{x}_i \quad \text{et } \hat{\pi}_\ell = \frac{n_\ell}{n}.$$

**Remarque.** *On utilise aussi parfois l'estimateur sans biais de la matrice de variance-covariance $\frac{n}{n-k}\boldsymbol{W}$.*

**Attribution.** La règle d'affection effective pour une observation $\boldsymbol{x}$ est finalement donnée par

$$\hat{y}(\boldsymbol{x}) = \underset{\ell=1\ldots k}{\operatorname{argmax}}\, \hat{s}_\ell(\boldsymbol{x})$$

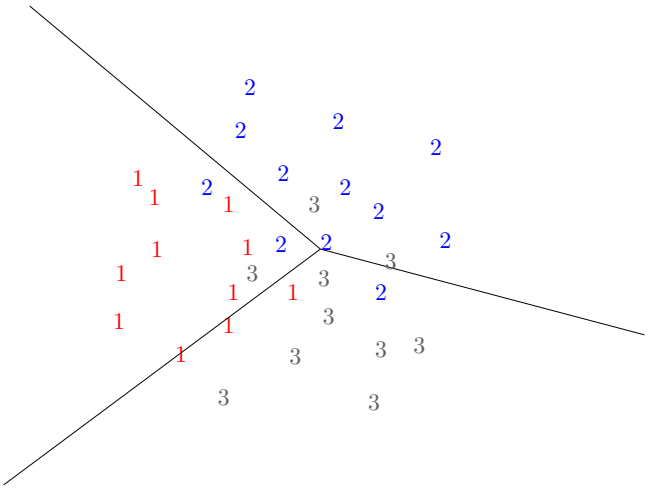
où

$$\hat{s}_\ell(\boldsymbol{x}) := x'\boldsymbol{W}^{-1}\boldsymbol{g}_\ell + \log\frac{n_\ell}{n} - \frac{1}{2}\boldsymbol{g}_\ell'\boldsymbol{W}^{-1}\boldsymbol{g}_\ell.$$

**Cas de deux groupes.** Si $k = 2$, l'espace $\mathbb{R}^2$ est séparé en deux zones dont la frontière est l'hyperplan affine d'équation

$$\log\frac{n_1}{n_2} - \frac{1}{2}(\boldsymbol{g}_1 + \boldsymbol{g}_2)'\boldsymbol{W}^{-1}(\boldsymbol{g}_1 - \boldsymbol{g}_2) + x'\boldsymbol{W}^{-1}(\boldsymbol{g}_1 - \boldsymbol{g}_2) = 0$$

**Zones de séparation.** De façon plus générale, la zone de séparation entre les régions d'attribution des classes $\ell$ et $h$ est la région $\mathcal{B}_{\ell,h} \subset \mathbb{R}^p$ définie par l'équation $\hat{s}_\ell(\boldsymbol{x}) = \hat{s}_h(\boldsymbol{x})$. Cette région $\mathcal{B}_{\ell,h}$ est un hyperplan affine car les fonctions discriminantes $\hat{s}_\ell$ sont linéaires :



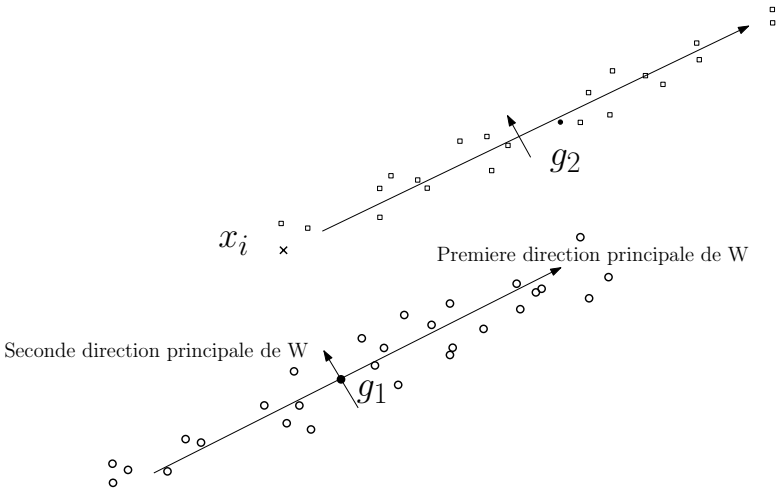

## 2.3 Métrique de Mahalanobis et ADL

Si les probabilités a priori sont égales, l'égalité (2) montre que la règle d'attribution revient à affecter une observation $\boldsymbol{x}_i$ au groupe pour lequel la quantité $(\boldsymbol{x} - \boldsymbol{g}_\ell)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{g}_\ell)$ est minimale. Cette quantité peut être interprétée comme une distance pour une métrique particulière : la métrique de Mahalanobis. Celle-ci est définie dans $\mathbb{R}^p$ par le produit scalaire

$$< \boldsymbol{u}, \boldsymbol{v} >_{\boldsymbol{W}} := \boldsymbol{u}'\boldsymbol{W}^{-1}\boldsymbol{v}.$$

Pour cette métrique, les points situés sur un ellipsoïde d'équation $(\boldsymbol{x} - \boldsymbol{g}_l)'\boldsymbol{W}^{-1}(\boldsymbol{x} - \boldsymbol{g}_l) = c$ sont tous équidistants du point $\boldsymbol{g}_l$. Cette normalisation par $\boldsymbol{W}^{-1}$ permet d'éviter que les directions relatives aux

grandes valeurs propres de $\boldsymbol{W}$ soient trop prépondérantes dans le calcul des distances. Par exemple, dans l'exemple ci-dessous, les deux nuages ont des orientations comparables et un étalement important selon la première direction de $\boldsymbol{W}$.




Le point $\boldsymbol{x}_i$ est plus proche de $\boldsymbol{g}_1$ que de $\boldsymbol{g}_2$ pour la métrique euclidienne mais en réalité il est plus naturel d'affecter ce point au groupe 2. En effet l'étalement dans la première direction de $\boldsymbol{W}$ est tel que certains points du groupe 2 sont à proximité de $\boldsymbol{x}_i$, alors que ce n'est pas le cas pour les points du groupe 1. Pour la métrique de Mahalanobis $\boldsymbol{x}_i$ est plus proche de $\boldsymbol{g}_2$ que de $\boldsymbol{g}_1$.

Pour mieux comprendre l'effet de cette métrique, considérons le cas où il n'y a qu'une seule classe, on a alors $\boldsymbol{W} = \boldsymbol{S}$. Soit $\boldsymbol{Z} = \mathbf{M}\boldsymbol{W}^{-1/2}$ le nuage renormalisé : la matrice de variance-covariance vaut l'identité et le nuage a une forme sphérique. On peut vérifier que

$$\mathrm{d}^2_{\boldsymbol{W}}(\boldsymbol{x}_i; \boldsymbol{x}_s) = \mathrm{d}^2(\boldsymbol{z}_i; \boldsymbol{z}_s)$$

où d désigne la distance euclidienne. La métrique de Mahalanobis revient à considérer la distance euclidienne pour un nuage associé pour lequel les variables sont non corrélées et de même variance 1.

Dans le cas de $k$ classes, $\boldsymbol{W}$ est la matrice de variance-covariance moyenne des $k$ nuages. La correspondance précédente entre $\mathrm{d}_{\boldsymbol{W}}$ et d n'est plus rigoureusement exacte car en général les matrices $\boldsymbol{S}_\ell$ ne coïncident pas exactement. Cependant si ces dernières ne diffèrent pas trop, à l'intérieur d'une même nuage la métrique de Mahalanobis revient en première approximation à considérer la distance euclidienne pour un nuage renormalisé (dont les matrices de variance-covariance sont des matrices identités).

## 2.4 Analyse discriminante quadratique (ADQ)

Contrairement à l'ADL, l'analyse discriminante quadratique autorise les matrices de variance-covariance des nuages à être différentes. Le logarithme des rapports entre probabilités a posteriori de deux classes $\ell$ et $h$ vérifie

$$\begin{aligned} \boldsymbol{x} \in \mathbb{R}^p, \quad \log \frac{P(Y = \ell \mid X = \boldsymbol{x})}{P(Y = h \mid X = \boldsymbol{x})} &= \log \frac{\pi_\ell}{\pi_h} + \log \frac{f_\ell(\boldsymbol{x})}{f_h(\boldsymbol{x})} \\ &= s_\ell(\boldsymbol{x}) - s_h(\boldsymbol{x}) \end{aligned}$$

où $s_\ell$ est la fonction discriminante du groupe $\ell$ définie par

$$s_\ell = \log \pi_\ell - \frac{1}{2} \log \det \boldsymbol{\Sigma}_\ell - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_\ell)' \boldsymbol{\Sigma}_\ell^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_l).$$

**Inférence.** Comme dans le cas de l'ADL, les quantités $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ et $\pi_1, \ldots, \pi_k$ peuvent être estimées par la méthode du maximum de vraisemblance :

**Proposition 4.** *En supposant que la loi conditionnelle de $(X|Y = \ell)$ est celle d'une loi normale multivariée p-dimensionnel (sans supposer ici que les k matrices de variances-covariances sont égales : hypothèse ADQ), les estimateurs du maximum de vraisemblance de $\boldsymbol{\mu}_\ell$, $\boldsymbol{\Sigma}_\ell$ et $\pi_\ell$ vérifient :*

$$\hat{\pi}_\ell = \frac{n_\ell}{n}, \quad \hat{\boldsymbol{\mu}}_\ell = \boldsymbol{g}_\ell \quad \text{et} \quad \widehat{\boldsymbol{\Sigma}}_\ell = \boldsymbol{S}_\ell.$$

**Remarque.** *On utilise aussi parfois les estimateurs sans biais des matrices de variance-covariance $\frac{n_\ell}{n_\ell - 1}\boldsymbol{S}_\ell$).*

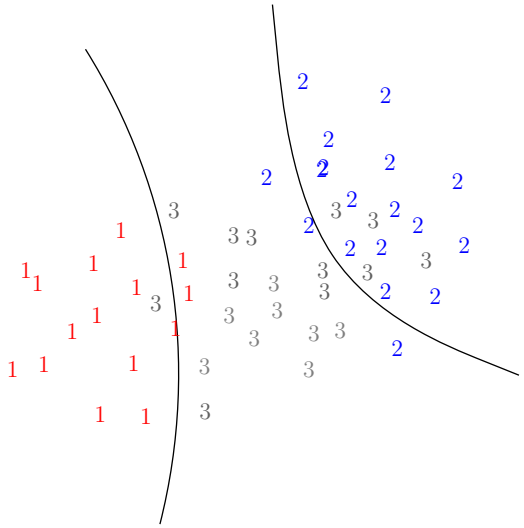
**Attribution.** La règle d'affection effective pour une observation $\boldsymbol{x}$ est comme précédemment donnée par

$$\hat{y}(\boldsymbol{x}) = \underset{\ell=1\ldots k}{\operatorname{argmax}}\, \hat{s}_\ell(\boldsymbol{x})$$

où

$$\hat{s}_\ell = \log \frac{n_\ell}{n} - \frac{1}{2}\log\det \boldsymbol{S}_\ell - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{g}_\ell)'\boldsymbol{S}_\ell^{-1}(\boldsymbol{x} - \boldsymbol{g}_\ell).$$

**Zones de séparation.** La zone de séparation entre les régions d'attribution des classes $\ell$ et $h$ est l'hypersurface de $\mathbb{R}^p$ définie par l'équation $\hat{s}_\ell(\boldsymbol{x}) = \hat{s}_h(\boldsymbol{x})$.



## 2.5 Choisir entre ADL et ADQ

Il est possible de construire un test sur l'égalité des matrices de variance-covariance à l'aide de la statistique :

$$Z := \left(1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right)\left[\left(\sum_{\ell=1\ldots k}\frac{1}{n_\ell - 1} - \frac{1}{n-k}\right)(n-k)\log\left|\frac{n}{n-k}\boldsymbol{W}\right| - \sum_{\ell=1\ldots k}(n_\ell - 1)\log\left|\frac{n_\ell}{n_\ell - 1}\boldsymbol{S}_\ell\right|\right].$$

On peut en effet montrer (admis) que sous l'hypothèse $H_0 : \boldsymbol{S}_1 = \ldots \boldsymbol{S}_k$ et sous de bonnes conditions, la statistique $Z$ converge vers une loi du $\chi^2$ à $\frac{p(p+1)(k-1)}{2}$ degrés de liberté. Cette propriété permet ainsi de construire le test de Box (voir par exemple [Anderson, 2002], chap 10).

Attention cependant : même si le test rejette $H_0$, l'ADQ ne donne pas nécessairement une meilleure classification que l'ADL car l'ADQ nécessite d'estimer beaucoup plus de coefficients que l'ADL.

Une stratégie parfois intéressante consiste à utiliser l'ADL en enrichissant la famille des variables explicatives de variables quadratiques $(\boldsymbol{x}^j)^2$ et de variables d'interaction $\boldsymbol{x}^j \times \boldsymbol{x}^{j'}$. Cette méthode est en effet moins « consommatrice » en paramètres. Dans tous les cas, on évaluera les erreurs de classement pour comparer les méthodes (voir plus loin).

### 2.6 Une version non paramétrique

Les estimateurs à noyau sont des estimateurs non paramétriques couramment utilisés en statistique. Ils permettent notamment d'estimer une densité sans hypothèse d'appartenance à une famille paramétrique de loi. On les définit par :

$$\boldsymbol{x} \in \mathbb{R}^p, \quad \hat{f}(\boldsymbol{u}) = \frac{1}{nh} \sum_{i=1\ldots n} \mathrm{K}\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)$$

où $h > 0$ est la fenêtre d'estimation et $\mathrm{K} : \mathbb{R}^p \mapsto \mathbb{R}^+$ est un noyau i.e. une fonction symétrique, à valeurs positives ou nulles et d'intégrale 1 (ex : noyau gaussien, $\mathrm{K} = \frac{1}{2}\ \mathbb{1}_{[-1,1]}$). On parle alors d'estimation non paramétrique.

Dans le contexte de l'analyse discriminante, on estime donc pour chaque groupe $\ell$ la densité jointe des variables explicatives par

$$\boldsymbol{x} \in \mathbb{R}^p, \quad \hat{f}_\ell(\boldsymbol{x}) = \frac{1}{n_\ell h} \sum_{i \in \mathrm{I}_\ell} \mathrm{K}\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right).$$

Comme pour les modèles gaussiens précédents, on utilise ensuite la formule de Bayes pour estimer $\widehat{P}(Y = \ell \mid X = \boldsymbol{x}) = \frac{\hat{f}_\ell(\boldsymbol{x})\hat{\pi}_\ell}{\sum_{j=1}^k \hat{\pi}_j \hat{f}_j(\boldsymbol{x})}$ et une observation est attribuée au groupe le plus probable selon la règle de Bayes.

## Références

[Anderson, 2002] Anderson, T. W. (2002). *An introduction to multivariate statistical analysis, Third edition.* Wiley, New Jersey.