Analyse discriminante

Christine Decaestecker & Marco Saerens

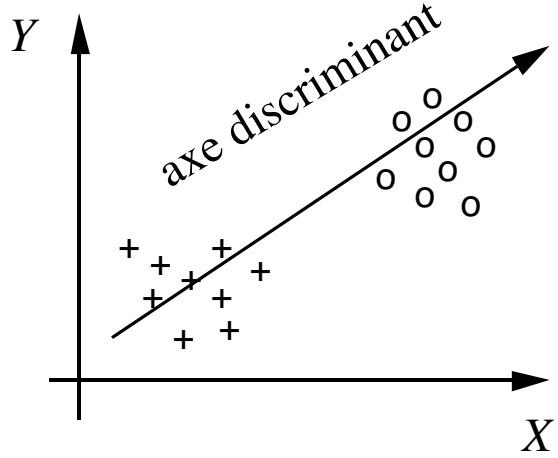
ULB & UCL

Analyse Discriminante

- **Particularités:** 2 formes/utilisations complémentaires:
  - ***méthode factorielle***: description "géométrique" de la séparation inter-classe (encore appelée analyse discriminante factorielle ou analyse discriminante linéaire de Fisher)
  - ***interprétation bayesienne***: classificateur bayesien (optimum au sens de la proba. de l'erreur) dans des conditions particulières pour les données! (encore appelée analyse discriminante décisionnelle, linéaire ou quadratique)
- **Restrictions:**
  Les variables descriptives X_1, X_2, ... X_p doivent être ***quantitatives***!
- **Analyse discriminante factorielle: principes de base:**
  Objectif: mettre en évidence des différences entre les classes c-à-d entre les observations appartenant à des classes différentes

  => description des liaisons entre la variable "classe" et les variables quantitatives: les q classes diffèrent-elles sur l'ensemble des variables numériques?

Méthode: déterminer un/des facteur(s), combinaison(s) linéaire(s) des variables descriptives, qui prenne(nt) des valeurs les + proches possible pour des éléments de la même classe, et les + éloignées possible entre éléments de classes différentes. (= facteurs discriminants)
Exemple:



- **Formulation:**

1) Décomposition de la matrice variance-covariance V

Ensemble des n observations $\mathbf{x}_i$ = un nuage de points, de centre de gravité $\mathbf{g}$ et de matrice variance-covariance $\mathbf{V}$.
Ce nuage est partagé en q ***sous-nuages*** par la variable "classe". Chaque sous-nuage (classe ω_k) d'effectif n_k est caractérisé par son centre de gravité (ou ***centroïde***) $\mathbf{g}_k$ et sa matrice variance-covariance $\mathbf{V}_k$.

V peut être décomposée en une somme de 2 matrices: **V = B + W**

avec **B** = matrice de variance inter-classe (B = "between")

= matrice variance-covariance pondérée des k centroïdes $\mathbf{g}_k$:

où $\mathbf{g}_k = (g_{k1}, g_{k2}, ..., g_{kj}, ..., g_{kp})^T$ et g_{kj} = moyenne de X_j dans ω_k

$$\mathbf{B}_{p\times p} = \frac{1}{n}\sum_{k=1}^{q} n_k \underbrace{(\mathbf{g}_k - \mathbf{g})(\mathbf{g}_k - \mathbf{g})^T}_{\text{matrice } \mathbf{C}^{(k)}\ (p\times p)} \quad \text{avec} \quad c_{jj'}^{(k)} = (g_{kj} - m_j)(g_{kj'} - m_{j'})$$

rend compte de la dispersion des centroïdes des classes autour du centre global **g**.

et **W** = matrice de variance intra-classe (W = "within")

= moyenne des k matrices variance-covariance des classes: $\mathbf{V}_k$

$$\mathbf{W}_{p\times p} = \frac{1}{n}\sum_{k=1}^{q} n_k \mathbf{V}_k$$

Généralisation de la relation classique unidimensionnelle valable pour toute variable X dont les valeurs sont regroupées par classe:

variance totale = moyenne des variances (intra) + variance des moyennes (inter)

Décomposition variance intra/inter classe

- La variance empirique totale s'écrit

$$
\begin{aligned}
\sigma^2 &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i-\mathbf{x}_j) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{g})^{\mathrm{T}}(\mathbf{x}_i-\mathbf{g}) \\
&= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in C(k)}(\mathbf{x}_i-\mathbf{g})^{\mathrm{T}}(\mathbf{x}_i-\mathbf{g}) \\
&= \frac{1}{n}\sum_{k=1}^{q}SS(k)
\end{aligned}
$$

Décomposition variance intra/inter classe

- Décomposition:

$$
\begin{aligned}
SS(k) &= \sum_{i \in C(k)} (\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} (\mathbf{x}_i - \mathbf{g}) \\
&= \sum_{i \in C(k)} (\mathbf{x}_i - \mathbf{g}(k) + \mathbf{g}(k) - \mathbf{g})^{\mathrm{T}} (\mathbf{x}_i - \mathbf{g}(k) + \mathbf{g}(k) - \mathbf{g}) \\
&= \sum_{i \in C(k)} \left(||\mathbf{x}_i - \mathbf{g}(k)||^2 + ||\mathbf{g}(k) - \mathbf{g}||^2 + 2 (\mathbf{x}_i - \mathbf{g}(k))^{\mathrm{T}} (\mathbf{g}(k) - \mathbf{g}) \right) \\
&= \sum_{i \in C(k)} \left(||\mathbf{x}_i - \mathbf{g}(k)||^2 + ||\mathbf{g}(k) - \mathbf{g}||^2 \right) \\
&= \underbrace{\sum_{i \in C(k)} ||\mathbf{x}_i - \mathbf{g}(k)||^2}_{\text{within}} + \underbrace{n(k) ||\mathbf{g}(k) - \mathbf{g}||^2}_{\text{between}}
\end{aligned}
$$

Décomposition variance intra/inter classe

- Et donc la variance totale s'écrit

$$
\begin{aligned}
\sigma^2 &= \frac{1}{n}\left[\sum_{k=1}^{q}\sum_{i\in C(k)}||\mathbf{x}_i-\mathbf{g}(k)||^2+\sum_{k=1}^{q}n(k)||\mathbf{g}(k)-\mathbf{g}||^2\right] \\
&= \frac{1}{n}\left[\sum_{k=1}^{q}n(k)\frac{1}{n(k)}\sum_{i\in C(k)}||\mathbf{x}_i-\mathbf{g}(k)||^2+\sum_{k=1}^{q}n(k)||\mathbf{g}(k)-\mathbf{g}||^2\right] \\
&= \frac{1}{n}\sum_{k=1}^{q}n(k)\left[\frac{1}{n(k)}\sum_{i\in C(k)}||\mathbf{x}_i-\mathbf{g}(k)||^2+||\mathbf{g}(k)-\mathbf{g}||^2\right] \\
&= \frac{1}{n}\sum_{k=1}^{q}n(k)\left[\sigma^2_{(w)}(k)+\sigma^2_{(b)}(k)\right]=\sigma^2_{(w)}+\sigma^2_{(b)}
\end{aligned}
$$

Projection sur un axe

- Nous projetons les observations sur un axe
- L'opérateur de projection s'écrit

$$\boldsymbol{\pi} = \mathbf{v}\mathbf{v}^{\mathrm{T}}$$

$$\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$$

Projection de la variance

- Pour la variance within:

$$
\begin{aligned}
\sigma^2_{\mathbf{v}(w)} &= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in C(k)} \left(\boldsymbol{\pi}\mathbf{x}_i - \boldsymbol{\pi}\mathbf{g}(k)\right)^{\mathrm{T}} \left(\boldsymbol{\pi}\mathbf{x}_i - \boldsymbol{\pi}\mathbf{g}(k)\right) \\
&= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in C(k)} \left(\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{x}_i - \mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{g}(k)\right)^{\mathrm{T}} \left(\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{x}_i - \mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{g}(k)\right) \\
&= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in C(k)} \left(\mathbf{x}_i - \mathbf{g}(k)\right)^{\mathrm{T}} \left(\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{v}\mathbf{v}^{\mathrm{T}}\right) \left(\mathbf{x}_i - \mathbf{g}(k)\right)
\end{aligned}
$$

Projection de la variance (suite)

$$
\begin{aligned}
\sigma^2_{\mathbf{v}(w)} &= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in C(k)} \left(\mathbf{x}_i - \mathbf{g}(k)\right)^{\mathrm{T}} \mathbf{v}\mathbf{v}^{\mathrm{T}} \left(\mathbf{x}_i - \mathbf{g}(k)\right) \\
&= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in C(k)} \mathbf{v}^{\mathrm{T}} \left(\mathbf{x}_i - \mathbf{g}(k)\right) \left(\mathbf{x}_i - \mathbf{g}(k)\right)^{\mathrm{T}} \mathbf{v} \\
&= \mathbf{v}^{\mathrm{T}} \left[\frac{1}{n}\sum_{k=1}^{q}\sum_{i\in C(k)} \left(\mathbf{x}_i - \mathbf{g}(k)\right) \left(\mathbf{x}_i - \mathbf{g}(k)\right)^{\mathrm{T}}\right] \mathbf{v} \\
&= \mathbf{v}^{\mathrm{T}} \left[\frac{1}{n}\sum_{k=1}^{q} n(k) \left(\frac{1}{n(k)}\sum_{i\in C(k)} \left(\mathbf{x}_i - \mathbf{g}(k)\right) \left(\mathbf{x}_i - \mathbf{g}(k)\right)^{\mathrm{T}}\right)\right] \mathbf{v} \\
&= \mathbf{v}^{\mathrm{T}} \left[\frac{1}{n}\sum_{k=1}^{q} n(k)\mathbf{W}_k\right] \mathbf{v} \\
&= \mathbf{v}^{\mathrm{T}}\mathbf{W}\mathbf{v}
\end{aligned}
$$

Projection de la variance

- Pour la variance between:

$$
\begin{aligned}
\sigma^2_{\mathbf{v}(b)} &= \frac{1}{n}\sum_{k=1}^{q} n(k)(\boldsymbol{\pi}\mathbf{g}(k) - \boldsymbol{\pi}\mathbf{g})^{\mathrm{T}}(\boldsymbol{\pi}\mathbf{g}(k) - \boldsymbol{\pi}\mathbf{g}) \\
&= \mathbf{v}^{\mathrm{T}}\left[\sum_{k=1}^{q} n(k)\frac{(\mathbf{g}(k) - \mathbf{g})\,(\mathbf{g}(k) - \mathbf{g})^{\mathrm{T}}}{n}\right]\mathbf{v} \\
&= \mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v}
\end{aligned}
$$

Projection de la variance

- Pour la variance totale:

$$\begin{aligned}
\sigma^2_{\mathbf{v}} &= = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\pi}\mathbf{x}_i - \boldsymbol{\pi}\mathbf{g})^{\mathrm{T}}(\boldsymbol{\pi}\mathbf{x}_i - \boldsymbol{\pi}\mathbf{g}) \\
&= \mathbf{v}^{\mathrm{T}}\left[\sum_{i=1}^{n}\frac{(\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})^{\mathrm{T}}}{n}\right]\mathbf{v} \\
&= \mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}
\end{aligned}$$

Décomposition de la variance projetée

- La variance empirique projetée se décompose également en within/between:

$$\sigma^2_{\mathbf{v}} = \sigma^2_{\mathbf{v}(w)} + \sigma^2_{\mathbf{v}(b)}$$

- Et donc

$$1 = \frac{\sigma^2_{\mathbf{v}(w)}}{\sigma^2_{\mathbf{v}}} + \frac{\sigma^2_{\mathbf{v}(b)}}{\sigma^2_{\mathbf{v}}}$$

$$0 < \frac{\sigma^2_{\mathbf{v}(b)}}{\sigma^2_{\mathbf{v}}} < 1$$

Problème d'optimisation

- Nous recherchons l'axe correspondant à la séparation maximale entre les classes

$$\max_{\mathbf{v}} \left(\frac{\sigma^2_{\mathbf{v}(b)}}{\sigma^2_{\mathbf{v}}} \right) = \max_{\mathbf{v}} \left(\frac{\mathbf{v}^{\mathrm{T}} \mathbf{B} \mathbf{v}}{\mathbf{v}^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{v}} \right)$$

- Et donc

$$\partial_{\mathbf{v}} \left(\frac{\mathbf{v}^{\mathrm{T}} \mathbf{B} \mathbf{v}}{\mathbf{v}^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{v}} \right) = 0$$

Problème d'optimisation

- Nous avons donc

$$2(\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v})\mathbf{B}\mathbf{v} - 2(\mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v})\boldsymbol{\Sigma}\mathbf{v} = 0$$

- D'où $\mathbf{B}\mathbf{v} = \left(\frac{\mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}}\right)\boldsymbol{\Sigma}\mathbf{v}$

- Et nous posons

$$0 < \lambda = \frac{\mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}} < 1$$

Problème aux valeurs/vecteurs propres

- Nous obtenons donc le problème aux valeurs/vecteurs propres

$$\boldsymbol{\Sigma}^{-1}\mathbf{B}\mathbf{v} = \lambda\mathbf{v}$$

$\boldsymbol{\Sigma}$ est aussi noté $\mathbf{V}$

- Il y a au plus (q – 1) valeurs propres non-nulles

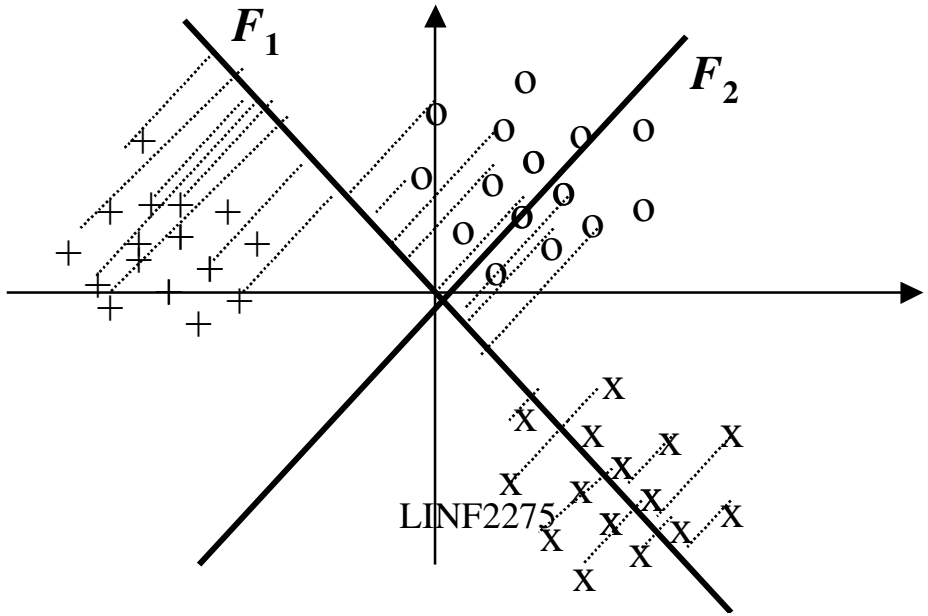
2) Recherche des facteurs discriminants

On travaille en **variables centrées => g** est ramené à l'origine!

Le 1er facteur discriminant (F_1) est une nouvelle variable, combinaison linéaire des variables descriptives (centrées), dont la variance inter-classe est maximum (ou, de façon équivalente la variance intra-classe est minimum).

Géométriquement: le 1er facteur détermine un axe dans le nuage de points (passant par l'origine) tel que les projections des points sur cet axe aient une variance inter-classe (variance des moyennes de classe) maximale.

Le 2eme facteur (F_2) est non corrélé (perpendiculaire) au 1er et de variance inter-classe max. Etc pour le 3eme ...



- **Propriétés**:
  - les facteurs sont entièrement déterminés par la matrice définie par: $\mathbf{V}^{-1}\,\mathbf{B}$ (vecteurs propres)
  - le nbre maximum de facteurs discriminants = $q - 1$
  - la part de variance inter-classe expliquée = variance inter / variance totale est décroissante entre les facteurs successifs.

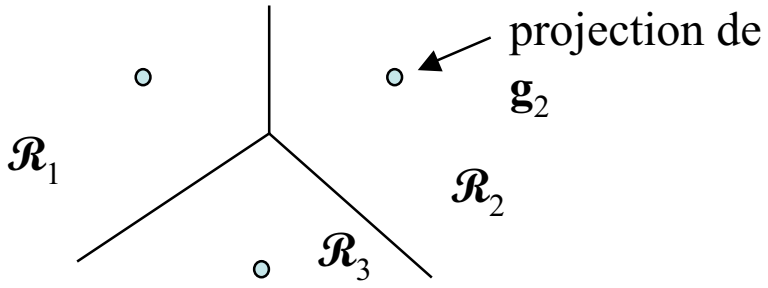
  Toutes ces propriétés s'expliquent par le fait que:

  une analyse **discriminante = ACP** sur le nuage des ***q* centroïdes,**
  pondérés par l'effectif des classes n_k,
  dans un espace $\Re^p$ avec $\mathbf{V}^{-1}$ comme métrique !

- **Représentation graphique:**
  - Si 2 groupes => 1 seul facteur = axe de projection où la séparation inter-classe est la mieux exprimée => coordonnées sur cet axe = scores discriminants.
  - Si + de 2 groupes => plan discriminant (F_1 et F_2) = plan de projection ou la variance inter-classe **B** (=> dispersion des centroïdes dans le plan) sera la mieux représentée!!

- **Interprétation des facteurs:**
  Comme en ACP: corrélations facteurs aux variables initiales
  + cercle des corrélations avec les 2 premiers facteurs ($q > 2$)

- **Analyse discriminante décisionnelle => méthode de classification:**
  1) règle géométrique (règle de Fisher):
  Les facteurs discriminants donnent la meilleure représentation de la séparation des q centroïdes de classe (dans un espace orthonormé).
  => pour un individu **x** projeté dans l'espace des facteurs: attribuer la classe dont le centroïde est le plus proche (au sens de la ***distance euclidienne***):
  => surfaces de séparation ***linéaires*** = hyperplans médians entre les centroïdes:



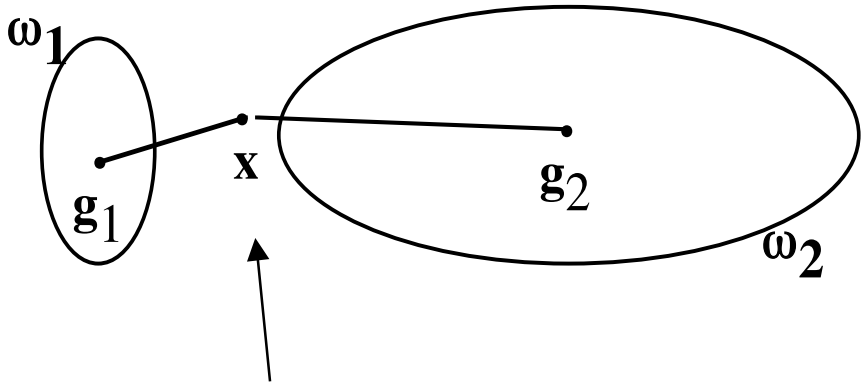
détermination de 3 régions de décision ($\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$) délimitant les points 'sensés' appartenir aux différentes classes

Traduction dans l'espace de départ (variables descriptives): allocation au centroïde $\mathbf{g}_k$ le plus proche au sens de la métrique $\mathbf{W}^{-1}$ (***distance de Mahalanobis***)

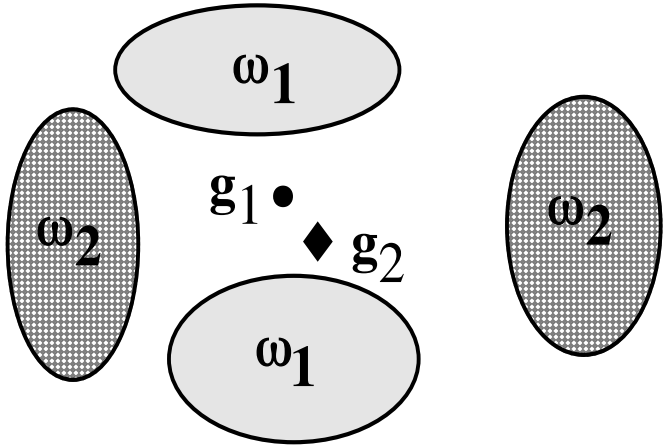
$$d^2_M(\mathbf{x}, \mathbf{g}_k) = (\mathbf{x} - \mathbf{g}_k)^T \mathbf{W}^{-1}(\mathbf{x} - \mathbf{g}_k)$$

Problèmes:

- La métrique $\mathbf{W}^{-1}$ est évaluée sur l'ensemble des données => problème si les classes ne sont pas de même "forme" (dispersion).
- une classe est représentée par son centroïde => problème si le centroïde n'est pas représentatif d'une classe (cas des classes non ellipsoïdales ou composées de sous-nuages différents => séparation fortement non linéaire).



x sera jugé plus proche de ω_1 que de ω_2



- **Justification: lien avec la classification bayesienne**
  On peut montrer que la règle de Fisher correspond à un classificateur bayesien (minimisation de la proba. de l'erreur) dans les conditions suivantes:
  - chaque classe suit une distribution gaussienne (multivariée) de même matrice variance-covariance **W** (les nuages de points ont la même 'forme'),
  - les classes sont équidistribuées: mêmes proba. *a priori*
    (très facilement généralisable si ce n'est pas le cas)

  En effet:
  Lorsque les distributions de classes sont gaussiennes de même matrice variance - covariance **V**, un classificateur bayesien définit les fonctions discriminantes $y_k(\mathbf{x})$ suivantes: ($\mathbf{x}$ alloué à ω_k si $y_k(\mathbf{x}) > y_j(\mathbf{x})$ pour tout $j \neq k$):

$$y_k(\mathbf{x}) = -\underbrace{(\mathbf{x} - \mathbf{g}_k)^T \mathbf{W}^{-1} (\mathbf{x} - \mathbf{g}_k)}_{} + 2 \ln (P(\omega_k))$$

  <=> $\mathbf{x}$ alloué à ω_k si $d^2_M(\mathbf{x}, \mathbf{g}_k) - 2 \ln (P(\omega_k))$ est ***minimum*** !!!
  règle de Fisher généralisée (dépendant des proba. *a priori*): favorise les classes fortement représentées !

Prise de décision Bayesienne dans le cas d'un mélange Gaussien

- Supposons que les observations de chaque classe sont générées par une Gaussienne:

$$\mathrm{P}(\mathbf{z}|y=\omega_k) = \frac{1}{(2\pi)^{p/2}|\mathbf{W}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\mathbf{g}(k))^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{z}-\mathbf{g}(k))\right]$$

- Où **z** est le vecteur des caractéristiques projeté sur les axes discriminants retenus

  = modèle paramétrique
- Dans ce cas-ci, nous supposons qu'il y a une matrice variance-covariance **W** égale pour toutes les classes

Prise de décision Bayesienne dans le cas d'un mélange Gaussien

- Nous devons calculer les probabilités à postériori

$$\begin{aligned} \mathrm{P}(y=\omega_i|\mathbf{z}) &= \frac{\mathrm{P}(y=\omega_i)\,\mathrm{P}(\mathbf{z}|y=\omega_i)}{\mathrm{P}(\mathbf{z})} \\ &= \frac{\mathrm{P}(\omega_i)\,\mathrm{P}(\mathbf{z}|y=\omega_i)}{\sum_{k=1}^{q}\mathrm{P}(\omega_k)\,\mathrm{P}(\mathbf{z}|y=\omega_k)} \end{aligned}$$

- Nous placons dès lors l'observation **z** dans la classe *k* telle que

$$\mathrm{P}(y=\omega_k|\mathbf{z}) > \mathrm{P}(y=\omega_i|\mathbf{z})$$

Prise de décision Bayesienne dans le cas d'un mélange Gaussien

- Ce qui est équivalent à

$$-\ln(\mathrm{P}(y=\omega_k|\mathbf{z}) < -\ln(\mathrm{P}(y=\omega_i|\mathbf{z}))$$

- Ou encore

$$(\mathbf{z}-\mathbf{g}(k))^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{z}-\mathbf{g}(k)) - 2\ln(\mathrm{P}(\omega_k)) + \ln(|\mathbf{W}|)$$
$$< (\mathbf{z}-\mathbf{g}(i))^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{z}-\mathbf{g}(i)) - 2\ln(\mathrm{P}(\omega_i)) + \ln(|\mathbf{W}|)$$

- Et donc comme la matrice $\mathbf{W}$ est commune,

$$(\mathbf{z}-\mathbf{g}(k))^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{z}-\mathbf{g}(k)) - 2\ln(\mathrm{P}(\omega_k))$$
$$< (\mathbf{z}-\mathbf{g}(i))^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{z}-\mathbf{g}(i)) - 2\ln(\mathrm{P}(\omega_i))$$

Généralisation au cas où les matrices variance-covariance $\mathbf{W}_k$ des classes ne sont pas égales: les fcts discriminantes du classif. bayesien deviennent:

$$y_k(\mathbf{x}) = -\ln|\mathbf{W}_k| - (\mathbf{x} - \mathbf{g}_k)^T\, \mathbf{W}_k^{-1}\, (\mathbf{x} - \mathbf{g}_k) + 2 \ln (P(\omega_k))$$

déterminant de la matrice

Dans ce cas, les ***surfaces de séparation*** entre 2 classes (définies par $y_k(\mathbf{x}) = y_j(\mathbf{x})$) ne sont ***plus linéaires*** => ***analyse discriminante quadratique***

En pratique:

La matrice $\mathbf{W}$, ou les matrices $\mathbf{W}_k$ doi(ven)t être estimée(s) à partir des exemples disponibles pour chaque classe, ***ainsi que*** les $P(\omega_k)$.

Lorsqu'on fait l'hypothèse d'égalité des matrices $\mathbf{W}_k$, la matrice $\mathbf{W}$ est obtenue par estimation 'poolée': $\mathbf{W}_{\text{pool}} = (n_1\, \mathbf{W}_1 + n_2\, \mathbf{W}_2 + ... + n_q\, \mathbf{W}_q)/n$ (n = effectif total)

L'usage et l'estimation de matrice particulière $\mathbf{W}_k$ demande que les effectifs de classe soient suffisamment importants ! Pour des faibles effectifs l'existence de $\mathbf{W}_k^{-1}$ n'est pas tjrs assurée, de même $|\mathbf{W}_k|$ peut être nul !