L'ANALYSE DISCRIMINANTE

Pierre-Louis GONZALEZ

ANALYSE DISCRIMINANTE

Prédire une variable qualitative à *k* classes
à l'aide de *p* prédicteurs

Deux aspects

- **Descriptif:** Quelles sont les combinaisons linéaires de variables qui permettent de séparer le mieux possible les ***k*** catégories ?

- **Décisionnel:** Un nouvel individu se présente pour lequel on connaît les valeurs des prédicteurs. Décider dans quelle catégorie il faut l'affecter

ANALYSE DISCRIMINANTE

Ensemble des méthodes utilisées pour prédire une variable qualitative à k catégories à l'aide de p prédicteurs.

EXEMPLES

Médecine Connaissant les symptômes présentés par un patient, peut-on porter un diagnostic sur sa maladie ?

Finance

- A partir des bilans d'une société, est-il possible d'estimer son risque de faillite à 2 ans ou 3 ans (scoring financier) ?

- Au moment d'une demande de prêt par un client, peut-on prévoir en fonction des caractéristiques du client, le risque de contentieux (credit scoring) ?

Pétrole

Au vu des analyses des carottes issues d'un forage, est-il possible de présumer de l'existence d'une nappe de pétrole ?

Télédétection

A partir de mesures par satellite des ondes réfléchies ou absorbées par le sol dans différentes longueurs d'onde, peut-on reconstituer automatiquement la nature du terrain étudié (forêt, sable, ville, mer...) ?

Marketing direct

Connaissant les caractéristiques d'un client, peut-on prévoir sa réponse à une offre de produit par courrier ?

Étude de textes

Interprétation d'une typologie

Quelques dates:

- Mahalanobis 1927
- Hotelling 1931
- Fisher 1936
- Rao 1950
- Anderson 1951
- Vapnik 1998

MÉTHODES GÉOMÉTRIQUES

Recherche des meilleures fonctions discriminantes

$$g\left(X_1, X_2 \ldots X_p\right)$$

LINÉAIRES

A.C.P. sur le nuage des centres de gravité des groupes munis de différentes métriques

NON LINÉAIRES

- **quadratique**
- **création de nouvelles variables** $f\left(X_1, X_2 \ldots X_p\right)$ **et application d'une méthode linéaire**
- **découpage en variables qualitatives et application d'une méthode sur variables qualitatives**

MÉTHODES PROBABILISTES

Estimation directe des probabilités d'appartenance aux groupes définis par *Y*

PARAMÉTRIQUE

- **modèle gaussien**
- **maximum de vraisemblance avec autre modèle paramétrique**

SEMI PARAMETRIQUE

- **régression logistique**

NON PARAMÉTRIQUE

- **Estimation de densité :**
- **Fonctions orthogonales**
- **Noyaux de Parzen**
- ***K* plus proches voisins**

Autres approches

- **Méthodes de type « boite noire » induisant le minimum d'erreurs de classement**
  - **Réseaux de neurones**
  - **SVM (Support Vecteur Machine)**

I. MÉTHODES GÉOMÉTRIQUES

1. Données - Notations

Les n individus $\underline{e_i}$ de l'échantillon constituent un nuage E, de $\mathbf{R}^p$ partagé en k sous-nuages : $E_1, E_2 \dots E_k$ de centres de gravité $\underline{g}_1, \underline{g}_2 \dots \underline{g}_k$

de matrices de variances $V_1, V_2 \dots V_k$

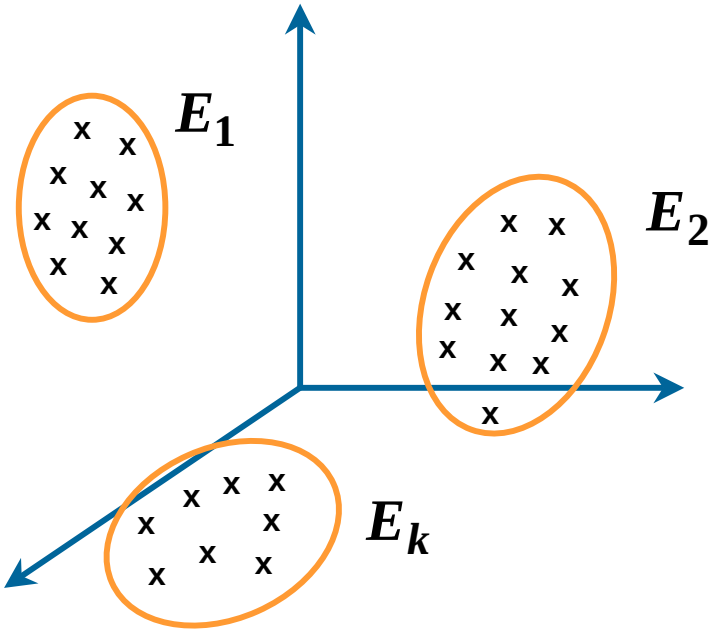
g = centre de gravité de E

V = matrice de variance de E

n individus e_i affectés des poids

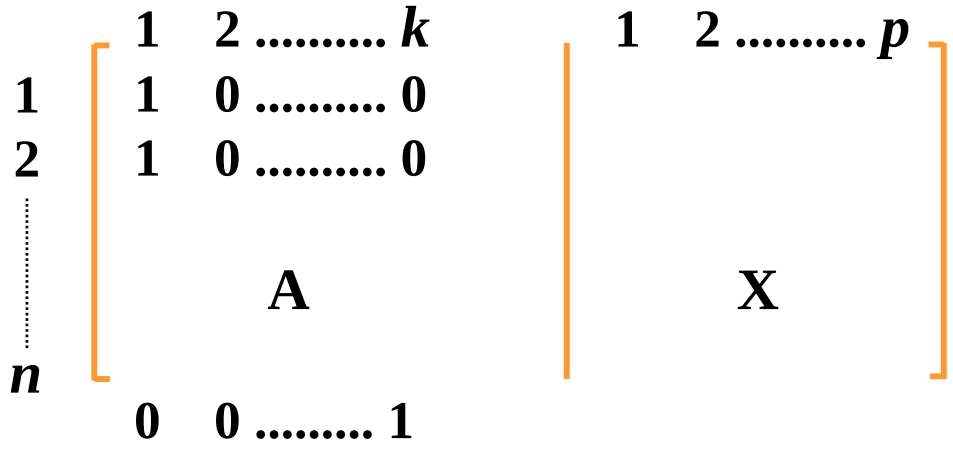
$p_1, p_2 \dots p_n$

rangés dans une matrice diagonale D



Notations matricielles

tableau de données



A Matrice des indicatrices de la variable qualitative à prédire

X Matrice des prédicteurs

$D_q = A'DA$ matrice diagonale des poids q_j des sous-nuages.

$(A'DA)^{-1}(A'DX)$ ses lignes sont les coordonnées des k centres de gravité $\underline{g}_1, \underline{g}_2 \cdots \underline{g}_k$

poids de la classe *j*

$$q_j = \sum_{e_i \in E_j} p_i$$

Centres de gravité

$$\underline{g}_j = \frac{1}{q_j} \sum_i p_i \underline{e}_i \quad \text{pour} \quad e_i \in E_j \qquad \underline{g} = \sum_{j=1}^{k} q_j \underline{g}_j$$

Matrice de variance-covariances de la classe Ej

$$V_j = \frac{1}{q_j} \sum_{\underline{e}_i \in E_j} p_i \left(\underline{e}_i - \underline{g}_j\right)\left(\underline{e}_i - \underline{g}_j\right)'$$

Matrice de variance interclasse : matrice de variance $\boldsymbol{B}$ des $\boldsymbol{k}$ centres de gravité affectés des poids $\boldsymbol{q_j}$:

$$\boldsymbol{B} = \sum_{j=1}^{k} q_j \left(\underline{g}_j - \underline{g}\right)\left(\underline{g}_j - \underline{g}\right)'$$

Matrice de variance intra-classe :

$$\boldsymbol{W} = \sum_{j=1}^{k} q_j V_j$$

En règle générale ***W*** inversible

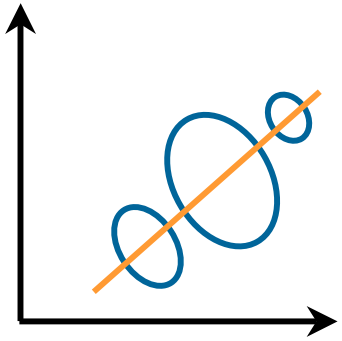
B non inversible (***k*** centres de gravité dans un sous-espace de dimension ***k-1*** de $\mathbf{R}^{p}$)

$$V = W + B$$

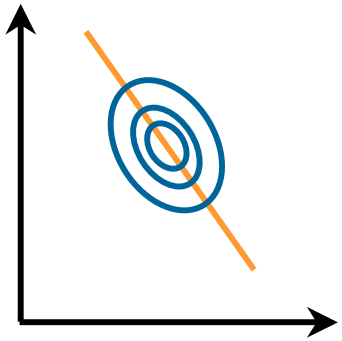
Variance totale = Moyenne des variances + Variance des moyennes

En analyse discriminante, on considère trois types de matrices

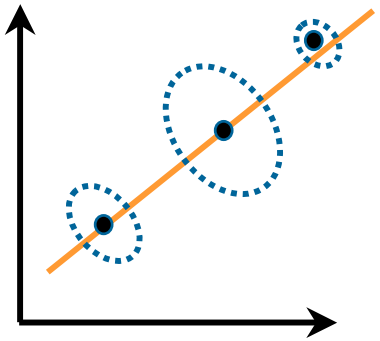
de variances-covariances et donc **trois types de corrélations**.



corrélation totale



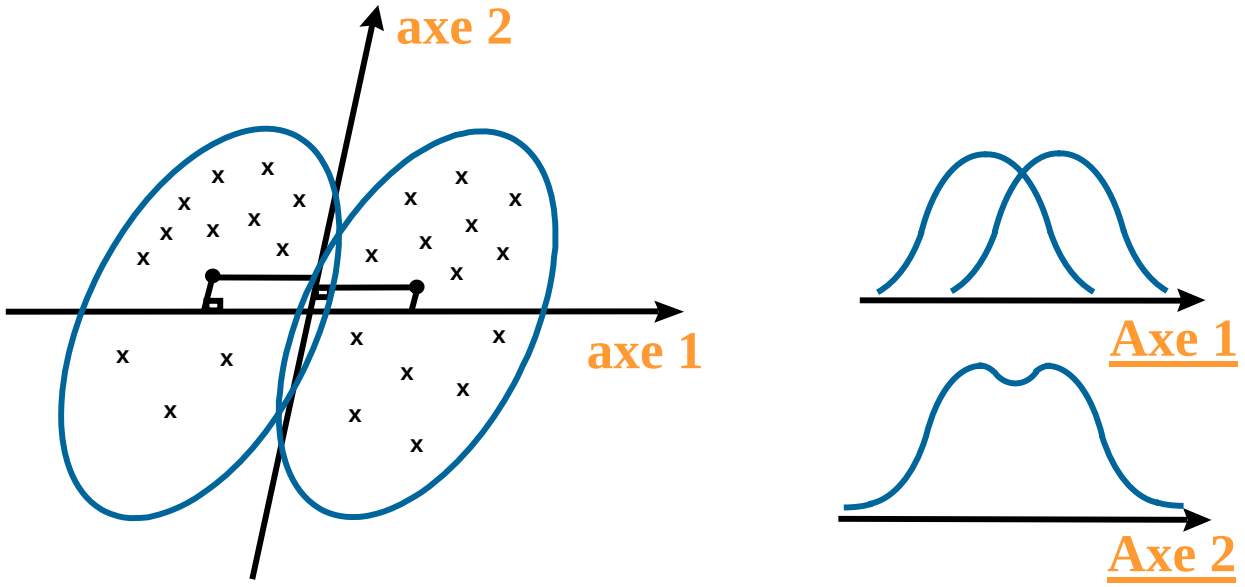
corrélation intra-classes



corrélation inter-classes

2. L'analyse factorielle discriminante (A.F.D.)

Elle consiste à chercher de nouvelles variables (les **variables discriminantes**) correspondant à des directions de $\mathbf{R}^p$ qui séparent le mieux possible en projection les k groupes d'observations.



L'axe 1 possède un bon pouvoir discriminant
L'axe 2 ne permet pas de séparer en projection les 2 groupes.

Supposons $\mathbf{R}^p$ muni d'une métrique M (calcul des distances)

$\underline{a}$ = **axe discriminant**

$\underline{u}$ = **facteur associé** $\qquad \underline{u} = M\underline{a}$

$X\underline{u}$ = **variable discriminante**

L'inertie du nuage des g_j projetés sur $\underline{a}$ doit être maximale.

La matrice d'inertie du nuage des $\underline{g}$ est MBM, l'inertie du nuage projeté sur $\underline{a}$ est $\underline{a}'MBM\underline{a}$ si $\underline{a}$ est M-normé à 1.

Il faut aussi qu'en projection sur $\underline{a}$, chaque sous-nuage reste bien groupé

donc que $\underline{a}'MV_JM\underline{a}$ soit faible pour $j = 1, 2 \ldots k$.

On cherchera donc à minimiser :

$$\sum_{j=1}^{k} q_j \underline{a}'MV_jM\underline{a} \quad \text{soit} \quad \underline{a}'MWM\underline{a}$$

Critère

La relation $\boldsymbol{V = B + W}$ entraîne que $\boldsymbol{MVM = MBM + MWM}$

donc : $\boldsymbol{\underline{a}' MVM \underline{a} = \underline{a}' MBM \underline{a} + \underline{a}' MWM \underline{a}}$

Maximiser le rapport de l'inertie inter-classe à l'inertie totale

$$\max_{\underline{a}} \frac{\underline{a}' MBM \underline{a}}{\underline{a}' MVM \underline{a}}$$

Ce maximum est atteint si $\boldsymbol{\underline{a}}$ est vecteur propre de $\boldsymbol{(MVM)^{-1}(MBM)}$ associé à sa plus grande valeur propre $\boldsymbol{\lambda_1}$

$$\boldsymbol{M^{-1}V^{-1}BM\underline{a} = \lambda \underline{a}}$$

A l'axe discriminant $\underline{a}$ est alors associé le facteur discriminant $\underline{u}$ tel que :

$$\underline{u} = M\,\underline{a}$$

On a alors : $V^{-1}B\underline{u} = \lambda\underline{u}$

Les facteurs discriminants, donc les variables discriminantes $X\,\underline{u}$ sont indépendantes de la métrique M.

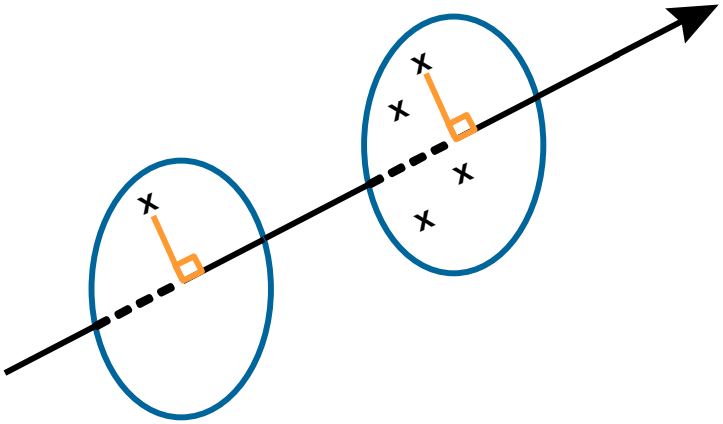
On choisira par commodité $M = V^{-1}$

$$\begin{cases} BV^{-1}\underline{a} = \lambda\underline{a} \\ V^{-1}B\underline{u} = \lambda\underline{u} \end{cases}$$

On a toujours $0 \le \lambda_1 \le 1$ car λ_1 est la quantité à maximiser.

Cas particuliers

Cas $\lambda_1 = 1$



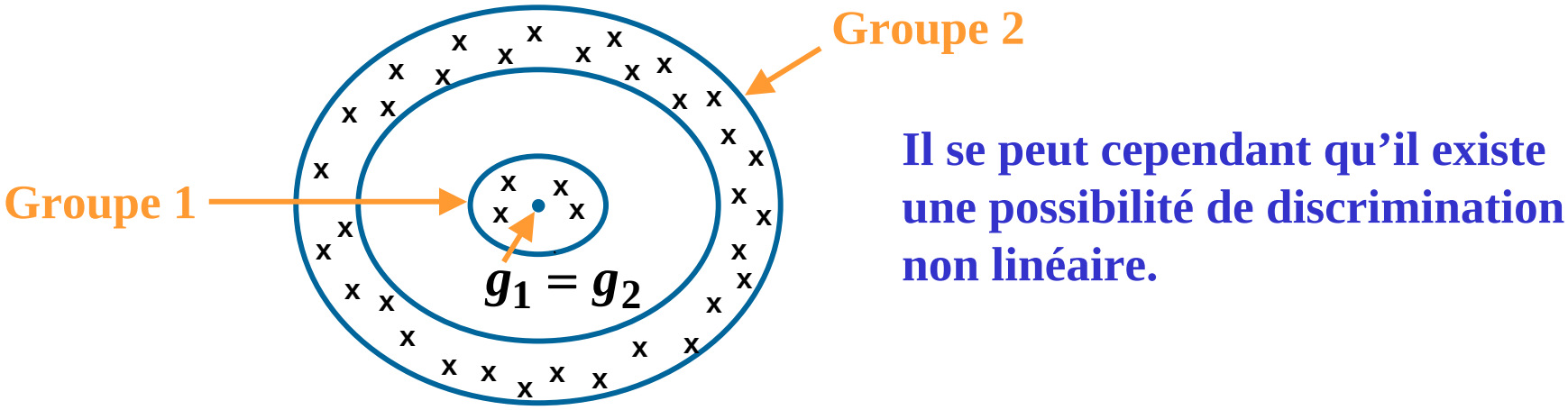
En projection sur $\underline{\boldsymbol{a}}$ les dispersions intra-classes sont nulles. Les $\boldsymbol{k}$ nuages sont donc chacun dans un hyperplan orthogonal à $\underline{\boldsymbol{a}}$.

Il y a discrimination parfaite si les centres de gravité se projettent en des points différents.

Cas $\lambda_1 = 0$

Le meilleur axe ne permet pas de séparer les centres de gravité g_i , c'est le cas où ils sont confondus.

Les nuages sont donc concentriques et aucune séparation linéaire n'est possible.

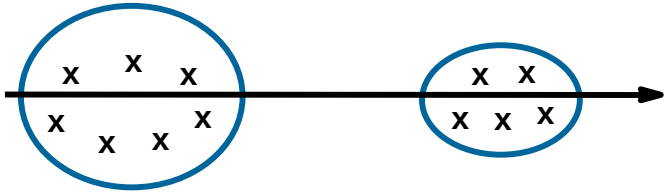


Il se peut cependant qu'il existe une possibilité de discrimination non linéaire.

La distance au centre permet ici de séparer les groupes, mais il s'agit d'une fonction quadratique des variables.

Autres propriétés

La valeur propre λ est une mesure pessimiste du pouvoir discriminant d'un axe.



$\lambda < 1$ mais les groupes sont bien séparés

Le nombre des valeurs propres non nulles, donc d'axes discriminants est égal à $k-1$ dans le cas habituel où $n > p > k$ et où les variables ne sont pas liées par des relations linéaires.

Remarque: Le cas de deux groupes

Il n'y a qu'une seule variable discriminante puisque $k-1=1$.

L'axe discriminant est alors nécessairement la droite reliant les deux centres de gravité $\underline{g}_2$ et $\underline{g}_1$: $\underline{a}=\underline{g}_1-\underline{g}_2$

Le facteur discriminant $\underline{u}$ vaut donc : $\underline{u}=V^{-1}\left(\underline{g}_1-\underline{g}_2\right)$

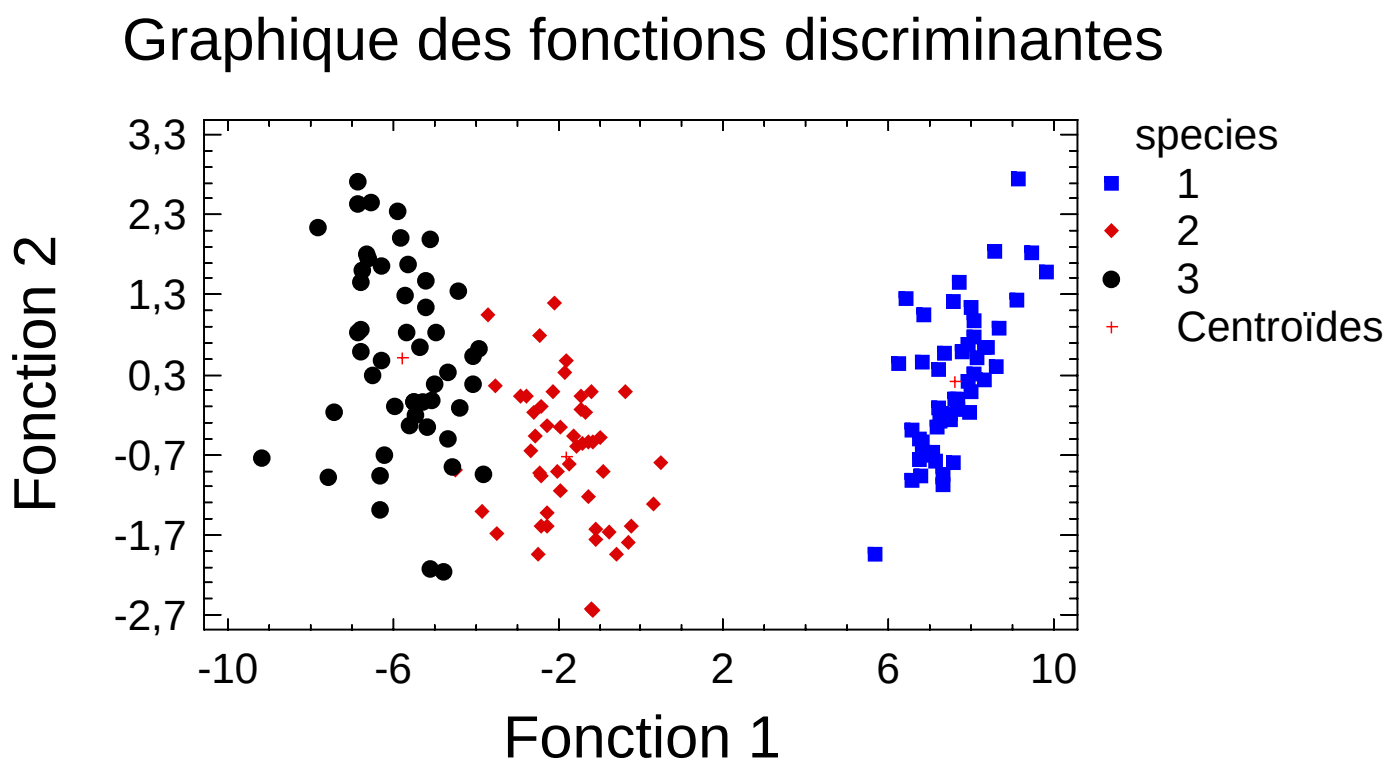
ou $\underline{u}=W^{-1}\left(\underline{g}_1-\underline{g}_2\right)$ qui lui est proportionnel

$W^{-1}\left(\underline{g}_1-\underline{g}_2\right)$ **est la fonction de Fisher (1936).**

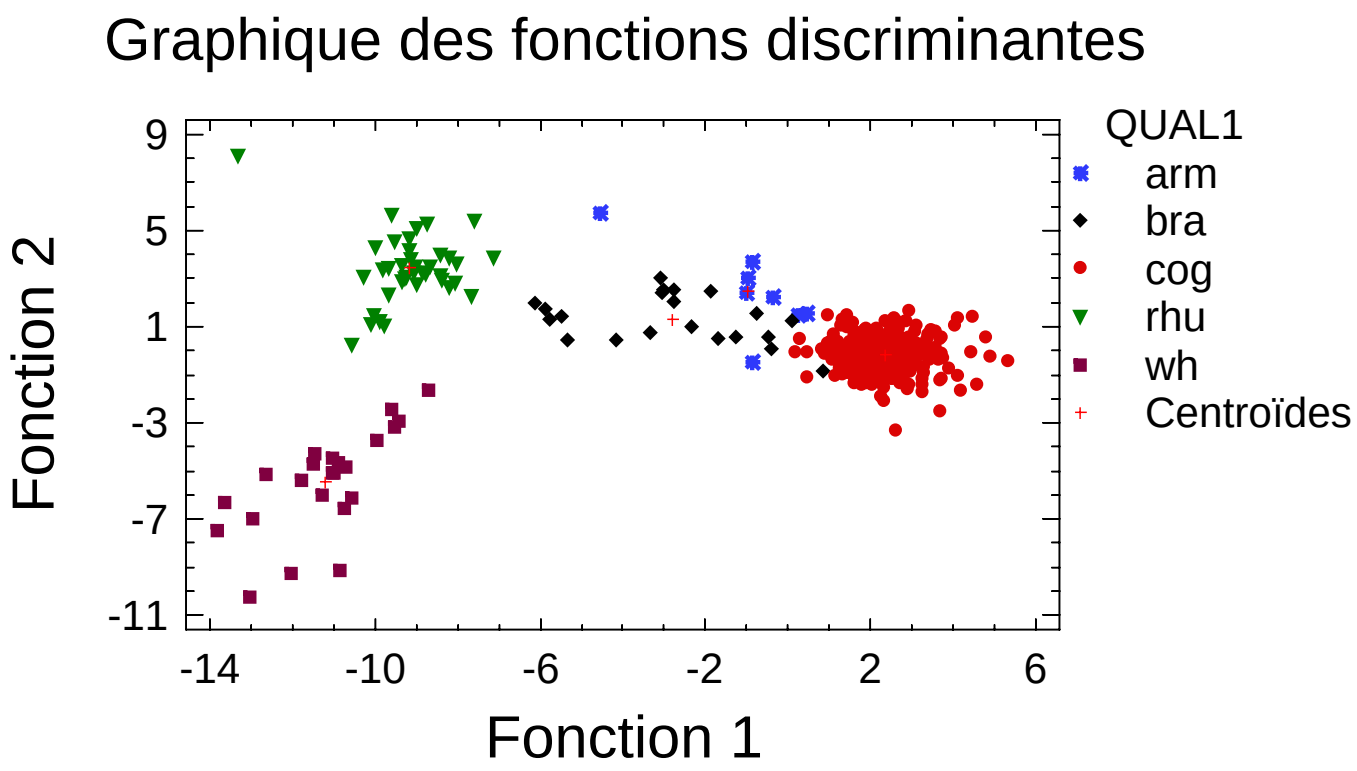
3. Exemples: Les iris de Fisher

		
Iris setosa	Iris versicolor	Iris virginica

Les iris de Fisher



Discrimination entre divers spiritueux à l'aide de dosages moléculaires



4. Une A.C.P. particulière

D'après les équations précédentes, l'analyse factorielle discriminante n'est autre que l'A.C.P. du nuage des k centres de gravité avec la métrique V^{-1} .

On en déduit que les variables discriminantes sont non corrélées deux à deux.

Dans le cas où il existe plusieurs axes discriminants $(k > 2)$ on peut utiliser les représentations graphiques usuelles de l'A.C.P. : cercle des corrélations...

5. Règles géométriques d'affectation

Ayant trouvé la meilleure représentation de la séparation en k groupes des n individus, on peut alors chercher à affecter une observation $\underline{e}$ à l'un des groupes.

La règle naturelle consiste à calculer les distances de l'observation à classer à chacun des k centres de gravité et à affecter selon la distance la plus faible. Métrique à utiliser ?

Règle de Mahalanobis Fisher

On utilise W^{-1}

$$d^2(\underline{e};\underline{g}_i) = (\underline{e}-\underline{g}_i)' W^{-1} (\underline{e}-\underline{g}_i)$$

6. Exemple: Qualité des vins de Bordeaux

Les données

	Température	Soleil	Chaleur	Pluie	Qualité
1	3064	1201	10	361	2
2	3000	1053	11	338	3
3	3155	1133	19	393	2
4	3085	970	4	467	3
5	3245	1258	36	294	1
6	3267	1386	35	225	1
7	3080	966	13	417	3
8	2974	1189	12	488	3
9	3038	1103	14	677	3
10	3318	1310	29	427	2
11	3317	1362	25	326	1
12	3182	1171	28	326	3
13	2998	1102	9	349	3
14	3221	1424	21	382	1
15	3019	1230	16	275	2
16	3022	1285	9	303	2
17	3094	1329	11	339	2
18	3009	1210	15	536	3
19	3227	1331	21	414	2
20	3308	1366	24	282	1
21	3212	1289	17	302	2
22	3361	1444	25	253	1
23	3061	1175	12	261	2
24	3478	1317	42	259	1
25	3126	1248	11	315	2
26	3458	1508	43	286	1
27	3252	1361	26	346	2
28	3052	1186	14	443	3
29	3270	1399	24	306	1
30	3198	1259	20	367	1
31	2904	1164	6	311	3
32	3247	1277	19	375	1
33	3083	1195	5	441	3
34	3043	1208	14	371	3

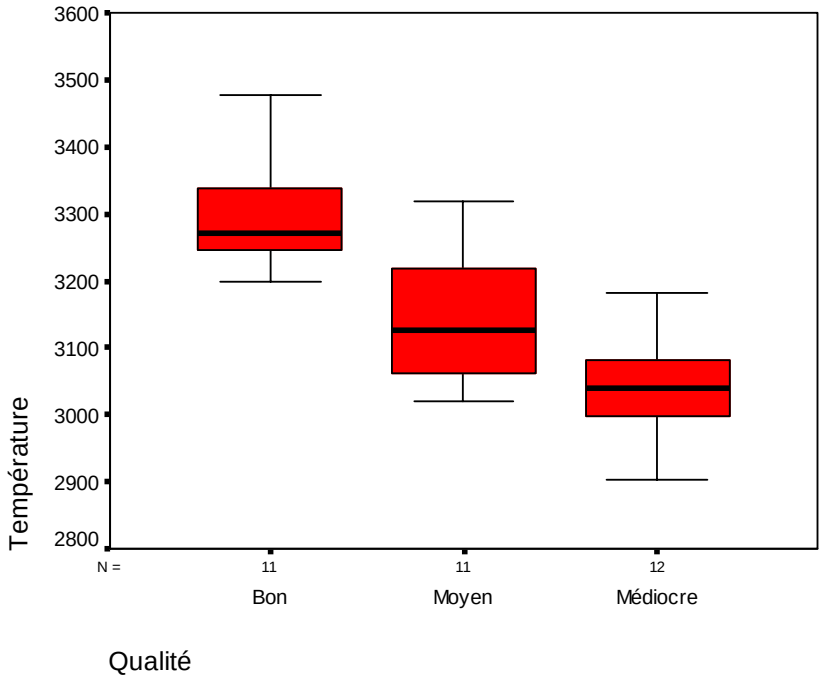
Analyse préalable

Température

Report

Température

Qualité	Mean	N	Std. Deviation
1	3306.36	11	92.06
2	3140.91	11	100.05
3	3037.33	12	69.34
Total	3157.88	34	141.18



Measures of Association

	Eta	Eta Squared
Température * Qualité	.799	.639

$$\text{Rapport de corrélation} = \eta^2 = \frac{\text{Between Groups Sum of Squares}}{\text{Total Sum of Squares}}$$

ANOVA Table

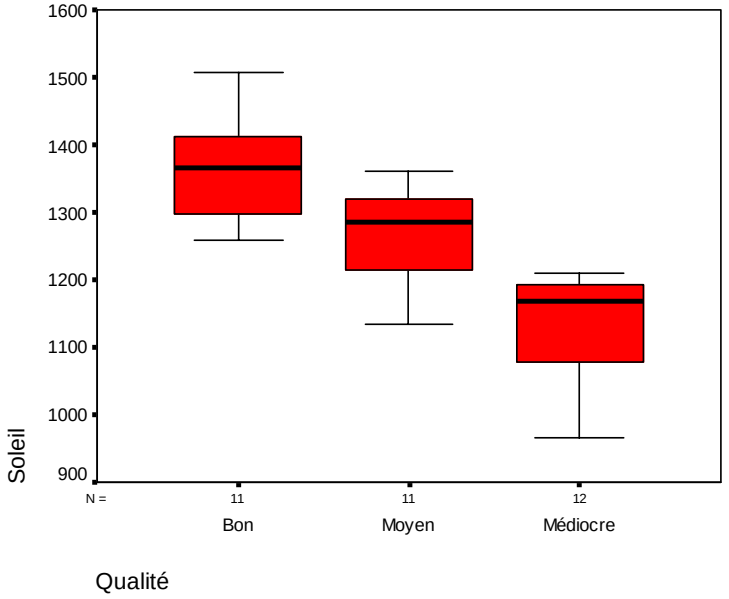
			Sum of Squares	df	Mean Square	F	Sig.
Température * Qualité	Between Groups	(Combined)	420067.4	2	210033.704	27.389	.000
	Within Groups		237722.1	31	7668.456		
	Total		657789.5	33			

Soleil

Report

Soleil

Qualité	Mean	N	Std. Deviation
Bon	1363.64	11	80.31
Moyen	1262.91	11	71.94
Médiocre	1126.42	12	88.39
Total	1247.32	34	126.62



Measures of Association

	Eta	Eta Squared
Soleil * Qualité	.786	.618

ANOVA Table

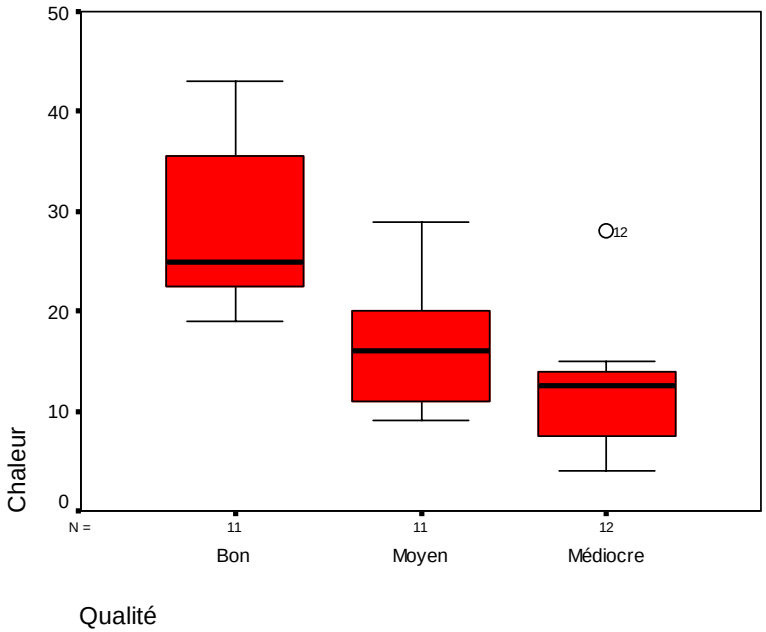
			Sum of Squares	df	Mean Square	F	Sig.
Soleil * Qualité	Between Groups	(Combined)	326909.1	2	163454.535	25.061	.000
	Within Groups		202192.4	31	6522.335		
	Total		529101.4	33			

Chaleur

Report

Chaleur

Qualité	Mean	N	Std. Deviation
Bon	28.55	11	8.80
Moyen	16.45	11	6.73
Médiocre	12.08	12	6.30
Total	18.82	34	10.02



Measures of Association

	Eta	Eta Squared
Chaleur * Qualité	.705	.497

ANOVA Table

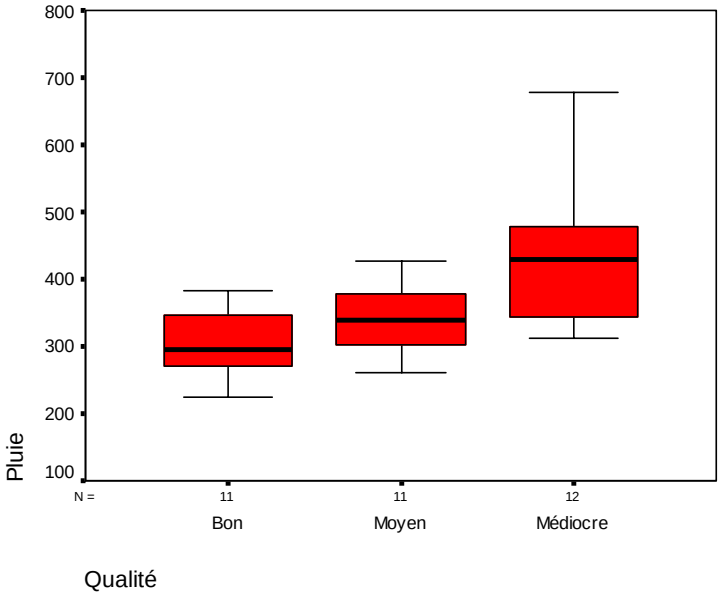
			Sum of Squares	df	Mean Square	F	Sig.
Chaleur * Qualité	Between Groups	(Combined)	1646.570	2	823.285	15.334	.000
	Within Groups		1664.371	31	53.689		
	Total		3310.941	33			

Pluie

Report

Pluie

Qualité	Mean	N	Std. Deviation
Bon	305.00	11	52.29
Moyen	339.64	11	54.99
Médiocre	430.33	12	104.85
Total	360.44	34	91.40



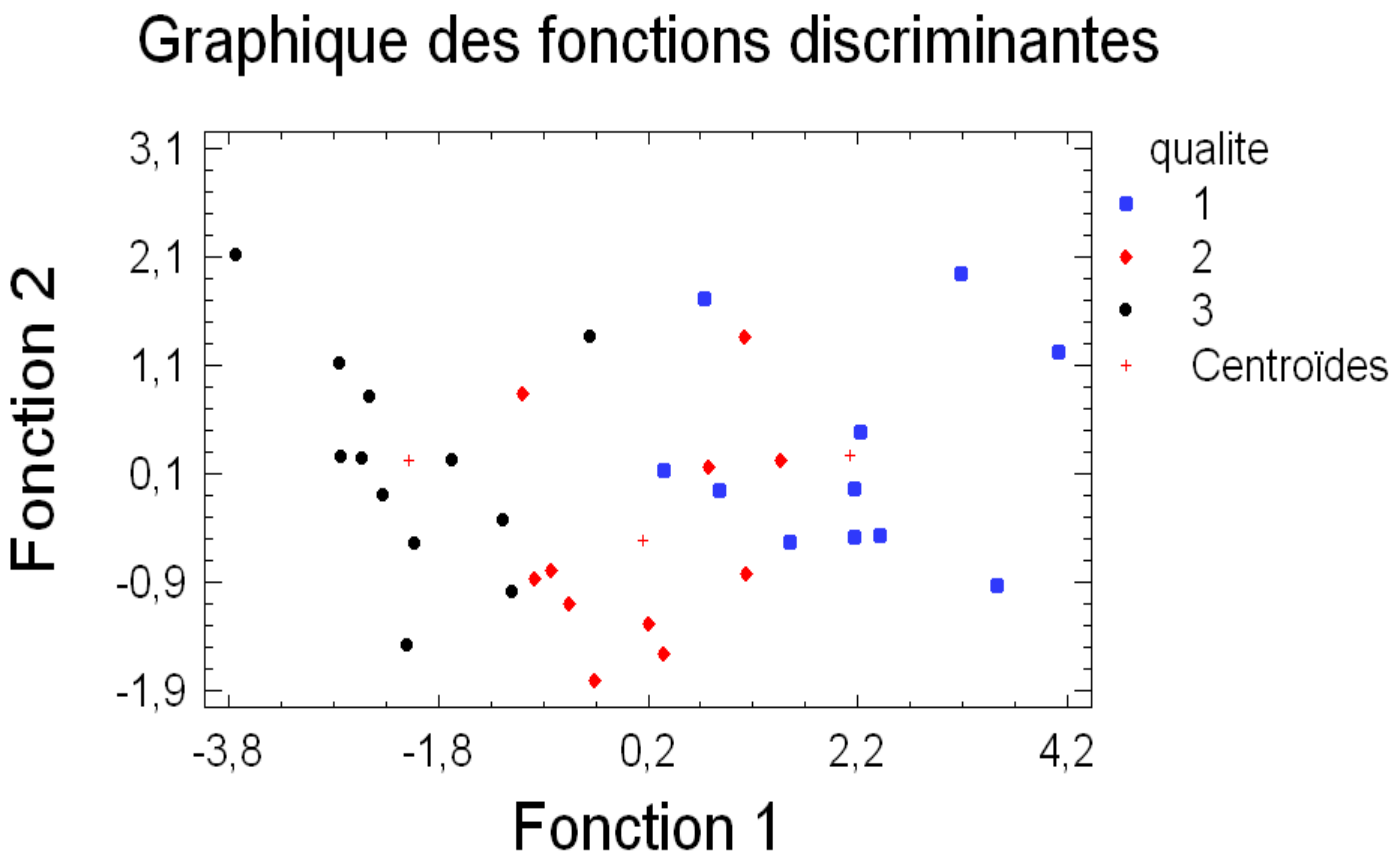
Measures of Association

	Eta	Eta Squared
Pluie * Qualité	.594	.353

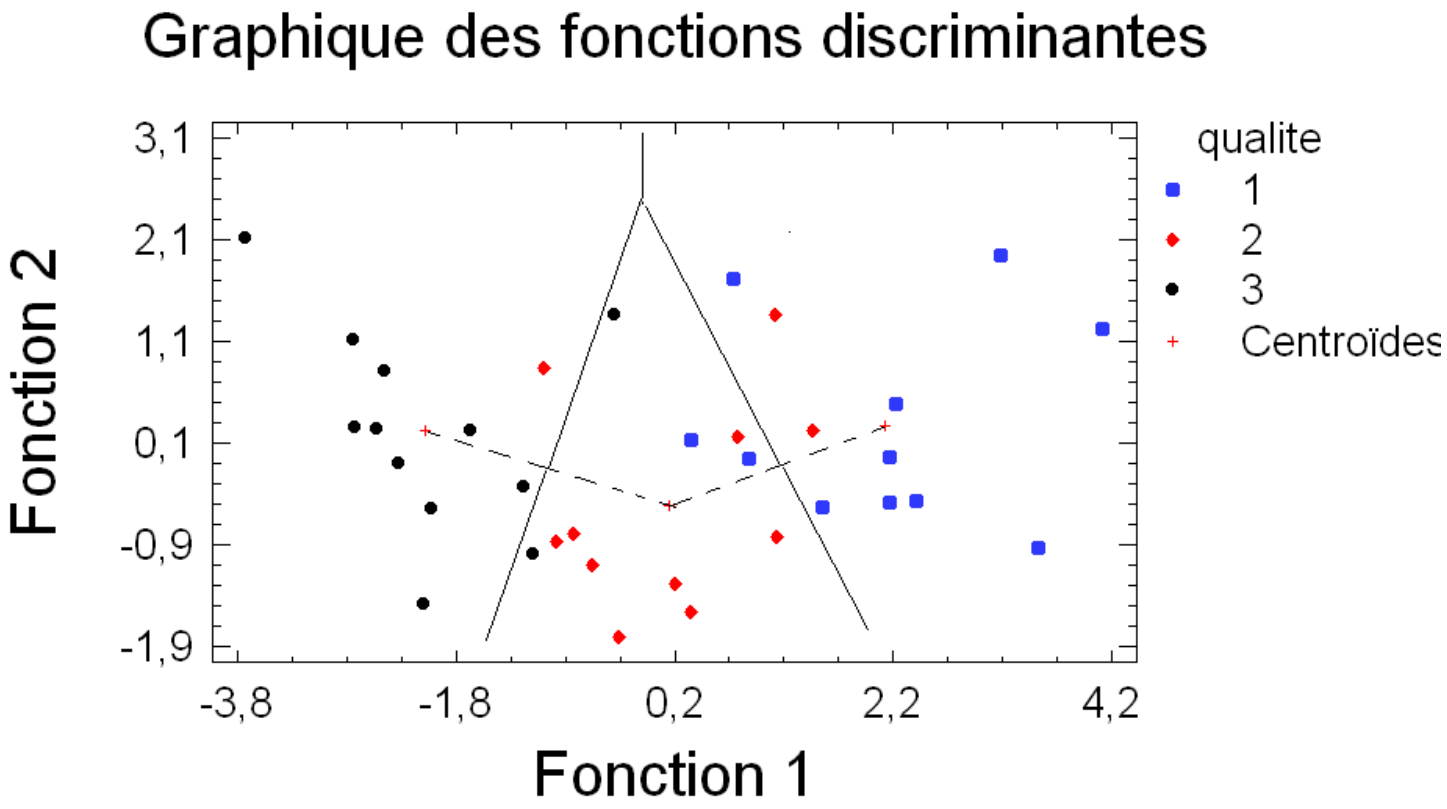
ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Pluie * Qualité	Between Groups	(Combined)	97191.170	2	48595.585	8.440	.001
	Within Groups		178499.2	31	5758.039		
	Total		275690.4	33			

Qualité des vins de Bordeaux



Qualité des vins de Bordeaux



Qualité des vins de Bordeaux: Pourcentage de bien classés

Tableau de classement

Observé qualite	Taille	Groupe Prévu qualité 1	2	3
1	11	9	2	0
		(81,82%)	(18,18%)	(0,00%)
2	11	2	8	1
		(18,18%)	(72,73%)	(9,09%)
3	12	0	2	10
		(0,00%)	(16,67%)	(83,33%)

Pourcentage d'observations bien classées: 79,41%

Qualité des vins de Bordeaux
Fonctions discriminantes

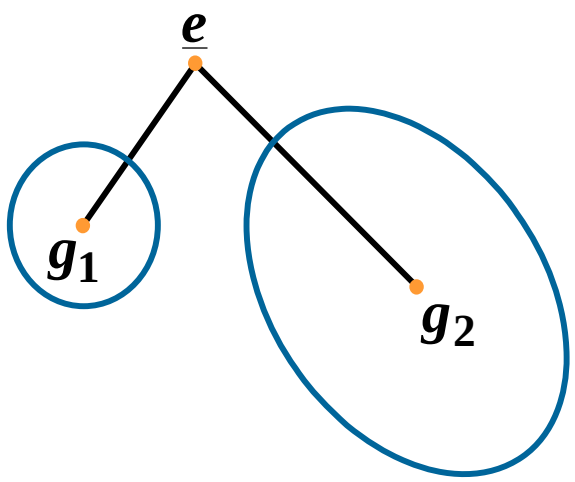
Coefficients des fonctions discriminantes pour qualité

Coefficients standardisés

	1	2
stemp	0,750126	-0,00405015
insol	0,547064	-0,430399
chaleur	-0,198237	0,935229
hpluies	-0,445097	0,468536

7. Insuffisance des règles géométriques

L'utilisation de la règle précédente conduit à des affectations incorrectes lorsque les dispersions des groupes sont très différentes entre elles : rien ne justifie alors l'usage de la même métrique pour les différents groupes.



$\underline{e}$ plus proche de g_1 que de g_2 au sens habituel.

Pourtant, il est plus naturel d'affecter $\underline{e}$ à la deuxième classe qu'à la première dont le pouvoir d'attraction est moindre.

Solution : métriques locales $\boldsymbol{M_i}$

Dans la plupart des cas, on choisit $\boldsymbol{M_i}$ proportionnel à $\boldsymbol{V_i^{-1}}$.

La question de l'optimalité d'une règle de décision géométrique ne peut cependant être résolue sans référence à un **modèle probabiliste**.

8. Remarques concernant la présentation de l'analyse discriminante dans les logiciels « américains »

8.1. Par ses liens avec l'analyse canonique, les auteurs de langue anglaise utilisent le terme : « ***ANALYSE DISCRIMINANTE CANONIQUE*** ».

On cherche la combinaison linéaire des variables qui a le plus grand coefficient de corrélation multiple avec la variable de classe.

- **Ce coefficient de corrélation est appelé première corrélation canonique.**

La valeur propre λ_1 (équation $\boldsymbol{V^{-1}B\underline{u} = \lambda_1 \underline{u}}$) est égale au carré de ce coefficient de corrélation.

- La variable définie par la combinaison linéaire est appelée la première composante canonique ou première variable canonique.

La deuxième variable canonique répond à deux critères :

- ne pas être corrélée avec la première,

- avoir le plus grand coefficient de corrélation multiple possible avec la variable de classe.

Ce processus peut être répété jusqu'au moment où le nombre de variables canoniques est égal au nombre de variables de départ ou au nombre de classes moins 1 s'il est plus petit.

8.2. Analyse de variance et métrique W^{-1}

S'il n'y avait qu'une seule variable explicative, on mesurerait l'efficacité de son pouvoir séparateur sur la variable de groupe au moyen d'une analyse de variance ordinaire à 1 facteur :

$$F = \frac{\text{Variance inter} / k-1}{\text{Variance intra} / n-k}$$

Comme il y a $\boldsymbol{p}$ variables, on peut rechercher la combinaison linéaire définie par des coefficients $\boldsymbol{u}$ donnant la valeur maximale pour la statistique de test, ce qui revient à maximiser :

$$\frac{\underline{u}' B \underline{u}}{\underline{u}' W \underline{u}}$$

La solution est donnée par l'équation :

$$W^{-1} B \underline{u} = \mu \underline{u} \quad \text{avec } \mu \text{ maximal}$$

Les vecteurs propres de $\boldsymbol{W^{-1}B}$ sont les mêmes que ceux de $\boldsymbol{V^{-1}B}$

avec $\mu = \dfrac{\lambda}{1-\lambda} \Leftrightarrow \lambda = \dfrac{\mu}{1+\mu}$

Les logiciels « américains » fournissent cette valeur propre μ :

si : $\mathbf{0 \leq \lambda \leq 1}$

on a en revanche : $\mathbf{0 \leq \mu \leq \infty}$

A ce point près, l'utilisation de V^{-1} ou de W^{-1} comme métrique est indifférente.

9. Analyse canonique discriminante et régression

L'analyse canonique discriminante, se réduit dans le cas de deux groupes à une régression multiple.

En effet après avoir centré, l'espace engendré par les deux indicatrices de la variable des groupes est de dimension ***1***.

Il suffit donc de définir une variable centrée $\boldsymbol{Y}$ ne prenant que les deux valeurs $\boldsymbol{a}$ et $\boldsymbol{b}$ sur les groupes ***1*** et ***2***.

$$\left(\boldsymbol{n_1 a + n_2 b = 0}\right)$$

On obtiendra alors un vecteur des coefficients de régression proportionnel à la fonction de Fisher pour un choix quelconque de $\boldsymbol{a}$.

IMPORTANT

On prendra garde au fait que les hypothèses habituelles de la régression ne sont pas vérifiées, bien au contraire :

Ici Y est non aléatoire

X l'est.

Ne pas utiliser, autrement qu'à titre indicatif, les statistiques usuelles fournies par un programme de régression.

10. Analyse discriminante sur variables qualitatives

$\boldsymbol{Y}$: variable de groupe

$\boldsymbol{\chi_1, \chi_2, \ldots \chi_p}$ variables explicatives à $\boldsymbol{m_1, m_2, \ldots m_p}$ modalités.

Exemples

- **Solvabilité d'emprunteurs auprès de banques**

Y → bon payeur
Y → mauvais payeur

χ_1 : sexe χ_2 : catégorie professionnelle

- **Risque en assurance automobile**

Y → bon conducteur
Y → mauvais conducteur

χ_1 : sexe χ_2 : tranche d'âge χ_3 : véhicule sportif ou non

- **Reclassement dans une typologie**

Y : classes

Caractéristiques du problème

- Grand nombre de prédicteurs qualitatifs
- Échantillons volumineux

Méthodes classiques inadaptées

- Analyse discriminante classiques : variables quantitatives
- Modèle log linéaire : trop de variables

D I S Q U A L

Méthode de discrimination fondée sur l'analyse factorielle

Prédicteurs qualitatifs

Estimer $P\left(Y = y / \chi_1 = x_1 \quad \chi_2 = x_2 \quad ...\right)$

- Approche multinomiale irréaliste

  P estimé par la fréquence

  $\prod_{i=1}^{k} m_i$ cases !

- Approche modèle

  Log-linéaire, linéaire, on néglige certaines interactions.

  Ex : $\left| n \;\; P\left(Y = y \,\middle|\, \chi_1 = i, \chi_2 = j, \chi_3 = k\right)\right.$

  $= \alpha_0 + \alpha_i + \beta_j + \sigma_k + \delta_{ij} + \varepsilon_{ik}$

Une méthode de discrimination sur variables qualitatives : la méthode DISQUAL

Les p prédicteurs sont p variables qualitatives $\chi_1\ \chi_2 \cdots \chi_p$ à $m_1\ m_2 \ldots m_p$ modalités.

1ère étape A.C.M. des variables $\chi_1\ \chi_2 \cdots \chi_m$

$\Leftrightarrow$ Analyse des correspondances du tableau disjonctif

$$X = \left(X_1 \middle| X_2 \middle| \ldots \middle| X_p\right)$$

2ème étape

On remplace les p variables qualitatives par les q coordonnées sur les axes factoriels

→ analyse discriminante sur ces q variables numériques Z_1 Z_2 ... Z_q

Facteur discriminant d = combinaison linéaire des Z_j qui sont des combinaisons linéaires des indicatrices.

3ème étape

Expression de ***d*** comme combinaison linéaire des indicatrices

$\Leftrightarrow$ attribuer à chaque catégorie de chaque variable une valeur numérique ou score.

Ceci revient donc à transformer chaque variable qualitative en une variable discrète à ***m*** valeurs (associées à chaque modalité).