Analyse discriminante, classification supervisée, scoring…

L'analyse factorielle discriminante

Discrimination sur variables qualitatives : scoring

Analyse discriminante probabiliste

Ce cours est basé sur celui de Gilbert Saporta

Bibliographie

Bardos: « Analyse discriminante », Dunod, 2001
Hastie, Tibshirani, Friedman : « The Elements of Statistical Learning », 2nd edition, Springer-Verlag, 2009
http://www.stanford.edu/~hastie/local.ftp/Springer/ESLII_print10.pdf
Nakache, Confais: « Statistique explicative appliquée », Technip, 2003
Thiria et al. :« Statistique et méthodes neuronales » Dunod, 1997
Thomas, Edelman,Crook: « Credit scoring and its applications », SIAM, 2002
Tufféry: « Data Mining et statistique décisionnelle »,4ème édition, Technip, 2012
Tufféry: «Étude de cas en statistique décisionnelle »,Technip, 2009
Vapnik : « Statistical Learning Theory », Wiley 1998

Introduction

But : étude et mise en évidence des liaisons entre une variable qualitative à k modalités à l'aide de p variables explicatives (numériques)

Observations multidimensionnelles réparties en k groupes définis a priori.

Méthode de classement, de discrimination

Autre terminologie: classification supervisée

Introduction

- **Double aspect:**

  descriptif

- Analyse factorielle discriminante, méthodes géométriques:

  description des liaisons entre les variables, recherche de combinaisons linéaires des variables explicatives qui expliquent au mieux les k modalités

- A l 'aide de représentations graphiques, plans factoriels

Introduction

- **Double aspect:**

  décisionnel, classement
- Analyse discriminante bayésienne, méthodes probabilistes

- prévision des modalités de la variable à expliquer à partir des valeurs prises pour les variables explicatives par un nouvel individu

Introduction

Domaine d'application

- *Médecine, météo, finance, RDF…*
- *Exemples :*
  - Pronostic des infarctus (J.P. Nakache)
    - *2 groupes : décès, survie (variables médicales)*
  - Iris de Fisher :
    - *3 espèces : 4 variables (longueur et largeur des pétales et sépales)*
  - Risque des demandeurs de crédit
    - *2 groupes : bons, mauvais (variables qualitatives)*

6

Plan du cours : $1^{ère}$ partie AFD

1. Données et notations
2. Exemple introductif
3. Réduction de dimension, axes et variables discriminantes.
4. Méthodes géométriques de classement..
5. Cas de 2 groupes

7

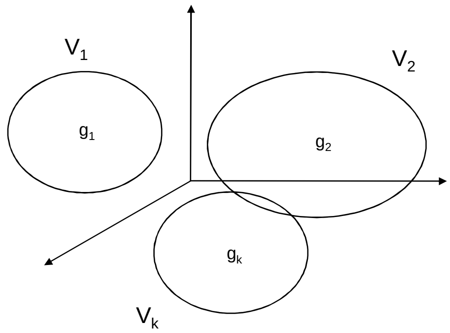
1. Données et notations

- Y variable à expliquer à k modalités partage la population en k sous populations, groupes, classes.
- On dispose d'un échantillon de n observations partagé en k classes de taille n_1 n_2…n_k
- 1 nuage de n points, k sous nuages de points de R^p
- p variables explicatives numériques

8

1. Données et notations

- g centre de gravité du nuage de n points
- V matrice de variance
- g_j centre de gravité et V_j matrice de variance du nuage des n_j points de la classe j



1. Données et notations exemple

Obs	ANNEE	TEMP	SOLEIL	CHAL	PLUIES	QUALITE
1	1924	3064	1201	10	361	2
2	1925	3000	1053	11	338	3
3	1926	3155	1133	19	393	2
4	1927	3085	970	4	467	3
5	1928	3245	1258	36	294	1
6	1929	3267	1386	35	225	1
7	1930	3080	966	13	417	3
8	1931	2974	1189	12	488	3
9	1932	3038	1103	14	677	3
10	1933	3318	1310	29	427	2
11	1934	3317	1362	25	326	1
12	1935	3182	1171	28	326	3
13	1936	2998	1102	9	349	3
14	1937	3221	1424	21	382	1
15	1938	3019	1230	16	275	2
16	1939	3022	1285	9	303	2
17	1940	3094	1329	11	339	2

Extrait du fichier de 34 vins bordeaux identifiés par l'année 4 variables quantitatives explicatives de la qualité des vins qui est la variable qualitative à 3 modalités:
1= bon
2=moyen
3= médiocre

Class Level Information				
QUALITE	Variable Name	Frequency	Weight	Proportion
1	_1	11	11.0000	0.323529
2	_2	11	11.0000	0.323529
3	_3	12	12.0000	0.352941

1. Données et notations exemple

QUALITE = 1					
Variable	**N**	**Sum**	**Mean**	**Variance**	**Standard Deviation**
TEMP	11	36370	3306	8474	92.0568
SOLEIL	11	15000	1364	6449	80.3060
CHAL	11	314.00000	28.54545	77.47273	8.8019
PLUIES	11	3355	305.00000	2735	52.2934

QUALITE = 3					
Variable	**N**	**Sum**	**Mean**	**Variance**	**Standard Deviation**
TEMP	12	36448	3037	4808	69.3389
SOLEIL	12	13517	1126	7813	88.3932
CHAL	12	145.00000	12.08333	39.71970	6.3024
PLUIES	12	5164	430.33333	10993	104.8456

QUALITE = 2					
Variable	**N**	**Sum**	**Mean**	**Variance**	**Standard Deviation**
TEMP	11	34550	3141	10009	100.0454
SOLEIL	11	13892	1263	5175	71.9409
CHAL	11	181.00000	16.45455	45.27273	6.7285
PLUIES	11	3736	339.63636	3023	54.9859

Total-Sample					
Variable	**N**	**Sum**	**Mean**	**Variance**	**Standard Deviation**
TEMP	34	107368	3158	19933	141.1843
SOLEIL	34	42409	1247	16033	126.6230
CHAL	34	640.00000	18.82353	100.33155	10.0166
PLUIES	34	12255	360.44118	8354	91.4016

11

1. Données et notations

Représentation des données

$$
\begin{array}{c}
\\ 1 \\ 2 \\ \\ i \\ \\ n
\end{array}
\begin{array}{c}
\begin{array}{cccc} 1 & 2 & \dots & k \end{array} \\
\left[\begin{array}{cccc}
0 & 1 & \dots & 0 \\
1 & 0 & \dots & 0 \\
 & & \dots & \\
0 & 0 & \dots & 1 \\
 & & & \\
1 & 0 & \dots & 0
\end{array}\right]
\end{array}
\quad
\begin{array}{c}
\begin{array}{cccc} 1 & 2 & j & p \end{array} \\
\left[\begin{array}{cccc}
X_1^1 & X_1^2 & X_1^j & X_1^p \\
 & & & \\
 & & & \\
X_i^1 & X_i^2 & X_i^j & X_i^p \\
 & & & \\
X_n^1 & X_n^2 & X_n^j & X_n^p
\end{array}\right]
\end{array}
$$

indicatrices des groupes — variables explicatives

Extrait du cours de Saporta

12

1.Données et notations

- Dispersion intergroupe - dispersion intra groupe.

W = matrice variance intra W $= 1/n \Sigma n_j V_j$

B = matrice variance inter B $= 1/n \Sigma n_i (g_i - g)(g_i - g)'$

V = W + B variance totale

1. Données et notations exemple

Between-Class Covariance Matrix, DF = 2				
Variable	TEMP	SOLEIL	CHAL	PLUIES
TEMP	18532.38566	15969.07046	1150.20683	-8284.08257
SOLEIL	15969.07046	14422.45897	962.28183	-7760.55100
CHAL	1150.20683	962.28183	72.64279	-487.05167
PLUIES	-8284.08257	-7760.55100	-487.05167	4287.84575

Pooled Within-Class Covariance Matrix, DF = 31				
Variable	TEMP	SOLEIL	CHAL	PLUIES
TEMP	7668.455523	1880.151515	461.320626	430.235582
SOLEIL	1880.151515	6522.334555	169.200635	-158.000978
CHAL	461.320626	169.200635	53.689394	-34.823069
PLUIES	430.235582	-158.000978	-34.823069	5758.039101

Total-Sample Covariance Matrix, DF = 33				
Variable	TEMP	SOLEIL	CHAL	PLUIES
TEMP	19933.01604	12734.85740	1223.40285	-5285.91622
SOLEIL	12734.85740	16033.37701	819.90731	-5478.90463
CHAL	1223.40285	819.90731	100.33155	-367.25312
PLUIES	-5285.91622	-5478.90463	-367.25312	8354.25401

Attention aux df

1. Données et notations exemple

Pooled Within-Class SSCP Matrix				
Variable	TEMP	SOLEIL	CHAL	PLUIES
TEMP	237722.1212	58284.6970	14300.9394	13337.3030
SOLEIL	58284.6970	202192.3712	5245.2197	-4898.0303
CHAL	14300.9394	5245.2197	1664.3712	-1079.5152
PLUIES	13337.3030	-4898.0303	-1079.5152	178499.2121

Between-Class SSCP Matrix				
Variable	TEMP	SOLEIL	CHAL	PLUIES
TEMP	420067.4082	361965.5971	26071.3547	-187772.5383
SOLEIL	361965.5971	326909.0700	21811.7215	-175905.8226
CHAL	26071.3547	21811.7215	1646.5700	-11039.8378
PLUIES	-187772.5383	-175905.8226	-11039.8378	97191.1702

Total-Sample SSCP Matrix				
Variable	TEMP	SOLEIL	CHAL	PLUIES
TEMP	657789.5294	420250.2941	40372.2941	-174435.2353
SOLEIL	420250.2941	529101.4412	27056.9412	-180803.8529
CHAL	40372.2941	27056.9412	3310.9412	-12119.3529
PLUIES	-174435.2353	-180803.8529	-12119.3529	275690.3824

Rappel introduction

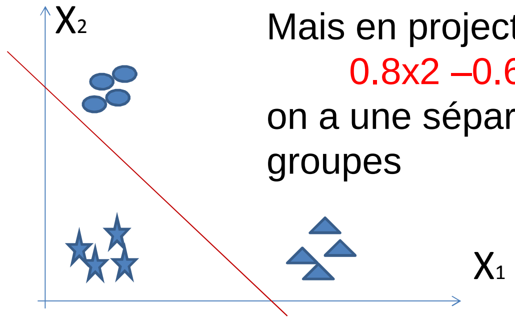
- Analyse factorielle discriminante a pour but la description, la visualisation des liaisons entre la variable qualitative à expliquer et les variables explicatives quantitatives.
- Elle répond à la question:

  Les groupes diffèrent-ils sur toutes les variables quantitatives?

- Cela passe par recherche de combinaisons linéaires des variables explicatives qui expliquent au mieux les k modalités. Les groupes sont alors visualisés à l 'aide de représentations graphiques, plans factoriels issus de ces combinaisons

2. Exemple introductif

- 3 groupes, 2 variables explicatives:
- Groupes 1 et 3 confondus sur l'axe 2
- Groupes 1 et 2 confondus sur l'axe 1



Mais en projection sur l'axe
0.8x2 –0.6x1
on a une séparation des 3 groupes

17

3. Réduction de dimension

Axes, facteurs, variables discriminantes

Approche directe

Rappel: V = W + B Totale = intra + inter

Recherche de combinaison linéaire : C= Xu

En supposant X centré, la variance de C se décompose en 2

$$Var(c) = c'Dc = u'X'Dxu = u'Vu = u'Wu + u'Bu$$

variance intra classe liée à la dispersion des observations des classes autour de leurs centres de gravité respectifs

variance inter classe liée à la dispersion des centres de gravité des classes autour de l'origine

18

3. Réduction de dimension

- Deux objectifs: *min u'Wu* *max u'Bu*

- Meilleur axe discriminant = max le rapport

- Compromis :

$$V = W + B$$

$$u' V u = \underbrace{u' W u}_{\min} + \underbrace{u' B u}_{\max}$$

$$\max \left(\frac{u' B u}{u' V u} \right) \quad ou \quad \left(\frac{u' B u}{u' W u} \right)$$

$$V^{-1} Bu = \lambda u \qquad W^{-1} Bu = \mu u$$

Interprétation en terme de pourcentage

19

3. Réduction de dimension

- Remarques mêmes vecteurs propres

$$\text{a) } V^{-1} Bu = \lambda u$$

$$Bu = \lambda V u$$

$$Bu = \lambda (W + B) u$$

$$(1-\lambda) Bu = \lambda W u$$

$$\text{b) } W^{-1} Bu = \frac{\lambda}{1-\lambda} u = \mu u$$

20

3. Réduction de dimension

Axes, facteurs, variables discriminantes

AFD= ACP particulière

ACP du nuage des g_i avec métrique V^{-1} ou W^{-1}

- En effet B est la matrice de variance des points g_j
- ACP : **axe** : a tq a'MVMa maximal

  a'MBMa maximal

  – Mais on a:
  – a'MVMa= a'MWMa+a'MBMa,
  – on prend comme critère le rapport inter/total

  a'MBMa/a'MVMa maximal

21

3. Réduction de dimension

– Solution (résultat général annulation dérivation vectorielle voir au tableau)

Axe a est vecteur propre de

$$(MVM)^{-1}MBM = M^{-1}V^{-1}BM$$

$$M^{-1}V^{-1}BMa = \lambda_1 a$$

$$V^{-1}BMa = \lambda_1 Ma$$

$$V^{-1}Bu = \lambda_1 u$$

Facteur u est vecteur propre de $V^{-1}B$ associé à la valeur propre maximale λ_1 (toujours inférieure à 1)

– **Facteur** u sont indépendants de M, donc les variables discriminantes aussi
– Par commodité on peut prendre $M= V^{-1}$ et on a $BV^{-1}a = \lambda a$ et $V^{-1}Bu=\lambda u$

22

3. Réduction de dimension

- AFD= ACP particulière
- **On peut prendre aussi**
  - **Métrique W^{-1} Mahalanobis**

- Définition:
  - u vecteur propre V^{-1} B est le facteur discriminant
  - la valeur propre est la mesure (pessimiste) de son pouvoir discriminant
- Les différents cas selon λ_1

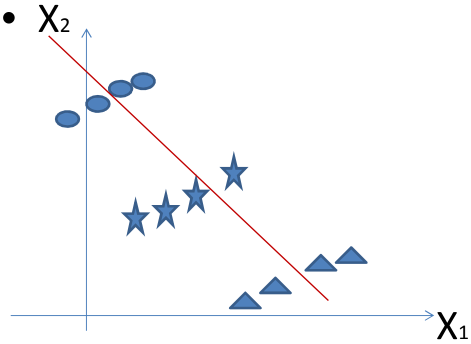
23

3. Réduction de dimension

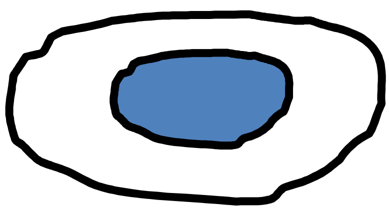
- $\lambda_1 = 1$ séparation parfaite,

pas de variance intra en projection

- X_2



$\lambda_1 = 0$ groupes concentriques,
pas de variance inter en projection
Pas de séparation linéaire possible
(distance au centre est fonction quadratique des variables)



0< λ_1 < 1 différents cas de séparation (qui peut être parfaite si les groupes bien éloignés)

24

3. Réduction de dimension

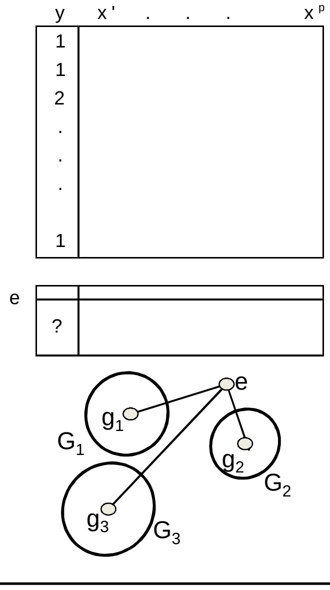
Nombre d'axes discriminants

- ACP des groupes : dimension de l'espace contenant les centres des groupes g_i
- Si n>p>k (cas fréquent), k-1 axes discriminants

  Exemple célèbre : Iris de Fisher
  - *K = 3 Setosa, Versicolor, Virginica*
  - *P=4 longueur pétale, longueur sépale, largeur pétale, largeur sépale*
  - *n_1=n_2=n_3 =50*

  Donc deux axes (voir exemple avec SAS)

4. Règles géométriques de classement



- Échantillon d'apprentissage: on a trouvé la meilleure représentation de la sépartion en k classes des n individus.
- e observation de groupe inconnu que l'on souhaite prévoir.

La règle géométrique de Mahalanobis-Fisher:

- e classé dans le groupe i tel que:

  d(e ; g_i) minimale

d est calculée avec la métrique W^{-1} (ou V^{-1})

4. Règles géométriques de classement

Utilisation des fonctions discriminantes

$$d^2(e\ ;\ g_i) = (e-g_i)'W^{-1}(e-g_i) = e'W^{-1}e - 2g'_iW^{-1}e + g'_iW^{-1}g_i$$

$$\min\ \mathrm{d}^2(e\ ;\ g_i) = \max\left(2g'_iW^{-1}e - \underbrace{g'_iW^{-1}g_i}_{\alpha_i}\right)$$

Linéaire par rapport aux coordonnées de e coeff beta + une constante alpha

k groupes $\Rightarrow$ k fonctions discriminantes

	1	2	....	k
1	α_1	α_2		α_k
X^1	β_{11}	β_{21}		β_{k1}
X^2				
X^p	β_{1p}	β_{2p}		β_{kp}

- On classe dans le groupe pour lequel la fonction est maximale.

27

5. Cas de deux groupes

- g_1 et g_2 sont sur une droite : 1 seul axe discriminant :

$$a = \alpha(g_1 - g_2)$$

- RAPPEL : en ACP axe a, facteur u = M a
- Combinaison discriminante proportionnelle à M (g_2 - g_1) = W^{-1} (g_2 - g_1) ou V^{-1} (g_2 - g_1)
- FONCTION DE FISHER :
- (vecteur de coefficients)
- Rque pour des raisons d'estimations
- On pondère W^{-1} par (n1+n2 -2)/ n1+n2

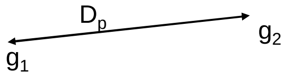
$$W^{-1}(g_2 - g_1) = W^{-1}\begin{pmatrix} \overline{X}_2^1 - \overline{X}_1^1 \\ \vdots \\ \overline{X}_2^p - \overline{X}_1^p \end{pmatrix}$$

28

Distance de MAHALANOBIS

Distance au sens de la métrique W^{-1}.

$$\boxed{D_p^2 = (g_1 - g_2)' W^{-1} (g_1 - g_2)}$$



distance entre les centres

1. pour p=1 : $$\frac{n_1 n_2}{n_1 + n_2}\left(\frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}}\right)^2 = \frac{n_1 n_2}{n_1 + n_2} D_1^2 \sim F(1; n_1 + n_2 - 2)$$

2. p quelconque : $$D_p^2 = (g_1 - g_2)' W^{-1} (g_1 - g_2)$$
$$D_p^2 = (g_1 - g_2)' W^{-1/2} \underbrace{W^{-1/2} (g_1 - g_2)}_{W^{-1/2} X}$$

- Standardisation

5. Cas de deux groupes

- **Une régression « incorrecte »**

  y à 2 valeurs (-1;+1) ou (0;1) ou (a;b) avec a=n/n_1 b=-n/n_2 (la moyenne de y=0)

$$\hat{\boldsymbol{\beta}} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

$$R^2 = \frac{D_p^2}{\frac{n(n-2)}{n_1 n_2} + D_p^2} \qquad D_p^2 = \frac{n(n-2)}{n_1 n_2} \frac{R^2}{1 - R^2}$$

- D_p distance de Mahalanobis entre groupes
- Incompréhensions et controverses (la fonction de Fisher pouvant être obtenue via une régression « bizarre » d'où une préférence pour la régression logistique)
- MAIS

5. Cas de deux groupes

Modèle linéaire usuel non valide : $y/\mathbf{X} \quad N(\mathbf{X\beta};\sigma^2\mathbf{I})$
(y non aléatoire ne prend que 2 valeurs, impossibilité de gaussienne)

en discriminante c'est l'inverse que l'on suppose :

$$\mathbf{X}/y=j \quad N_p(\boldsymbol{\mu}_j;\boldsymbol{\Sigma})$$

En conséquences:

- *Pas de test,*
- *pas d'erreurs standard sur les coefficients*

31

5. Cas de deux groupes

Fonctions de classement et fonction de Fisher

On classe dans G_1 si:

Fonction de classement du groupe 1

Fonction de classement du groupe 2

$$2g_1^{'}W^{-1}e-g_1^{'}W^{-1}g_1 > 2g_2^{'}W^{-1}e-g_2^{'}W^{-1}g_2$$

$$d = (g_1-g_2)'W^{-1}e > \tfrac{1}{2}(g_1^{'}W^{-1}g_1-g_2^{'}W^{-1}g_2)$$

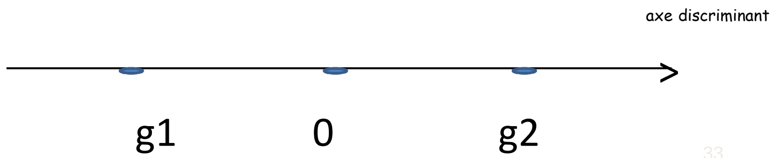
- Fonction de Fisher > c un seuil
- Score de Fisher: $(g_1-g_2)'W^{-1}e-\tfrac{1}{2}(g_1^{'}W^{-1}g_1-g_2^{'}W^{-1}g_2)$

32

5. Cas de deux groupes

- **Interprétation géométrique**
- Projection sur la droite des centres avec la métrique W^{-1}
- d = coordonnée (valeur de la fonction discriminante)
- si d > seuil on affecte à g2

  si d < seuil on affecte à g1

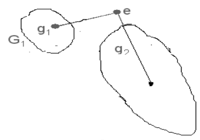


33

Règles géométriques de classement

Insuffisance des règles géométriques

- Pas optimal si les variances des groupes sont très différentes :groupe 2 plus attractif mais affectation au groupe 1



D'où nécessité de méthodes probabilistes (voir troisième partie)

De plus les variables explicatives doivent être quantitatives (deuxième partie: discrimination sur qualitatives)

34

Deuxième partie: Discrimination sur variables qualitatives et scoring

1. Le problème
2. Disqual

II.1 Discrimination sur variables qualitatives- le problème

Y variable de groupe

$X_1, X_2, \dots, X_p$ Variables explicatives à $m_1, m_2, \dots, m_p$ modalités

Exemples

- Solvabilité d'emprunteurs auprès de banques

  Y : bon payeur / mauvais payeur

X_1: sexe, X_2: catégorie professionnelle etc.

- Risque en assurance automobile

  Y : bon conducteur (pas d'accidents) / mauvais conducteur

X_1: sexe, X_2: tranche d'âge, X_3: véhicule sportif ou non ...

- Reclassement dans une typologie

  Y numéro de groupe

II.1 Discrimination sur variables qualitatives -le problème

- Deux idées équivalentes :
  - Transformer les variables qualitatives explicatives en variables quantitatives.
    Donner des valeurs numériques (notes ou scores) aux modalités de façon optimale: maximiser la distance de Mahalanobis dans R^p Quantification optimale
  - Travailler sur le tableau disjonctif des variables explicatives
- *Une réalisation : Passage par l'intermédiaire d'une analyse des correspondances multiples.*

$$\begin{pmatrix} X_1 & & X_2 & & & \\ 0 & 1 & 1 & 0 & 0 & \\ . & & & & & \cdots \\ . & & & & & \\ . & & & & & \end{pmatrix}$$

37

II.1 Discrimination sur variables qualitatives -le problème

- **Quantification :** Transformer une variable qualitative en une variable numérique et se ramener au cas précédent (1ère partie).
- Exemple : État matrimonial de 7 individus
- Quantification : $\begin{pmatrix} C \\ C \\ M \\ M \\ M \\ V \\ D \end{pmatrix} \xrightarrow{\substack{C = Célibataire \\ M = Marié \\ V = Veuf \\ D = Divorcé}} \begin{pmatrix} a_1 \\ a_1 \\ a_2 \\ a_2 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}$
- Infinités de solutions pour les a_i

38

II.1 Discrimination sur variables qualitatives -le problème

- **Quantification :** X tableau disjonctif, x vecteur des quantifications pour les individus, a vecteur des coefficients a_i on a:
- **x = Xa**

- Une quantification est simplement une combinaison linéaire des indicatrices

39

II.1 Discrimination sur variables qualitatives -le problème

La fonction de Fisher est une combinaison linéaire des variables quantifiées

$$s = \sum_{I=1}^{p} \alpha_i \tilde{X}_i$$

$$\tilde{X}_i = \sum_{j=1}^{m_i} \beta_j 1_j$$

- S est une combinaison linéaire des ($m_1 + m_2 + \ldots + m_p$) indicatrices des variables

40

II.1 Discrimination sur variables qualitatives -le problème

Analyse discriminante sur p variables qualitatives à modalités est équivalente à

Analyse discriminante avec p prédicteurs numériques (les indicatrices) mais il a p relations linéaires entre indicatrices X n'est pas de plein rang: $\text{rank}(X)=\Sigma m_i - p$

donc la matrice W n'est pas inversible

Plusieurs solutions à ce problème

41

II.1 Discrimination sur variables qualitatives -le problème

- Solution classique: éliminer une indicatrice par prédicteur (GLM , LOGISTIC de SAS)
- Disqual (Saporta, 1975):
  - ADL effectuée sur une sélection de facteurs de l'ACM de X. Analogue de la régression sur composantes principales
  - Composantes sélectionnées de manière experte selon inertie et pouvoir discriminant

42

II.2 DISQUAL

$1^{ère}$ étape

- Analyse des correspondances du tableau des prédicteurs.

$$X = \begin{array}{c} \\ \\ 1 \\ 2 \\ \\ \\ \\ \\ n \end{array} \begin{array}{c} \text{Profession} \\ \begin{array}{cccc} P_1 & P_2 & P_3 & P_4 \end{array} \\ \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ & . & & \\ & . & & \\ & . & & \\ & . & & \\ & & & \end{array} \right. \end{array} \begin{array}{c} \text{Logement} \\ \begin{array}{cc} \text{Prop.} & \text{Loc.} \end{array} \\ \left| \begin{array}{cc} 0 & 1 \\ 1 & 0 \\ . & \\ . & \\ . & \\ . & \\ & \end{array} \right| \end{array} \cdots$$

variables indicatrices

$$Z = \begin{array}{c} 1 \\ 2 \\ . \\ . \\ . \\ n \end{array} \begin{pmatrix} z^1 & . \ . \ . & z^k \end{pmatrix}$$

- *k variables numériques : garder les coordonnées factorielles les plus discriminantes*

II.2 DISQUAL

$2^{ème}$ étape

- Analyse discriminante linéaire (Fisher). Score $\mathbf{s} = \sum_{j=1}^{k} d_j \, \mathbf{z}^j$
- Score = combinaison linéaire des coordonnées factorielles= combinaison linéaire des indicatrices des catégories
- Coefficients = grille de notation
- $\mathbf{z^j = Xu^j}$ $\mathbf{u^j}$: coordonnées des catégories sur l'axe n°j

$$s = \sum_{j=1}^{k} d_j X u^j = X \underbrace{\sum_{j=1}^{k} d_j u^j}_{\text{grille de score}}$$

$$\begin{pmatrix} . \\ d_j \\ . \end{pmatrix} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = \begin{pmatrix} . \\ \dfrac{\bar{z}_1^j - \bar{z}_2^j}{V(\mathbf{z}^j)} \\ . \end{pmatrix}$$

II.2 DISQUAL

Sélection des axes

- Selon l'ordre de l'ACM
  - % d'inertie
- Selon le pouvoir discriminant
  - Student sur 2 groupes,F sur k groupes
- Régularisation, contrôle de la VC dimension

45

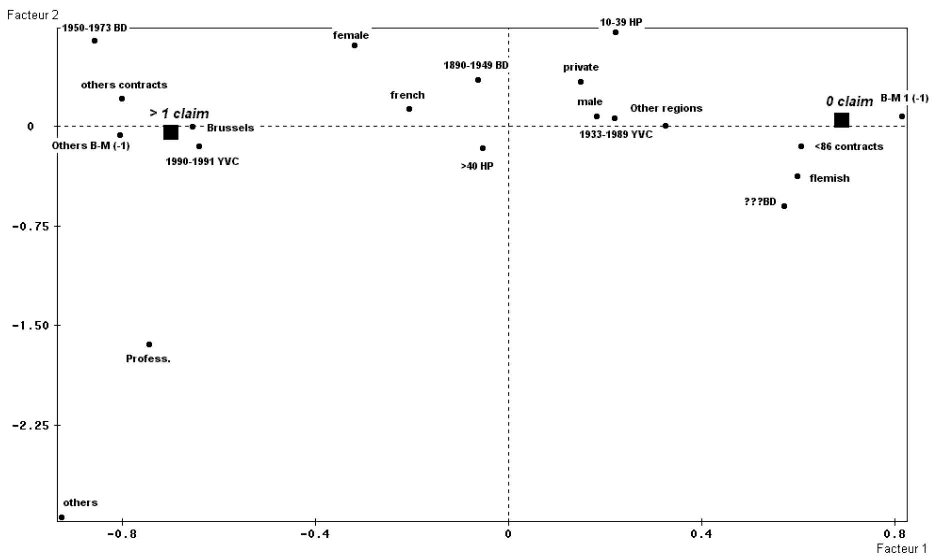
II.2 DISQUAL

Exemple assurance (SPAD)

- 1106 contrats automobile belges:
- 2 groupes: « 1 bons», « 2 mauvais »
- 9 prédicteurs: 20 catégories
  - Usage (2), sexe (3), langue (2), age (3), région (2), bonus-malus (2), puissance (2), durée (2), age du véhicule (2)

46

ACM



II.2 DISQUAL

Exemple assurance (SPAD)

ADL de Fisher sur les composantes

FACTEURS	CORRELATIONS	COEFFICIENTS
1 F 1	0.719	6.9064
2 F 2	*0.055*	*0.7149*
3 F 3	-0.078	-0.8211
4 F 4	*-0.030*	*-0.4615*
5 F 5	0.083	1.2581
6 F 6	*0.064*	*1.0274*
7 F 7	*-0.001*	*0.2169*
8 F 8	0.090	1.3133
9 F 9	-0.074	-1.1383
10 F 10	-0.150	-3.3193
11 F 11	*-0.056*	*-1.4830*
CONSTANTE		0.093575

R2 = 0.57923 F = 91.35686
D2 = 5.49176 T2 = 1018.69159

Score= 6.90 F1 - 0.82 F3 + 1.25 F5 + 1.31 F8 - 1.13 F9 - 3.31 F10

II.2 DISQUAL

Exemple assurance (SPAD)

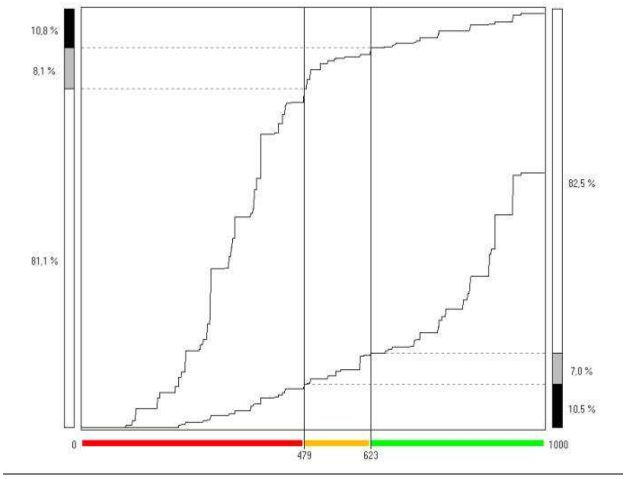
- **scores normalisés**
  - Echelle de 0 à 1000
  - Transformation linéaire du score et du seuil

  (voir détail au tableau)

Grille de score (« scorecard »)

CATEGORIES	COEFFICIENTS DISCRIMINANT FUNCTION	TRANSFORMED COEFFICIENTS (SCORE)
2 . Use type		
USE1 - Profess.	-4.577	0.00
USE2 - private	0.919	53.93
4 . Gender		
MALE - male	0.220	24.10
FEMA - female	-0.065	21.30
OTHE - companies	-2.236	0.00
5 . Language		
FREN – French	-0.955	0.00
FLEM - flemish	2.789	36.73
24 . Birth date		
BD1 - 1890-1949 BD	0.285	116.78
BD2 - 1950-1973 BD	-11.616	0.00
BD? - ???BD	7.064	183.30
25 . Region		
REG1 - Brussels	-6.785	0.00
REG2 – Other regions	3.369	99.64
26 . Level of bonus-malus		
BM01 - B-M 1 (-1)	17.522	341.41
BM02 - Others B-M (-1)	-17.271	0.00
27 . Duration of contract		
C<86 - <86 contracts	2.209	50.27
C>87 - others contracts	-2.913	0.00
28 . Horsepower		
HP1 - 10-39 HP	6.211	75.83
HP2 - >40 HP	-1.516	0.00
29 . year of vehicle construction		
YVC1 - 1933-1989 YVC	3.515	134.80
YVC2 - 1990-1991 YVC	-10.222	0.00

Fonctions de répartition



Extension: Cas des prédicteurs numériques

- Si prédicteurs numériques (taux d'endettement, revenu…)
- Découpage en classes
  – Avantages, détection des liaisons non linéaires
  – Prise en compte des interactions

Extension: Cas des prédicteurs numériques

- Amélioration considérable de l'efficacité du score

  Rappel : $Score = f_1(x_1) + f_2(x_2) + \ ...$

  Modèle additif **sans** interaction

- *Exemple : État matrimonial et nombre d'enfants.*

2 catégories $(M_1 \quad M_2)$ 3 catégories $(E_1 \quad E_2 \quad E_3)$

variable croisée à 6 catégories

$$
\begin{array}{c}
 \\
1 \\
2 \\
. \\
. \\
. \\
 \\
n
\end{array}
\begin{array}{c}
(M_1E_1 \ M_1E_2 \qquad\qquad M_2E_3) \\
\begin{pmatrix}
1 & 0 & . & . & . & 0 \\
0 & 1 & . & . & . & 0 \\
. & & & & & \\
. & & & & & \\
. & & & & & \\
. & & & & & \\
. & & & & &
\end{pmatrix}
\end{array}
$$

53