Analyse des données:
Les méthodes factorielles

Prof. Mohamed El Merouani

Introduction:

- L'analyse des données est une des branches les plus vivantes de la statistique.
- Les principales méthodes de l'analyse des données se séparent en deux groupes:
  - Les méthodes de classification,
  - Les méthodes factorielles.

Les méthodes de classification:

- Elles visent à réduire la taille de l'ensemble des individus en formant des groupes homogènes d'individus ou de variables.
- Ces groupes on les appelle aussi des classes, ou familles, ou segments, ou clusters.
- La classification est appelée aussi Segmentation ou Clustering ou…

Les méthodes factorielles:

- Parmi les méthodes descriptives ou non-supervisées du Datamining, on trouve les méthodes factorielles de l'Analyse des données.
- les méthodes factorielles consistent en la projection sur un espace de dimension inférieure pour obtenir une visualisation de l'ensemble des liaisons entre variables tout en minimisant la perte de l'information.

Les méthodes factorielles:

- Elles cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques.
- **Si on travaille avec un tableau de variables numériques, on utilisera l'analyse en composantes principales,**
- Si on travaille avec des variables qualitatives, on utilisera l'analyse des correspondances.
- Les liens entre deux groupes de variables peuvent être traités par l'analyse canonique.

Les méthodes factorielles:

Les méthodes factorielles regroupent :

- **L'ACP :** L'analyse en composantes principales
- **L' AFC :** L'analyse factorielle des correspondances

L'ACP

- L'ACP (Hotelling, 1933) a pour objectif de réduire le nombre de données, souvent très élevé, d'un tableau de données représenté, algébriquement, comme une matrice et, géométriquement comme un nuage de points.
- L'ACP consiste en l'étude des projections des points de ce nuage sur un axe (axe factoriel ou principal), un plan ou un hyperplan judicieusement déterminé.
- Mathématiquement, on obtiendrait le meilleur ajustement du nuage par des sous-espaces vectoriels.

- Soit un tableau de données ayant p lignes et q colonnes:

colonnes / *lignes*	1	....	J	...	*q*
1	x_{11}	...	x_{1j}	...	x_{1q}
i	x_{i1}	...	x_{ij}	...	x_{iq}
p	x_{p1}	...	x_{pj}	...	x_{pq}

- On représente ce tableau sous forme d'une matrice notée X de type (p,q).

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2q} \\ \vdots & & \vdots & & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{iq} \\ \vdots & & & \vdots & & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pj} & \cdots & x_{pq} \end{pmatrix}$$

L'ACP

- Algébriquement, il s'agit de chercher les valeurs propres maximales de la matrice des données et par conséquent ses vecteurs propres associés qui représenteront ces sous-espaces vectoriels (axes factoriels ou principales).

Procédure de l'ACP:

- On cherche X' la transposée de la matrice X.
- On détermine les valeurs propres de la matrice symétrique X'X.
- Soient $\lambda_1, \lambda_2, \ldots, \lambda_q$ ces valeurs propres.
- On les classe $\lambda_1>\lambda_2>\lambda_3>\lambda_4>\ldots$.
- Alors $X'X=A\Lambda A^{-1}$ où

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_q \end{pmatrix}$$

Procédure de l'ACP:

- D'après les propriétés de la trace des matrices; on a:

$$tr\left(X'X\right)=tr\left(A\Lambda A^{-1}\right)=tr\left(AA^{-1}\Lambda\right)=tr\Lambda$$

- Soit $tr\left(X'X\right)=\lambda_1+\lambda_2+\cdots+\lambda_q$
- En raison des valeurs numériques décroissantes de $\lambda_1, \lambda_2, \ldots$, la somme des premiers valeurs propres représente, souvent, une proportion importante de la trace de X'X.

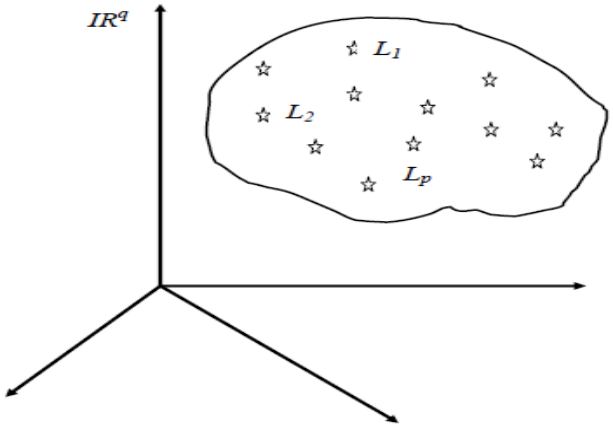
Procédure de l'ACP:

- Ainsi, dans la pratique on peut se limiter à trouver les premiers valeurs propres λ_1, λ_2, …, λ_s avec s assez inférieur à q.
- L'information perdue est alors relativement faible.
- On pratique s=3 (trois premiers valeurs propres les plus grands)

Procédure de l'ACP:

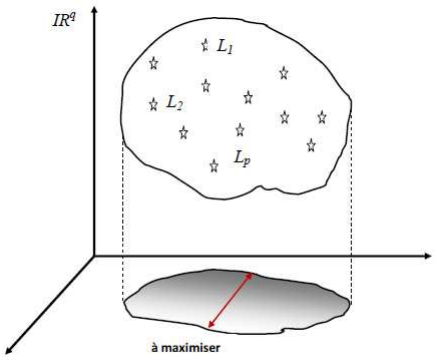
- Les valeurs propres trouvés étant simples, les espaces propres associés aux vecteurs propres seront des droites vectorielles (on les appelles des axes factoriels ou des facteurs).
- D'un point de vue général, L'ACP nous a permit de traiter un très grand nombre de données (matrice) pour identifier un nombre relativement restreint de données (axes factoriels)

- **<u>Géométriquement,</u>** on représente le tableau comme un nuage de points.



L'ACP géométriquement:

- Lors de la projection, le nuage peut être déformé est donc serait différent de réel, alors les méthodes d'ajustement consistent en minimiser cette possible déformation et ce en maximisant les distances projetées.



Distance ou métrique utilisée:

- Soient L_m et L_n deux points de IR^q:

$$L_m = \left(x_{m1}, x_{m2}, \cdots, x_{mj}, \cdots, x_{mq}\right)$$

$$L_n = \left(x_{n1}, x_{n2}, \cdots, x_{nj}, \cdots, x_{nq}\right)$$

- La distance euclidienne (classique) entre ces points est:

$$d\left(L_m, L_n\right) = \sqrt{\sum_{j=1}^{q}\left(x_{mj} - x_{nj}\right)^2}$$

Distance ou métrique utilisée:

- Ou bien

$$d^2\left(L_m, L_n\right) = \left(x_{m1} - x_{n1}\right)^2 + \cdots + \left(x_{mj} - x_{nj}\right)^2 + \cdots + \left(x_{mq} - x_{nq}\right)^2$$

- Les points L_m et L_n sont encore plus proches lorsque la somme précédente est plus petite.
- Si les différents coordonnées des points L ne se mesurent pas avec les mêmes unités, la distance d sera la somme des termes de « poids » très différents.

Distance ou métrique utilisée:

- Pour éviter ce problème des unités, on va centrer auparavant les vecteurs colonnes de la matrice X.
- Le tableau des données centrés Y est :

$$Y = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1q} - \bar{x}_q \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2q} - \bar{x}_q \\ \vdots & & \ddots & \vdots \\ x_{p1} - \bar{x}_1 & x_{p2} - \bar{x}_2 & \cdots & x_{pq} - \bar{x}_q \end{pmatrix}$$

L'ACP normé:

- On s'intéresse à étudier la matrice des variances-covariances V au lieu de la matrice X de départ.
- La matrice V est une matrice de type carrée d'ordre q de terme général v_{kl} égal à:

$$v_{kl} = \frac{1}{p}\sum_{i=1}^{p}(y_{ik} - \bar{y}_k)(y_{il} - \bar{y}_l) = \frac{1}{p}\sum_{i=1}^{p}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$$

$$v_{kl} = \frac{1}{p}\sum_{i=1}^{p}(x_{ik}\,x_{il} - \bar{x}_k\,\bar{x}_l)$$

- La matrice V des variances-covariances est telle que $$V = \frac{1}{p} Y' Y$$
- On peut aussi considérer la matrice Z des données centrées et normés d'éléments z_{ij}

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

- Avec

$$\bar{x}_j = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad ; \quad \sigma_j = \sqrt{\frac{1}{p} \sum_{i=1}^{p} \left(x_{ij} - \bar{x}_j\right)^2}$$

Matrice centrée normée:

- Donc, la matrice des données centrées et normées sera:

$$Z = \begin{pmatrix} \frac{x_{11} - \bar{x}_1}{\sigma_1} & \frac{x_{12} - \bar{x}_2}{\sigma_2} & \cdots & \frac{x_{1q} - \bar{x}_q}{\sigma_q} \\ \frac{x_{21} - \bar{x}_1}{\sigma_1} & \frac{x_{22} - \bar{x}_2}{\sigma_2} & \cdots & \frac{x_{2q} - \bar{x}_q}{\sigma_q} \\ \vdots & & \ddots & \vdots \\ \frac{x_{p1} - \bar{x}_1}{\sigma_1} & \frac{x_{p2} - \bar{x}_2}{\sigma_2} & \cdots & \frac{x_{pq} - \bar{x}_q}{\sigma_q} \end{pmatrix}$$

- A partir de cette matrice, on définit la matrice Γ des corrélations entre les q variables prises deux à deux:

$$\Gamma = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1q} \\ \rho_{21} & 1 & \cdots & \rho_{2q} \\ \vdots & & \ddots & \vdots \\ \rho_{q1} & \cdots & \cdots & 1 \end{pmatrix}$$

- Γ résume la structure des dépendances linéaires entre les q variables et on a $\Gamma = \frac{1}{p} Z'Z$

Procédure de l'ACP normé:

- On extrait les valeurs propres les plus grands λ_1, λ_2, …, de la matrice V des variances-covariances ou de la matrice Γ des corrélations.
- En pratique, on arrête l'extraction des valeurs propres lorsque la somme des s valeurs propres que l'on a déterminés représente un pourcentage satisfaisant de la variance.

Procédure de l'ACP normé:

- On détermine les vecteurs propres associés aux valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_s$
- Ce sont les axes factoriels
- Dans la majorité des cas, ne sont prise en considération que les deux, les trois, ou les quatre premiers axes factoriels.
- Les axes factoriels sont perpendiculaires et ne sont pas corrélés entre eux.

Nombre d'axes à retenir :

Les critères les plus utilisables sont les suivantes :

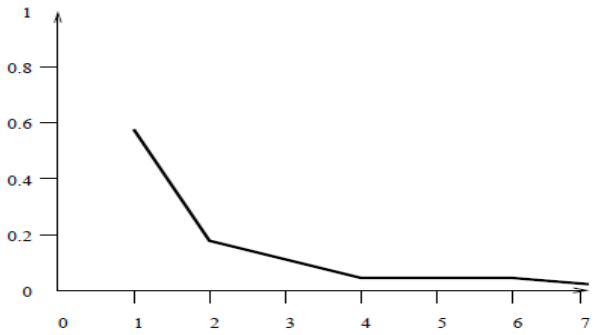
1°) **Interprétation des axes :** On retient que les axes que l'on peut attribuer une forme d'interprétation économique, par exemple, soit directement, soit en terme des variables avec lesquelles ils sont très corrélés.

2°) **Critère de Kaiser** (variables centrées et réduites) : On ne retient que les axes associés à valeurs propres supérieurs à 1, c'est-à-dire dont la variance est supérieure à celle des variables d'origine.

- Une autre interprétation est que la moyenne des valeurs propres étant 1, on ne garde que celles qui sont supérieures à cette moyenne.

Nombre d'axes à retenir :

3°) **Éboulis des valeurs propres** :On cherche un « coude » dans le graphe des valeurs propres et on ne conserve que les valeurs jusqu'au ce « coude ».



Qualités et défauts de l'ACP :

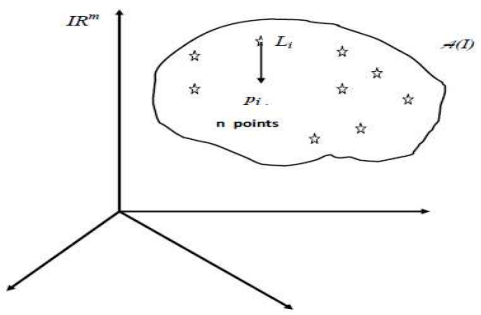
- D'un point de vue technique, ce procédé a pour objet l'étude de la structure de la matrice des variances-covariances ou de la matrice des corrélations.
- Mais, le procédé est imparfait dans la mesure que le nuage est déformé par la projection, même si cette dernière est la plus idéale possible. Certains points sont plus altérés que d'autres par la transformation.
- L'inconvénient majeur réside dans l'interprétation des axes. Parfois, l'explication est évidente et fait que l'analyse en composantes principales soit redondante ; ou bien elle est contingente pour l'analyste et dans ce dernier cas elle n'apporte pas des renseignements très convaincant es pour l'analyse économétrique postérieure.

L'AFC

- L'AFC a pour objet le traitement de l'information contenue dans un tableau appelé de contingence ou de dépendance, relatif à deux ensembles de nature quelconque, en relation par moyen d'un processus naturel ou expérimental plus ou moins bien connu.
- Les données sont ici pondérées. Les fréquences de répétitions s'interprète facilement en termes de probabilités.

L'AFC

- Le tableau de dépendance peut être ainsi représenté dans un espace approprié par un nuage de points affectés de probabilités.



- Considérons un tableau à double entrée :

Ensemble J (paramèttres) / *Ensemble I (individus)*	1	....	J	...	*m*
1	x_{11}	...	x_{1j}	...	x_{1m}
i	x_{i1}	...	x_{ij}	...	x_{im}
n	x_{n1}	...	x_{nj}	...	x_{nm}

- Dans le cas qualitatif, le tableau précédent se présente sous la forme d'un tableau des uns et des zéros (suivant si l'individu i possède ou non le paramètre j).
- La probabilité associée au terme x_{ij} est:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}}$$

J / I	1	....	i	...	m	Total
1	p_{11}	...	p_{1j}	...	p_{1m}	$p_{1.}$
⋮						
i	p_{i1}	...	p_{ij}	...	p_{im}	$p_{i.}$
⋮						
n	p_{n1}	...	p_{nj}	...	p_{nm}	$p_{n.}$
Total	$p_{.1}$		$p_{.j}$		$p_{.m}$	1

33

- Où les probabilités marginales sont:

$$p_{i\bullet} = \sum_{j=1}^{m} p_{ij} \quad avec \quad i = 1, \cdots, n$$

$$p_{\bullet j} = \sum_{i=1}^{n} p_{ij} \quad avec \quad j = 1, \cdots, m$$

- qui vérifient les propriétés:

$$\sum_{i=1}^{n} p_{i\bullet} = 1 \quad et \quad \sum_{j=1}^{m} p_{\bullet j} = 1$$

C'est quoi « les correspondances »?

- Lorsque les variables sont **quantitatives**, on fait une étude de **corrélation**.
- Mais, lorsqu'on a aussi des variables **qualitatives**, on doit faire une étude des **correspondances**.

35

Indépendance?

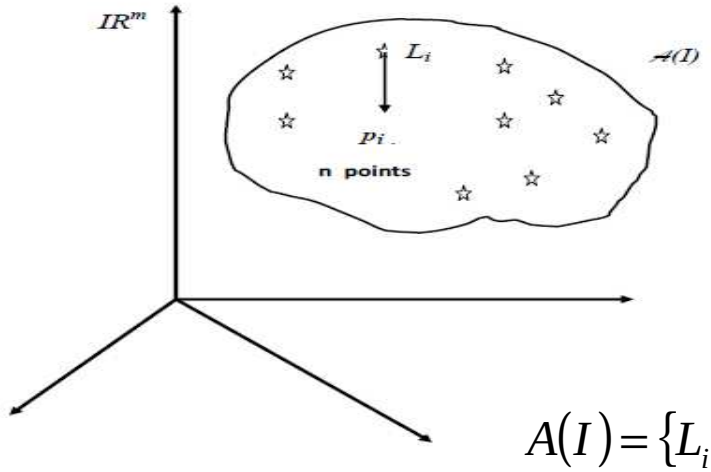
- Probabilités conditionnelles, dans ce cas:

$$\frac{p_{ij}}{p_{i\cdot}} = p_{\cdot j} \Leftrightarrow \frac{p_{ij}}{p_{\cdot j}} = p_{i\cdot}$$

- Formule d'indépendance:

$$p_{ij} = p_{i\cdot} \times p_{\cdot j}$$

36



$$A(I) = \{L_i ; p_{i.}\}$$

$$L_i = \left(\frac{p_{i1}}{p_{i.}}, \frac{p_{i2}}{p_{i.}}, \ldots, \frac{p_{ij}}{p_{i.}}, \ldots, \frac{p_{im}}{p_{i.}} \right)$$

37

Distance du χ^2

- Pour deux individus quelconques *i* et *i'*:

$$d^2(L_i, L_{i'}) = \sum_j \frac{1}{p_{.j}} \left(\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}} \right)^2$$

- Pourquoi une telle distance?

38

Pourquoi la distance du χ^2?

- La distance euclidienne ne prend pas compte complètement de tous les caractères étudiés!
- Il a été alors proposé de modifier la distance euclidienne en tenant compte des écarts entre deux probabilités de deux individus d'avoir un caractère en donnant de l'importance aux probabilités que l'individu ait tous les caractères étudiés.
- Cela, donc, par multiplication par l'inverse de la probabilité d'avoir tous les caractères.

39

Pourquoi la distance du χ^2?

- Aussi, parce que la distance du χ^2 a une propriété qui s'appelle «**la propriété d'équivalence distributionnelle** » et que la distance euclidienne ne vérifie pas!
- Si deux colonnes j et j' de J correspond au même ligne i, il est logique de les regrouper en une seule de probabilité ($p_{ij}+p_{ij'}$), il faut alors que cette opération ne modifie pas les distances entre les i.

40

Pourquoi la distance du χ^2?

- Plus généralement, la distance du χ^2 est égale à la distance euclidienne entre:

$$\left(\frac{p_{i1}}{p_{i\cdot}\sqrt{p_{\cdot 1}}}, \frac{p_{i2}}{p_{i\cdot}\sqrt{p_{\cdot 2}}}, \cdots, \frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}, \cdots, \frac{p_{im}}{p_{i\cdot}\sqrt{p_{\cdot m}}} \right)$$

$$\left(\frac{p_{i'1}}{p_{i'\cdot}\sqrt{p_{\cdot 1}}}, \frac{p_{i'2}}{p_{i'\cdot}\sqrt{p_{\cdot 2}}}, \cdots, \frac{p_{i'j}}{p_{i'\cdot}\sqrt{p_{\cdot j}}}, \cdots, \frac{p_{i'm}}{p_{i'\cdot}\sqrt{p_{\cdot m}}} \right)$$

41

- Ce sont les points qu'on a noté M_i dans le cours

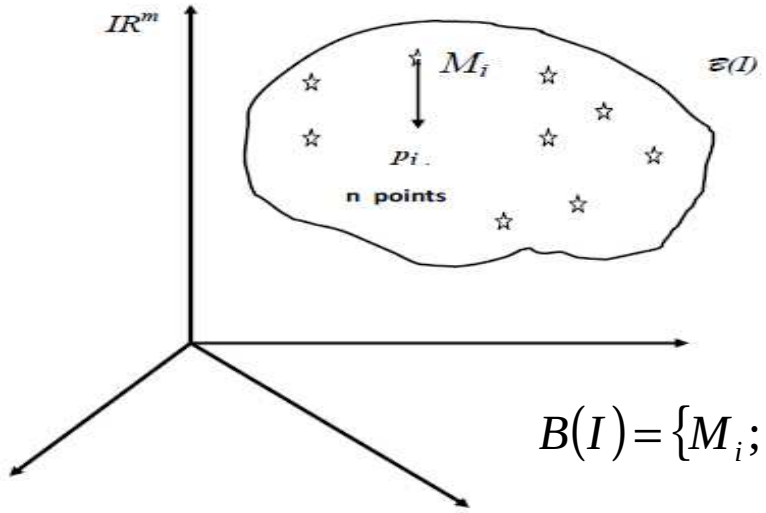
$$M_i = (\beta_{i1}, \beta_{i2}, \cdots, \beta_{ij}, \cdots, \beta_{im})$$

- Avec

$$\beta_{ij} = \frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}$$

- $p_{i\cdot}$ étant toujours la pondération

42



$$B(I) = \{M_i ; p_{i\cdot}\}$$

43

- Ainsi la distance du χ2 entre deux points M_i et $M_{i'}$ est:

$$d^2(M_i, M_{i'}) = \sum_j (\beta_{ij} - \beta i'j)^2$$

44

Projection du nuage $\mathcal{B}(I)$ sur un axe:

- On projet orthogonalement le nuage $\mathcal{B}(I)$ sur un axe (espace vectoriel de dim 1) de vecteur unitaire u, de telle façon que l'information perdue soit minime.
- Comme en ACP, ce qui revient à
  max $u'Wu$, sous la condition $u'u$=1, avec W est la matrice des variances-covarainces de $\mathcal{B}(I)$.
- Ce qui revient à trouver la valeur propre la plus grande λ_{max} de W.

45

Matrice des variances-covariances W:

- La matrice des variances-covariances W du nuage $\mathcal{B}(I)$ relativement à un paramètre j est:

$$W = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & & \ddots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mm} \end{pmatrix}$$

46

Matrice des variances-covariances W:

- La variance v_{jj} caractérise la dispersion du nuage tout au long de l'axe j:

$$v_{jj} = \sum_i p_{i\cdot} \left(\beta_{ij} - \sqrt{p_{\cdot j}}\right)^2$$

- La covariance v_{jk} est

$$v_{jk} = \sum_i p_{i\cdot} \left(\beta_{ij} - \sqrt{p_{\cdot j}}\right)\left(\beta_{ik} - \sqrt{p_{\cdot k}}\right)'$$

47

Matrice des variances-covariances W:

- Soit encore, en remplaçant β_{ij} par sa valeur:

$$v_{jk} = \sum_i \left(\frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot} p_{\cdot k}}} \right)\left(\frac{p_{ik} - p_{i\cdot} p_{\cdot k}}{\sqrt{p_{i\cdot} p_{\cdot k}}} \right)'$$

- Posons $\dfrac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot} p_{\cdot k}}} = r_{ij} \;\; ; \;\; i = 1, \cdots, n \;\; , \;\; j = 1, \cdots, m$

48

Matrice des variances-covariances W:

$$\left(r_{ij}\right)_{\substack{1\le i\le n \\ 1\le j\le m}} = R$$

$$W = R'R$$

où R' est la transposée de R.

• Maximiser $u'Wu$ revient à maximiser $u'R'Ru$ sous la condition $u'u=1$, c'est-à-dire déterminer les vecteurs propres associés aux valeurs propres de la matrice $R'R$.

49

Variabilité totale du nuage $\mathcal{B}(I)$:

- On appelle la variabilité totale du nuage $\mathcal{B}(I)$, la trace de la matrice W:

$$V_B = tr(W) = \sum_j v_{jj}$$

- On parle aussi de la variabilité totale du nuage projeté $\mathcal{G}(I)$ qui sera $Vc=\lambda_{max}$

50

Variabilité expliquée du nuage $\mathcal{B}(I)$:

- La partie de variabilité expliquée par la projection de $\mathcal{B}(I)$, sur u est alors:

$$\delta = \frac{Vc}{V_B}$$

- Soit encore:

$$\delta = \frac{\lambda_{\max}}{tr(W)}$$

51

Projection du nuage $\mathcal{B}(I)$ sur un plan:

- Comme en ACP, les vecteurs propres de W s'appellent « axes factoriels » du nuage.
- La détermination des axes factoriels se fait en diagonalisant la matrice symétrique W.
- En pratique, on se contente des valeurs propres les plus grands.

52

Recherche des facteurs:

- Les points du nuage $\mathcal{G}(I)$ possèdent un nombre réduit de coordonnées dans le référentiel formé les axes factoriels.
- Ces coordonnées sont les valeurs de nouvelles variables qui s'appellent: Facteurs.
- Le premier facteur correspond aux coordonnées sur le premier axe factoriel.

53

Recherche des facteurs:

- On peut démontrer que les facteurs sont non-corrélés entre eux et s'expriment comme combinaisons linéaires des données.
- Réciproquement, les données ont des coefficients qui sont des combinaisons linéaires des facteurs.
- Ainsi, à partir des facteurs, il est possible de reconstruire un tableau de données avec une minime perte d'information, c'est-à-dire obtenir un tableau plus facilement accessible à l'analyse descriptive.

54

Proximité en IR^m et en IR^n:

- On a vu précédemment les proximités entre n points de IR^m.
- Par des calculs symétriques, on peut étudier les proximités de m points de IR^n.
- Sauf qu'il existe des relations entre les facteurs de IR^m et les facteurs de IR^n.
- Il est alors possible de représenter, sur le même graphique, dans le plan des deux premiers axes factoriels, les proximités entre les individus et les proximités entre les caractères.
- Cette simultanéité de représentation suggère parfois une interprétation économique, sociale, politique,…des axes factoriels.

55

Inconvénients et avantages de l'AFC

- Les inconvénients sont les défauts de toute analyse factorielle: déformation inévitable du nuage durant la projection et la signification ou interprétation des axes.
- L'avantage essentiel réside dans l'étude des caractères qualitatifs.

56