ANALYSE DES DONNEES

Master MSS semestre1

Année universitaire2021-2022

A.RAHMOUNE

Département de Maths

Faculté des sciences

Université de Boumerdes M'hamed Bougara

UMBB

Novembre 2021

Support du cours

Chapitre 1

On distingue deux grandes parties complémentaires en Analyse des données (AdD):

I: Analyse Factorielle (en abréviation **AF):** C'est une méthode de **réductions** des données, projections sur des axes bien choisis (Axes factoriels)

II.Classification automatique (CA): C'est une méthode de **groupement** de données par classes (arboréscence)

Les différentes méthodes de l'analyse factorielle sont:

1. Analyse Factorielle générale (AFG)d'un nuage de points quelconques: La démarche théorique de toutes analyses.
2. Analyse en composantes principales (ACP)(On distingue 3 varaiantes)
3. Analyse factorielle des correspondances(AFC): La plus puissante dans les applications, et la plus intéressante du point de vue logique.
4. Analyse des correspondances multiples (ACM), généralise en quelque sorte de l'AFC.
5. Analyse facorielle déscriminante (AFD).
6. Analyse Canonique simple (ACS)- Utilité plus que théorique que pratique.
7. Analyse canonique généralisée (ACG): La plus généraliste, de cette méthode on peut retrouver toutes les méthodes citées plus haut.

Concernant les méthodes des CA, on peut citer:

1. Les Méthodes Hierarchiques Ascendantes et déscendanes(CHA, CHD).
2. Les méthodes de partionnements: Méthode des nuées dynamique (Méthode des centres mobiles)
3. Les méthodes classées dans un espace métrique quelconques.
4. Les méthodes classées à partir d'observations qualitatives.

Remarques:

Les mot clés de **I: Analyse Factorielle (AF) sont:**

Le nuage de points (individus, variables), les masses affectées aux points (pondération); la métrique, pour les<< **inputs**>>

Pour les<< **outputs**>> sont: Les axes d'inertie-Axes factorielles-,les coordonnées des points sur ces axes-composantes principales-, et divers indications aidants à l'interprétation.

D'une méthode d'analyse factorielle à l'autre, seuls varient les <<inputs>>

Le but et la démarche essentielle de l Analyse des Données est:

Consentir une perte en information* afin d'obtenir un gain en signification.**

Ilustrons cette idée par un exemple:

Dans une étude d'une série statistique(p.ex: Etude des revenus d'une population de 1000 ménages), on peut remplaçer un diagramme en batôn (données discrètes)par un histogramme (données groupées en classes).

Cette opération- Le groupement des données en classes- nous fait **perdre** l'information et la précision (faire l'hypothèse d'uniformité à l'intérieur de chaque classe), en contre partie, nous fait **gagner** en lecture, lisibilité, et facilité d'interprétation.

Certains mots: *Information, et signification* doivent être bien précisés, en effet:

L' information doit être utilisée dans le sens de la théorie de l'information (Voir Information au sens de **Kullback**)

Analyse Factorielle générale(AFG) et analyse en composantes principales(ACP)

Notes historiques et bibliographiques

L'analyse de données au départ était déstinée à répondre aux besoins des Sciences humaines et plus spécialement la psychologie (Spearman-psychologue américain, 1928- s'est penché sur l'étude des profils psychologiques des individus, tel que: Déterminer une variable explicative cachée dans des bateries de tests présentées aux sujets, détecter le caractère dominant d'un sujet,trouver des corrélations entre variables,...,etc.), Burt (connu par ses tableaux, ..)

Remarquer la terminologie de l'analyse des données est issue du langage des psychologues: profil, caractère,...,etc.

L'école française durant les années soixante sous légide de J.P. Benzécri a développé cette partie de statistique exploratoire en mettant l'accent sur l'aspect géométrique (centre de gravité, inertie,...,etc)

Aprés la diffusion de la micro-informatique et la vulgarisation de l'aspect logiciel, l'analyse des données a pris de l'ampleur et la diversité.

Actuellement elle touche la quasi totalité des domaines: les domaines de la biologie (biométrie, épidimiologie) la météorologie, l'économie, la gestion(marketing: Analyse d'enquêtes d'opinion et segmentation des marchés-objet de la classification hiearchique-

Succinctement:

Les méthodes d'analyse des données concernent les domaines où une masse importante, voire grandiose de données et de variables sont en jeu.

Le recours à la statistique déscriptive uni et bidimensionnelle (représentations graphiques et calcul des caractéristiques) ne suffisent guère pour répondre aux divers questions, tels que:

Quelle est la variable dominante?.

Quel est le profil d'un individu vérifiant certaine modalité?

et ne peut nullement mettre en évidence l'aspect structurel important tels que les liaisons, les proximités entre les variables, corrélation,...,etc.

La Théorie de L'analyse des données est divisée en deux grandes parties complémentaires, les méthodes d'analyse factorielle et les méthodes de classification.

Si La première est une méthode de **réduction** des données initiales, basée sur le concept de **projection.**

La deuxième est celle de **groupement** par classes homogènes (arboréscence).

Comme la projection sur des espaces (Axe facoriels) nous fait perdre certaine objectivité dans l'interprétation des résultas, on complète généralement notre étude par les méthodes de classification automatique

Commençons par la déscription de l'analyse générale d'un nuage de points quelconques

Remaquer:

les principales méthodes d'analyse factorielle (ACP, AFC),...,etc) en découlent

L'analyse en composantes principales (ACP): C'est la première méthode d'analyse des données apparue, elle traite les données croisant: individus et variable quantitatives ou qualitatives rendues quantitatives par codage numérique (Cette notion de codage sera expliciter plus tard)

L'analyse factorielle des correspondances (AFC) :C'est la plus importante méthode, privilégiée par l'école française.

- Remarques

  les différentes méthodes de l'analyse factorielle (leurs applications) dépend de **la nature**

des données traitées et de leurs **représentations sous forme de tableaux**

Analyse générale d'un nuage de points quelconques .

On désigne par *l'analyse générale:*

*L*es résultats classiques qui sont à la base des diverses méthodes d'analyse plus spécialisées. la plupart des méthodes(ACP, AFCorréspondance,...etc) peuvent se ramener au modèle général via une transformation sur les données (matrice de base) tel que (centrer, réduire-normaliser-les observations,...,etc)

Cette méthode générale est basée sur la notion de projection sur des espaces métriques muni de leurs produit scalaire(tel que l'espace enclédien $\mathbb{R}^n$muni du produit scalaire usuel) sur les propriétes des matrices symétriques définies positives(type: ${}^tXX$).

L'objective est d'exhiber(trouver)un sous espace de dimension minimum engendré par une base (formée par des axes qu'on appellera axes factoriels) tel que le nuage ramené à cette base est visible-facile à interpréter tout en minimisant la perte d'information (Principe de toute analyse statistique), cett opération est appelée **ajustement.**

Ajustement par un sous espace de $\mathbb{R}^p$

Pour simplifier, plaçons nous dans $\mathbb{R}^p$muni d'un produit scalaire usuelle-métrique M= I_p- ,cherchons à décrire la position par rapport à l'origine -$0_{\mathbb{R}^p}$−et la forme du nuage dans un espace de dimension aussi faible que possible.

Commençons donc par chercher la droite F_1 passant par l'origine de vecteur directeur u$\in \mathbb{R}^p$ qui ajuste au mieux le nuage.

Ainsi F_1=D_u = $\left\{\alpha u, \text{ avec } \|u\|_{I_p} = 1, \alpha \in \mathbb{R}\right\}$ droite engendrée par le vecteur u (avec u unitaire)

La projection du vecteur $\mathrm{X}_i =^t (x_{i1}, ..., x_{ip})$ où (i=1,...,n)sur la droite $\mathrm{F}_1 = D_u$ s'écrit: $\mathrm{p}_{r_u}(\mathrm{X}_i) = \alpha.u$ où $\alpha =< u, X_i >_{I_p} =^t uX_i$

$$^t uX_i = (u_1, ..., u_p) \begin{pmatrix} \mathrm{x}_{i1} \\ . \\ . \\ \mathrm{x}_{ip} \end{pmatrix} = \sum_{j=1}^{p} u_j x_{ij}$$ i étant fixé(l'individu n°i)

Or le carré de la distance à l'origine d'un point (individu) du nuage qui est unequantité fixe se décompose en carré de la projection sur F_1 soit $\mathrm{p}_r(\mathrm{u})$ et en carré de la distance à F_1 notons la d(u)(Le fameux Théoréme de Pythagore)

Il est clair que:

Rendre minimum d(u) est équivalent à rendre maximum $\mathrm{p}_u(\mathrm{u})$.

Maximiser la somme des carrés des projections (pour tous les individus): C'est rendre maximumS_1^2 où

$$S_1^2 = \sum_{i=1}^{n} \left\{ ^t uX_i \right\}^2$$

Or $\left\{ ^t uX_i \right\}^2 = \left\{ ^t uX_i \right\}^t \left\{ ^t uX_i \right\} = \left\{ ^t uX_i \right\} \left\{ ^t X_i u \right\} =^t u(X_i^t X_i)u$ remarquer le transposé d'un scalair est lui même.

la quantité à maximiser s'écrit alors

$$S_1^2 =^t u \left\{ \sum_{i=1}^{n} X_i^t X_i \right\} u =^t u^t XXu$$

avec la contrainte ^{t}uu=1(vecteur unitaire)

Remarquer $^t\mathrm{X}_i = (x_{i1}, ..., x_{ip})$ et $^t\mathrm{X}$=($^t\mathrm{X}_1, ..., ^t X_n$)

Résolution du système: $\begin{cases} \max_u S_1^2 = \max_u \left\{ ^t u \left\{ \sum_{i=1}^{n} X_i^t X_i \right\} u \right\} = \max_u \left\{ ^t u^t XXu \right\} \\ \text{Avec la contrainte } ^t uu = 1 \end{cases}$

Une des méthodes pour résoudre le système est celle ci:

faisons intervenir l'opérateur de Lagrange l avec comme multiplicateur λ

$$l(u) = u^t XXu - \lambda(^t uu - 1)$$

Dérivons $l(u)$ par rapport u : $\frac{dl(u)}{du} = 2^t XXu - 2\lambda u$

On a:

$$\frac{dl(u)}{du} = 0 \Leftrightarrow^t XXu = \lambda u$$

Cela signifie que:

u est un vecteur propre de la matrice ${}^tXX$ de valeur propre λ.

Le maximum cherché est donc une valeur propre de ${}^tXX$ (la plus grande)

En effet:

${}^tXXu = \lambda u$ prémultiplions par ^{t}u, cela donne: ${}^tu{}^tXXu = \lambda {}^tuu$ sous la contrainte ${}^tuu = 1$

maximiser ^{t}u^{t}XXu ça revient à exhiber la plus grande λ.

Si nous cherchons l'espace à deux dimensions qui ajuste le mieux le nuage, on doit chercher une deuxième droite de direction v passant par l'origine et maximise $v{}^t$X^{t}Xv où v étant unitaire et orthogonale au vecteur u (^{t}vu=0: Le produit scalair des $\mathbb{R}^p$est nul)

On a la relation matricielle v${}^tX{}^t$Xv-2$\lambda v - \beta u = 0$ (généralisation du lagrangien avec deux contraintes)

λ et β étant les multiplicateurs de Lagrange, prémultiplions par ${}^tu$

et que ^{t}vu=0 il vient alors ^{t}XXt=λv

v sera donc le second vecteur propre, associé à la seconde valeur propre de ^{t}XX, en généralisant,

on obtient le résultat suivant:

Une base orthonormée du sous-espace vectorielle à q dimensions ajustant au sens des moindres carrées le nuage- maximisations des projections(p_r)ou minimisation des(d)- est constituée par:

Les q vecteurs propres correspondants aux q plus grandes valeurs propres de la matrice symétrique ${}^tX\ X$.

Ajustement par un sous espace de $\mathbb{R}^n$

Dans l'espace $\mathbb{R}^n$ on dispose de p (variables) X^j

Soit w le vecteur des cosinus dircteurs d'une droite G_1 passant par l'origine, la condition

que G_1 ajuste au mieux le nuage de $\mathbb{R}^n$ est: la somme des carrées des projections- ${}^t\mathrm{w}\mathrm{X}^j-$soit maximale, qui se traduit par:

$\max_v \sum_{j=1}^{p} \left\{ {}^t w X^j \right\}^2$

or

$$\sum_{j=1}^{p} \left\{ {}^t w X^j \right\}^2 = \sum_{j=1}^{p} \left\{ {}^t w X^{jt} X^j w \right\}$$

Soit

$$S_1^2 = \sum_{j=1}^{p} \left\{ {}^t w X^{jt} X^j w \right\} =^t w \sum_{j=1}^{p} \left\{ X^{jt} X^j \right\}^t w =^t w X^t X w$$

Le vecteur cherché **w à n** composantes rend donc maximum la forme quadratique $\mathrm{S}_1^2 = {}^t w X^t X w$avc la contrainte ${}^t\mathrm{vv}=1$

d'aprés ce qui précède w est un vecteur propre de la matrice $(X\ {}^tX)_{n\times n}$ (remarquer la dimension: $(n \times p) \times (p \times n) = n \times n$) relatif à la plus grande valeur propre.

Relation entre l'étude sur $\mathbb{R}^p$ et $\mathbb{R}^n$

Trouvons les relations entre le vecteur $\mathrm{w}_{(n)}$(matrice $\mathrm{X}^t\mathrm{X}$) et $\mathrm{u}_{(p)}$ (matrice ${}^t\mathrm{XX}$)

On a la relation:

w_q est le qu-ièm vecteur propre de la matrice X^tXde valeur propre $\lambda_q : X^tXw_q = \lambda_q w_q$ (*)

il y'a m valeurs propres non nulles, m étant le rang de $\mathrm{X}^t\mathrm{X}$, donc le rang de X (voir propriéte du rang)

E prémultiplions par ${}^t\mathrm{X}$ l'égalité: $X^tXw_q = \lambda_q w_q$

$${}^t\mathrm{X}\ X({}^tXw_q) = \lambda_q({}^tXw_q)$$

or cette égalité signifie: le vecteur(${}^tXw_q$) est un vecteur propre de${}^t\mathrm{XX}$ relative à λ_q

A chaque vecteur propre w_q (q$\leq m$) de $\mathrm{X}\ {}^tX$ correspond donc un vecteur propre

$$u_q = {}^tXw_q$$

de ${}^t\mathrm{XX}$ relative à la même valeur propre .

Toute valeur propre non nulle de X^tX est donc valeur propre de ${}^tXX$ (et réciproquement)et ainsi les vecteurs propres correspondants sont liés par la relation $u_q = {}^t Xw_q$ avec ^{t}ww=1 (remarquer les dimensions: u_q est un vecteur de p lignes)

Les vecteurs propres sont cependant difinis à un coefficient prés: si w_q est unitaire alors $u_q(={}^tXw_q)$ n'est pas forcément unitaire , en effet: ${}^tu_qu_q={}^t({}^tXw_q)({}^tXw_q) = {}^t w_qX^tX\ w_q = \lambda_q$ d'aprés (*)

Exprimons w_q en fonction de u_q en prémultipliant $u_q = {}^tXw_q$ par X $\Longrightarrow Xu_q = X^tXw_q$ d'aprés ${}^tu_qu_q=\lambda_q$

donc

$$w_q = \frac{1}{\lambda_q} Xu_q$$

Afin de donner à ces relations une forme symétrique

remplaçons l'égalité $u_q = {}^tXw_q$ par

$$\begin{aligned} u_q &= \frac{1}{\sqrt{\lambda_q}}\ {}^tXw_q \qquad \text{où q=1,...,m} \\ i-e &: \text{On rend les vecteurs } u_q \text{unitaires} \end{aligned}$$

On a dans ce cas

$$\begin{aligned} w_q &= \frac{1}{\sqrt{\lambda_q}}\ Xu_q \qquad \text{où q=1,....,m} \\ &\text{avec } u_q, w_q \text{ unitaires} \end{aligned}$$

La coordonnée d'un point-individu- de $\mathbb{R}^p$ sur l'axe factoriel F_q de cosinus directeur $u_q$est ${}^tu_qX_i$ (ou encore ${}^tX_iu_q$)

les coordonnées des n individus-nuage de l'espace $\mathbb{R}^p$-sont donc:

$$({}^tX_1u_q, {}^tX_2u_q, ..., {}^tX_nu_q)$$

Ce sont donc les composantes du vecteur

$$Xu_q$$

c'est à dire au coefficient $\frac{1}{\sqrt{\lambda_q}}$ prés, les composantes de w_q d'aprés la relation $w_q = \frac{1}{\sqrt{\lambda_q}} X u_q$.

De même , les composantes des p variables de $\mathbb{R}^n$ sur l'axe factoriel G_q de cosinus directeur $w_q$sont:

$$({}^tX^1 w_q, {}^t X^2 w_q, ..., {}^t X^p w_q)$$

Ce sont donc les composantes du vecteur

$${}^tX w_q$$

et au même coefficient $\frac{1}{\sqrt{\lambda_q}}$ prés, celles du vecteur u_q d'aprés l'égalité: $u_q = \frac{1}{\sqrt{\lambda_q}} {}^tX w_q$.

Résumé

Les cosinus directeurs du qu-ième axe factoriel dans un espace: Sont les coordonnées des points sur le qu-ième axe factoriel de **l'autre** espace, multiplié par

$$\frac{1}{\sqrt{\lambda_q}}.$$

Reconstitution du tableau X des données initiales

Un répère formé par les s premiers axes factoriels permet de reconstituer les positions des points avec une précision dépendant du rapport

$$\frac{\lambda_1 + ... + \lambda_s}{\sum_{i=1}^{m} \lambda_i} \quad \text{où s} \leq m$$

(on suppose qu'il y'a m valeurs propres)

Remarque

$$\sum_{i=1}^{m} \lambda_i = Trace {}^tXX = \sum_j \sum_i x_{ij}^2$$

Si on veut reconstituer les valeurs numériques initiales

La relation

$w_q = \dfrac{1}{\sqrt{\lambda_q}} X u_q \Leftrightarrow X u_q = \sqrt{\lambda_q} w_q$ posmultiplions les deux membres par ${}^t\mathrm{u}_q$: $X u_q \, {}^t\mathrm{u}_q = \sqrt{\lambda_q} v_q^t \mathrm{u}_q$

faisons la somme sur l'indice q: $X \sum_{q=1}^{p} \left\{u_q^t u_q\right\} = \sum_{q=1}^{p} \sqrt{\lambda_q} w_q^t \mathrm{u}_q$

La quantité $\sum_{q=1}^{p} \left\{u_q^t u_q\right\} = I_{p\times p}$ la matrice unite de $\mathbb{R}^p$ (c'est le produit de la matrice orthogonale par sa transposée, qui est aussi son inverse.d'où:

$$\begin{aligned} X_{n\times p} &= \sum_{q=1}^{p} \sqrt{\lambda_q} w_q^t u_q \\ \text{Remarquer les dimensions} &: n \times p = (n \times 1)(1 \times p) \end{aligned}$$

On a donc une reconstitution approchée du tableau X (données initiales)

En se limitant aux **s premièrs** axes factoriels, et en négligeant les termes $\sqrt{\lambda_{s+1}}, ..., \sqrt{\lambda_p}$ on a:

$$X_{n\times p} = \sum_{q=1}^{s} \sqrt{\lambda_q} w_q^t u_q$$

On remplace ainsi les n×p nombres du tableau X par s×$(n + p + 1)$ nombres

Si X est tableau de 1000×80 (1000 individus sur lesquels on fait passer 80 variables) et si s=10 (on se contente des dix premiérs facteurs), on remplace 80000 données par 10810 nombres(méthode de réduction de données)

En pratique on ne cherche pas à retrouver intégralement le tableau initiale mais un tableau approché.

L'analyse en composantes principales (ACP)

L'ACP s'applique aux donnée numériques **quantitatives**, présentées sous forme de tableaux:

Individus×variables

Elle comporte 3 variantes :ACP générale, ACP centrée réduite et ACP des rangs. on présente par l'ACP centrée réduite dite ACP faite sur la matrice de corrélation, c'est la plus courante.

Nature des données:

Les données traitées par l'ACP doivent être quantitatives (modalités mesurables, à titre d'exemples: le nombre d'enfants, l'âge, le poids, la taille,...,etc.)

Ces variables quantitatives sont isolées **(discrètes)** ou groupées en classes **(continues)**

Il convient de noter les données suivant leurs rôle dans l'analyse, ainsi on distingue **les données actives,** celles qui s'impliquent dant toutes les phases de l'analyse: De la détermination de l'espace sur lequel on projette jusqu'a la visualisation et l'interprétation des résultats , les données **passives (illustrative ou encore supplémentaires)** elle interviennent uniquement dans la phase finale de visualisation et permettent de mieux visualiser et de bien caractériser des groupes d'individus et les relations entre les variables initiales.

Présentation:

Tableau des données.

Les données soumises à l'ACP sont présentées sous la forme:

Tableau croisant n individus et p variables

X (Matrice des données)

Individus/Variables	X^1	X^2....	X^j....	X^{p-1}	X^p
1	x_{11}	x_{12}...	x_{1j}.....		x_{1p}
.....	.....	.....	....	.....	.....
i	x_{i1}	x_{i2}....	x_{ij} notée parfois (x_i^j)..		x_{ip}
.....	.....	.....	.....	.....	.....
n	x_{n1}	x_{n2}....	x_{nj}....	.....	x_{np}

Ainsi x_{ij} désigne l'observation de la j-ième variable faite sur le i-ièm individu (parfois notée

x_i^j, remarquer le numéro de l'individu i est en bas)

Les calculs statistiques (évaluation des paramètres) peuvent effectués pour **chaque colonne,**quant aux lignes si les variables sont **homogènes exprimés dans la même unité**, c'est affirmatif, exemple chiffre d'affaire durant les p années, revenu durant les p mensualités, sinon, si les variables sont héterogènes et héteroclites, unité différentes, c'est négatif, exemple pour une région déterminée (individu) il est abérant de parler d'un nombre moyen de ses caractéristiques(variables) tel que la superficie, le nombre d'habitants, infrastrucure urbaine(le nombre de logements), nombre de lignes téléphoniques.

Comme la perception visuelle humaine est dans un espace n'exédant pas 3 dimensions ($\mathbb{R}$:La droite réelle, $\mathbb{R}^2$:le plan, et l'espace: $\mathbb{R}^3$) dés lors que le nombre des données est au dela, des méthodes d'étude s'imposent, l'ACP réduite en est un.

Présentation de l'ACP réduite

La présentation des données dans l'espace:

Décomposons l'ensemble en deux représentations: Celle des individus et celles des variables

Le tabeau des données décrit ci -dessus les n individus à l'aide des p variables

Chaque individu i peut donc être représenté par un point de l'espace à **p dimensions** appelé **espace des variables,** l'ensemble des points représentés dans cet espace est appelé **nuage des individus**

La figure ci-dessous correspond à un nuage des individus

avec espace des variables à 3 dimensions(variable1,2,3)

les coordonnées de i dans cet espace sont (x_{i1}, x_{i2}, x_{i3})

De même chaque variable X_j peut être représentée par un point dans un espace à **n dimensions appelé espace des individus,** l'ensemble de ces points forme **le nuage des variables**

La figure ci-dessous représente la visualisation de la variable X_j dans l'espace de 3 individus

Les coordonnées sont(x_{1j}, x_{2j}, x_{3j})

L'analyse des individus permet la recherche des ressemblances, d'eventuelle existence homogéinité, de corrélation.

Quant à l'analyse des variables dont le principe consiste outre à determiner celles qui sont liées et celles qui s'opposent, et d'éventuelles corrélation.

Remarquer la correspondance entre les deux analyses est trés forte, en effet: Il s'agit d'une lecture différente du même texte(les données, le corpus statistique).Chaque individu est décrit par une variable et chaque variable est affecté à un individu. Autrement dit, ils'agit d'une traduction doublée de la même information.

Pour parer à l'héteroginité des variables, la solution consiste à les centrer et à les réduire.

Transformation des données (centrage et normalisation)

Chaque variable $\mathrm{X}^j = {}^t(\mathrm{x}_{1j}, ..., x_{ij}, ..., x_{nj})$ est remplaçée par $\mathrm{X}^{*j} = {}^t(\mathrm{x}^*_{1j}, ..., x^*_{ij}, ..., x^*_{nj})$ avec

$$x^*_{ij} = \frac{x_{ij} - \overline{X}^j}{\sigma_j}$$

Les variables se retrouvent ainsi toutes exprimées en nombre d'écart types, ce qui permet de considerer chacune d'elles indépendamment du choix de son unité.

Analyse du nuage des individus

Individus/Variables	X^1	X^2....	X^j....	X^{p-1}	X^p
1	x_{11}	x_{12}...	x_{1j}.....		x_{1p}
.....	.....	.....	....	.....	.....
i	x_{i1}	x_{i2}....	x_{ij}		x_{ip}
.....	.....	.....	.....	.....	.....
n	x_{n1}	x_{n2}....	x_{nj}....	.....	x_{np}

X:

Point moyen ou centre de gravité: G= ${}^t(\overline{X}^1, ..., \overline{X}^p)$

On a avec $\mathbf{1}_n = {}^t(1, ..., 1)$

$$G = {}^t X D_p \mathbf{1}_n$$

Remarquer l'égalité des dimensions: $p \times 1 = (p \times n) \times (n \times n) \times (n \times 1)$

Soit la matrice (données centrées)

$$Y = X - 1^t G$$

Matrice variance covariance $V_{p\times p}$=$\left[cov(X^j, X^{j'})_{j,j'=1,\ldots,p}\right]$ V s'écrit:

$$V = {}^t Y D_p Y$$

où la matrice de poids $D_p = \begin{bmatrix} p_1 & \ldots\ldots & 0 \\ . & . & \\ 0 & \ldots\ldots & p_n \end{bmatrix}$ où $p_i > 0$,et $\sum_{i=1}^n p_i = 1$

généralement les n individus sont choisis aléatoirement ($p_i = \frac{1}{n}$ d'où $D_p = \frac{1}{n} I_n$ avc $I_n$matrice unité de $\mathbb{R}^n$)

Dans ce cas

$$V = \frac{1}{n}{}^t YY \quad \text{avec} \quad Y = X - 1^t G$$

Recherche des axes factoriels u_q dans $\mathbb{R}^p$

Comme le vecteur propre de la matrice variance covariance V=$\frac{1}{n}{}^t YY$ correspond à la quièmevaleur propre λ_q

u_q est donc vecteur propre de ${}^t YY$ relative à la valeur propre $\mathbf{n}\times\lambda_q$

Les formules précédantes de l'analyse générale deviennent:

$w_q = \frac{1}{\sqrt{n\lambda_q}} Y u_q$.La i-ème composante de w_q vaut $w(i)_q = \frac{1}{\sqrt{n\lambda_q}}{}^t Y_i u_q$.