Cours d'analyse de données

Jean-Marc Lasgouttes
Jean-Marc.Lasgouttes@inria.fr

Magistère de Finance de Paris 1, 2è année

http://www-roc.inria.fr/~lasgoutt/ana-donnees/

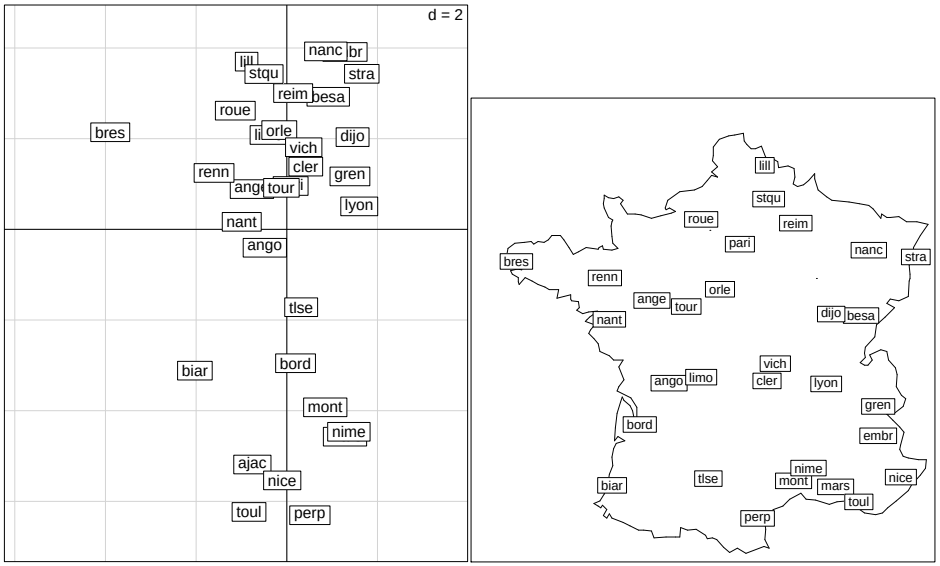
Introduction

qu'est-ce que l'analyse de données ?

Exemple : la température en France

	janv	fev	mars	avri	mai	juin	juil	aout	sept	oct	nov	dec
ajac	7.7	8.7	10.5	12.6	15.9	19.8	22.0	22.2	20.3	16.3	11.8	8.7
ange	4.2	4.9	7.9	10.4	13.6	17.0	18.7	18.4	16.1	11.7	7.6	4.9
ango	4.6	5.4	8.9	11.3	14.5	17.2	19.5	19.4	16.9	12.5	8.1	5.3
besa	1.1	2.2	6.4	9.7	13.6	16.9	18.7	18.3	15.5	10.4	5.7	2.0
biar	7.6	8.0	10.8	12.0	14.7	17.8	19.7	19.9	18.5	14.8	10.9	8.2
bord	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2
bres	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16.0	14.7	12.0	9.0	7.0
cler	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6
dijo	1.3	2.6	6.9	10.4	14.3	17.7	19.6	19.0	15.9	10.5	5.7	2.1
embr	0.5	1.6	5.7	9.0	13.0	16.4	18.9	18.3	15.3	10.1	4.6	0.5
gren	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3
lill	2.4	2.9	6.0	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5
limo	3.1	3.9	7.4	9.9	13.3	16.8	18.4	17.8	15.3	10.7	6.7	3.8
lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1
mars	5.5	6.6	10.0	13.0	16.8	20.8	23.3	22.8	19.9	15.0	10.2	6.9
mont	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10.0	6.5
nanc	0.8	1.6	5.5	9.2	13.3	16.5	18.3	17.7	14.7	9.4	5.2	1.8
nant	5.0	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5
nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16.0	11.5	8.2
nime	5.7	6.8	10.1	13.0	16.6	20.8	23.6	22.9	19.7	14.6	9.8	6.5
orle	2.7	3.6	6.9	9.8	13.4	16.6	18.4	18.2	15.6	10.9	6.6	3.6
pari	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16.0	11.4	7.1	4.3
perp	7.5	8.4	11.3	13.9	17.1	21.1	23.8	23.3	20.5	15.9	11.5	8.6
reim	1.9	2.8	6.2	9.4	13.3	16.4	18.3	17.9	15.1	10.3	6.1	3.0
renn	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4
roue	3.4	3.9	6.8	9.5	12.9	15.7	17.6	17.2	15.0	11.0	6.8	4.3
stqu	2.0	2.9	6.3	9.2	12.7	15.6	17.4	17.4	15.0	10.5	6.1	3.1
stra	0.4	1.5	5.6	9.8	14.0	17.2	19.0	18.3	15.1	9.5	4.9	1.3
toul	8.6	9.1	11.2	13.4	16.6	20.2	22.6	22.4	20.5	16.5	12.6	9.7
tlse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5
tour	3.5	4.4	7.7	10.6	13.9	17.4	19.1	18.7	16.2	11.7	7.2	4.3
vich	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16.0	11.0	6.6	3.4

La température en France (2)



Individus et variables

Population groupe ou ensemble d'*individus* que l'on analyse.

Recensement étude de tous les individus d'une population donnée.

Sondage étude d'une partie seulement d'une population appelée *échantillon*.

Variables ensemble de caractéristiques d'une population.
- *quantitatives* : nombres sur lesquels les opérations usuelles (somme, moyenne,...) ont un sens ; elles peuvent être *discrètes* (ex : nombre d'éléments dans un ensemble) ou *continues* (ex : prix, taille) ;
- *qualitatives* : appartenance à une catégorie donnée ; elles peuvent être *nominales* (ex : sexe, CSP) ou *ordinales* quand les catégories sont ordonnées (ex : très résistant, assez résistant, peu résistant).

L'analyse de données

But synthétiser, structurer l'information contenue dans des données multidimensionnelles (n individus, p variables).

Deux groupes de méthodes
- *méthodes de classification* : réduire la taille de l'ensemble des individus en formant des groupes homogènes ;
- *méthodes factorielles* : réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques.

Deux types de méthodes factorielles
- *analyse en composantes principales* : variables numériques ;
- *analyse des correspondances* : variables qualitatives.

But du cours

Méthodes couvertes par le cours
- analyse en composantes principales (ACP) ;
- analyse (factorielle) des correspondances (AFC) ;
- analyse des correspondances multiples (ACM).

Compétences recherchées
- comprendre les fondements mathématiques des méthodes ;
- savoir interpréter les tables et graphiques issus de ces méthodes ;
- être capable de mener soi-même une telle étude.

Ce que ce cours n'est pas

Un cours de mathématiques financières il n'y a pas de modèles probabilistes de processus financiers (cours de bourse...).

Un cours de statistique inférentielle il ne sera presque pas question ici de tests, d'estimateurs, de prévision statistique.

Un cours orienté « utilisateur » on cherche à la fois à savoir utiliser les méthodes d'analyse de données, et à comprendre les fondements mathématiques de ces méthodes.

Un cours appliqué aux données financières ce cours est avant tout un cours de méthode ; la plupart des exemples abordés ne seront pas issus de cette application.

Un cours « pratique » Les contraintes d'effectif et de matériel ne permettent pas d'effectuer des travaux pratiques.

Outils utilisés

Algèbre linéaire les données sont vues de manière *abstraite* comme un nuage de points dans un espace vectoriel ; les notions suivantes doivent être bien comprises

- *vecteurs* : produits scalaires, décomposition selon une base
- *matrices* : addition, multiplication, transposée, trace
- *valeurs* et *vecteurs propres* : définition, propriétés
- *métriques* : définitions des distances dans un espace vectoriel par une norme, lien avec le produit scalaire

Attention : les étudiants sont supposés maîtriser le calcul matriciel et la notion de valeur propre ; les TD et examens comporteront du calcul matriciel !

Théorie des probabilités on utilisera quand même quelques tests statistiques.

Statistiques et probabilités

Une approche différente Les probabilités reposent sur un modèle de données et font en général des hypothèses simplificatrices. Ici, on utilisera plus des considérations *géométriques*.

3 liens possibles

- les données statistiques sont empruntes d'une forme de variabilité liée aux erreurs de mesures ; on peut modéliser cette erreur par une variable aléatoire ;
- on constate souvent que la répartition d'une variable est proche d'une loi de probabilités connue ;
- surtout, quand des données sont issue d'un sondage, on peut considérer que ce sont des tirages d'une variable aléatoire. Quand les échantillons sont assez grands, on connaît des lois limites.

Références

Ces références sont données à titre indicatif ; *aucun livre n'est demandé pour ce cours*.

Base du cours Gilbert Saporta, *Probabilités, analyse des données et statistique*, 2nde édition, Technip, 2006.

Version plus simple Jean-Marie Bouroche et Gilbert Saporta, *L'analyse des données*, Que Sais-je ?, Presses Universitaires de France, 2002.

Logiciel de traitement de données Les tables et graphiques présentés dans le cours et les TD sont produits par le logiciel R (à l'aide du paquetage `ade4`). R est un logiciel libre (et donc gratuit) disponible pour Windows, Mac OS X et Linux à l'adresse `http://www.r-project.org`.

Archives de ce cours cours, TD avec corrigé, données sont disponibles à
`http://www-roc.inria.fr/~lasgoutt/ana-donnees/`

Partie I

variables quantitatives : analyse en composantes principales

Description de données quantitatives

Définition On appelle « variable » un vecteur $\mathbf{x}$ de taille n. Chaque coordonnée x_i correspond à un individu. On s'intéresse ici à des valeurs numériques.

Poids Chaque individu peut avoir un poids p_i, tel que $p_1 + \cdots + p_n = 1$, notamment quand les individus n'ont pas la même importance (échantillons redressés, données regroupées,...). On a souvent $p = 1/n$.

Représentation histogramme en découpant les valeurs de la variable en classes ; ou alors « boîte à moustache ».

Résumés on dispose d'une série d'indicateurs qui ne donne qu'une vue partielle des données : effectif, moyenne, médiane, variance, écart type, minimum, maximum, étendue, 1er quartile (25% inférieurs), 4ème quartile (25% supérieurs), ... Ces indicateurs mesurent principalement la tendance centrale et la dispersion.

On utilisera principalement la moyenne, la variance et l'écart type.

Moyenne arithmétique

Définition On note

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

ou pour des données pondérés

$$\bar{x} = \sum_{i=1}^{n} p_i x_i.$$

Propriétés la moyenne arithmétique est une mesure de *tendance centrale* qui dépend de toutes les observations et est sensible aux valeurs extrêmes. Elle est très utilisée à cause de ses bonnes propriétés mathématiques.

Variance et écart-type

Définition la *variance* de $\mathbf{x}$ est définie par

$$s_{\mathbf{x}}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ ou } s_{\mathbf{x}}^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$$

L'*écart-type* $s_{\mathbf{x}}$ est la racine carrée de la variance.

Propriétés La variance satisfait la formule suivante

$$s_{\mathbf{x}}^2 = \sum_{i=1}^{n} p_i x_i^2 - (\bar{x})^2$$

La variance est « la moyenne des carrés moins le carré de la moyenne ». L'écart-type, qui a la même unité que $\mathbf{x}$, est une mesure de *dispersion*.

Mesure de liaison entre deux variables

Définitions la covariance observée entre deux variables $\mathbf{x}$ et $\mathbf{y}$ est

$$s_{\mathbf{xy}} = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \bar{x}\bar{y}.$$

et le *coefficient de r de Bravais-Pearson* ou coefficient de corrélation est donné par

$$r_{\mathbf{xy}} = \frac{s_{\mathbf{xy}}}{s_{\mathbf{x}} s_{\mathbf{y}}} = \frac{\sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} p_i (y_i - \bar{y})^2}}.$$

Ces deux grandeurs sont symétriques : $s_{\mathbf{xy}} = s_{\mathbf{yx}}$ et $r_{\mathbf{xy}} = r_{\mathbf{yx}}$.

Propriétés du coefficient de corrélation

Borne On a toujours (inégalité de Cauchy-Schwarz)

$$-1 \leq r_{\mathbf{xy}} \leq 1.$$

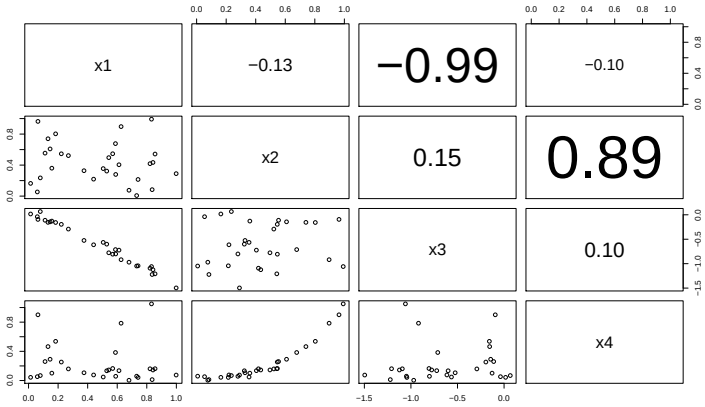
Variables liées $|r_{\mathbf{xy}}| = 1$ si et seulement si $\mathbf{x}$ et $\mathbf{y}$ sont linéairement liées :

$$ax_i + by_i = c, \text{ pour tout } 1 \leq i \leq n.$$

En particulier, $r_{\mathbf{xx}} = 1$.

Variables décorrélées si $r_{\mathbf{xy}} = 0$, on dit que les variables sont *décorrélées*. Cela ne veut pas dire qu'elles sont indépendantes !

Le coefficient de corrélation par l'exemple



Interprétation on a 4 variables numériques avec 30 individus. Les variables 1 et 2 sont indépendantes ; les variables 1 et 3 ont une relation linéaire ; les variables 2 et 4 ont une relation non-linéaire.

Que signifie une corrélation linéaire ?

Qu'est ce qui est significatif ? si on a assez de données, on peut considérer qu'une corrélation supérieure à $0,5$ est forte, et une corrélation entre $0,3$ et $0,5$ est moyenne.

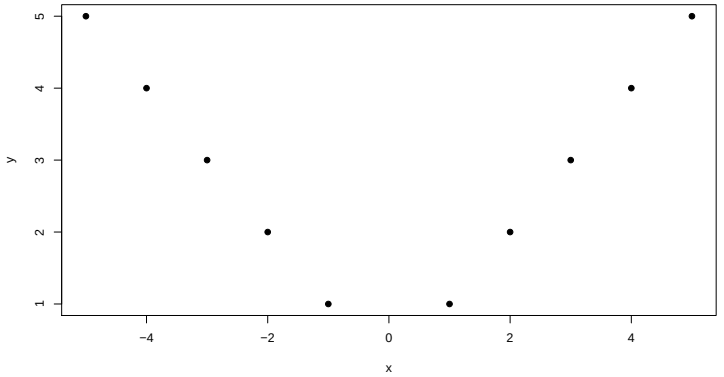
Une corrélation égale à un indique que les deux variables sont équivalentes.

Qu'est-ce que cela veut dire ? une corrélation significative indique une liaison entre deux variables, mais pas nécessairement un lien de causalité. Exemple :

Le nombre de pompiers présents pour combattre un incendie est corrélé aux dégâts de l'incendie.

Mais *ce ne sont pas les pompiers qui causent les dégâts*.

Que signifie une décorrélation ?



Pour ces deux variables, on a $r = 0$.

Rappels : notation matricielle

Matrice tableau de données carré ou rectangulaire, noté par un lettre majuscule grasse (ex : $\mathbf{X}$).

Vecteur matrice à une seule colonne, noté par une lettre minuscule grasse (ex : $\mathbf{x}$).

Cas particuliers matrice identité à n lignes et n colonnes et vecteur unité de dimension n :

$$\mathbf{I}_n = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}, \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Transposition de matrice échange des lignes et des colonnes d'une matrice ; on note $\mathbf{M}'$ la transposée de $\mathbf{M}$.

Tableau de données

On note x_i^j la valeur de la j-ème variable pour le i-ème individu. Pour n individus et p variables, on a le tableau

$$\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^p) = \begin{bmatrix} x_1^1 & x_1^2 & & & \cdots & x_1^p \\ x_2^1 & x_2^2 & & & & \\ & & & \vdots & & \\ & & \cdots & x_i^j & & \\ \vdots & & & & \ddots & \\ x_n^1 & & & & & x_n^p \end{bmatrix}.$$

$\mathbf{X}$ est une matrice rectangulaire à n lignes et p colonnes.

Vecteurs variable et individu

Variable Une colonne du tableau

$$\mathbf{x}^j = \begin{bmatrix} x_1^j \\ \vdots \\ x_i^j \\ \vdots \\ x_n^j \end{bmatrix}$$

Individu Une ligne du tableau

$$\mathbf{e}_i' = (x_i^1, \dots, x_i^j, \dots, x_i^p)$$

La matrice des poids

Définition on associe aux individus un poids p_i tel que

$$p_1 + \cdots + p_n = 1$$

et on représente ces poids dans la matrice diagonale de taille n

$$\mathbf{D} = \begin{bmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{bmatrix}.$$

Cas uniforme tous les individus ont le même poids $p_i = 1/n$ et $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$.

Point moyen et tableau centré

Point moyen c'est le vecteur $\mathbf{g}$ des moyennes arithmétiques de chaque variable :

$$\mathbf{g}' = (\bar{x}^1, \ldots, \bar{x}^p),$$

où

$$\bar{x}^j = \sum_{i=1}^{n} p_i x_i^j.$$

On peut aussi écrire $\mathbf{g} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$.

Tableau centré il est obtenu en centrant les variables autour de leur moyenne

$$y_i^j = x_i^j - \bar{x}^j$$

ou, en notation matricielle,

$$\mathbf{Y} = \mathbf{X} - \mathbf{1}_n\mathbf{g}' = (\mathbf{I} - \mathbf{1}_n\mathbf{1}_n'\mathbf{D})\mathbf{X}$$

Matrice de variance-covariance

Définition c'est une matrice *carrée* de dimension p

$$\mathbf{V} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & & & \\ \vdots & & \ddots & \\ s_{p1} & & & s_p^2 \end{bmatrix},$$

où s_{kl} est la covariance des variables $\mathbf{x}^k$ et $\mathbf{x}^\ell$ et s_j^2 est la variance de la variable $\mathbf{x}^j$

Formule matricielle

$$\mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X} - \mathbf{g}\mathbf{g}' = \mathbf{Y}'\mathbf{D}\mathbf{Y}.$$

Matrice de corrélation

Définition Si l'on note $r_{k\ell} = s_{k\ell}/s_k s_\ell$, c'est la matrice $p \times p$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & & \\ \vdots & & \ddots & \\ r_{p1} & & & 1 \end{bmatrix},$$

Formule matricielle $\mathbf{R} = \mathbf{D}_{1/\mathbf{s}}\mathbf{V}\mathbf{D}_{1/\mathbf{s}}$, où

$$\mathbf{D}_{1/\mathbf{s}} = \begin{bmatrix} \frac{1}{s_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_p} \end{bmatrix}$$

L'analyse de composantes principales (ACP)

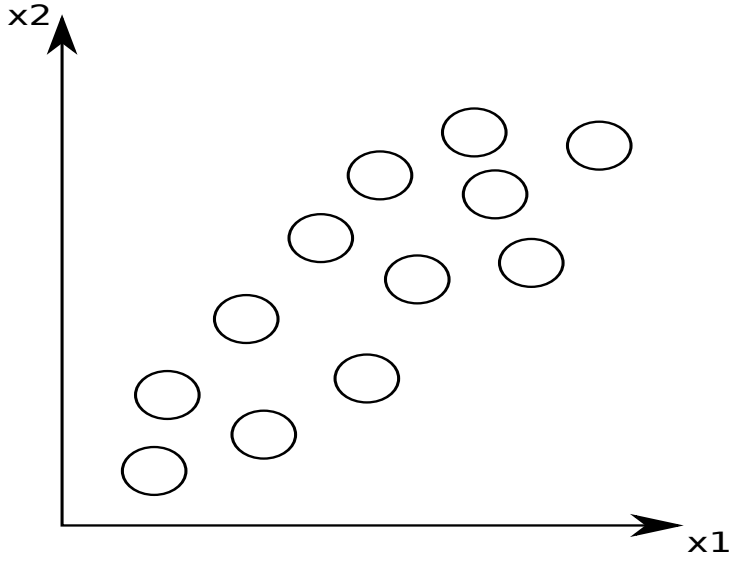
Contexte chaque individu est considéré comme un point d'un espace vectoriel F de dimension p. L'ensemble des individus est un *nuage* de points dans F et $\mathbf{g}$ est son *centre de gravité*.

Principe on cherche à réduire le nombre p de variables tout en préservant au maximum la structure du problème.

Pour cela on projette le nuage de points sur un sous-espace de dimension inférieure.

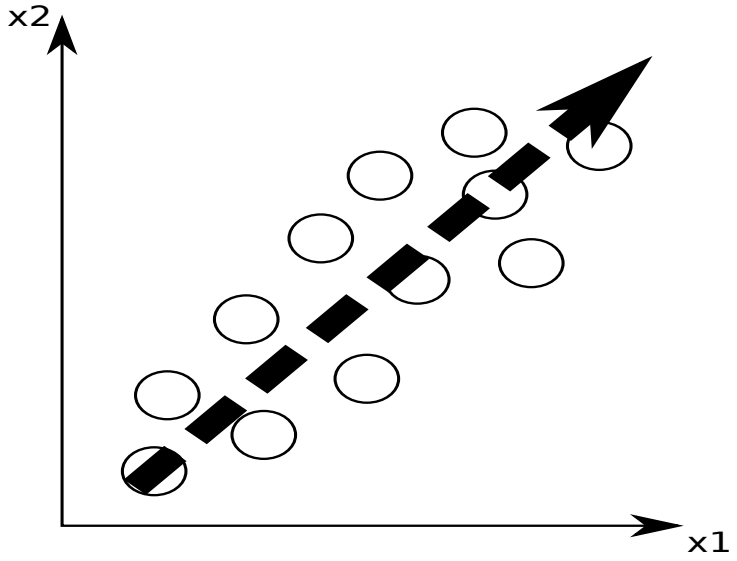
Exemple en dimension 2

On veut passer de 2 variables à 1 seule.



Exemple en dimension 2 (suite)

On cherche la direction qui différencie le plus les points entre eux.



Distance entre individus

Motivation afin de pouvoir considérer la structure du nuage des individus, il faut définir une distance, qui induira une géométrie.

Distance euclidienne classique la distance la plus simple entre deux points de $\mathbb{R}^p$ est définie par

$$d^2(\mathbf{u},\mathbf{v}) = \sum_{j=1}^{p} (u_j - v_j)^2 = \|\mathbf{u}-\mathbf{v}\|^2$$

Généralisation simple on donne un poids $m_j > 0$ à la variable j

$$d^2(\mathbf{u},\mathbf{v}) = \sum_{j=1}^{p} m_j (u_j - v_j)^2$$

Utiliser ce poids est équivalent à multiplier la coordonnée j par $\sqrt{m_j}$

Métrique

Définition soit $\mathbf{M} = \mathrm{diag}(m_j)$, où $m_1, \dots, m_p$ sont des réels strictement positifs. On pose

$$\|\mathbf{u}\|_{\mathbf{M}}^2 = \mathbf{u}'\mathbf{M}\mathbf{u} = \sum_{j=1}^{p} m_j u_j^2,$$
$$d_{\mathbf{M}}^2(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|_{\mathbf{M}}^2.$$

Espace métrique il est défini par le produit scalaire

$$\langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{M}} = \mathbf{u}'\mathbf{M}\mathbf{v} = \sum_{j=1}^{p} m_j u_j v_j.$$

On notera que $\|\mathbf{u}\|_{\mathbf{M}}^2 = \langle \mathbf{u},\mathbf{u}\rangle_{\mathbf{M}}$.

Orthogonalité on dit que $\mathbf{u}$ et $\mathbf{v}$ sont $\mathbf{M}$-orthogonaux si $\langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{M}} = 0$.

Propriétés du produit scalaire

Le produit scalaire est commutatif

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} = \langle \mathbf{v}, \mathbf{u} \rangle_{\mathbf{M}}$$

Le produit scalaire est linéaire

$$\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle_{\mathbf{M}} = \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} + \langle \mathbf{u}, \mathbf{w} \rangle_{\mathbf{M}},$$
$$\langle \mathbf{u}, \lambda \mathbf{v} \rangle_{\mathbf{M}} = \lambda \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} \quad \text{pour tout } \lambda \in \mathbb{R}.$$

Identité remarquable

$$\|\mathbf{u} + \mathbf{v}\|^2_{\mathbf{M}} = \|\mathbf{u}\|^2_{\mathbf{M}} + \|\mathbf{v}\|^2_{\mathbf{M}} + 2\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}}$$

Le cas de la métrique $\mathbf{D}_{1/\mathbf{s}^2}$

Pourquoi cette métrique ?
- pour que les distances soient indépendantes des unités de mesure
- pour qu'elles ne privilégient pas les variables dispersées.

Équivalence avec les données réduites on a $\mathbf{D}_{1/\mathbf{s}^2} = \mathbf{D}_{1/\mathbf{s}} \mathbf{D}_{1/\mathbf{s}}$ et donc

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{D}_{1/\mathbf{s}^2}} = \langle \mathbf{D}_{1/\mathbf{s}} \mathbf{u}, \mathbf{D}_{1/\mathbf{s}} \mathbf{v} \rangle.$$

Travailler avec la métrique $\mathbf{D}_{1/\mathbf{s}^2}$ est équivalent à diviser chaque variable par son écart-type et à utiliser la métrique $\mathbf{I}$.

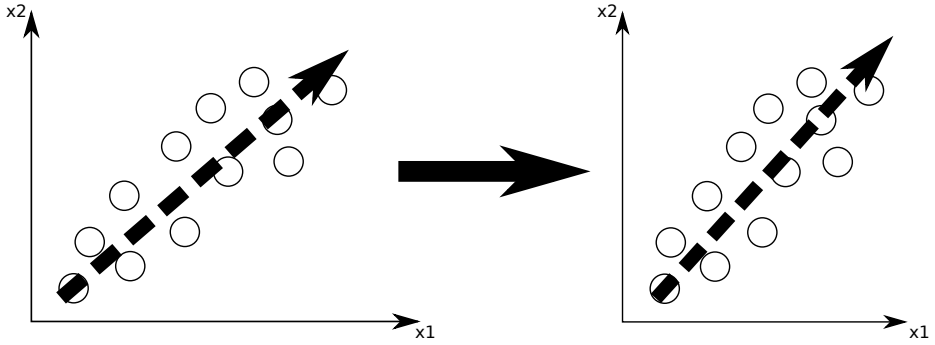
Données centrées réduites c'est le tableau $\mathbf{Z}$ contenant les données

$$z_i^j = \frac{x_i^j - \bar{x}^j}{s_j},$$

qui se calcule matriciellement comme $\mathbf{Z} = \mathbf{Y}\mathbf{D}_{1/\mathbf{s}}$.

Utilisation des métriques

Utiliser une métrique est donc équivalent à « tordre » les données, par exemple pour les rendre comparables



Exemple utiliser la métrique réduite est équivalent à travailler sur les données centrées réduites $\mathbf{Z} = \mathbf{Y}\mathbf{D}_{1/\mathbf{s}}$.

Inertie

Définition l'inertie en un point $\mathbf{v}$ du nuage de points est

$$I_{\mathbf{v}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{v}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{v})' \mathbf{M} (\mathbf{e}_i - \mathbf{v}).$$

Inertie totale c'est $I_{\mathbf{g}}$, qui est la plus petite inertie possible, puisque

$$I_{\mathbf{v}} = I_{\mathbf{g}} + \|\mathbf{v} - \mathbf{g}\|^2_M$$

Autres relations $I_{\mathbf{g}}$ mesure la moyenne des carrés des distances entre les individus

$$2I_{\mathbf{g}} = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \|\mathbf{e}_i - \mathbf{e}_j\|^2_{\mathbf{M}}.$$

L'inertie totale est aussi donnée par la trace de la matrice $\mathbf{MV}$

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{MV}),$$

la trace d'une matrice étant la somme de ses éléments diagonaux.

Métriques particulières

Métrique usuelle $\mathbf{M} = \mathbf{I}_p$ correspond au produit scalaire usuel et

$$I_{\mathrm{g}} = \mathsf{Tr}(\mathbf{V}) = \sum_{j=1}^{p} s_i^2$$

Métrique réduite obtenue quand $\mathbf{M} = \mathbf{D}_{1/\mathbf{s}^2} = \mathbf{D}_{1/\mathbf{s}}^2$

$$I_{\mathrm{g}} = \mathsf{Tr}(\mathbf{D}_{1/\mathbf{s}^2}\mathbf{V}) = \mathsf{Tr}(\mathbf{D}_{1/\mathbf{s}}\mathbf{V}\mathbf{D}_{1/\mathbf{s}}) = \mathsf{Tr}(\mathbf{R}) = p.$$

L'analyse de composantes principales (version 2)

Principe on cherche à projeter le nuage de points sur un espace F_k de dimension $k < p$.

Critère on veut que la moyenne des carrés des distances entre les points projetés soit maximale (elle est toujours plus petite que pour le nuage original).

Pour cela on cherche F_k, sous espace de dimension k de F_p, tel que l'inertie du nuage projeté sur F_k soit maximale.

Rappels : valeurs propres et vecteurs propres

Définition un vecteur $\mathbf{v} \neq \mathbf{0}$ de taille p est un *vecteur propre* d'une matrice $\mathbf{A}$ de taille $p \times p$ s'il existe $\lambda \in \mathbb{C}$ telle que

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

λ est une *valeur propre* de $\mathbf{A}$ associée à $\mathbf{v}$.

Domaine En général, les vecteurs propres et valeurs propres sont complexes ; dans tous les cas qui nous intéressent, ils seront réels.

Interprétation des vecteurs propres ce sont les directions dans lesquelles la matrice agit.

Interprétation des valeurs propres c'est le facteur multiplicatif associé à une direction donnée.

Valeurs et vecteurs propres : un exemple concret

La matrice

$$\begin{pmatrix} 5 & 1 & -1 \\ 2 & 4 & -2 \\ 1 & -1 & 3 \end{pmatrix}$$

a pour vecteurs propres

$$\mathbf{v}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},\ \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},\ \mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

On vérifie facilement que les valeurs propres associées sont

$$\lambda_1 = 2,\ \lambda_2 = 4,\ \lambda_3 = 6.$$

Valeurs et vecteurs propres : cas particuliers

Matrice nulle sa seule valeur propre est 0, et tout vecteur est vecteur propre.

Matrice identité tout vecteur est vecteur propre de $\mathbf{I}$ avec valeur propre 1, puisque $\mathbf{Iv} = \mathbf{v}$.

Matrice diagonale si $\mathbf{D}_\lambda$ est une matrice diagonale avec les coefficients $\lambda_1, \ldots, \lambda_p$, alors le i-ème vecteur coordonnée est vecteur propre de $\mathbf{D}_\lambda$ associé à la valeur propre λ_i.

L'action d'une matrice diagonale est de multiplier chacune des coordonnées d'un vecteur par la valeur propre correspondante.

Matrice diagonalisable c'est une matrice dont les vecteurs propres forment une *base* de l'espace vectoriel : tout vecteur peut être représenté de manière unique comme combinaison linéaire des vecteurs propres. Une matrice de taille $p \times p$ qui a p valeurs propres réelles distinctes est diagonalisable dans $\mathbb{R}$.

Quelques matrices diagonalisables

Matrice symétrique une matrice symétrique réelle ($\mathbf{A}' = \mathbf{A}$) possède une base de vecteurs propres orthogonaux et ses valeurs propres sont réelles

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \text{ si } i \neq j, \qquad \text{et } \lambda_i \in \mathbb{R}.$$

Matrice $\mathbf{M}$-symétrique une matrice $\mathbf{M}$-symétrique réelle ($\mathbf{A}'\mathbf{M} = \mathbf{MA}$) possède une base de vecteurs propres $\mathbf{M}$-orthogonaux et ses valeurs propres sont réelles

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle_{\mathbf{M}} = 0 \text{ si } i \neq j, \qquad \text{et } \lambda_i \in \mathbb{R}.$$

Matrice définie positive c'est une matrice symétrique dont les valeurs propres sont strictement positives

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \text{ si } i \neq j, \qquad \text{et } \lambda_i > 0.$$

Analyse de $\mathbf{VM}$

Valeurs propres la matrice $\mathbf{VM}$ est $\mathbf{M}$-symétrique : elle est donc diagonalisable et ses valeurs propres $\lambda_1, \ldots, \lambda_p$ sont réelles.

Axes principaux d'inertie ce sont les p vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_p$ tels que

$$\mathbf{VMa}_k = \lambda_k \mathbf{a}_k, \qquad \text{avec } \langle \mathbf{a}_k, \mathbf{a}_\ell \rangle_{\mathbf{M}} = 1 \text{ si } k = \ell, \, 0 \text{ sinon.}$$

Ils sont $\mathbf{M}$-orthonormaux.

Signe des valeurs propres les valeurs propres de $\mathbf{VM}$ sont positives et on peut les classer par ordre décroissant

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0.$$

Idée du lien avec l'inertie on sait que $Tr(\mathbf{VM}) = \lambda_1 + \cdots + \lambda_p$. Si on ne garde que les données relatives à $\mathbf{a}_1, \ldots, \mathbf{a}_q$, on gardera l'inertie $\lambda_1 + \cdots + \lambda_q$, et c'est le mieux qu'on puisse faire.

Résultat principal

Théorème principal (Admis)

1. *Si F_k est le sous-espace de dimension k portant l'inertie principale, alors*

$$F_{k+1} = F_k \oplus f_{k+1},$$

   où f_{k+1} est le sous espace de dimension 1 $\mathbf{M}$-orthogonal à F_k portant l'inertie maximale : les solutions sont « emboîtées » ;

2. *F_k est engendré par les k vecteurs propres de $\mathbf{VM}$ associés aux k plus grandes valeurs propres.*

Interprétation du théorème l'ACP sur $k+1$ variables est obtenue par ajout d'une variable d'inertie maximale à l'ACP sur k variables. Il n'est pas nécessaire de refaire tout le calcul.

Les composantes principales

Coordonnées des individus supposons que $\mathbf{e}_i - \mathbf{g} = \sum_{\ell=1}^{p} c_{i\ell}\mathbf{a}_\ell$, alors

$$\langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k \rangle_{\mathbf{M}} = \sum_{\ell=1}^{p} c_{i\ell} \langle \mathbf{a}_\ell, \mathbf{a}_k \rangle_{\mathbf{M}} = c_{ik}$$

La coordonnée de l'individu centré $\mathbf{e}_i - \mathbf{g}$ sur l'axe principal $\mathbf{a}_k$ est donc donné par la projection $\mathbf{M}$-orthogonale

$$c_{ik} = \langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k \rangle_{\mathbf{M}} = (\mathbf{e}_i - \mathbf{g})'\mathbf{M}\mathbf{a}_k.$$

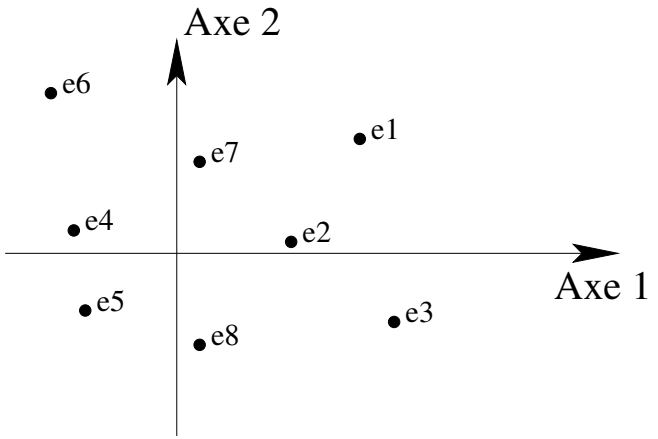
Composantes principales ce sont les variables $\mathbf{c}_k$ de taille n définies par

$$\mathbf{c}_k = \mathbf{Y}\mathbf{M}\mathbf{a}_k.$$

Chaque $\mathbf{c}_k$ contient les coordonnées des projections $\mathbf{M}$-orthogonales des individus centrés sur l'axe défini par les $\mathbf{a}_k$.

Représentation des individus dans un plan principal

Qu'est-ce que c'est ? pour deux composantes principales $\mathbf{c}_1$ et $\mathbf{c}_2$, on représente chaque individu i par un point d'abscisse c_{i1} et d'ordonnée c_{i2}.



Quand ? Elle est utile quand les individus sont discernables.

Propriétés des composantes principales

Moyenne arithmétique les composantes principales sont centrées :

$$\bar{c}_k = \mathbf{c}_k'\mathbf{D}\mathbf{1}_n = \mathbf{a}_k'\mathbf{M}\mathbf{Y}'\mathbf{D}\mathbf{1}_n = 0$$

car $\mathbf{Y}'\mathbf{D}\mathbf{1}_n = \mathbf{0}$ (les colonnes de $\mathbf{Y}$ sont centrées).

Variance la variance de $\mathbf{c}_k$ est λ_k car

$$\begin{aligned} V(\mathbf{c}_k) &= \mathbf{c}_k'\mathbf{D}\mathbf{c}_k = \mathbf{a}_k'\mathbf{M}\mathbf{Y}'\mathbf{D}\mathbf{Y}\mathbf{M}\mathbf{a}_k \\ &= \mathbf{a}_k'\mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a}_k = \lambda_k \mathbf{a}_k'\mathbf{M}\mathbf{a}_k = \lambda_k. \end{aligned}$$

Covariance de même, pour $k \neq \ell$,

$$\operatorname{cov}(\mathbf{c}_k, \mathbf{c}_\ell) = \mathbf{c}_k'\mathbf{D}\mathbf{c}_\ell = \cdots = \lambda_\ell \mathbf{a}_k'\mathbf{M}\mathbf{a}_\ell = 0.$$

Les composantes principales ne sont pas corrélées entre elles.

Facteurs principaux

Définition on associe à un axe principal $\mathbf{a}_k$ le facteur principal $\mathbf{u}_k = \mathbf{M}\mathbf{a}_k$ de taille p. C'est un vecteur propre de $\mathbf{M}\mathbf{V}$ car

$$\mathbf{M}\mathbf{V}\mathbf{u}_k = \mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a}_k = \lambda_k \mathbf{M}\mathbf{a}_k = \lambda_k \mathbf{u}_k$$

Calcul en pratique, on calcule les $\mathbf{u}_k$ par diagonalisation de $\mathbf{M}\mathbf{V}$, puis on obtient les $\mathbf{c}_k = \mathbf{Y}\mathbf{u}_k$. Les $\mathbf{a}_k$ ne sont pas intéressants.

La valeur d'une variable $\mathbf{c}_k$ pour l'individu $\mathbf{e}_i$ est donc

$$c_{ik} = (\mathbf{e}_i - \mathbf{g})'\mathbf{u}_k = \sum_{j=1}^{p} y_i^j u_{kj}$$

où $\mathbf{u}_j' = (u_{i1}, \ldots, u_{ip})$.

Formules de reconstruction

Il est possible de reconstruire le tableau centré $\mathbf{Y}$ à partir des composantes principales et des facteurs principaux

$$\mathbf{Y} = \sum_{k=1}^{p} \mathbf{c}_k \mathbf{a}_k' = \sum_{k=1}^{p} \mathbf{c}_k \mathbf{u}_k' \mathbf{M}^{-1}.$$

Preuve il suffit de calculer

$$\left(\sum_{k=1}^{p} \mathbf{c}_k \mathbf{a}_k' \right) \mathbf{M} \mathbf{a}_\ell = \sum_{k=1}^{p} \mathbf{c}_k \mathbf{a}_k' \mathbf{M} \mathbf{a}_\ell = \mathbf{c}_\ell = \mathbf{Y} \mathbf{M} \mathbf{a}_\ell.$$

Comme $\mathbf{M}$ est inversible et que les $\mathbf{a}_k$ forment une base, on obtient $\mathbf{Y}$.

Approximation si on prend les k premiers termes seulement, on obtient la meilleure approximation de $\mathbf{Y}$ par une matrice de rang k au sens des moindres carrés (théorème de Eckart-Young).

L'ACP sur les données centrées réduites

Matrice de variance-covariance c'est la matrice de corrélation car

$$\mathbf{Z}'\mathbf{D}\mathbf{Z} = \mathbf{D}_{1/\mathrm{s}}\mathbf{Y}'\mathbf{D}\mathbf{Y}\mathbf{D}_{1/\mathrm{s}} = \mathbf{D}_{1/\mathrm{s}}\mathbf{V}\mathbf{D}_{1/\mathrm{s}} = \mathbf{R}.$$

Métrique on prend la métrique $\mathbf{M} = \mathbf{I}_p$.

Facteurs principaux ce sont les p vecteurs propres orthonormés de $\mathbf{R}$,

$$\mathbf{R}\mathbf{u}_k = \lambda_k \mathbf{u}_k, \text{ avec } \langle \mathbf{u}_k, \mathbf{u}_\ell \rangle = 1 \text{ si } k = \ell, \, 0 \text{ sinon.}$$

dont les valeurs propres sont classés par valeur propre décroissante

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0$$

Composantes principales elles sont données par $\mathbf{c}_k = \mathbf{Z}\mathbf{u}_k$.

Nombre d'axes à retenir

Dimension de l'espace des individus L'ACP visant à réduire la dimension de l'espace des individus, on veut conserver aussi peu d'axes que possible. Il faut pour cela que les variables d'origine soient raisonnablement corrélées entre elles.

Les seuls critères utilisables sont empiriques.

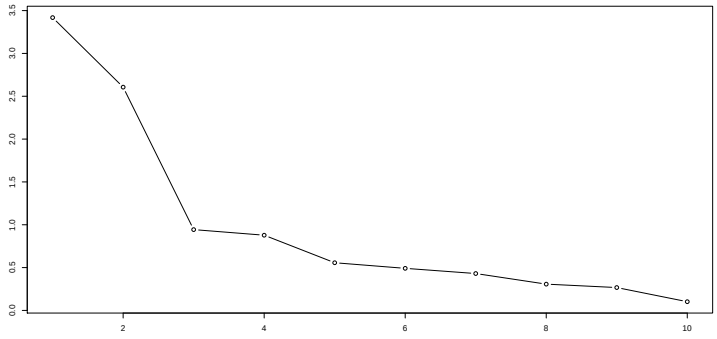
Interprétation des axes on s'efforce de ne retenir que des axes à propos desquels une forme d'interprétation est possible (soit directement, soit en terme des variables avec lesquels ils sont très corrélés). On donnera des outils à cet effet plus loin dans le cours.

Critère de Kaiser (variables centrées-réduites) on ne retient que les axes associés à des valeurs propres supérieures 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 (suite)

Éboulis des valeurs propres on cherche un « coude » dans le graphe des valeurs propres



L'espace des variables

Métrique $\mathbf{D}$ il faut munir l'espace des variables d'une métrique raisonnable. On choisit toujours la métrique $\mathbf{D}$ des poids :

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{D}} = \mathbf{x}'\mathbf{D}\mathbf{y}, \qquad \|\mathbf{x}\|^2_{\mathbf{D}} = \mathbf{x}'\mathbf{D}\mathbf{x}.$$

Interprétation pour deux variables *centrées* $\mathbf{x}$ et $\mathbf{y}$, on a

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{D}}, \qquad V(\mathbf{x}) = \|\mathbf{x}\|^2_{\mathbf{D}},$$
$$\operatorname{cor}(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{D}}}{\|\mathbf{x}\|_{\mathbf{D}} \|\mathbf{y}\|_{\mathbf{D}}} = \cos(\widehat{\mathbf{x}\mathbf{y}}).$$

Exemple les vecteurs $\mathbf{c}_k/\sqrt{\lambda_k}$ forment une base $\mathbf{D}$-orthonormale

$$\left\langle \frac{\mathbf{c}_k}{\sqrt{\lambda_k}}, \frac{\mathbf{c}_\ell}{\sqrt{\lambda_\ell}} \right\rangle_{\mathbf{D}} = \operatorname{cor}(\mathbf{c}_k, \mathbf{c}_\ell) = \begin{cases} 1, & \text{si } k = \ell, \\ 0, & \text{sinon.} \end{cases}$$

Corrélation entre composantes et variables initiales

Quand on travaille sur les variables centrées-réduites, la corrélation entre une composante principale $\mathbf{c}_k$ et une variable $\mathbf{z}^j$ est

$$r(\mathbf{z}^j, \mathbf{c}_k) = \frac{\operatorname{cov}(\mathbf{z}^j, \mathbf{c}_k)}{\sqrt{V(\mathbf{c}_k)}} = \frac{(\mathbf{z}^j)'\mathbf{D}\mathbf{c}_k}{\sqrt{\lambda_k}}$$

et donc le vecteur des corrélations de $\mathbf{c}_k$ avec $\mathbf{Z}$ est

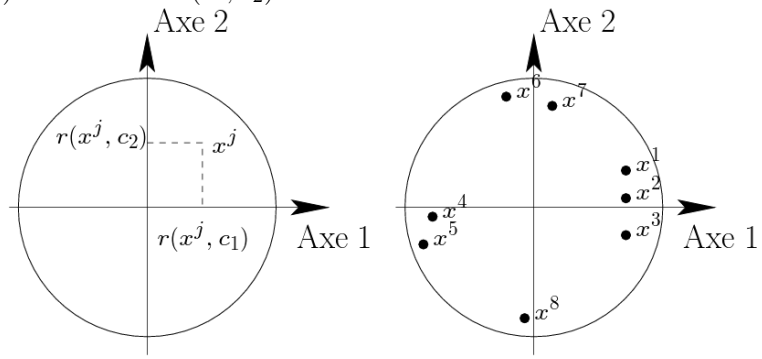
$$\mathbf{r}(\mathbf{Z}, \mathbf{c}_k) = (r(\mathbf{z}^1, \mathbf{c}_k), \ldots, r(\mathbf{z}^p, \mathbf{c}_k))' = \frac{\mathbf{Z}'\mathbf{D}\mathbf{c}_k}{\sqrt{\lambda_k}}.$$

Comme $\mathbf{Z}'\mathbf{D}\mathbf{c}_k = \mathbf{Z}'\mathbf{D}\mathbf{Z}\mathbf{u}_k = \mathbf{R}\mathbf{u}_k = \lambda_k \mathbf{u}_k$, on a finalement

$$\mathbf{r}(\mathbf{Z}, \mathbf{c}_k) = \sqrt{\lambda_k}\mathbf{u}_k.$$

Le cercle des corrélations

Qu'est-ce que c'est ? c'est une représentation où, pour deux composantes principales, par exemple $\mathbf{c}_1$ et $\mathbf{c}_2$, on représente chaque variable $\mathbf{z}^j$ par un point d'abscisse $r(\mathbf{z}^j, \mathbf{c}_1)$ et d'ordonnée $r(\mathbf{z}^j, \mathbf{c}_2)$.



Effet « taille » cela arrive quand toutes les variables sont corrélées positivement avec la première composante principale. Cette composante est alors appelée facteur de « taille », la seconde facteur de « forme ».

Le cercle des corrélations (suite)

Pourquoi un cercle ? comme les $\mathbf{c}_k/\sqrt{\lambda_k}$ forment une base $\mathbf{D}$-orthonormale,

$$\mathbf{z}^j = \sum_{k=1}^{p} \left\langle \frac{\mathbf{c}_k}{\sqrt{\lambda_k}}, \mathbf{z}^j \right\rangle_{\mathbf{D}} \frac{\mathbf{c}_k}{\sqrt{\lambda_k}} = \sum_{i=1}^{p} r(\mathbf{c}_k, \mathbf{z}^j) \frac{\mathbf{c}_k}{\sqrt{\lambda_k}}$$

et donc

$$\|\mathbf{z}^j\|^2_{\mathbf{D}} = 1 = \sum_{k=1}^{p} r^2(\mathbf{c}_k, \mathbf{z}^j).$$

Les points sont bien à l'intérieur d'un cercle de rayon 1.

Interprétation

- les points sont la projection orthogonale dans $\mathbf{D}$ des variables dans le plan défini par les composantes principales $\mathbf{c}_1$ et $\mathbf{c}_2$.
- Il ne faut interpréter la proximité des points que s'ils sont proches de la circonférence.

Contribution d'un individu à une composante

Définition On sait que $V(\mathbf{c}_k) = \lambda_k = \sum_{i=1}^{n} p_i c_{ik}^2$. La contribution de l'individu i à la composante k est donc

$$\frac{p_i c_{ik}^2}{\lambda_k}$$

Interprétation la contribution d'un individu est importante si elle excède d'un facteur α le poids p_i de l'individu concerné, c'est-à-dire

$$\frac{p_i c_{ik}^2}{\lambda_k} \geq \alpha p_i,$$

ou de manière équivalente

$$|c_{ik}| \geq \sqrt{\alpha \lambda_k}$$

Choix de α selon les données, on se fixe en général une valeur de l'ordre de 2 à 4, que l'on garde pour *tous* les axes

Individus sur-représentés

Qu'est-ce que c'est ? c'est un individu qui joue un rôle trop fort dans la définition d'un axe, par exemple

$$\frac{p_i c_{ik}^2}{\lambda_k} > 0,25$$

Effet il « tire à lui » l'axe k et risque de perturber les représentations des autres points sur les axes de rang $\geq k$. Il est donc surtout problématique sur les premiers axes. Un tel individu peut être le signe de données erronées.

Solution on peut le retirer de l'analyse et le mettre en « individu supplémentaire ».

Qualité globale de la représentation

Calcul de l'inertie on se souvient que $I_{\mathrm{g}} = \mathrm{Tr}(\mathbf{VM})$; comme la trace d'une matrice est la somme de ses valeurs propres, on a

$$I_{\mathrm{g}} = \lambda_1 + \lambda_2 + \cdots + \lambda_p.$$

Définition la qualité de la représentation obtenue par k valeurs propres est la proportion de l'inertie expliquée

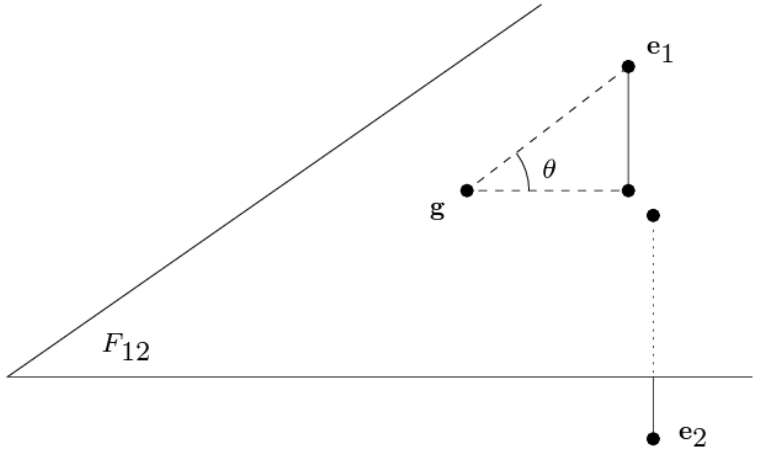
$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Si par exemple $\lambda_1 + \lambda_2$ est égal 90% de I_{g}, on en déduit que le nuage de points est aplati autour du premier plan principal.

Utilisation cette valeur sert seulement à évaluer la projection retenue, pas à choisir le nombre d'axes à garder.

Qualité locale de la représentation

But on cherche à déterminer si le nuage de points est très aplati par la projection sur les sous-espaces principaux. Dans ce cas, deux individus éloignés pourraient artificiellement sembler proches les uns des autres.



Angle entre un individu et un axe principal

Il est défini par son cosinus carré. Le cosinus de l'angle entre l'individu centré i et l'axe principal k est

$$\cos(\widehat{\mathbf{e}_i, \mathbf{a}_k}) = \frac{\langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k \rangle_{\mathbf{M}}}{\|\mathbf{e}_i - \mathbf{g}\|_{\mathbf{M}}}.$$

car les $\mathbf{a}_k$ forment une base orthonormale. Comme $\langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k \rangle_{\mathbf{M}} = c_{ik}$,

$$\cos^2(\widehat{\mathbf{e}_i, \mathbf{a}_k}) = \frac{c_{ik}^2}{\sum_{k=1}^{p} c_{ik}^2}.$$

Cette grandeur mesure la qualité de la représentation de l'individu i sur l'axe principal $\mathbf{a}_j$.

Angle entre un individu et un sous-espace principal

C'est l'angle entre l'individu et sa projection orthogonale sur le sous-espace. La projection de $\mathbf{e}_i - \mathbf{g}$ sur le sous-espace F_q, $q \leq p$, est $\sum_{k=1}^{q} c_{ik}\mathbf{a}_k$, et donc

$$\cos^2(\widehat{\mathbf{e}_i, F_q}) = \frac{\sum_{k=1}^{q} c_{ik}^2}{\sum_{k=1}^{p} c_{ik}^2}.$$

La qualité de la représentation de l'individu i sur le plan F_q est donc la somme des qualités de représentation sur les axes formant F_q. Il est significatif quand le point $\mathbf{e}_i$ n'est pas trop près de $\mathbf{g}$.

Critères Un $\cos^2$ égal à $0,9$ correspond à un angle de 18 degrés. Par contre, une valeur de $0,5$ correspond à un angle de 45 degrés ! On peut considérer par exemple les valeurs supérieures à $0,80$ comme correctes.

Variables supplémentaires quantitatives

Motivation les composantes principales étant définies pour maximiser les contributions, le fait que les corrélations obtenues soient proches de 1 peut ne pas être significatif. Par contre, une corrélation forte entre une composante principale et une variable n'ayant pas participé à l'analyse est très significative.

Méthode on « met de coté » certaines variables pour qu'elles ne soient pas utilisées dans l'analyse (on diminue donc la dimension de $\mathbf{R}$ en enlevant des lignes et des colonnes). On cherche ensuite à savoir si elles sont liées à un axe donné.

Corrélation on calcule la corrélation de la variable avec les composantes principales et on la place dans le cercle des corrélations. Si $\hat{\mathbf{z}}$ est le vecteur centré-réduit correspondant à cette variable, on calcule

$$\operatorname{cor}(\hat{\mathbf{z}}, \mathbf{c}_k) = \frac{\operatorname{cov}(\hat{\mathbf{z}}, \mathbf{c}_k)}{\sqrt{V(\mathbf{c}_k)}} = \frac{\langle \hat{\mathbf{z}}, \mathbf{c}_k \rangle_{\mathbf{D}}}{\sqrt{\lambda_k}} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} p_i \hat{z}_i c_{ik}.$$

On peut éventuellement utiliser un test statistique pour déterminer si une corrélation est significative.

Variables supplémentaires qualitatives

Représentation on peut représenter par des symboles différents les individus de chaque catégorie sur les axes principaux.

Valeur-test on considère les $\hat{n}$ individus ayant une certaine caractéristique (homme, femme...) et la coordonnée $\hat{c}_k$ de leur barycentre sur la k-ième composante principale. La valeur-test est

$$\hat{c}_k \sqrt{\frac{\hat{n}}{\lambda_k}} \sqrt{\frac{n-1}{n-\hat{n}}}.$$

Quand $\hat{n}$ est assez grand, elle est significative si sa valeur absolue est supérieure à 2 ou 3.

Idée du calcul Si les $\hat{n}$ individus étaient pris au hasard, $\hat{c}_k$ serait une variable aléatoire centrée (les $\mathbf{z}$ sont de moyenne nulle) et de variance $\frac{\lambda_k}{\hat{n}} \frac{n-\hat{n}}{n-1}$ car le tirage est sans remise.

Individus supplémentaires

Méthode on « met de coté » certains individus pour qu'ils ne soient pas utilisées dans l'analyse (ils ne sont pas pris en compte dans le calcul des covariances). On cherche ensuite à savoir si ils sont liés à un axe donné.

Cas des individus sur-représentés on peut décider d'utiliser ces points en individus supplémentaires, en particulier quand les points constituent un échantillon et ne présentent pas d'intérêt en eux-mêmes.

Représentation on les ajoute à la représentation sur les plans principaux. Pour calculer leur coordonnée sur un axe fixé, on écrit

$$\hat{c}_k = \langle \hat{\mathbf{z}}, \mathbf{u}_k \rangle = \sum_{j=1}^{p} \hat{z}^j u_{kj},$$

où les $\hat{z}^j$ sont les coordonnées centrées-réduites d'un individu supplémentaire $\hat{\mathbf{z}}$.

Ces individus peuvent servir d'échantillon-test pour vérifier les hypothèses tirées de l'ACP sur les individus actifs.

L'ACP en trois transparents (1)

Données les données représentent les valeurs de p variables mesurées sur n individus ; les individus peuvent avoir un poids. En général (et dans ce résumé), on travaille sur des données centrées réduites $\mathbf{Z}$ (on retranche la moyenne et on divise par l'écart type).

Matrice de corrélation c'est la matrice $\mathbf{R}$ de variance-covariance des variables centrées réduites. Elle possède p valeurs propres $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.

Inertie totale c'est la moitié de la moyenne des distances au carré entre les individus ; elle mesure l'étendue du nuage de points. C'est la grandeur qu'on cherche à garder maximale et elle peut s'écrire

$$I_{\mathbf{g}} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = p.$$

Facteurs principaux $\mathbf{u}_k$ ce sont des vecteurs propres orthonormés de $\mathbf{R}$ associés aux λ_k : $\mathbf{R}\mathbf{u}_k = \lambda_k \mathbf{u}_k$. Leur j-ième composante (sur p) u_{kj} est le poids de la variable j dans la composante k.

Composantes principales $\mathbf{c}_k$ ce sont les vecteurs $\mathbf{Z}\mathbf{u}_k$ de dimension n. Leur i-ième coordonnée c_{ik} est la valeur de la composante k pour l'individu i. Les $\mathbf{c}_k$ sont décorrélées et leur variance est $V(\mathbf{c}_k) = \lambda_k$.

L'ACP en trois transparents (2)

Nombre d'axes on se contente en général de garder les axes *interprétables* de valeur propre supérieure à 1 (critère de Kaiser).

Cercle des corrélations il permet de visualiser comment les variables sont corrélées (positivement ou négativement) avec les composantes principales. À partir de là, on peut soit trouver une signification physique à chaque composante, soit montrer que les composantes séparent les variables en paquets.

Représentation des individus pour un plan principal donné, la représentation des projections des individus permet de confirmer l'interprétation des variables. On peut aussi visualiser les individus aberrants (erreur de donnée ou individu atypique).

Contribution d'un individu à une composante c'est la part de la variance d'une composante principale qui provient d'un individu donné. Si cette contribution est supérieur de 2 à 4 fois au à son poids, l'individu définit la composante. Si elle est très supérieure aux autres, on dit qu'il est *sur-représenté* et on peut avoir intérêt à mettre l'individu en donnée supplémentaire.

L'ACP en trois transparents (3)

Qualité globale de la représentation c'est la part de l'inertie totale $I_{\mathbf{g}}$ qui est expliquée par les axes principaux qui ont été retenus. Elle permet de mesurer la précision et la pertinence de l'ACP.

Qualité de la représentation d'un individu elle permet de vérifier que tous les individus sont bien représentés par le sous-espace principal choisi ; elle s'exprime comme le carré du cosinus de l'angle entre l'individu et sa projection orthogonale.

Individus supplémentaires quand un individu est sur-représenté sur un des premiers axes, on peut le supprimer de l'analyse et le réintroduire dans la représentation comme individu supplémentaire.

Variables supplémentaires quantitatives certaines variables peuvent être mises de coté lors de l'ACP et reportées séparément sur le cercle des corrélation.

Variables supplémentaires qualitatives elles peuvent être représentées sur la projection des individus, et leur liaison aux axes est donnée par les valeurs-test.