Analyse statistique multivariée

STAT-S-401

Professeure: Catherine Dehon
Bâtiment R42 - bureau R42.6.204
e-mail: cdehon@ulb.ac.be

Assistantes: Afrae Hassouni et
Joséa Ronrotiana Rasoafaraniaina

Master en ingénieur de gestion
(UE du bloc 1)

Année 2018-2019

AVERTISSEMENT

Ce syllabus a été rédigé dans le but de faciliter la prise de notes pendant le cours théorique.
La mise à jour du présent syllabus sera faite via le cours théorique et l'université virtuelle (UV).
Il est bien entendu que l'examen portera sur l'ensemble de la matière vue au cours théorique (des éléments pourraient être ajoutés oralement au cours) ainsi que la matière des travaux pratiques.

A savoir ....

• Buts du cours

1. Traiter et décrire l'information contenue dans des grandes bases de données.
2. Choisir de manière appropriée les différentes méthodes d'analyse multivariée.
3. Interpréter correctement les graphiques et résultats fournis par un logiciel statistique (R).
4. Résoudre des problèmes avec données réelles.

Transformer des données en information en:

- réduisant la dimension et en simplifiant la structure
- groupant les individus et les variables
- analysant les dépendances.

$\Rightarrow$ Différentes méthodes permettant de visualiser et résumer l'information contenue dans des bases de données de grande dimension.

• Méthode d'enseignement et support

Théorie : Cours ex cathedra. Syllabus de théorie contenant la copie des transparents projetés (et commentés) au cours disponible sur l'UV.

Exercices: Les séances d'exercices sont organisées en salle informatique.

• Méthode d'évaluation

L'examen écrit est organisé durant la session de janvier. L'examen comporte une partie théorique et une partie pratique, sans interruption.

Une feuille A4 recto-verso écrite à la main est autorisée pendant l'examen.

Examen pratique optionnel, par groupe de 2, en salle informatique (décembre) sur 4 points.

PLAN DU COURS

Introduction à la détection des valeurs aberrantes

Utiliser des estimateurs et des distances robustes pour mettre en évidence les individus ayant des comportements différents de la majorité des données.

Analyse en composantes principales (ACP)

Technique de représentation et de réduction d'un ensemble de variables numériques.

Analyse des correspondances binaires (ACOBI)

Méthode adaptée à l'étude des éventuelles relations existantes entre 2 variables qualitatives (étude des tableaux de contingence).

Analyse des correspondances multiples (AFCM)

Généralisation de l'analyse des correspondances. Permet de décrire les relations entre plusieurs variables qualitatives.

Analyse des Corrélations Canoniques

Caractériser "au mieux" les relations linéaires entre 2 ensembles de variables quantitatives.

Analyse discriminante

Techniques destinées à classer (affecter à des classes préexistantes) des individus caractérisés par un certain nombre de variables.

Méthodes de classification

Méthodes permettant de grouper les individus ou variables suivant des critères de proximité.

REFERENCES

Casin, P. (1999), *Analyse des données et des panels de données*, De Boeck Université, Paris.

Dehon, C. , Droesbeke, J-J. et Vermandele C. (2014), *Eléments de statistique*, Bruxelles, Editions de L'Unviversité de Bruxelles.

Johnson, R.A., and Wichern, D.W. (1992), *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

Lebart, L., Morineau, A., Piron, M. (2000), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

McClave, J.T., Benson, P.G., and Sincich, T. (2001), *Statistics for Business and Economics*, Prentice Hall, New Jersey.

Tenenhaus, M. (1996), *Méthodes statistiques en gestion*, Dunod, Paris.

Chapitre 1

Background mathematics

1.1 Matrix calculus

A is a matrix with n line and p column :

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{kp} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nj} & \cdots & a_{np} \end{pmatrix} = (a_{ij})$$

where a_{ij} $(i \in \{1, ..., n\}; j \in \{1, ..., p\})$ gives the element line i and column j

It can be regarded as a point in $IR^{n \times p}$

A is called a square matrix if n=p

Transpose of a matrix

The transpose $\mathbf{A}'$ of an $n \times p$ matrix $\mathbf{A} = (a_{ij})$ is the $p \times n$ matrix whose ij-th element is a_{ji}

Example:

If $\mathbf{A} = \begin{pmatrix} 1 & 3 & -1 \\ 4 & 1 & 2 \end{pmatrix}$, then $\mathbf{A}' = \begin{pmatrix} 1 & 4 \\ 3 & 1 \\ -1 & 2 \end{pmatrix}$.

- It follows that:

$$(\mathbf{A}')' = \mathbf{A}$$

- The square matrix $\mathbf{A}_{K \times K}$ is **symmetric** if $\mathbf{A}' = \mathbf{A}$, it is to say that $a_{kl} = a_{lk} \forall k, l \in \{1, \ldots, K\}$.

Multiplication

The product of $\mathbf{A}$ and $\mathbf{B}$ is possible only if the number of columns of $\mathbf{A}$ is equal to the number of lines of $\mathbf{B}$. Then the product $\mathbf{A}_{K\times L} = (a_{kl})$ with $\mathbf{B}_{L\times H} = (b_{lh})$ is given by $\mathbf{C}_{K\times H} = (c_{kh})$ where

$$c_{kh} = \sum_{l=1}^{L} a_{kl} b_{lh} \qquad k = 1, \ldots, K; h = 1, \ldots, H.$$

• Properties: Let $\mathbf{A}_{m\times n}$, $\mathbf{B}_{n\times p}$, $\mathbf{C}_{p\times q}$, $\mathbf{D}_{n\times p}$, $\mathbf{E}_{n\times n}$ and $\mathbf{F}_{n\times n}$

$$\begin{aligned}
(\mathbf{AB})\mathbf{C} &= \mathbf{A}(\mathbf{BC}) \\
\mathbf{A}(\mathbf{B}+\mathbf{D}) &= \mathbf{AB}+\mathbf{AD} \\
(\mathbf{B}+\mathbf{D})\mathbf{C} &= \mathbf{BC}+\mathbf{DC} \\
\mathbf{EF} &\neq \mathbf{FE}
\end{aligned}$$

• The square matrix $\mathbf{A}_{K\times K}$ is **idempotent** if $\mathbf{A}^2 = \mathbf{A}$

• $\mathbf{A}_{K\times K}$ is **orthogonal** if $\mathbf{A}'\mathbf{A} = \mathbf{I}$

The rank of a matrix

Q vectors of same dimension $\mathbf{y}_1, \ldots, \mathbf{y}_Q$ are said to be linearly independent if

$$\sum_{q=1}^{Q} \alpha_q \mathbf{y}_q = \mathbf{0}$$

is verified only for $\alpha_1 = \alpha_2 = \ldots = \alpha_Q = 0$

Let $\mathbf{A}$ be an $n \times p$ matrix.

- The column rank is the maximum number of linearly independent columns.
- The row rank is the maximum number of linearly independent rows.
- The two ranks are equal and it is called the rank and denoted by: $r(\mathbf{A})$.

$$\Rightarrow r(\mathbf{A}) \leq min(n, p)$$

The determinant of $\mathbf{A}_{K\times K}$

The determinant of a squared matrix $\mathbf{A}_{K\times K}$ is a scalar, noted by $|\mathbf{A}|$, given by:

- $K=1$: if $\mathbf{A}=a$, then $|\mathbf{A}|=a$;
- $K=2$: if $\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, then $|\mathbf{A}| = a_{11}a_{22}-a_{21}a_{12}$;
- $K=3$: si $\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$, then

$$\begin{aligned} |\mathbf{A}| &= a_{11}a_{22}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32} \\ &\quad -a_{11}a_{23}a_{32}-a_{13}a_{22}a_{31}-a_{12}a_{21}a_{33}; \end{aligned}$$

- If $K>3$ then

$$|\mathbf{A}| = \sum_{l=1}^{K} a_{kl}A_{kl} \qquad k\in\{1,\ldots,K\}$$

where $A_{kl}=(-1)^{k+l}|\mathbf{M}_{kl}|$ with $\mathbf{M}_{kl}$ the squared sub-matrix of $\mathbf{A}$ without line k and column l

The trace of $\mathbf{A}_{K\times K}$

The trace of a square $K \times K$ matrix $\mathbf{A}$ is the sum of its diagonal elements:

$$tr(A) = \sum_{i=1}^{K} a_{ii}$$

Example:

$$A = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix} \Longrightarrow tr(A) = 3 + 2 = 5$$

- Properties: Let $\mathbf{A}_{m\times m}$, $\mathbf{B}_{m\times m}$

$$\begin{aligned} tr(\mathbf{A}+\mathbf{B}) &= tr(\mathbf{A}) + tr(\mathbf{B}) \\ tr(\lambda\mathbf{A}) &= \lambda tr(\mathbf{A}) \quad \lambda \text{ is a scalar} \\ tr(\mathbf{A}') &= tr(\mathbf{A}) \\ tr(\mathbf{AB}) &= tr(\mathbf{BA}) \end{aligned}$$

Quadratic forms

Let $\mathbf{x}$ be $K \times 1$ vector and $\mathbf{A}$ an $K \times K$ symmetric matrix, then the double sums of the form:

$$F(x_1, x_2, \ldots, x_K) = \sum_{i=1}^{K} \sum_{j=1}^{K} x_i x_j a_{ij} = \mathbf{x}'\mathbf{A}\mathbf{x}$$

can be written as this product of matrix, called a quadratic form in x:

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_K \end{pmatrix} \begin{pmatrix} a_{11} & \ldots & a_{1K} \\ a_{21} & \ldots & a_{2K} \\ \ldots & \ldots & \ldots \\ a_{K1} & \ldots & a_{KK} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_K \end{pmatrix}$$

We say that $\mathbf{A}$ is:

- positive definite if $\mathbf{x}'\mathbf{A}\mathbf{x} > 0 \quad \forall \mathbf{x} \neq 0$
- positive semidefinite if $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0 \quad \forall \mathbf{x} \neq 0$
- negative definite if $\mathbf{x}'\mathbf{A}\mathbf{x} < 0 \quad \forall \mathbf{x} \neq 0$
- negative semidefinite if $\mathbf{x}'\mathbf{A}\mathbf{x} \leq 0 \quad \forall \mathbf{x} \neq 0$

1.2 Geometric point of view in IR^P

Consider the column-vector

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \cdots & a_P \end{pmatrix}'.$$

Geometrically $\mathbf{a}$ can be represent in IR^P by line segment OA from the origin O to the point A with coordinate given by vector $\mathbf{a}$.

$OE_1, OE_2, \ldots, OE_p$ are the vectors defining IR^P associated with

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \ldots, \mathbf{e}_P = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Then for an observation A in IR^P with associated vector $\mathbf{a} = \begin{pmatrix} a_1 & a_2 & \cdots & a_P \end{pmatrix}'$

$$OA = a_1 OE_1 + a_2 OE_2 + \ldots + a_p OE_P$$

• The scalar product $< OA, OB >$ between two vectors is defined by :

$$\begin{aligned} < OA, OB > &= \mathbf{a}'\mathbf{b} = (a_1, \ldots, a_P)(b_1, \ldots, b_P)' \\ &= \sum_{p=1}^{P} a_p b_p \end{aligned}$$

• The euclidean norm $\|OA\|$ measures the length of the vector :

$$\|OA\|^2 = < OA, OA > = \mathbf{a}'\mathbf{a} = \sum_{p=1}^{P} a_p^2$$

A unit vector is a vector with unit length.

• The euclidean distance $d(A, B)$ between two points A and B is defined by:

$$\begin{aligned} d^2(A, B) &= \|AB\|^2 = \|OA - OB\|^2 \\ &= \sum_{p=1}^{P} (a_p - b_p)^2 \end{aligned}$$

$\Rightarrow d(O, A) = \|OA\|$

• The cosine of the angle between vectors OA and OB is defined by:

$$\cos(OA, OB) = \frac{< OA, OB >}{\|OA\| \|OB\|}$$

The vectors OA and OB are orthogonal iff

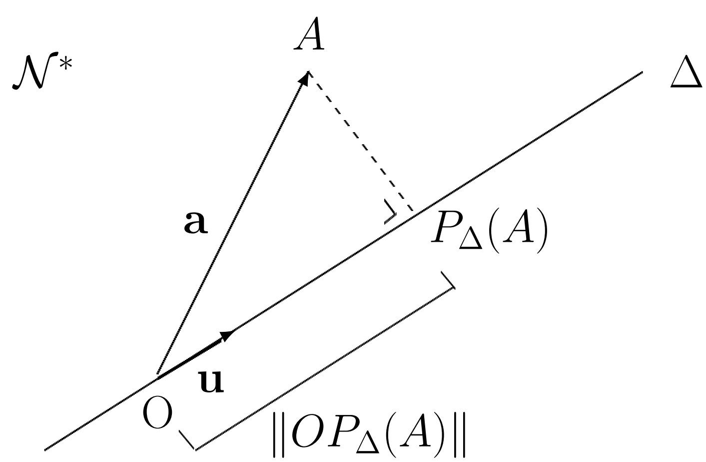
$$\cos(OA, OB) = \cos(\pm 90^\circ) = 0$$

It is to say iff

$$< OA, OB >= \mathbf{a}'\mathbf{b} = \sum_{p=1}^{P} a_p b_p = 0$$

1.2.1 Orthogonal projection in IR^1

Orthogonal projection of observation A in IR^P on the axis Δ that is passing through the origin:



The direction Δ is generated by the unit vector OU noted for simplicity by $\mathbf{u}$ with coordinates $\mathbf{u} = (u_1, \ldots, u_P)$.

The point $P_\Delta(A)$ is given by the orthogonal projection of A on the subspace Δ.

It is the nearest point on Δ to the point A. This means that u and $AP_\Delta(A)$ are orthogonal:

$$\cos(\alpha) = \frac{\|OP_\Delta(A)\|}{\|OA\|}$$

Moreover, since $\cos(\alpha) = \frac{<OA,u>}{\|OA\|}$, we obtain that:

$$\|OP_\Delta(A)\| =< OA, u >= \sum_{p=1}^{P} a_p u_p$$

1.2.2 Orthogonal projection in a subspace IR^H

• A normalized orthogonal system $u_1, \ldots, u_H$ is such that:

$$\begin{aligned} \|u_h\| &= 1 && \forall h \in \{1, \ldots, H\} \\ < u_h, u_l > &= 0 && \forall h \neq l \in \{1, \ldots, H\} \end{aligned}$$

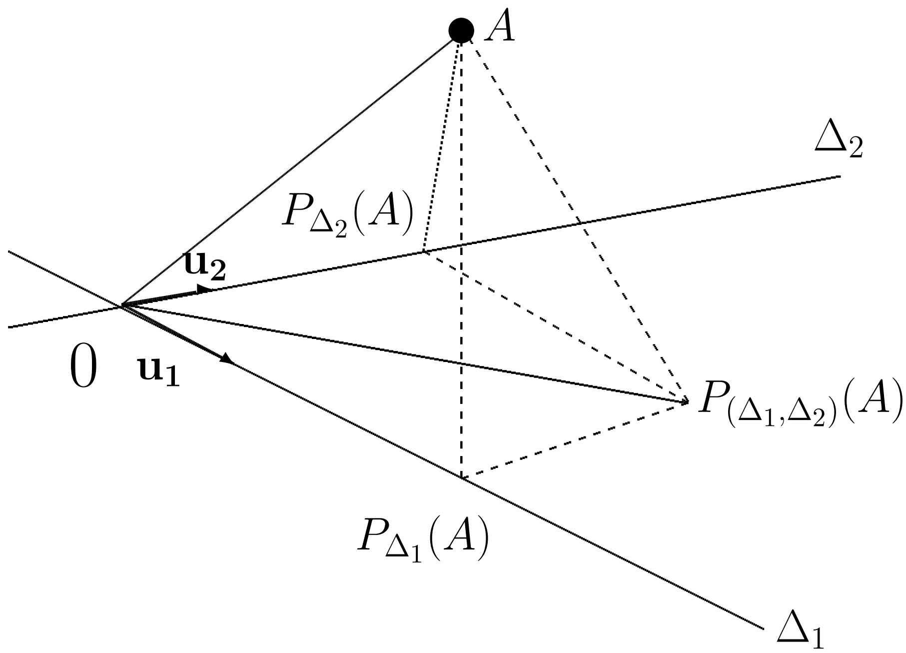
• These vectors generate a subspace of IR^P called L which is of dimension H. This subspace contains all the linear combinations:

$$\sum_{h=1}^{H} \alpha_h u_h$$

• The orthogonal projection of observation A in IR^P on the subspace L is given by $P_L(A) \in L$. Among all the points in the subspace L, this point is the closest to A. It is given by:

$$OP_L(A) = \sum_{h=1}^{H} < OA, u_h > u_h$$

$$\|OP_L(A)\|^2 = \sum_{h=1}^{H} < OA, u_h >^2$$



1.3 Eigenvalues and eigenvectors

Let

- $\mathbf{A}$ be a matrix of dimension $P \times P$
- $\mathbf{u}$ be a column vector of dimension $P \times 1$

• Transformation of space IR^P by $\mathbf{A}$:

$$\mathbf{A} : IR^P \longrightarrow IR^P : \mathbf{u} \longrightarrow \mathbf{A}\mathbf{u}$$

• $\mathbf{u}$ is an eigenvector (non null) of $\mathbf{A}$ associated with eigenvalue λ iff:

$$\begin{aligned} \mathbf{A}\mathbf{u} &= \lambda\mathbf{u} \\ \Rightarrow \mathbf{A}\mathbf{u} - \lambda\mathbf{u} &= 0 \\ \Rightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{u} &= 0 \end{aligned}$$

• λ is an eigenvalue of $\mathbf{A}$ iff

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

Comments:

- If $\mathbf{u}$ is an eigenvector of $\mathbf{A}$ associated with λ, then $\alpha\mathbf{u}$ ($\forall\alpha \in IR_0$) is also an eigenvector associated with same same eigenvalue
- The equation
$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$
can have no real solution. In this case, the transformation of IR^P by the matrix $\mathbf{A}$ has no fixed direction
- Each matrix $\mathbf{A}$ has at most P distinct eigenvalues
- If two real eigenvalues are the same $\Longrightarrow$ there exists a plane of eigenvectors
- Eigenvectors associated with distinct eigenvalues are linearly independent
- Let $\lambda_1, \ldots, \lambda_P$ be the eigenvalues of $\mathbf{A}$: $\sum_{p=1}^{P} \lambda_p = trace(\mathbf{A})$ et $\prod_{p=1}^{P} \lambda_p = det(\mathbf{A})$

Comments:

- A real symmetric matrix has only real eigenvalues
- A singular matrix has at least one eigenvalues zero
- A symmetric matrix is positive definite if and only if all its eigenvalues are positive
- A symmetric matrix is positive semidefinite if and only if all its eigenvalues are non-negative
- In practice, we take the eigenvectors $u_1, \ldots, u_P$ in order to have an orthonormal basis. Therefore, A can be written as follows:

$$A = \sum_{p=1}^{P} \lambda_p \underline{\mathrm{u}}_p \underline{\mathrm{u}}_p'$$

The particular case of the correlation matrix

The correlation matrix ($P \times P$) is given by

$$\mathbf{R} = \frac{1}{n}(X^*)'X^*$$

where X^* ($n \times P$) is the matrix of standardized data

- $\mathbf{R}$ is positive semidefinite:

$$\begin{aligned} \underline{x}'R\underline{x} &= \frac{1}{n}\underline{x}'(X^*)'X^*\underline{x} = \frac{1}{n}(X^*\underline{x})'X^*\underline{x} \\ &= \frac{1}{n}\|X^*\underline{x}\|^2 \geq 0 \quad \forall \underline{x} \neq 0 \end{aligned}$$

- $\mathbf{R}$ is positive definite iff the columns are linearly independent (the matrix X^* is of rank P)

- The number of non zero eigenvalues is equal to the rank of $\mathbf{R}$

1.4 Références

- Magnus, J.R., Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley Series in Probability and Statistics, England.

Chapitre 2

A short introduction on robust statistics

2.1 Why robust statistics ?

• Develop procedures (in estimation, in testing problem, in regression, in time series, . . .) that are valid (bias, efficiency) under small deviations from the underlying model

“All models are wrong, but some are useful.”
(Box, 1979)

$$\Downarrow$$

- Robustness: Find the structure fitting the majority of the data.

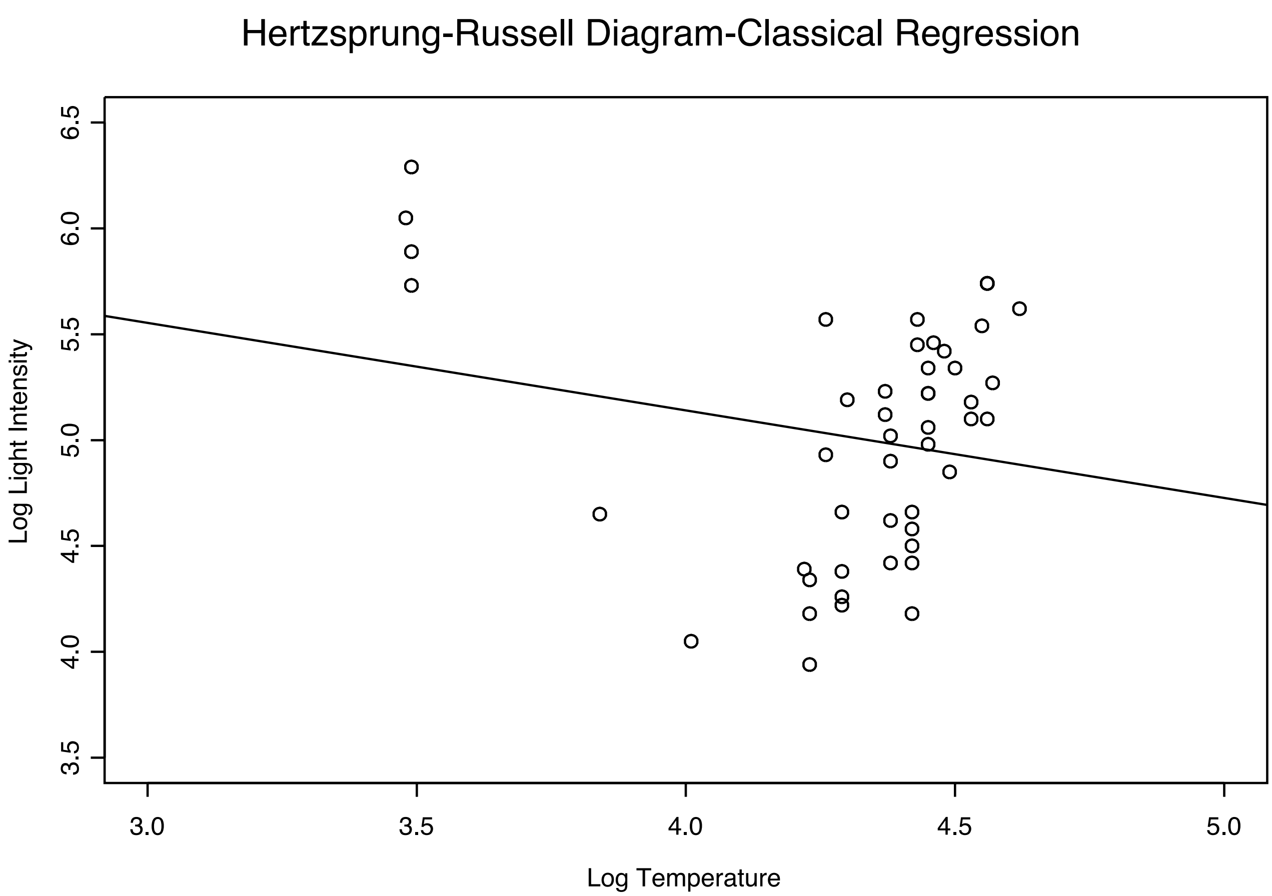
- Diagnostics: Identify outliers and sub-structure in the sample

- Robust methods are needed in explanatory analysis (data mining)

- Robust methods allows to control the weight of outliers (leverage points) in the statistical procedure

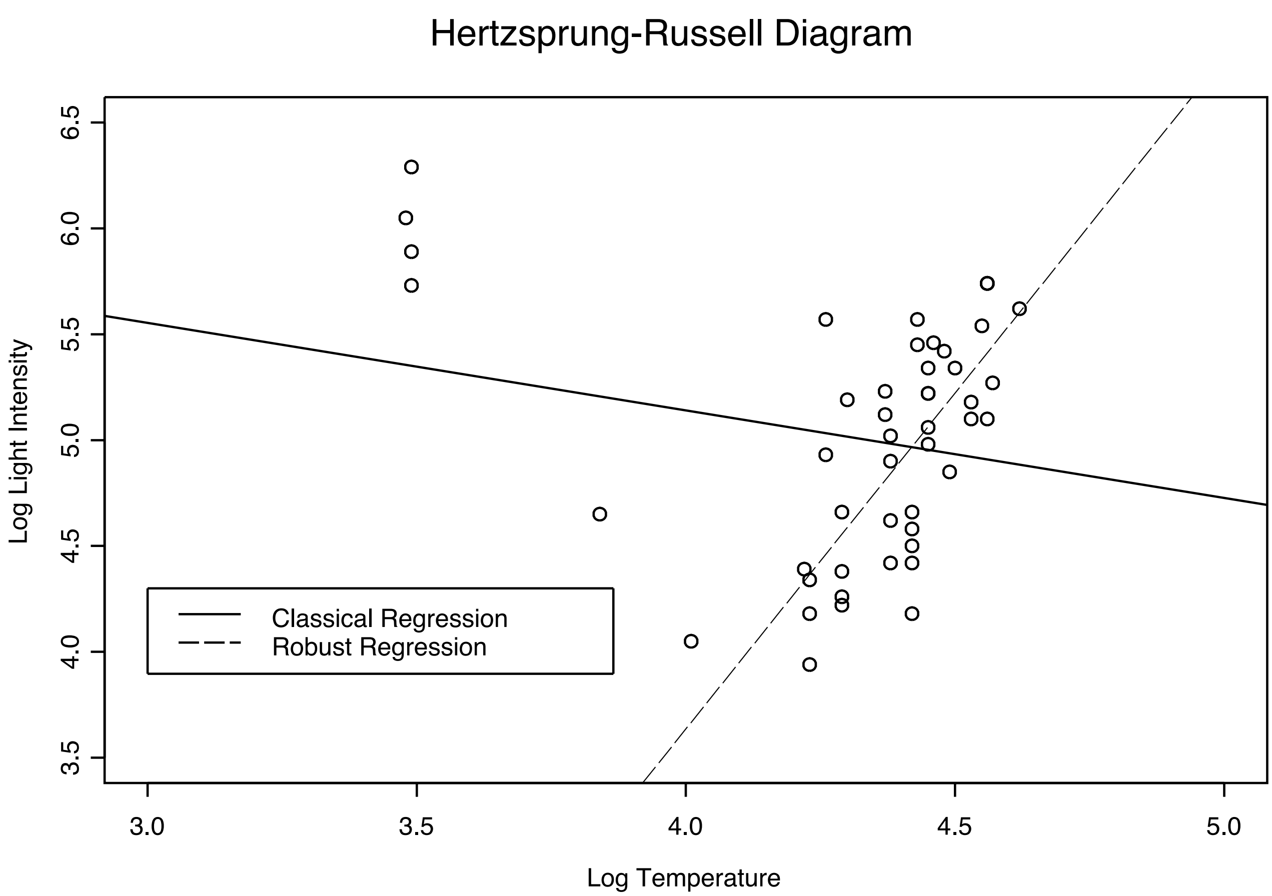
• Regression and Multivariate Analysis are used in many fields. But classical methods are very vulnerable to the presence of outliers

• Example of Simple Regression - Astronomy Data: 43 stars (the majority) are in the direction of Gygnus but 4 stars are called giants.



• Regression and Multivariate Analysis are used in many fields. But classical methods are very vulnerable to the presence of outliers

• Example of Simple Regression - Astronomy Data: 43 stars (the majority) are in the direction of Gygnus but 4 stars are called giants.



To perform the analysis:

• Inclusion of outliers using classical methods $\Rightarrow$ fallacious results

• Two-step procedure: Detection of outliers in the first step, and classical methods applied to the "clean sample" (exclusion of outliers) $\Rightarrow$ need detection of outliers

• Robust Methods:
1) Valid results for the majority of the data
2) Detection of outliers

Parametric, non-parametric and robust statistics

Robust statistics is an extension of parametric statistics: Statistics model: (χ, β, P)

Parametric hypothesis: $P \in \{P_\theta | \theta \in \Theta\}$

Non-parametric hypothesis: $P \in \{$ large family of distributions $\}$

Robust hypothesis P is "close" to one element of $\{P_\theta | \theta \in \Theta\}$

Important remarks

• Robust statistics doesn't replace classical one

• The two-step procedure, where classical methods are used in the second step after having deleted outliers, requires robust methods

• The word "robust" is used in various context, with different meaning.

New concept linked to robustness

The bias and the efficiency are well-known in statistics but robust statistics need new "measures":

- Influence function (IF): local stability
- Breakdown point: global validity
- Maxbias curve : a theoritical summary

Important: Trade-off between robustness and efficiency

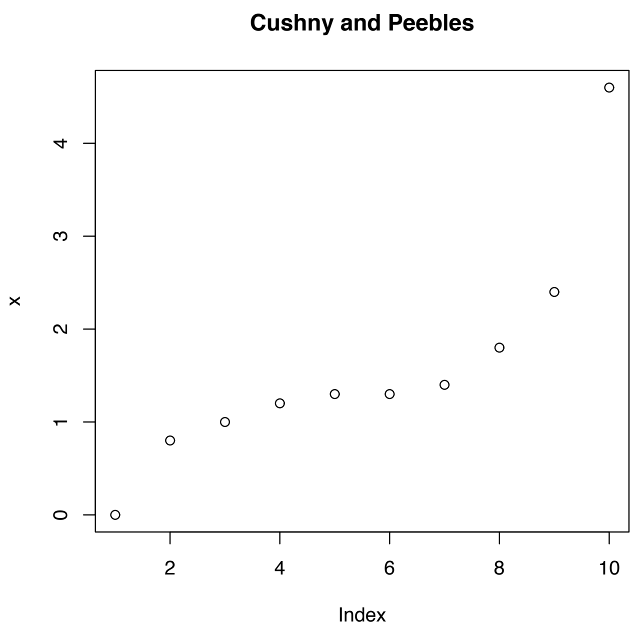
Example: Cushny and Peebles

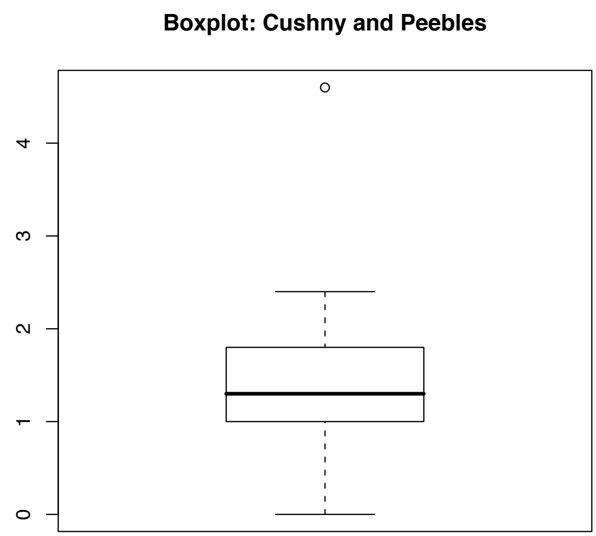
2.2 Detection

- Cushny and Peebles reported the results of a clinical trial of the effect of various drug on duration of sleep:

Sample: {0,0.8,1,1.2,1.3,1.3,1.4,1.8,2.4,4.6}

The last observation 4.6 seems to be outlier relatively to the other nine observation.





The rejection rule: The 3 σ rule

- If $X \sim N(\mu, \sigma^2)$, it is well known that:

$$P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.999$$

- Tchebyshev's rule (valid for all distribution):

  at least$(1 - \frac{1}{k^2})$of observations $\in (\mu \pm k\sigma)$

Example: if $k = 3$ at least 89% of observations $\in (\mu \pm 3\sigma)$

But μ and σ are unknown !!!!

Classical rule: an observation x_i is considered as an outliers if

$$x_i \notin (\bar{x} \pm 3s) = (-2.11; 5.27)$$

PROBLEM: MASKING EFFECT !!!!

The robust 3 σ rule

An observation x_i is considered as an outliers if

$$
\begin{aligned}
x_i &\notin [med(x) - 3MAD(x), med(x) + 3MAD(x)] \\
&\notin (-0.48, 3.08)
\end{aligned}
$$

A robust estimator of scale is given by the median absolute deviation MAD, which is the median of the n distances to the median:

$$MAD(x) = c\, med(|x_i - med(x)|)$$

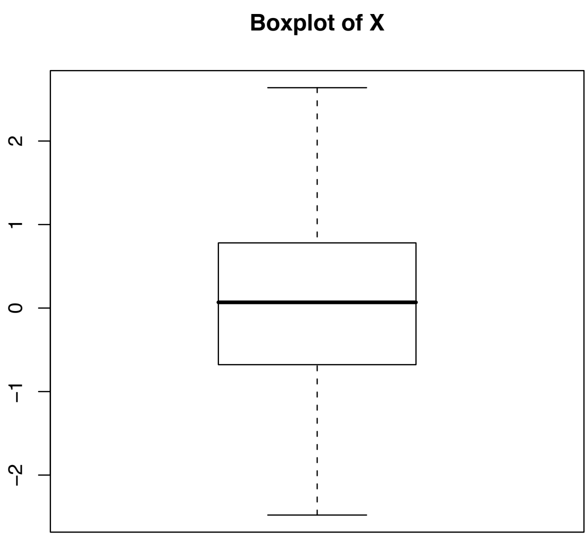
where $c = \frac{1}{\Phi^{-1}(3/4)}$ in order to obtain Fisher consistency at the normal distribution.

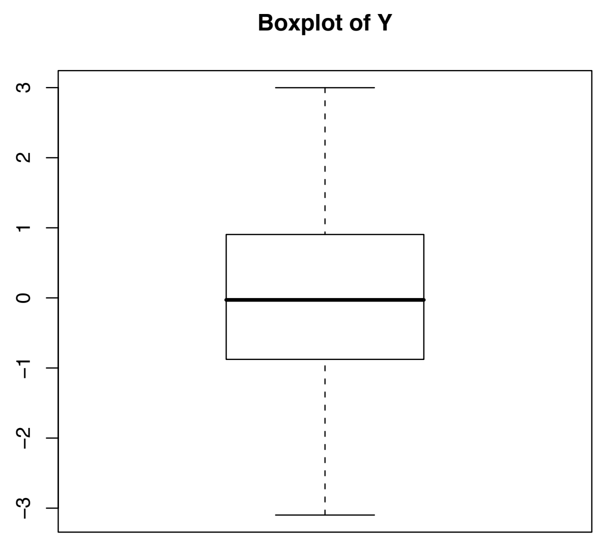
The rejection rule estimation is then given by:

$$\frac{0 + 0.8 + 1.0 + 1.2 + 1.3 + 1.4 + 1.8 + 2.4}{9} = 1.24$$

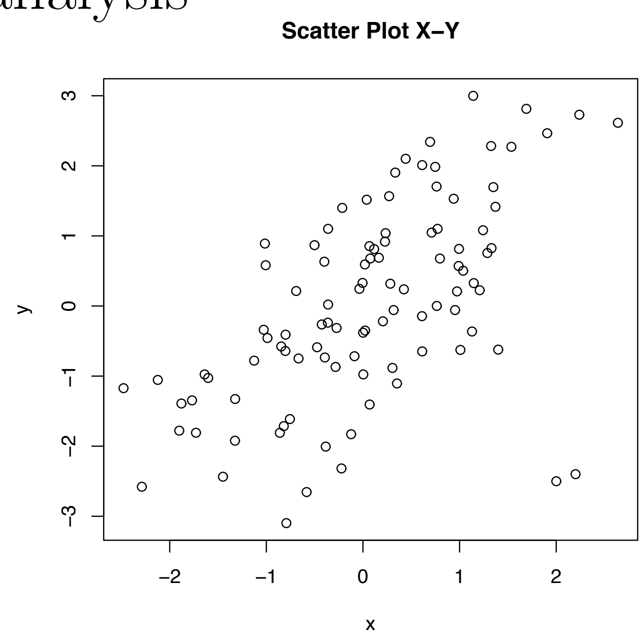
Bivariate simulated example

Univariate analysis





Bivariate analysis



Outliers in two-dimension space but not in in a single one dimensional space

Multivariate example

Stack loss (Rousseeuw & Leroy, 1987)

i	x_1	x_2	x_3	y	i	x_1	x_2	x_3	y
1	80	27	89	42	12	58	17	88	13
2	80	27	88	37	13	58	18	82	11
3	75	25	90	37	14	58	19	93	12
4	62	24	87	28	15	50	18	89	8
5	62	22	87	18	16	50	18	86	7
6	62	23	87	18	17	50	19	72	8
7	62	24	93	19	18	50	19	79	8
8	62	24	93	20	19	50	20	80	9
9	58	23	87	15	20	56	20	82	15
10	58	18	80	14	21	70	20	91	15
11	58	18	89	14					

x_1: air flow, x_2: cooling water inlet temperature, x_3: acide concentration

y: stack loss, defiend as the percentage of ingoing ammonia that escapes unabsorbed (response).

BUT: It is not possible to visualize all information in one figure

Mahalanobis distances

Let X be the matrix of data of dimension $n \times p$

Let x_i be the vector of dimension $p \times 1$

Classical Mahalanobis distances are defined by:

$$MD_i = \sqrt{((x_i - T(X))'C(X)^{-1}(x_i - T(X)))}$$

where $T(X)$ is the mean vector:

$$T(X) = \frac{1}{n}\sum x_i$$

and $C(X)$ is the empirical covariance matrix:

$$C(X) = \frac{1}{n}\sum((x_i - T(X))(x_i - T(X)))'$$

$T(X)$ and $C(X)$ are not robust

$$\Downarrow$$

MASKING EFFECT

Robust Multivariate estimators

Let b be a constant and A ($p \times p$) a non-singuliar matrix

Let $X = \{x_1, \ldots, x_n\}$,

$Y = \{x_1 + b, \ldots, x_n + b\} = X + b$,

$Z = AX + b$

Equivariance for the location estimator $T(X)$:

- Translation equivariant:$= T(Y) = T(X) + b$
- Affine equivariant:$= T(Z) = AT(X) + b$

Equivariance for the covariance estimator $C(X)$:

- Translation invariant:$= C(Y) = C(X)$
- Affine equivariant:$= C(Z) = A'C(X)A$

Generalization of the univariate median

The median is an univariate location estimator with $BDP = 50\%$ which is defined by the minimization problem:

$$med(x) = argmin_t \sum_{i=1}^{n} |x_i - t|$$

• First proposition: the L_1 estimator minimizes $\sum_{i=1}^{n} \|x_i - T\|$

Problem: not affin equivariant

• Second proposition: the coordinatewise median:

$$T = (med_i x_{i1}, \ldots, med_i x_{ip})$$

Problem: For $p \geq 3$ the coordinatewise median is not always in the convex hull of the sample

Several propositions of affine equivariant estimators

- Multivariate M-estimateurs (Maronna, 76)
- Convex Peeling (Barnett, 76; Bennington, 78)
- Ellipsoid Peeling (Titterington, 78; Hebling, 83)
- Iterative Trimming (Gnanadesikan and Kettering, 78)
- Generalized median (Oja, 83)
- ...

PROBLEM:
all these estimators have a BDP$\leq \frac{1}{p+1}$

$$\Downarrow$$

BDP decreases when the dimension increases !!!!

Stahel-Donoho estimator

Stahel (1981) and Donoho (1982) proposed the first affine equivariant estimators for which the BDP is of 50%.

It is based on the concept of outlyingness:

$$u_i = \sup_{\|v\|=1} \frac{|x_i v' - \text{median}_j(x_j v')|}{\text{median}_l \, |x_l v' - \text{median}_j(x_j v')|}$$

Reweighted classical estimators with weights given by $w(u_i)$:

$$T(x) = \frac{\sum_i w(u_i) x_i}{\sum_i w(u_i)}$$

$$C(x) = \frac{\sum_i w(u_i)(x_i - T(x))(x_i - T(x))'}{\sum_i w(u_i)}$$

Minimum Covariance Determinant (MCD)

Suppose that $p = 2$ for simplicity: $Z = (X, Y) \in IR^2$, with

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{bmatrix} \Longrightarrow \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The generalized variance defined as:

$$\det(\Sigma) = \sigma_X^2 \sigma_Y^2 - \sigma_{YX}^2$$

can be seen as a generalization of the variance.

$T(X)$: mean of the 50% points of X for which the determinant of the empirical covariance matrix is minimal;
$C(X)$: given by the same covariance matrix, multiplied by a factor to obtain consistency

Properties:

- affin equivariant
- BDP$= 50\%$
- asymptotic normality (Butler et Jhun, 1988)

S-estimators

• Classical estimators (t_n, C_n) can be obtained by minimizing $det(C)$ under the constraint:

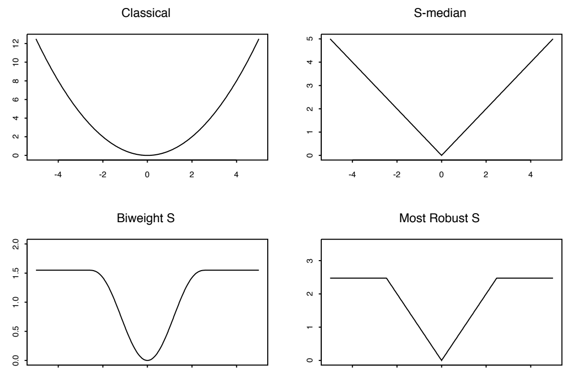
$$\frac{1}{n}\sum_{i=1}^{n}(\sqrt{(x_i - t)'C^{-1}(x_i - t)})^2 = p$$

$\forall (t, C) \in R^P \times PSD(p)$ where $PSD(p)$ is the set of all symmetric and positive definite matrix of dimension$(p \times p)$

• S-estimators (t_n, C_n) can be obtained by minimizing $det(C)$ under the constraint:

$$\frac{1}{n}\sum_{i=1}^{n}\rho(\sqrt{(x_i - t)'C^{-1}(x_i - t)}) \leq b$$

$\forall (t, C) \in R^P \times PSD(p)$



Robust distances

$$RD_i = \sqrt{((x_i - T(X))'C(X)^{-1}(x_i - T(X)))}$$

where $T(X)$ is a robust multivariate estimator of location and $C(X)$ is a robust estimator of the covariance matrix

Idea: Represent graphically the robust distances. Outliers can be detected by large distances.

How to find the cutoff ?? Suppose that

$$X \sim N_p(\mu, \Sigma), \text{ then}$$
$$\Sigma^{-1/2}(X - \mu) \sim N(0, I)$$

It follows that $((x_i - \mu)'\Sigma^{-1}(x_i - \mu))$ is the sum of p independent standardized normal squared

$$\Downarrow$$

$$((x_i - \mu)'\Sigma^{-1}(x_i - \mu)) \sim \chi^2_p$$

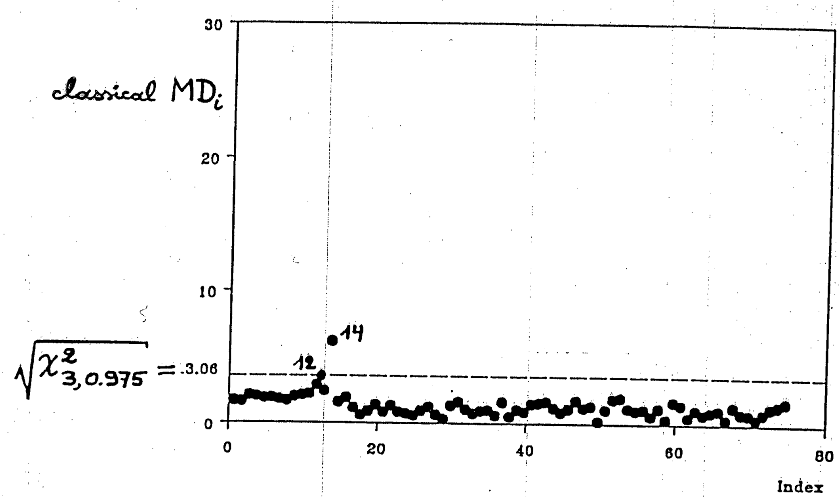
The cut-off will be then approximated by the squared root of the 0.975 quantile of the χ^2_p

13

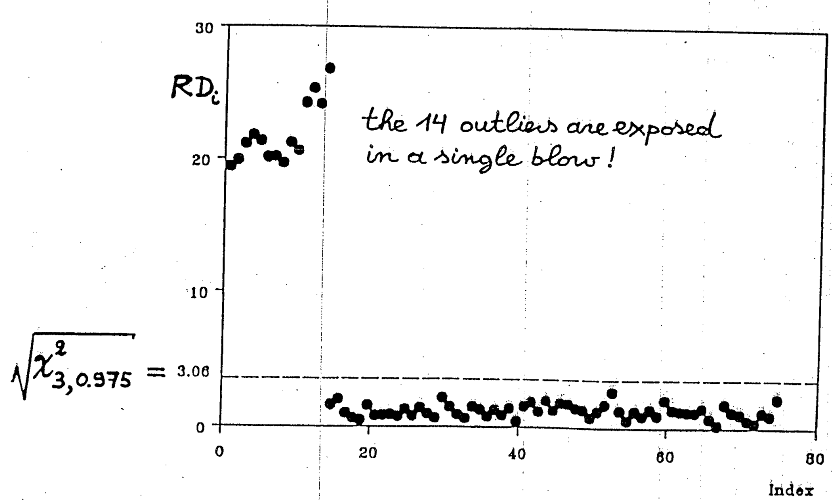
Data of Hawkins-Bradu-Kass (1984), $n = 75$, $p = 3$

Known that cases 1 to 14 are outliers!

(12 and 14 mask all others)



distances based on MVE:



2.2.1 References

Droesbeke, J-J , Saporta, G. , Thomas-Agnan, Ch. (2015), *Méthodes robustes en statistique*, Editions TECHNIP, Paris.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics*, John Wiley and Sons, NY.

Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.-P. (2009), *Robust Methods in Biostatistics*, Chichester, UK: John Wiley Sons.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley and Sons.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics*, John Wiley and Sons, NY.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outliers Detection*, John Wiley and Sons, NY.

Chapitre 3

Principal Component Analysis (PCA)

3.1 Introduction

- Basic tools **to reduce the dimension** of a multivariate data matrix
- Descriptive technique using **geometrical approach** to reduce the dimension

The output consists of:

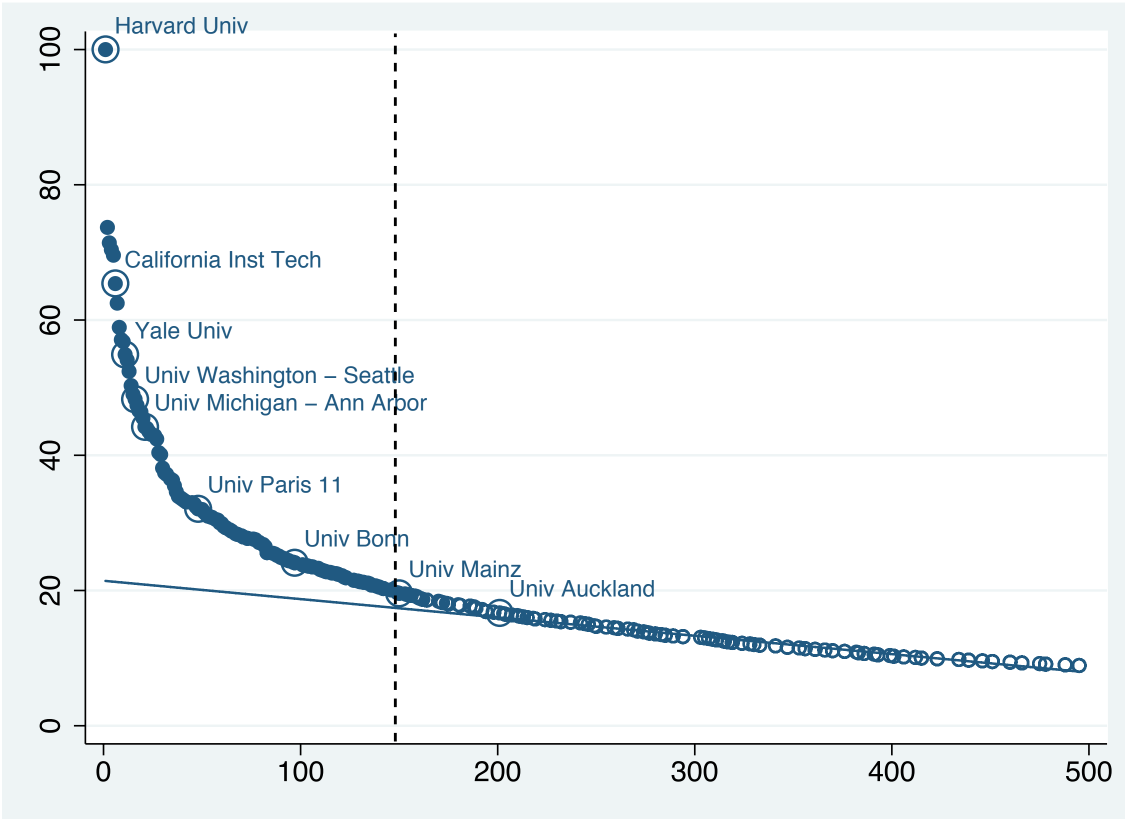
- graphical representation of individuals showing similarities and dissimilarities
- graphical representation of variables based on correlations

3.1.1 Example: Academic Ranking of World Universities (2007)

Question: Can a single "indicator" accurately sum up research excellence ?

- **Alumni (10%)**: Alumni recipients of the Nobel prize or the Fields Medal;
- **Award (20%)**: Current faculty Nobel laureates and Fields Medal winners;
- **HiCi (20%)**: Highly cited researchers in 21 broad subject categories;
- **N&S (20%)**: Articles published in Nature and Science;
- **PUB (20%)**: Articles in the Science Citation Index-expanded, and the Social Science Citation Index;
- **PCP (10%)**: The weighted score of the previous 5 indicators divided by the number of full-time academic staff members.

Case study on the TOP 50 (Overall score relative to rank)



Universités	Variables					
	Alumni	Award	HiCi	N&S	SCI	Size
1. Harvard Univ.	100	100	100	100	100	73
2. Stanford Univ.	42	78.7	86.1	69.6	70.3	65.7
3. Univ. California, Berkeley	72.5	77.1	67.9	72.9	69.2	52.6
4. Univ. Cambridge	93.6	91.5	54	58.2	65.4	65.1
5. Massachusetts Inst. Tech. (MIT)	74.6	80.6	65.9	68.4	61.7	53.4
6. California Inst. Tech.	55.5	69.1	58.4	67.6	50.3	100
7. Columbia Univ.	76	65.7	56.5	54.3	69.6	46.4
8. Princeton Univ.	62.3	80.4	59.3	42.9	46.5	58.9
9. Univ. Chicago	70.8	80.2	50.8	42.8	54.1	41.3
10. Univ. Oxford	60.3	57.9	46.3	52.3	65.4	44.7

Universités	**Variables**					
	Alumni	Award	HiCi	N&S	SCI	Size
11. Yale Univ.	50.9	43.6	57.9	57.2	63.2	48.9
12. Cornell Univ.	43.6	51.3	54.5	51.4	65.1	39.9
13. Univ. California, Los Angeles	25.6	42.8	57.4	49.1	75.9	35.5
14. Univ. California, San Diego	16.6	34	59.3	55.5	64.6	46.6
15. Univ. Pennsylvania	33.3	34.4	56.9	40.3	70.8	38.7
16. Univ. Washington, Seattle	27	31.8	52.4	49	74.1	27.4
17. Univ. Wisconsin, Madison	40.3	35.5	52.9	43.1	67.2	28.6
18. Univ. California, San Francisco	0	36.8	54	53.7	59.8	46.7
19. Johns Hopkins Univ.	48.1	27.8	41.3	50.9	67.9	24.7
20. Tokyo Univ.	33.8	14.1	41.9	52.7	80.9	34
21. Univ. Michigan, Ann Arbor	40.3	0	60.7	40.8	77.1	30.7
22. Kyoto Univ.	37.2	33.4	38.5	35.1	68.6	30.6
23. Imperial Coll. London	19.5	37.4	40.6	39.7	62.2	39.4
24. Univ. Toronto	26.3	19.3	39.2	37.7	77.6	44.4
25. Univ. Coll. London	28.8	32.2	38.5	42.9	63.2	33.8
26. Univ. Illinois, Urbana Champaign	39	36.6	44.5	36.4	57.6	26.2
27. Swiss Fed. Inst. Tech. - Zurich	37.7	36.3	35.5	39.9	38.4	50.5
28. Washington Univ., St. Louis	23.5	26	39.2	43.2	53.4	39.3
29. Northwestern Univ.	20.4	18.9	46.9	34.2	57	36.9
30. New York Univ.	35.8	24.5	41.3	34.4	53.9	25.9
31. Rockefeller Univ.	21.2	58.6	27.7	45.6	23.2	37.8
32. Duke Univ.	19.5	0	46.9	43.6	62	39.2
33. Univ. Minnesota, Twin Cities	33.8	0	48.6	35.9	67	23.5
34. Univ. Colorado, Boulder	15.6	30.8	39.9	38.8	45.7	30
35. Univ. California, Santa Barbara	0	35.3	42.6	36.2	42.7	35.1
36. Univ. British Columbia	19.5	18.9	31.4	31	63.1	36.3
37. Univ. Maryland, Coll. Park	24.3	20	40.6	31.2	53.3	25.9
38. Univ. Texas, Austin	20.4	16.7	46.9	28	54.8	21.3
39. Univ. Paris VI	38.4	23.6	23.4	27.2	54.2	33.5
40. Univ. Texas Southwestern Med. Center	22.8	33.2	30.6	35.5	38	31.9
41. Vanderbilt Univ.	19.5	29.6	31.4	23.8	51	36
42. Univ. Utrecht	28.8	20.9	27.7	29.9	56.6	26.6
43. Pennsylvania State Univ. - Univ. Park	13.2	0	45.1	37.7	58	23.7
44. Univ. California, Davis	0	0	46.9	33.1	64.2	30
45. Univ. California , Irvine	0	29.4	35.5	28	48.9	32.1
46. Univ. Copenhagen	28.8	24.2	25.7	25.2	51.4	31.7
47. Rutgers State Univ., New Brunswick	14.4	20	39.9	32.1	44.8	24.2
48. Univ. Manchester	25.6	18.9	24.6	28.3	56.9	28.4
49. Univ. Pittsburgh, Pittsburgh	23.5	0	39.9	23.6	65.6	28.5
50. Univ. Southern California	0	26.8	37.1	23.4	52.7	25.9

Univariate and bivariate analysis

The first step of all statistical analysis is the univariate and bivariate analysis

- Univariate statistics

Statistiques	Alumni (X_1)	Award (X_2)	HiCi (X_3)	N&S (X_4)	SCI (X_5)	Size (X_6)
Mean	34.09	36.10	46.62	43.09	60.10	38.63
Median	38.80	32	44.80	40.10	61.85	35.30
Min	0	0	23.40	23.40	23.20	21.30
Max	100	100	100	100	100	100
Variance	525.74	625.57	207.82	217.51	156.63	212.33

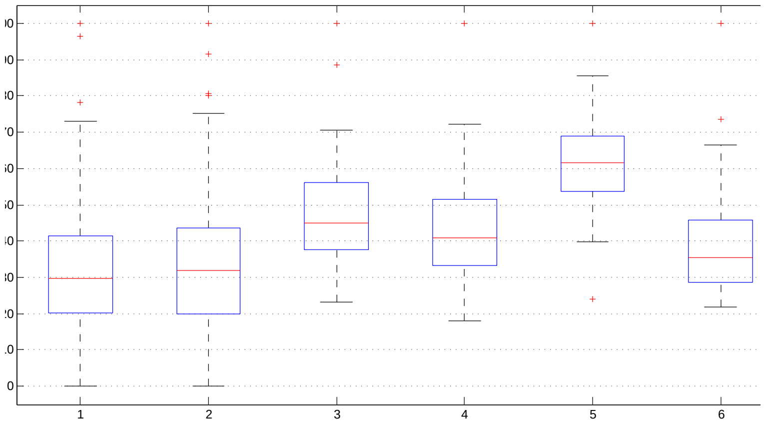
- Correlation matrix:

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.75 & 0.56 & 0.68 & 0.40 & 0.58 \\ 0.75 & 1.00 & 0.59 & 0.73 & 0.09 & 0.74 \\ 0.56 & 0.59 & 1.00 & 0.84 & 0.60 & 0.60 \\ 0.68 & 0.73 & 0.84 & 1.00 & 0.49 & 0.74 \\ 0.40 & 0.09 & 0.60 & 0.49 & 1.00 & 0.16 \\ 0.58 & 0.74 & 0.60 & 0.74 & 0.16 & 1.00 \end{pmatrix}.$$

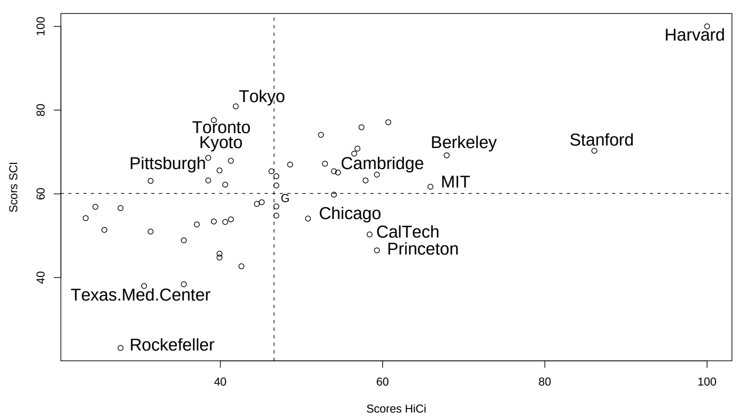
Variables are positively correlated $\rightarrow$ size factor

Graphics

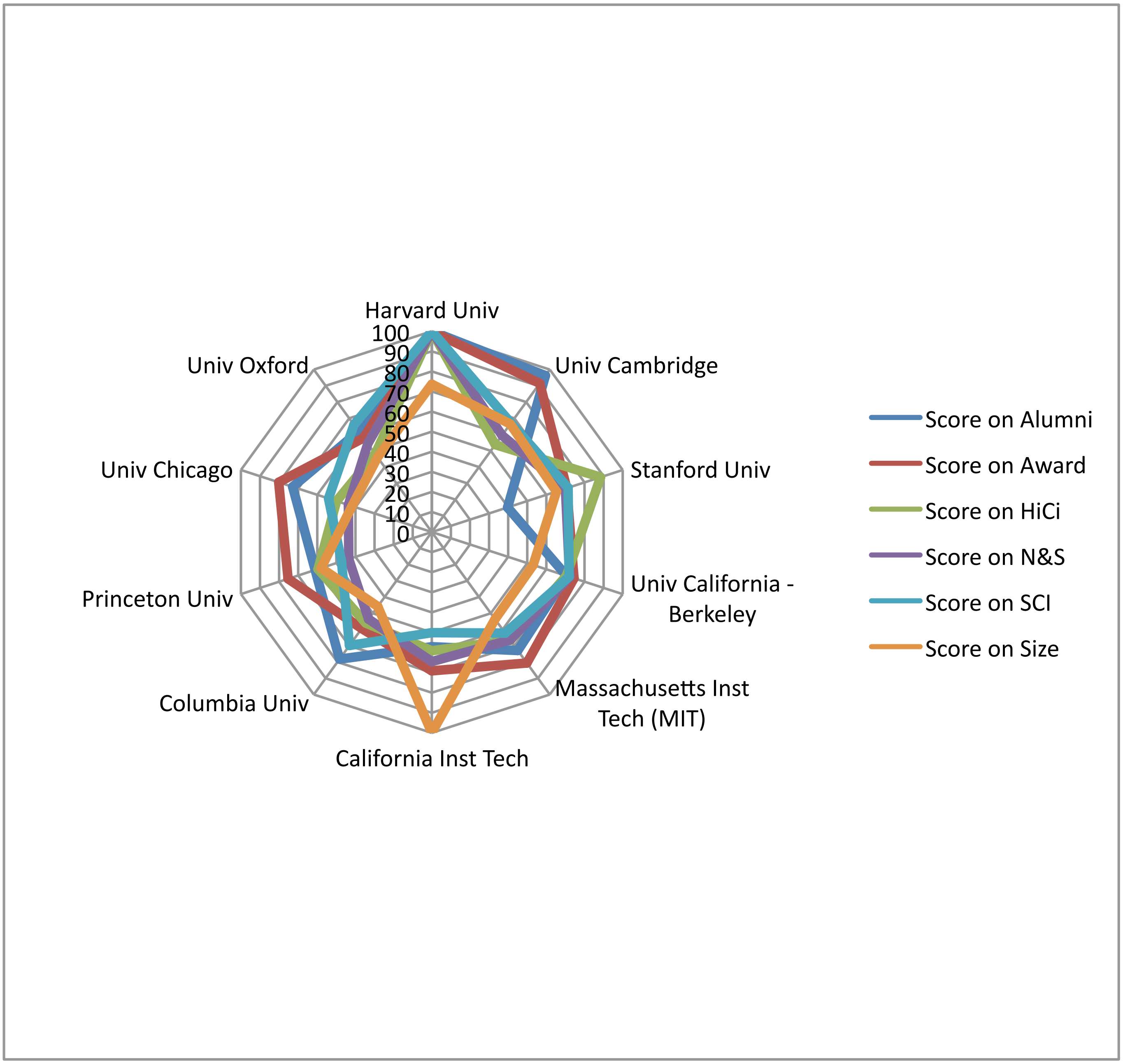
- Univariate graphs - Boxplot to detect outliers



- Scatterplots to detect bivariate structure



• Radar type of graph based on TOP 10 to detect multivariate structure



Visualization is not easy when the data contains a large number of individuals

3.1.2 The geometric point of view

Data matrix X $(n \times p)$ is composed of n observations (or individuals) and p variables.

	X_1	$\dots$	X_p	$\dots$	X_P	
1	x_{11}	$\dots$	x_{1p}	$\dots$	x_{1P}	$\longrightarrow \underline{\mathrm{x}}_1'$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
i	x_{i1}	$\dots$	x_{ip}	$\dots$	x_{iP}	$\longrightarrow \underline{\mathrm{x}}_i'$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
n	x_{n1}	$\dots$	x_{np}	$\dots$	x_{nP}	$\longrightarrow \underline{\mathrm{x}}_n'$
$Mean$	$\bar{x}_1$	$\dots$	$\bar{x}_p$	$\dots$	$\bar{x}_P$	
$Variance$	s_1^2	$\dots$	s_p^2	$\dots$	s_P^2	
	$\downarrow$		$\downarrow$		$\downarrow$	
	$\underline{\mathrm{v}}_1$	$\dots$	$\underline{\mathrm{v}}_p$	$\dots$	$\underline{\mathrm{v}}_P$	

Examples:

- ARWU scores of universities on research variables
- indicators of corruption on countries, …

• Cloud of n points in IR^P :

Proximity between two individuals (observations) reflects a similar behavior on the p variables

• Cloud of p points in IR^n :

Proximity between two variables reflects a similar behavior on the n individuals

BUT ... when n or/and p are large (larger than 2 or 3), we cannot produce interpretable graphs of these clouds of points

Develop methods to reduce the dimension without loosing too much information, the information about the variation and structure of clouds in both spaces

• Simplest way of dimension reduction:
Take just one variable - Not a very reasonable approach

• Alternative method:
Consider the simple average - All the element are considered with equal importance

• Other solution:
Use a weighted average with fixed weights - Choice of weight is arbitrary

Example: ARWU (2007)

- Take only the variable measuring the number of articles published in Nature and Science
- Summarize the 6 variables using the mean
- Use the weights proposed by the "rankers"

Question:

How to project the point cloud onto a space of lower dimension without loosing too much information?

How to construct new uncorrelated variables $\Phi_1, \Phi_2, \ldots, \Phi_M$ (where M is small) summarizing in the best way the structure of the initial point cloud ?

These new variables will be given as a weighted average, but how to choose the optimal weights?

The new variables will be called "*principal components*"

Several criteria exist in the literature to obtain "*principal components*":

- Inertia criteria (Pearson, 1901).

  This point of view is based on geometric approach facilitating the understanding and the interpretation of output.

  Moreover correspondence analysis for qualitative variables is a generalization of this method.

  This approach is extensively used in french textbooks and software

- Correlation and Variance criteria (Hotelling, 1931).

  Methods used in several english textbooks and software.

3.2 The geometric approach of Pearson

3.2.1 The n-dimensional point cloud

Each individual i denoted as I_i in IR^P is associated with vector $\underline{\mathrm{x}}_i = (x_{i1}, \ldots, x_{iP})'$

$\Longrightarrow$ Cloud of n points: $\aleph = \{I_1, \ldots, I_n\}$.

• Center of gravity G of $\aleph$:

$$\underline{\mathrm{g}} = (\bar{x}_1, \ldots, \bar{x}_P)'$$

In the example on ranking where the variables are Alumni, Award, HiCi, N&S, SCI and PCP, G characterize an university with mean profile :

$$\underline{\mathrm{g}} = (34.09, 36.10, 46.62, 43.09, 60.10, 38.63)'$$

• The total inertia is the dispersion of the cloud $\aleph$ around the gravity center G

$$\begin{aligned}
I(\aleph, G) &= \frac{1}{n}\sum_{i=1}^{n} d^2(I_i, G) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{p=1}^{P}(x_{ip} - \bar{x}_p)^2\right) \\
&= \sum_{p=1}^{P}\left(\frac{1}{n}\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)^2\right) \\
&= \sum_{p=1}^{P} s_p^2
\end{aligned}$$

$\Longrightarrow$ The total inertia is the sum of variances

For the ranking example:

$$\begin{aligned} I(\aleph, G) &= 525.7 + 625.6 + 207.8 \\ &\quad +217.5 + 156.6 + 212.3 \\ &= 1945.5 \end{aligned}$$

The largest part of the total inertia is due to the "Nobels" variables

$\Longrightarrow$ The choice of units has clearly an impact.

• Solution: Normalize the PCA

PCAn is independent of the choice of units because it uses the standardized variables:

$$x^*_{ip} = \frac{x_{ip} - \bar{x}_p}{s_p} \quad \forall i \in \{1, \ldots, n\}; p \in \{1, \ldots, P\}$$

Data matrix X^* of standardized observations

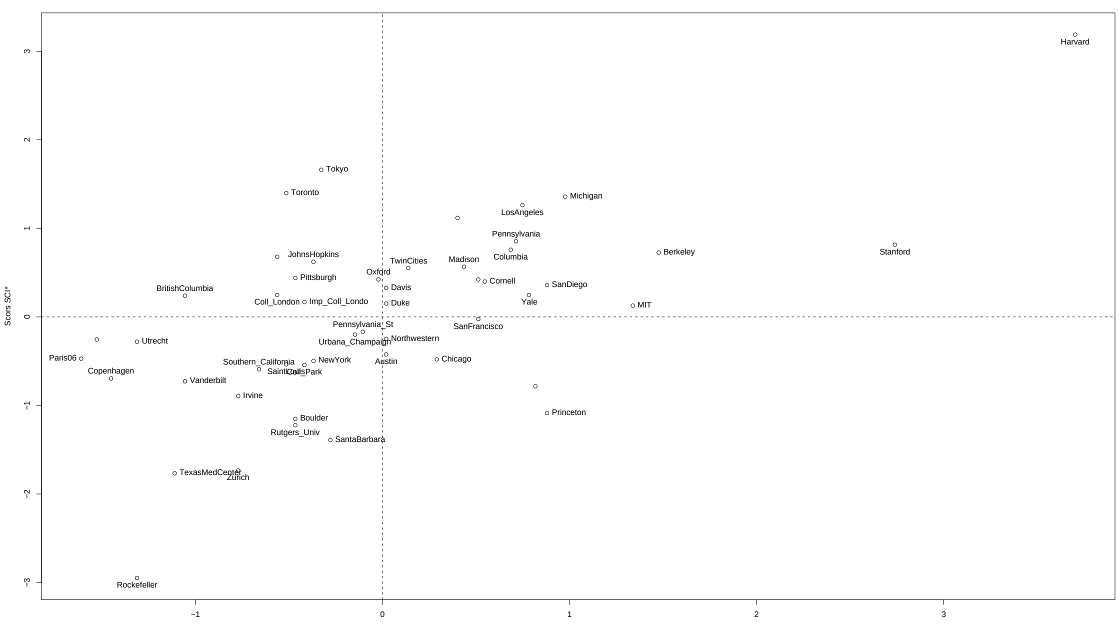
$\Longrightarrow$ Point cloud $\aleph^* = \{I^*_1, \ldots, I^*_n\}$

$\Longrightarrow$ Center of gravity G is the origin of IR^P

$\Longrightarrow$ Total inertia: $I(\aleph^*, O) = P$

Example ARWU (2007) on two variables:

Universités	**Variables**	
	X_1^* (HiCi*)	X_2^* (SCI*)
1. Harvard Univ.	3.70	3.19
2. Stanford Univ.	2.74	0.81
3. Univ. California, Berkeley	1.48	0.73
4. Univ. Cambridge	0.51	0.42
5. Massachusetts Inst. Tech. (MIT)	1.34	0.13
⋮	⋮	⋮
31. Rockefeller Univ.	−1.31	−2.95
⋮	⋮	⋮
49. Univ. Pittsburgh, Pittsburgh	−0.47	0.44
50. Univ. Southern California	−0.66	−0.59
Moyenne	0	0
Variance	1	1



3.2.2 First principal component

Projection of $\aleph^* = \{I_1^*, \ldots, I_n^*\} \in IR^P$ on a subspace of dimension one (IR^1)

First projecting direction

Find a projecting direction Δ_1 to adjust in "a better way" the point cloud $\aleph^*$

$$\Downarrow$$

Minimize the loss of information measured by the inertia of cloud $\aleph^*$ around this direction :

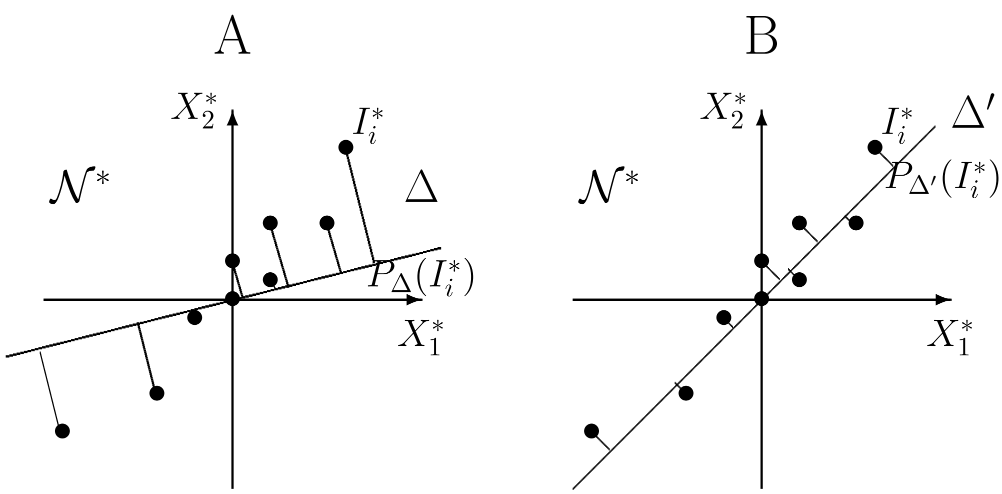
$$I(\aleph^*, \Delta_1) = \frac{1}{n} \sum_{i=1}^{n} d^2(I_i^*, P_{\Delta_1}(I_i^*))$$

where $P_{\Delta_1}(I_i^*)$ is the orthogonal projection of I_i^* on the direction Δ_1

PROBLEM:

Find the direction Δ_1 passing through the origin such that:

$$I(\aleph^*, \Delta_1) = \min_{\Delta \text{through } O} I(\aleph^*, \Delta)$$



Direction Δ_1 is called the first principal axis

Let u_1 be the vector of norm 1 associated to the direction Δ_1:

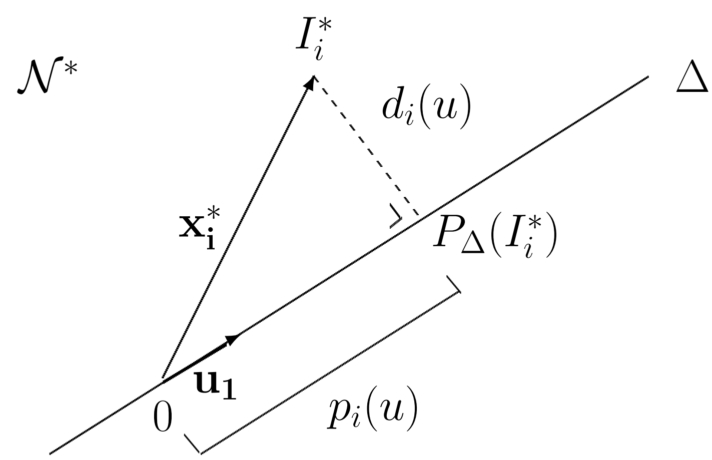
$$\underline{\mathrm{u}}_1 = (u_{1,1}, \ldots, u_{1,P})'$$

More generally let u be the vector of norm 1 from the origin associated to the direction Δ:

$$\underline{\mathrm{u}} = (u_1, \ldots, u_P)'$$

RESOLUTION :

IR^P



Let:

$$d_i(u) = \|I_i^* P_\Delta(I_i^*)\|$$
$$p_i(u) = \|OP_\Delta(I_i^*)\|$$

Find the vector u_1 of norm 1 such that :

$$u_1 = \underset{u \text{ st } \|u\|=1}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} d_i^2(u)$$

By Pythagora's theorem:

$$\|OI_i^*\|^2 = p_i(u)^2 + d_i(u)^2$$

Then

$$u_1 = \underset{u \text{ st } \|u\|=1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} d_i^2(u)$$

is equivalent to

$$u_1 = \underset{u \text{ st } \|u\|=1}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} p_i^2(u)$$

Using the scalar product:

$$p_i(u) =< u, OI_i^* >= \underline{\mathrm{u}}'\underline{\mathrm{x}}_i^* = \sum_{p=1}^{P} u_p x_{ip}^*$$

it follows that:

$$u_1 = \underset{u \text{ st } u'u=1}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} (\underline{\mathrm{u}}'\underline{\mathrm{x}}_i^*)^2.$$

Using matrices in the formulation:

$$\begin{aligned}\sum_{i=1}^{n} (\underline{\mathrm{u}}'\underline{\mathrm{x}}_i^*)^2 &= \sum_{i=1}^{n} \underline{\mathrm{u}}'\underline{\mathrm{x}}_i^*(\underline{\mathrm{x}}_i^*)'\underline{\mathrm{u}} = \underline{\mathrm{u}}' \left(\sum_{i=1}^{n} \underline{\mathrm{x}}_i^*(\underline{\mathrm{x}}_i^*)' \right) \underline{\mathrm{u}} \\ &= \underline{\mathrm{u}}'(X^*)'X^*\underline{\mathrm{u}}\end{aligned}$$

We have a optimization problem under constraint:

Maximizing $\frac{1}{n}\underline{u}'(X^*)'X^*\underline{u}$

under the constraint $\underline{u}'\underline{u} = 1$

$\Longrightarrow$ To solve this problem, we introduce the Lagrange function:

$$L(\underline{u}, \lambda) = \frac{1}{n}\underline{u}'(X^*)'X^*\underline{u} - \lambda(\underline{u}'\underline{u} - 1)$$

The solution of this problem is given by the resolution of a system of $P+1$ equations:

$$\begin{cases} \frac{\partial}{\partial u_1}L = & 0 \\ \ldots = & \ldots \\ \frac{\partial}{\partial u_P}L = & 0 \\ \frac{\partial}{\partial \lambda}L = & 0 \end{cases}$$

The last equation gives the constraint

Let derive componentwise: $u_p \, \forall p \in \{1, \ldots, P\}$:

$$
\begin{aligned}
\frac{\partial}{\partial u_p} L &= \frac{\partial}{\partial u_p}\left(\frac{1}{n}\underline{u}'(X^*)'X^*\underline{u} - \lambda(\underline{u}'\underline{u} - 1)\right) \\
&= \frac{\partial}{\partial u_p}\left(\frac{1}{n}\sum_{i=1}^{n}(\underline{u}'x_i^*)^2 - \lambda(\sum_{l=1}^{P} u_l^2 - 1)\right) \\
&= \frac{\partial}{\partial u_p}\left(\frac{1}{n}\sum_{i=1}^{n}(\sum_{l=1}^{P} u_l x_{il}^*)^2 - \lambda(\sum_{l=1}^{P} u_l^2 - 1)\right) \\
&= \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{P} u_l x_{il}^*\right) x_{ip}^* - 2\lambda u_p
\end{aligned}
$$

Putting together the P first equations leads to:

$$\begin{bmatrix} \frac{\partial}{\partial u_1}L \\ \dots \\ \frac{\partial}{\partial u_p}L \\ \dots \\ \frac{\partial}{\partial u_P}L \end{bmatrix} = 2 \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{P} u_l x_{il}^*\right) x_{i1}^* - \lambda u_1 \\ \dots \\ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{P} u_l x_{il}^*\right) x_{ip}^* - \lambda u_p \\ \dots \\ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{P} u_l x_{il}^*\right) x_{iP}^* - \lambda u_P \end{bmatrix}$$

$$= \ 2 \left(\frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} x_{i1}^* \\ \dots \\ x_{ip}^* \\ \dots \\ x_{iP}^* \end{bmatrix} (x_i^*)'\underline{u} - \lambda \underline{u} \right)$$

$$= \ 2(\frac{1}{n}\sum_{i=1}^{n} x_i^*(x_i^*)'\underline{u} - \lambda\underline{u})$$

$$= \ 2(\frac{1}{n}(X^*)'X^*\underline{u} - \lambda\underline{u})$$

The system of $P+1$ equations is then equivalent to the following system:

$$\begin{cases} \frac{1}{n}(X^*)'X^*\underline{u} & = \lambda\underline{u} \\ \underline{u}'\underline{u} & = 1 \end{cases}$$

SOLUTION: The first principal axis Δ_1 through the origin is given by the eigenvector u_1 of the correlation matrix $R = \frac{1}{n}(X^*)'X^*$ of variables X_p ($p \in \{1, \ldots, P\}$) associated with the largest eigenvalue λ_1.

Remarks:

- $\lambda = \lambda\underline{u}'\underline{u} = \frac{1}{n}\underline{u}'(X^*)'X^*\underline{u}$
- All the eigenvectors are orthogonal
- All eigenvalues are positive or null
- The number of strictly positive eigenvalues is given by the rank of X^*

Example ARWU (2007):

Eigenvalues and eigenvectors of $\mathbf{R}$

Valeurs propres	Vecteurs propres	Alumni (X_1)	Award (X_2)	HiCi (X_3)	N&S (X_4)	SCI (X_5)	PCP (X_6)
3.94	$\mathbf{u}_1$	0.42	0.42	0.44	0.47	0.26	0.41
1.09	$\mathbf{u}_2$	−0.08	−0.42	0.27	0.06	0.79	−0.34
0.47	$\mathbf{u}_3$	0.76	0.19	−0.37	−0.23	0.16	−0.40
0.26	$\mathbf{u}_4$	−0.11	0.34	0.49	0.14	−0.32	−0.71
0.13	$\mathbf{u}_5$	−0.13	−0.01	−0.54	0.80	0.02	−0.21
0.12	$\mathbf{u}_6$	−0.45	0.70	−0.24	−0.24	0.43	−0.01

$u_1 = (0,42; 0.42; 0.44; 0.47; 0.26; 0.41)'$ and

$\lambda_1 = 3.94$

The norm of u_1

$$\|u_1\| = \sum_{p=1}^{P} u_{1,p}^2 = 0.42^2 + \ldots + 0.41^2 = 1$$

is indeed equal to one

First principal component

Orthogonal projection of point cloud $\aleph^*$ on the axis Δ_1:

$$P_{\Delta_1}(\aleph) = \{P_{\Delta_1}(I_1^*), \ldots, P_{\Delta_1}(I_n^*)\}$$

Coordinate of project point $P_{\Delta_1}(I_i^*)$ define the values of the n individuals on the new variable Φ_1. This variable, the best compromise to summarize the information in dimension one, is called the first principal component:

$$\begin{aligned}\phi_{i1} &= \|OP_{\Delta_1}(I_i^*)\| =< u_1, OI_i^* > \\ &= \underline{\mathrm{u}}_1' \underline{\mathrm{x}}_i^* = \sum_{p=1}^{P} u_{1,p} x_{ip}^*\end{aligned}$$

Let Φ_1 be the vector that contains the n coordinates on the first principal component

$$\Phi_1 = X^* \underline{\mathrm{u}}_1$$

The first principal component is a linear combination of the initial variables, it is to say a weighted average.

Example: ARWU (2007)

$$\begin{aligned}\Phi_1 &= (0.42) * Alumni^* + (0.42) * Award^* \\ &+ (0.44) * HiCi^* + (0.47) * NS^* \\ &+ (0.26) * SCI^* + (0.41) * PCP^*\end{aligned}$$

University	**First axis**		
	Φ_1	CTR_{Δ_1}	$\cos^2$
1. Harvard Univ.	7.50	0.29	0.95
2. Stanford Univ.	3.88	0.08	0.84
3. Univ. California, Berkeley	3.57	0.06	0.96
4. Univ. Cambridge	3.58	0.07	0.78
5. Massachusetts Inst. Tech. (MIT)	3.33	0.06	0.92
6. California Inst. Tech.	3.61	0.07	0.53
7. Columbia Univ.	2.34	0.03	0.82
8. Princeton Univ.	1.93	0.02	0.44
9. Univ. Chicago	1.48	0.01	0.36
10. Univ. Oxford	1.41	0.01	0.71
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Properties of Φ_1

- Φ_1 is centered (weighted mean of centered variables):

$$\begin{aligned}\bar{\Phi}_1 &= \frac{1}{n}\sum_{i=1}^{n}\phi_{i1} = \frac{1}{n}\sum_{i=1}^{n}\sum_{p=1}^{P}u_{1,p}x_{ip}^{*} \\ &= \sum_{p=1}^{P}u_{1,p}\frac{1}{n}\sum_{i=1}^{n}x_{ip}^{*} = \sum_{p=1}^{P}u_{1,p}\bar{x}_p^{*} = 0\end{aligned}$$

- The variance of Φ_1 is equal to λ_1:

$$\begin{aligned}s_{\Phi_1}^2 &= \frac{1}{n}\sum_{i=1}^{n}(\phi_{i1}-\bar{\phi}_1)^2 = \frac{1}{n}\sum_{i=1}^{n}\phi_{i1}^2 = \frac{1}{n}\Phi_1'\Phi_1 \\ &= \frac{1}{n}\underline{u}_1'(X^*)'X^*\underline{u}_1 = \underline{u}_1'\frac{1}{n}(X^*)'X^*\underline{u}_1 \\ &= \underline{u}_1'\lambda_1\underline{u}_1 = \lambda_1\underline{u}_1'\underline{u}_1 = \lambda_1\end{aligned}$$

• The variance of Φ_1 is equal to the inertia of the point cloud projected on Δ_1 :

$$s^2_{\Phi_1} = \frac{1}{n}\sum_{i=1}^{n} \phi_{i1}^2 = \frac{1}{n}\sum_{i=1}^{n} \|OP_{\Delta_1}(I_i^*)\|^2$$
$$= I(P_{\Delta_1}(\aleph^*), O)$$

• Correlation between X_p and Φ_1 is given by

$$r_{X_p,\Phi_1} = \sqrt{\lambda_1} u_{1,p}$$

Indeed, the associated covariance is given by

$$s_{X_p^*,\Phi_1} = \frac{1}{n}\sum_{i=1}^{n} x_{ip}^* \phi_{i1} \quad \forall p \in \{1, \dots, P\}$$

It follows that

$$\begin{bmatrix} s_{X_1^*,\Phi_1} \\ \dots \\ s_{X_p^*,\Phi_1} \\ \dots \\ s_{X_P^*,\Phi_1} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_{i1}^*\phi_{i1} \\ \dots \\ \frac{1}{n}\sum_{i=1}^n x_{ip}^*\phi_{i1} \\ \dots \\ \frac{1}{n}\sum_{i=1}^n x_{iP}^*\phi_{i1} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}(\underline{\mathrm{v}}_1^*)'\Phi_1 \\ \dots \\ \frac{1}{n}(\underline{\mathrm{v}}_p^*)'\Phi_1 \\ \dots \\ \frac{1}{n}(\underline{\mathrm{v}}_P^*)'\Phi_1 \end{bmatrix}$$

$$= \frac{1}{n}\begin{bmatrix} (\underline{\mathrm{v}}_1^*)' \\ \dots \\ (\underline{\mathrm{v}}_p^*)' \\ \dots \\ (\underline{\mathrm{v}}_P^*)' \end{bmatrix}\Phi_1 = \frac{1}{n}(X^*)'\Phi_1 = \frac{1}{n}(X^*)'X^*\underline{\mathrm{u}}_1$$

$$= \lambda_1\underline{\mathrm{u}}_1$$

Leading to :

$$s_{X_p^*,\Phi_1} = \lambda_1 u_{1,p} \quad \forall p \in \{1,\dots,P\}$$

Hence,

$$r_{X_p,\Phi_1} = r_{X_p^*,\Phi_1} = \frac{s_{X_p^*,\Phi_1}}{s_{\Phi_1}} = \frac{\lambda_1 u_{1,p}}{\sqrt{\lambda_1}} = \sqrt{\lambda_1} u_{1,p}$$

Example: ARWU (2007)

r_{X_k,Φ_h}	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6
Alumni	0.83	−0.09	−0.52	0.06	0.05	0.16
Award	0.84	−0.44	−0.13	−0.17	0.01	−0.24
HiCi	0.86	0.29	0.26	−0.25	0.19	0.08
N&S	0.94	0.06	0.16	−0.07	−0.29	0.08
SCI	0.51	0.82	−0.11	0.16	−0.01	−0.15
Size	0.81	−0.35	0.28	0.36	0.075	0.00

Φ_1 is positively correlated with all the variables

The proximity of Φ_1 with all the initial variables is given by:

$$\frac{1}{P}\sum_{p=1}^{P} r^2_{X_p,\Phi_1} = \frac{1}{P}\sum_{p=1}^{P}\lambda_1 u^2_{1,p} = \frac{\lambda_1}{P}\sum_{p=1}^{P} u^2_{1,p} = \frac{\lambda_1}{P}$$
$$= \frac{3.94}{6} = 66\%$$

Global quality of the first principal component

Using the decomposition of total inertia, we capture the percentage of information taking into account by the first principal component:

$$\|OI_i^*\|^2 = \|OP_{\Delta_1}(I_i^*)\|^2 + \|I_i^* P_{\Delta_1}(I_i^*)\|^2$$

$$\begin{aligned}\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\|OI_i^*\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\|OP_{\Delta_1}(I_i^*)\|^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}\|I_i^* P_{\Delta_1}(I_i^*)\|^2\end{aligned}$$

$$\Rightarrow I(\aleph^*, O) = I(P_{\Delta_1}(\aleph^*), O) + I(\aleph^*, \Delta_1)$$

"Total inertia = inertia explained by Δ_1
+ residual inertia"

$\rightarrow$ Global quality is given by $\frac{\lambda_1}{P}$

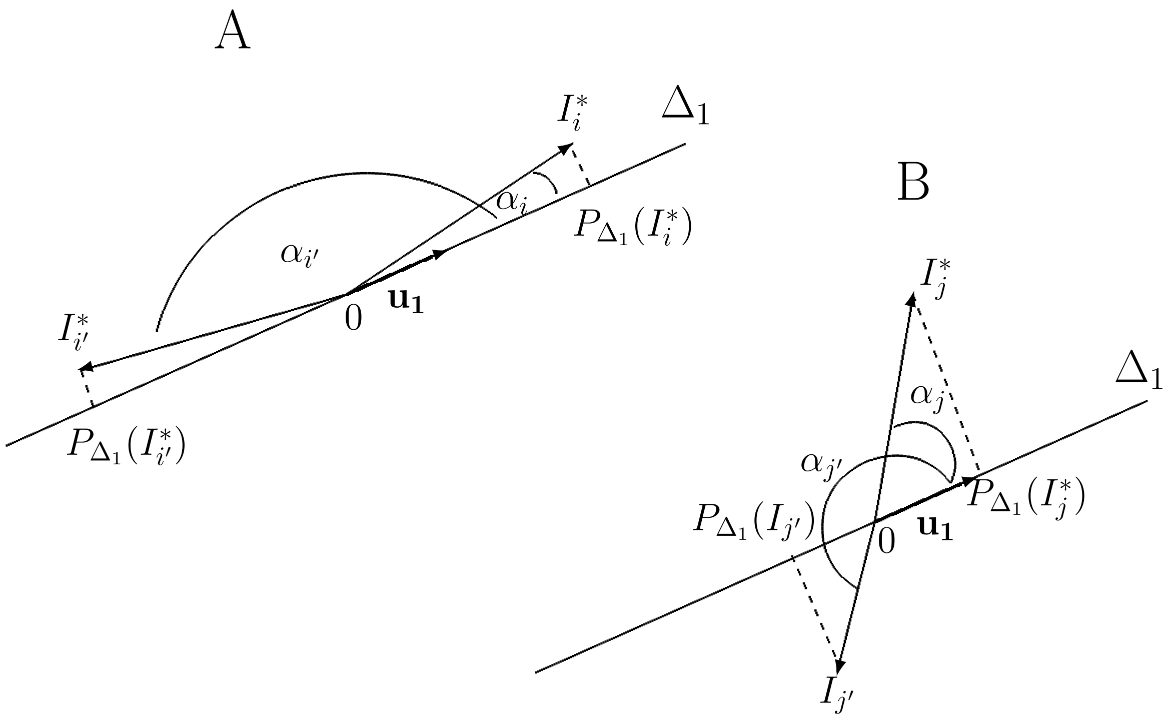
Example: ARWU (2007) $\frac{\lambda_1}{P} = \frac{3.94}{6} = 66\%$

Quality of the representation of each individual on the first axis

The quality of the representation of each individuals I_i^* on the axis Δ_1 is measured by the squared cosines of the angle between the vector OI_i^* and the axis Δ_1:

$$\begin{aligned}\cos^2(OI_i^*, \Delta_1) &= \cos^2(OI_i^*, OP_{\Delta_1}(I_i^*)) \\ &= \frac{\|0P_{\Delta_1}(I_i^*)\|^2}{\|0I_i^*\|^2} = \frac{\phi_{i1}^2}{\|0I_i^*\|^2}.\end{aligned}$$

The representation of individual i is satisfying on the first axis if $\cos^2(OI_i^*, \Delta_1)$ is close to 1.



Example: ARWU (2007)

$$\|OI^*_{Harvard}\|^2 = d^2(O, I^*_{Harvard}) = (3.70)^2 + (3.19)^2 + \ldots = 59.21$$

$$\Rightarrow \cos^2(OI^*_{Harvard}, \Delta_1) = \frac{(7.50)^2}{59.21} = 0.95$$

Contribution of each individual on the construction of the first axis

Note that :

$$\lambda_1 = I(P_{\Delta_1}(\aleph^*), O) = s^2_{\Phi_1} = \frac{1}{n}\sum_{i=1}^{n} \phi_{i1}^2$$

The contribution of each individual i on the variance Φ_1 is then given by

$$CTR_{\Delta_1}(i) = \frac{\frac{1}{n}\phi_{i1}^2}{\lambda_1}$$

Each contribution gives a percentage since

$$\sum_{i=1}^{n} CTR_{\Delta_1}(i) = 1$$

Interpretation: One individual is important in the construction of the first axis if its contribution is large. The construction of the first principal component is based essentially on individuals far away from the center of gravity.

Universities	**First axis**			**Second axis**		
	Φ_1	CTR_{Δ_1}	$\cos^2$	Φ_2	CTR_{Δ_2}	$\cos^2$
1. Harvard Univ.	7.50	0.29	0.95	1.65	0.05	0.05
2. Stanford Univ.	3.88	0.08	0.84	0.13	0.00	0.00
3. Univ. California, Berkeley	3.57	0.06	0.96	−0.06	0.00	0.00
4. Univ. Cambridge	3.58	0.07	0.78	−1.23	0.03	0.09
5. Massachusetts Inst. Tech. (MIT)	3.33	0.06	0.92	−0.67	0.01	0.04
6. California Inst. Tech.	3.61	0.07	0.53	−2.35	0.10	0.23
7. Columbia Univ.	2.34	0.03	0.82	0.00	0.00	0.00
8. Princeton Univ.	1.93	0.02	0.44	−1.94	0.07	0.44
9. Univ. Chicago	1.48	0.01	0.36	−1.24	0.03	0.26
10. Univ. Oxford	1.41	0.01	0.71	−0.24	0.00	0.02
11. Yale Univ.	1.58	0.01	0.92	0.04	0.00	0.00
12. Cornell Univ.	1.07	0.01	0.87	0.18	0.00	0.02
13. Univ. California, Los Angeles	0.71	0.00	0.20	1.21	0.03	0.57
14. Univ. California, San Diego	0.74	0.00	0.22	0.49	0.00	0.10
15. Univ. Pennsylvania	0.40	0.00	0.13	0.89	0.01	0.62
16. Univ. Washington, Seattle	0.14	0.00	0.01	1.37	0.03	0.82
17. Univ. Wisconsin, Madison	0.16	0.00	0.02	0.79	0.01	0.58
18. Univ. California, San Francisco	0.17	0.00	0.01	0.09	0.00	0.00
19. Johns Hopkins Univ.	−0.03	0.00	0.00	0.83	0.01	0.32
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31. Rockefeller Univ.	−1.13	0.01	0.11	−2.99	0.16	0.77
32. Duke Univ.	−0.80	0.00	0.25	0.78	0.01	0.24
33. Univ. Minnesota, Twin Cities	−1.07	0.01	0.31	1.40	0.04	0.53
34. Univ. Colorado, Boulder	−1.31	0.01	0.64	−0.70	0.01	0.18
35. Univ. California, Santa Barbara	−1.44	0.01	0.46	−0.98	0.02	0.21
36. Univ. British Columbia	−1.41	0.01	0.72	0.25	0.00	0.02
37. Univ. Maryland, Coll. Park	−1.51	0.01	0.92	0.01	0.00	0.00
38. Univ. Texas, Austin	−1.65	0.01	0.76	0.39	0.00	0.04
39. Univ. Paris VI	−1.61	0.01	0.59	−0.56	0.01	0.07
40. Univ. Texas Southwestern Med. Center	−1.63	0.01	0.52	−1.48	0.04	0.43
41. Vanderbilt Univ.	−1.71	0.01	0.76	−0.72	0.01	0.13
42. Univ. Utrecht	−1.76	0.02	0.83	−0.08	0.00	0.00
43. Pennsylvania State Univ., Univ. Park	−1.67	0.01	0.68	0.85	0.01	0.17
44. Univ. California, Davis	−1.70	0.01	0.55	1.16	0.02	0.26
45. Univ. California, Irvine	−1.97	0.02	0.79	−0.59	0.01	0.07
46. Univ. Copenhagen	−1.88	0.02	0.77	−0.64	0.01	0.09
47. Rutgers State Univ., New Brunswick	−1.91	0.02	0.83	−0.46	0.00	0.05
48. Univ. Manchester	−1.94	0.02	0.83	−0.12	0.00	0.00
49. Univ. Pittsburgh, Pittsburgh	−1.80	0.02	0.66	1.02	0.02	0.21
50. Univ. Southern California	−2.21	0.02	0.86	−0.15	0.00	0.00

3.2.3 Second principal component

Second projecting direction

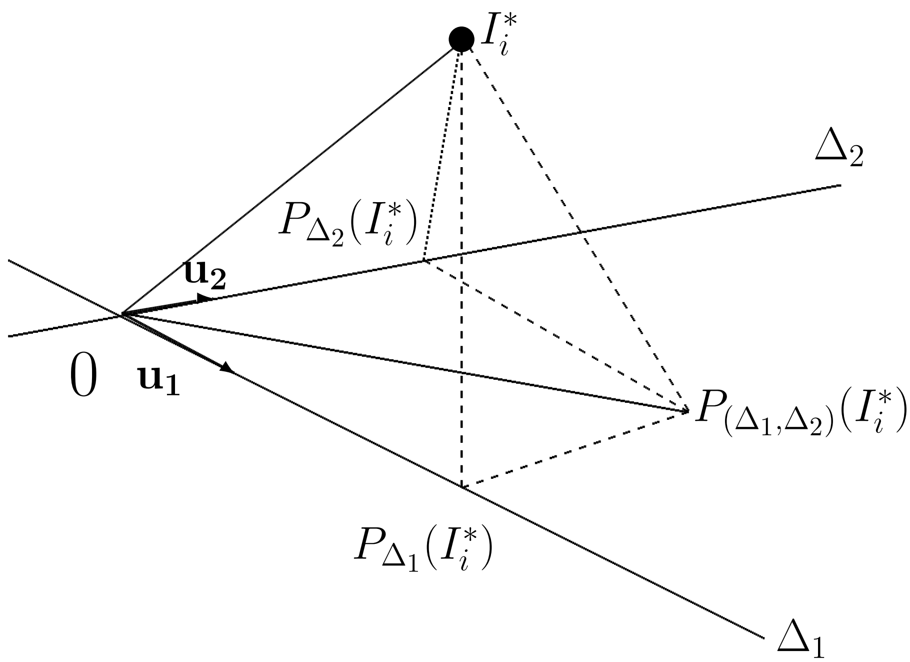
The second projecting axis Δ_2 is

- an axis through the origin of IR^P (the gravity center of point cloud $\aleph^*$)
- orthogonal to Δ_1
- minimizing the residual inertia $I(\aleph^*, (\Delta_1, \Delta_2))$

In practice, we can show that Δ_2 is given by the direction u_2, eigenvector with unitary norm of the correlation matrix R associated with the second largest eigenvalue λ_2.

The sub-space (Δ_1, Δ_2) of dimension 2 is called the first principal plan.

- *Decomposition of the total inertia*



Let:

- $P_{\Delta_1}(I_i^*)$ the orthogonal projection of I_i^* on the axis Δ_1
- $P_{\Delta_2}(I_i^*)$ the orthogonal projection of I_i^* on the axis Δ_2
- $P_{(\Delta_1,\Delta_2)}(I_i^*)$ the orthogonal projection of I_i^* on the axis (Δ_1, Δ_2).

By Pythagora's theorem:

$$\|0I_i^*\|^2 = \|0P_{(\Delta_1,\Delta_2)}(I_i^*)\|^2 + \|I_i^* P_{(\Delta_1,\Delta_2)}(I_i^*)\|^2$$

Moreover

- $P_{\Delta_1}(I_i^*)$ is the orthogonal projection of $P_{(\Delta_1,\Delta_2)}(I_i^*)$ on the axis Δ_1
- $P_{\Delta_2}(I_i^*)$ is the orthogonal projection of $P_{(\Delta_1,\Delta_2)}(I_i^*)$ on the axis Δ_2,

$$\begin{aligned}
\Longrightarrow \|0I_i^*\|^2 = \|0P_{\Delta_1}(I_i^*)\|^2 \; &+ \; \|0P_{\Delta_2}(I_i^*)\|^2 \\
&+ \; \|I_i^* P_{(\Delta_1,\Delta_2)}(I_i^*)\|^2 \\
&\Downarrow \\
\Longrightarrow \frac{1}{n}\sum_{i=1}^{n} \|0I_i^*\|^2 = \frac{1}{n}\sum_{i=1}^{n} & \quad \|0P_{\Delta_1}(I_i^*)\|^2 + \frac{1}{n}\sum_{i=1}^{n} \|0P_{\Delta_2}(I_i^*)\|^2 \\
&+ \; \frac{1}{n}\sum_{i=1}^{n} \|I_i^* P_{(\Delta_1,\Delta_2)}(I_i^*)\|^2 \\
&\Downarrow \\
I(\aleph^*, 0) = I(P_{\Delta_1}(\aleph^*), 0) \; &+ \; I(P_{\Delta_2}(\aleph^*), 0) + I(\aleph^*, (\Delta_1, \Delta_2)).
\end{aligned}$$

Second principal component

Orthogonal projection of point cloud $\aleph^*$ on the axis Δ_2:

$$P_{\Delta_2}(\aleph^*) = \{P_{\Delta_2}(I_1^*), \ldots, P_{\Delta_2}(I_n^*)\}$$

In the same way that for the first direction, define:

$$\phi_{i2} = \|0P_{\Delta_2}(I_i^*)\| \qquad \forall i = 1, \ldots, n$$

where ϕ_{i2} gives the value of individual i on the second principal component Φ_2

The second principal component is also a weighted average of initial variables

$$\begin{aligned} \phi_{i2} &= < u_2, 0I_i^* > \\ &= \underline{\mathrm{u}}_2' \underline{\mathrm{x}}_i^* \\ &= \sum_{p=1}^{P} u_{2,p} x_{ip}^*. \end{aligned}$$

Let Φ_2 be the vector that contains the n coordinate on the first principal component $\Phi_2 = (\phi_{12}, \ldots, \phi_{n2})'$:

$$\Phi_2 = X^* u_2.$$

The second new variable Φ_2 is a linear combination of the initial variables $X_1^*, \ldots, X_P^*$:

$$\Phi_2 = \sum_{p=1}^{P} u_{2,p} X_p^*.$$

Example: ARWU (2007)

$$\begin{aligned} \Phi_2 &= -0.08 * Alumni^* - 0.42 * Award^* \\ &+ 0.27 * HiCi^* + 0.06 * NS^* \\ &+ 0.79 * SCI^* - 0.34 * PCP^* \end{aligned}$$

The second component discriminates between in one hand Nobel prize (Award) and size (PCP), and in the other hand the volume of publication (SCI and HiCi) (to be verified with correlation matrix)

Properties of Φ_2

- Φ_2 has zero mean (exercise)
- Φ_2 has a variance equal to λ_2 (exercise)

  It follows that

$$\begin{aligned}\lambda_2 = s^2_{\Phi_2} &= \frac{1}{n}\sum_{i=1}^{n}\phi_{i2}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\|0P_{\Delta_2}(I_i^*)\|^2 \\ &= I(P_{\Delta_2}(\aleph^*), 0).\end{aligned}$$

- The correlation between Φ_1 and Φ_2 is equal

to zero:

$$\begin{aligned}s_{\Phi_1,\Phi_2} &= \frac{1}{n}\sum_{i=1}^{n}\phi_{i1}\phi_{i2} \\ &= \frac{1}{n}\Phi_1'\Phi_2 = \frac{1}{n}u_1'(X^*)'X^*u_2 \\ &= u_1'\lambda_2 u_2 = \lambda_2 u_1' u_2 = 0\end{aligned}$$

$$\Longrightarrow r_{\Phi_1,\Phi_2} = 0.$$

• Correlation between the second component and initial variables (exercise):

$$r_{X_p,\Phi_2} = \sqrt{\lambda_2} u_{2,p} \qquad \forall p = 1, \ldots, P.$$

Example: ARWU (2007)

r_{X_k,Φ_h}	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6
Alumni	0.83	−0.09	−0.52	0.06	0.05	0.16
Award	0.84	−0.44	−0.13	−0.17	0.01	−0.24
HiCi	0.86	0.29	0.26	−0.25	0.19	0.08
N&S	0.94	0.06	0.16	−0.07	−0.29	0.08
SCI	0.51	0.82	−0.11	0.16	−0.01	−0.15
Size	0.81	−0.35	0.28	0.36	0.075	0.00

$$\Downarrow$$

Φ_2 discriminates, for universities with globally the same level on Φ_1, 2 behaviors:

• Volume of publication dominates the number of Nobel prize : $\phi_{\{Michigan,2\}} = 2.10$,

• Nobel prizes dominates the score on the volume of publication: $\phi_{\{Rockfeller,2\}} = -2.99$

Global quality of the second principal component

Percentage of inertia explained by Δ_2:

$$\frac{\lambda_2}{P}$$

Percentage of inertia explained by the first principal plan (Δ_1, Δ_2):

$$\frac{\lambda_1 + \lambda_2}{P}$$

Example: ARWU (2007)

Δ_2 explains $\frac{1.09}{6} = 18.17\%$ of total inertia

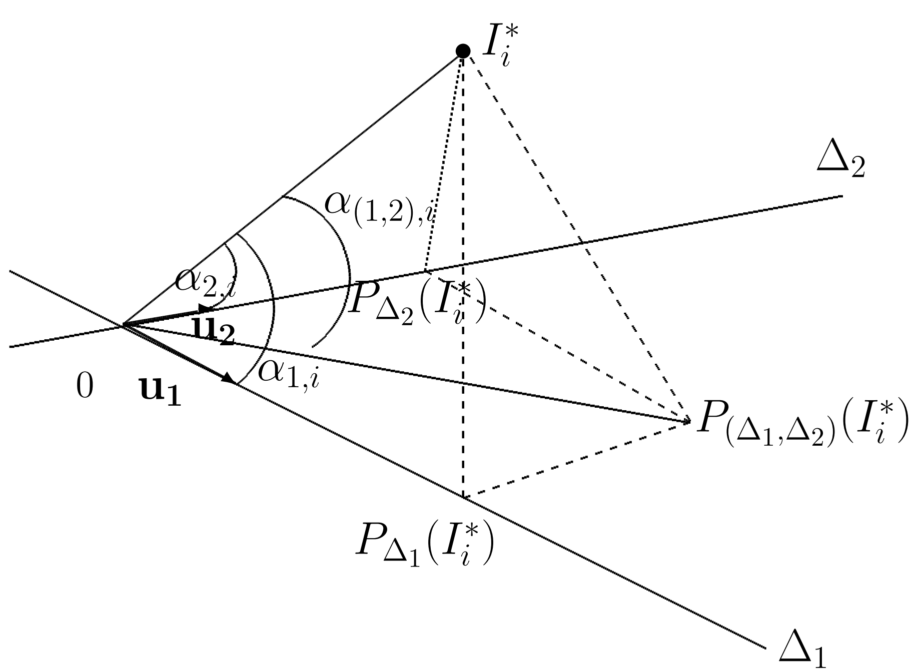
$$\Downarrow$$

Then (Δ_1, Δ_2) explains $\frac{3.94+1.09}{6} = 83.83\%$ of total inertia

Quality of the representation of each individual on the second axis

Quality of representation of each point I_i^* on the axis Δ_2 is measured by the squared cosines of angle between the vector OI_i^* and the direction Δ_2 :

$$\cos^2(OI_i^*, \Delta_2) = \frac{\|0P_{\Delta_2}(I_i^*)\|^2}{\|0I_i^*\|^2} = \frac{\phi_{i2}^2}{\|0I_i^*\|^2}.$$



Quality of representation of each point I_i^* on the plan (Δ_1, Δ_2) is measured by the squared cosines of angle between the vector OI_i^* and the plan (Δ_1, Δ_2) :

$$\begin{aligned}
\cos^2(OI_i^* \quad , \quad (\Delta_1, \Delta_2)) &= \frac{\|0P_{(\Delta_1,\Delta_2)}(I_i^*)\|^2}{\|0I_i^*\|^2} \\
&= \frac{\|0P_{(\Delta_1)}(I_i^*)\|^2 + \|0P_{(\Delta_2)}(I_i^*)\|^2}{\|0I_i^*\|^2} \\
&= \frac{\phi_{i1}^2 + \phi_{i2}^2}{\|0I_i^*\|^2} \\
&= \cos^2(OI_i^*, \Delta_1) + \cos^2(OI_i^*, \Delta_2).
\end{aligned}$$

Contribution of each individual on the construction of the second axis Δ_2

Note that:

$$\lambda_2 = I(P_{\Delta_2}(\aleph^*), 0) = s^2_{\Phi_2} = \frac{1}{n}\sum_{i=1}^{n} \phi^2_{i2},$$

The contribution of each individual i on the variance Φ_2 is given by:

$$CTR_{\lambda_2} = \frac{\frac{1}{n}\phi^2_{i2}}{\lambda_2}.$$

Universities	**First axis**			**Second axis**		
	Φ_1	CTR_{Δ_1}	$\cos^2$	Φ_2	CTR_{Δ_2}	$\cos^2$
1. Harvard Univ.	7.50	0.29	0.95	1.65	0.05	0.05
2. Stanford Univ.	3.88	0.08	0.84	0.13	0.00	0.00
3. Univ. California, Berkeley	3.57	0.06	0.96	−0.06	0.00	0.00
4. Univ. Cambridge	3.58	0.07	0.78	−1.23	0.03	0.09
5. Massachusetts Inst. Tech. (MIT)	3.33	0.06	0.92	−0.67	0.01	0.04
6. California Inst. Tech.	3.61	0.07	0.53	−2.35	0.10	0.23
7. Columbia Univ.	2.34	0.03	0.82	0.00	0.00	0.00
8. Princeton Univ.	1.93	0.02	0.44	−1.94	0.07	0.44
9. Univ. Chicago	1.48	0.01	0.36	−1.24	0.03	0.26
10. Univ. Oxford	1.41	0.01	0.71	−0.24	0.00	0.02
11. Yale Univ.	1.58	0.01	0.92	0.04	0.00	0.00
12. Cornell Univ.	1.07	0.01	0.87	0.18	0.00	0.02
13. Univ. California, Los Angeles	0.71	0.00	0.20	1.21	0.03	0.57
14. Univ. California, San Diego	0.74	0.00	0.22	0.49	0.00	0.10
15. Univ. Pennsylvania	0.40	0.00	0.13	0.89	0.01	0.62
16. Univ. Washington, Seattle	0.14	0.00	0.01	1.37	0.03	0.82
17. Univ. Wisconsin, Madison	0.16	0.00	0.02	0.79	0.01	0.58
18. Univ. California, San Francisco	0.17	0.00	0.01	0.09	0.00	0.00
19. Johns Hopkins Univ.	−0.03	0.00	0.00	0.83	0.01	0.32
⋮	⋮	⋮	⋮	⋮	⋮	⋮
31. Rockefeller Univ.	−1.13	0.01	0.11	−2.99	0.16	0.77
32. Duke Univ.	−0.80	0.00	0.25	0.78	0.01	0.24
33. Univ. Minnesota, Twin Cities	−1.07	0.01	0.31	1.40	0.04	0.53
34. Univ. Colorado, Boulder	−1.31	0.01	0.64	−0.70	0.01	0.18
35. Univ. California, Santa Barbara	−1.44	0.01	0.46	−0.98	0.02	0.21
36. Univ. British Columbia	−1.41	0.01	0.72	0.25	0.00	0.02
37. Univ. Maryland, Coll. Park	−1.51	0.01	0.92	0.01	0.00	0.00
38. Univ. Texas, Austin	−1.65	0.01	0.76	0.39	0.00	0.04
39. Univ. Paris VI	−1.61	0.01	0.59	−0.56	0.01	0.07
40. Univ. Texas Southwestern Med. Center	−1.63	0.01	0.52	−1.48	0.04	0.43
41. Vanderbilt Univ.	−1.71	0.01	0.76	−0.72	0.01	0.13
42. Univ. Utrecht	−1.76	0.02	0.83	−0.08	0.00	0.00
43. Pennsylvania State Univ., Univ. Park	−1.67	0.01	0.68	0.85	0.01	0.17
44. Univ. California, Davis	−1.70	0.01	0.55	1.16	0.02	0.26
45. Univ. California, Irvine	−1.97	0.02	0.79	−0.59	0.01	0.07
46. Univ. Copenhagen	−1.88	0.02	0.77	−0.64	0.01	0.09
47. Rutgers State Univ., New Brunswick	−1.91	0.02	0.83	−0.46	0.00	0.05
48. Univ. Manchester	−1.94	0.02	0.83	−0.12	0.00	0.00
49. Univ. Pittsburgh, Pittsburgh	−1.80	0.02	0.66	1.02	0.02	0.21
50. Univ. Southern California	−2.21	0.02	0.86	−0.15	0.00	0.00

3.2.4 Extended dimensions

The h^{th} projecting axis Δ_h is

- an axis passing through the origin of IR^P (the gravity center of point cloud $\aleph^*$)
- orthogonal to $\Delta_1, \ldots, \Delta_{h-1}$
- minimizing the residual inertia

In practice, we can show that Δ_h is given by the direction u_h which is the eigenvector (with unitary norm) of the correlation matrix R that is associated with the h^{th} largest eigenvalue λ_h.

It is clear that if h is equal to the rank of X^*, the data cloud $\aleph^*$ is contained in the subspace generated by $\{u_1, \ldots, u_h\}$ and the reduction mechanism can stop.

Orthogonal projection of point cloud $\aleph^*$ on the axis Δ_h:

$$P_{\Delta_h}(\aleph^*) = \{P_{\Delta_h}(I_1^*), \ldots, P_{\Delta_h}(I_n^*)\}$$

In the same way that for other directions, define:

$$\phi_{ih} = \|0P_{\Delta_h}(I_i^*)\| \qquad \forall i = 1, \ldots, n$$

where ϕ_{ih} gives the value of individual i on the principal component Φ_h

The principal component is also a weighted average of the initial variables

$$\begin{aligned}
\phi_{ih} &= < u_h, 0I_i^* > \\
&= \underline{\mathrm{u}}'_h \underline{\mathrm{x}}_i^* \\
&= \sum_{p=1}^{P} u_{h,p} x_{ip}^*.
\end{aligned}$$

Properties of Φ_h

- Φ_h has zero mean (exercise)
- Φ_h has a variance equal to λ_h (exercise)
- Correlation between $\Phi_l (l \in \{1, \ldots, h-1\}$ and Φ_h is equal to zero:

$$
\begin{aligned}
s_{\Phi_l,\Phi_h} &= \frac{1}{n}\sum_{i=1}^{n} \phi_{il}\phi_{ih} \\
&= \frac{1}{n}\Phi_l'\Phi_h = \frac{1}{n}u_l'(X^*)'X^*u_h \\
&= u_l'\lambda_h u_h = \lambda_h u_l' u_h = 0 \\
&\Longrightarrow r_{\Phi_l,\Phi_h} = 0.
\end{aligned}
$$

- Correlation between the h^{th} component and the initial variables (exercise):

$$r_{X_p,\Phi_h} = \sqrt{\lambda_h} u_{h,p} \qquad \forall p = 1, \ldots, P.$$

Correlations and eigenvectors

By linear algebra:

$$R = \frac{1}{n}(X^*)'X^* = \sum_{h=1}^{H} \lambda_h u_h u_h'.$$

Then, for each $p \neq l \in \{1, \ldots, P\}$:

$$r_{X_p,X_l} = \sum_{h=1}^{H} \lambda_h u_{h,p} u_{h,l}.$$

Question: How many principal components needed?

Stopping rules for determining the number of principal components:

- Classical rule based on τ_h, the percentage of variance explained by the first h principal components, $h \in \{1, \ldots, H\}$:
$$\tau_h = \frac{\lambda_1 + \ldots + \lambda_h}{\lambda_1 + \ldots + \lambda_H} = \frac{\lambda_1 + \ldots + \lambda_h}{P}.$$
If τ is big enough (close to one), h is the number of factors to choose. But this rule is rather subjective.

- Keep principal component Φ_h iff $\lambda_h > 1$ (mean of eigenvalues).

- Examine the scree s plot that shows the fraction of total variance in the data explained by each principal component

3.2.5 Graphical representations

The principal components are used to represent graphically individuals and variables

Map of individuals

Projection of the data cloud $\aleph^*$ on the first principal plan (Δ_1, Δ_2):

$$\Downarrow$$

$\forall i = 1, \ldots, n$ the projection $P_{(\Delta_1,\Delta_2)}(I_i*)$ of individual I_i* on the first plan has coordinates

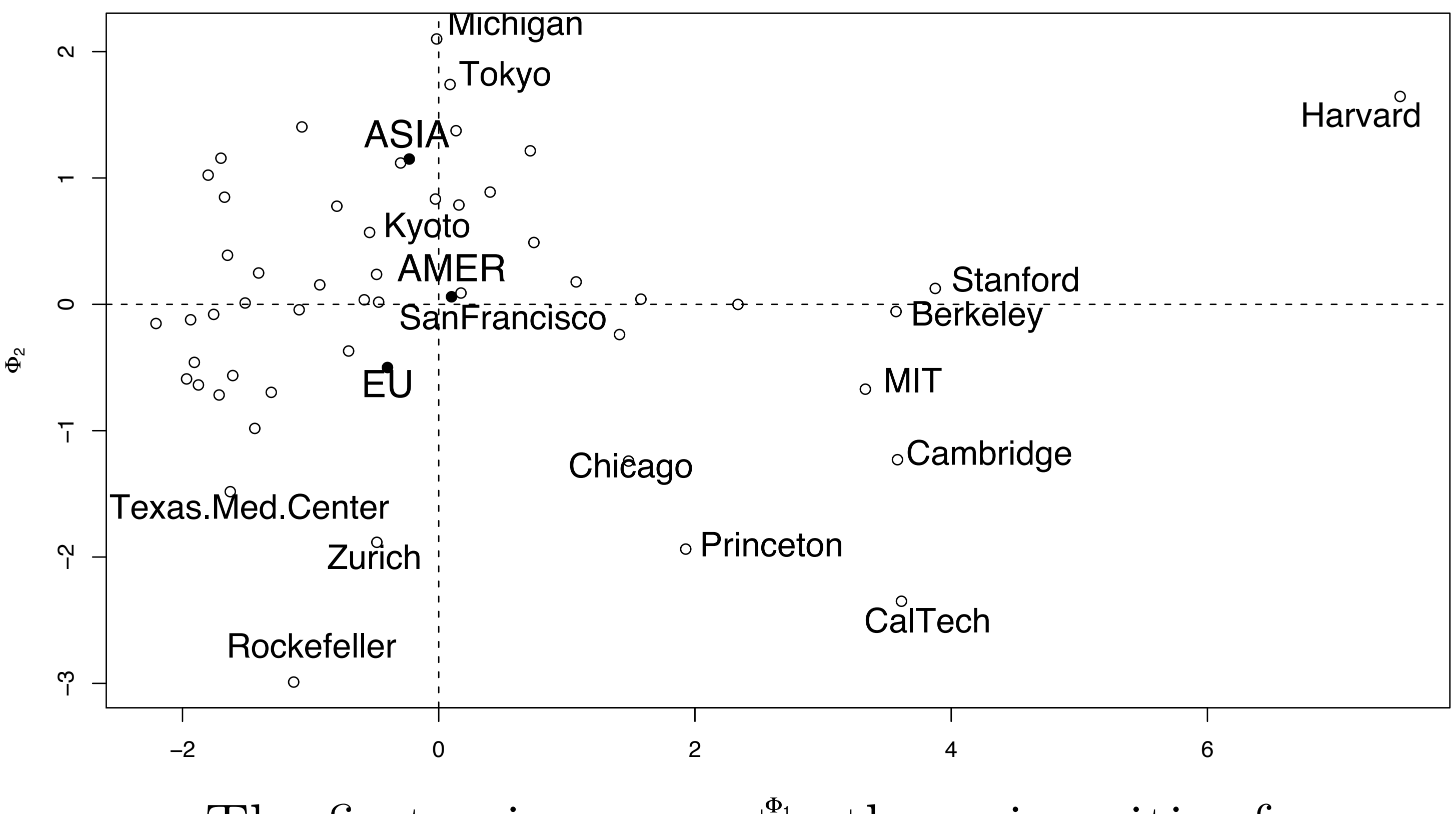
$$(\phi_{i1}, \phi_{i2})$$

on the axis Δ_1 and Δ_2.

This graph makes the interpretation of axis easier as well as the comparison between individuals

Example: ARWU (2007)

Well represented individuals can be interpreted



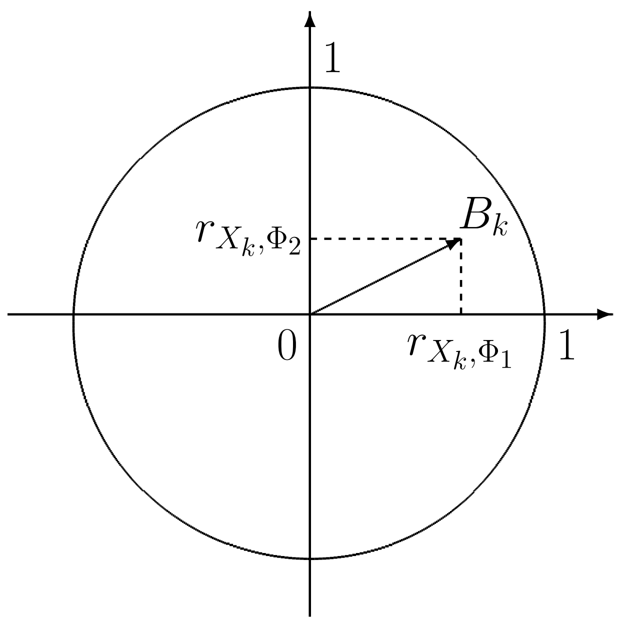
- The first axis segregates the universities from the less quality to the best quality in terms on research
- The second axis discriminates between "volume of publication" and "Nobel prizes"
- Harvard seems to be an outlier

If the principal plan is not sufficient, (Δ_1, Δ_3) and (Δ_2, Δ_3) plans can also be analyzed

Correlations circle

Representation of variables is based on the projection of the cloud of p variables X^* in IR^n on the principal components. The coordinate on the fisrt principal plan are

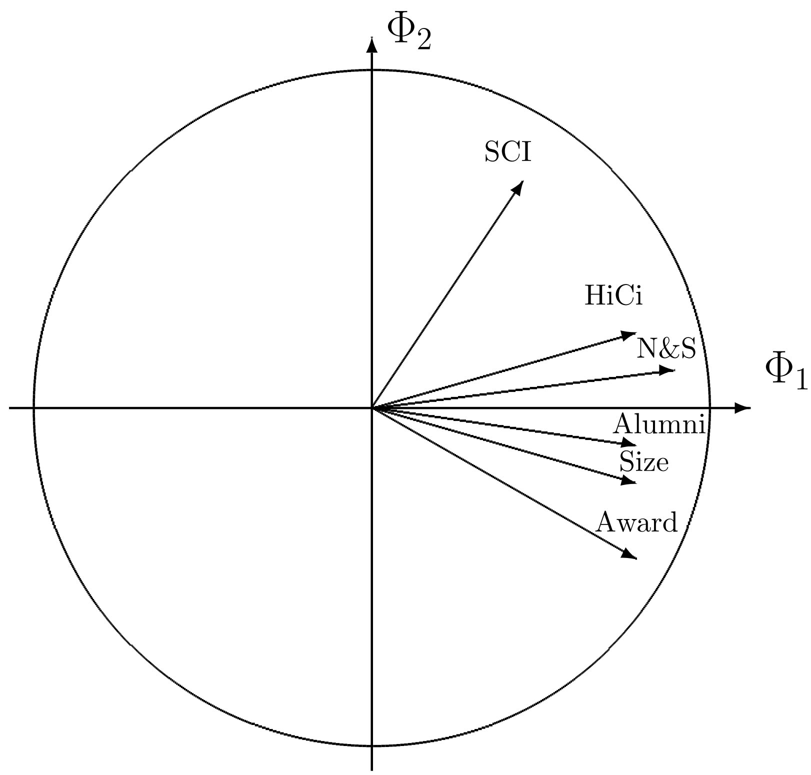
$$B_p = (r_{X_p,\Phi_1}, r_{X_p,\Phi_2}).$$



This graph makes it easier to visualize

- correlations between old and new variables
- the quality of the representation of X_p given by the norm of the vector $0B_p$

Example: ARWU (2007)



- All variables have a good quality of representation in IR^2
- The first principal component is positively correlated with all variables (quality factor)
- The second principal component discriminates between "Volume" and "Prizes" $\Longrightarrow$ type of research quality

3.3 Additional variables or individuals

• Additional individuals i_s

- Step 1: Standardize the coordinate of new individual i_s using mean and standard deviation calculated on active individuals

- Step 2: Project new standardize individual on principal axis:

$$\begin{aligned}
\phi_{i_s 1} &= \sum_{p=1}^{P} u_{1,p} x^*_{i_s p} \\
\phi_{i_s 2} &= \sum_{p=1}^{P} u_{2,p} x^*_{i_s p} \\
&\text{etc}
\end{aligned}$$

- Step 3: Project this observation on the first plan.

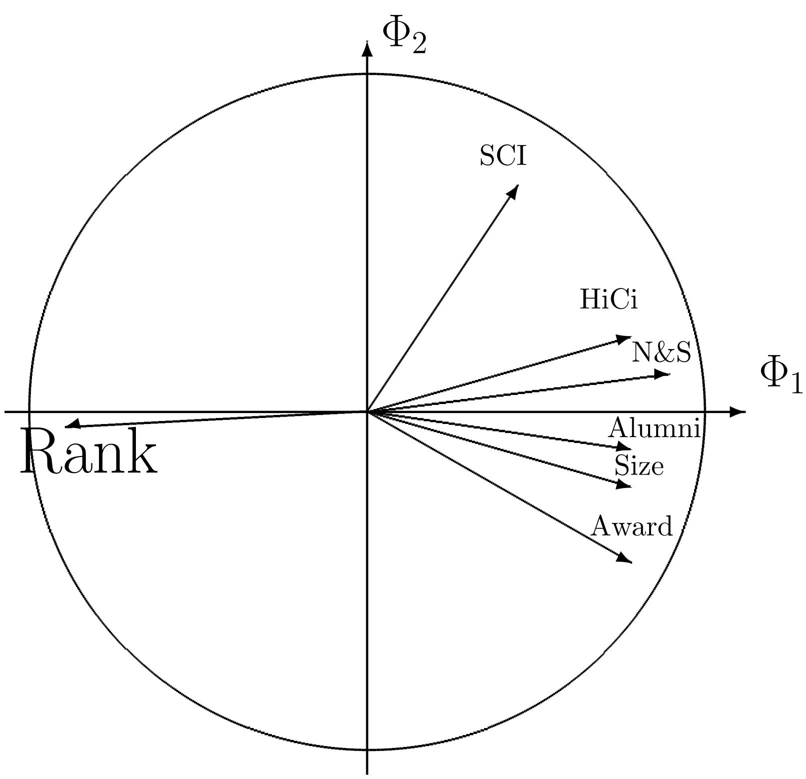
• Additional continuous variable X_s

The information on the additional continuous variable X_s will be given by the correlations circle where the coordinates are

$$r_{X_s,\Phi_1} \quad \text{and} \quad r_{X_s,\Phi_2}$$

Example: ARWU (2007)

Representation of the ranking given in Shanghai ranking



• Additional qualitative variable X_s

If the variable is qualitative, the correlation can not be used

$$\Downarrow$$

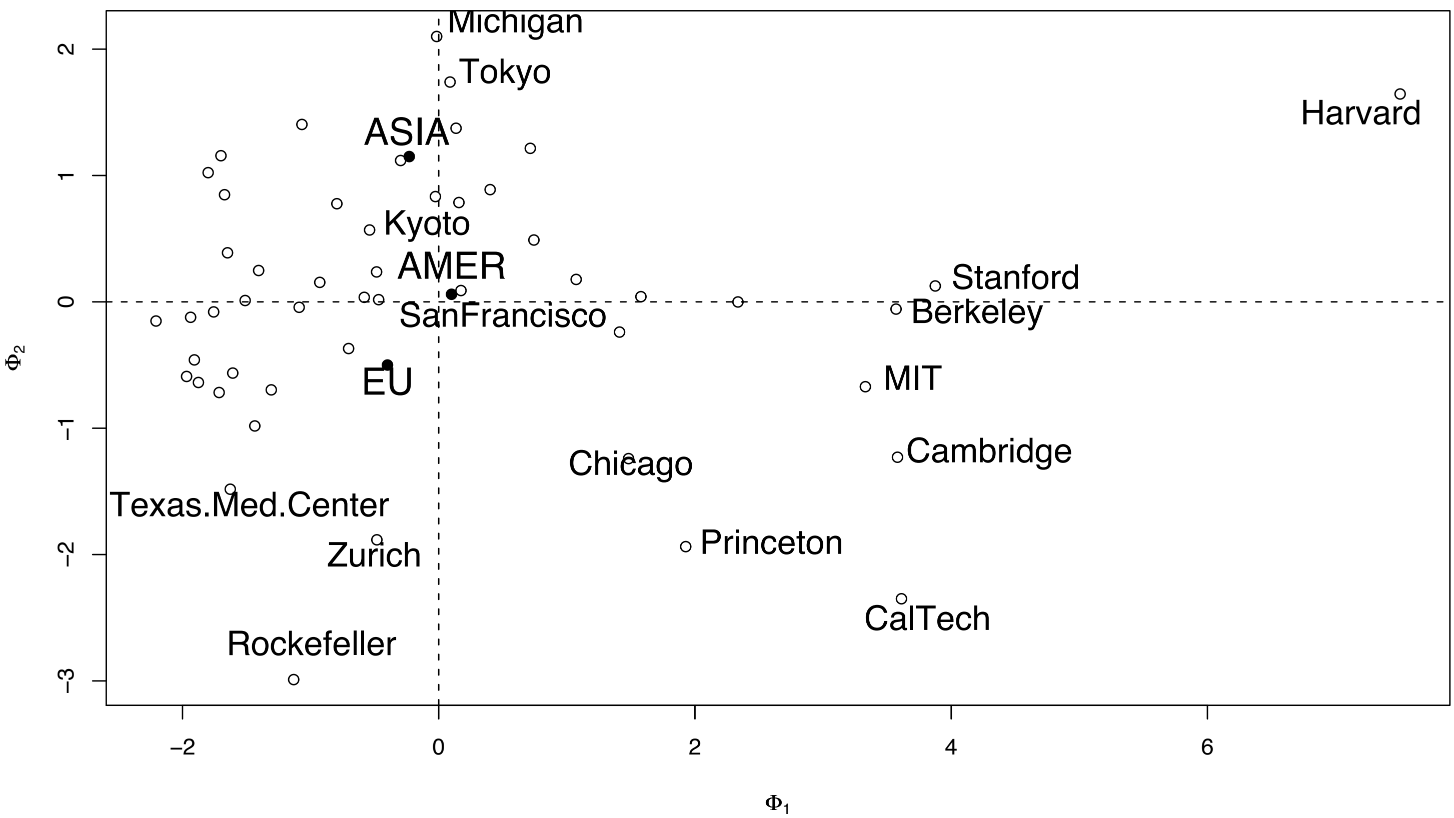
Create K groups individuals formed by the K categories of X_s

Then project the K mean individuals on the map of individuals

Note that if the variable is ordinal, you can link the mean individuals by the way of a line

Example: ARWU (2007)

Representation of groups of individuals : european, asian and US universities



- US universities is a little bit better than the two others
- European universities perform better in terms of Nobel prizes
- Asian universities perform better in terms of volume of publications

3.4 ACP following Hotelling

These procedures seem to be less complex but are less intuitive from a geometrical point of view

Correlation criteria

Find J new standardized uncorrelated variables $Z_1, \ldots, Z_J$ such that the following criteria is maximized:

$$\sum_{j=1}^{J}[\frac{1}{P}\sum_{p=1}^{P} r^2_{X_p,Z_j}].$$

It is possible to prove that the maximum is reached by reducing the principal principal components

$$Z_j = \Phi_j^* = \frac{\Phi_j}{\sqrt{\lambda_j}}$$

and the maximum is given by $\frac{\lambda_1+\ldots+\lambda_J}{P}$.

Variance criteria

Find J new uncorrelated variables $Z_1, \ldots, Z_J$ such that

$$Z_j = \sum_{p=1}^{P} \nu_{j,p} X_p$$

where the vectors

$$\nu_j = (\nu_{j,1}, \ldots, \nu_{j,P})'$$

maximize the following criteria

$$\sum_{j=1}^{J} s^2_{Z_j}.$$

- The maximum is given by

$$\lambda_{\nu_1} + \ldots + \lambda_{\nu_J}$$

- The maximum is reached for orthogonal eigenvectors of covariance matrix

- If the standardized variables are used, then $Z_j = \Phi_j$ and the maximum is given by $\lambda_1 + \ldots + \lambda_J$

QUANTIFYING ACADEMIC EXCELLENCE, WHAT DO THE SHANGHAI RANKING MEASURE ?

C. Dehon, A. McCathie & V. Verardi

Université libre de Bruxelles, ECARES - CKE

September 2009

- Increased competition in Higher Education

$$\Downarrow$$

emergence of multiple rankings

- The most widely reported university rankings are:
  - Academic Ranking of World Universities (ARWU - Shanghai)
  - THES-QS Ranking (Times Higher Education)

- We choose the ARWU: objective choice of variables and greater transparency

$\Rightarrow$ OUR AIM: to find the underlying factors measured by ARWU

SHANGHAI RANKING (ARWU): VARIABLES AND WEIGHTS

- **Alumni (10%)**: Alumni recipients of the Nobel prize or the Fields Medal;
- **Award (20%)**: Current faculty Nobel laureates and Fields Medal winners;
- **HiCi (20%)**: Highly cited researchers in 21 broad subject categories;
- **N&S (20%)**: Articles published in Nature and Science;
- **PUB (20%)**: Articles in the Science Citation Index-expanded, and the Social Science Citation Index;
- **PCP (10%)**: The weighted score of the previous 5 indicators divided by the number of full-time academic staff members..

http://www.arwu.org/rank/2008/ranking2008.htm

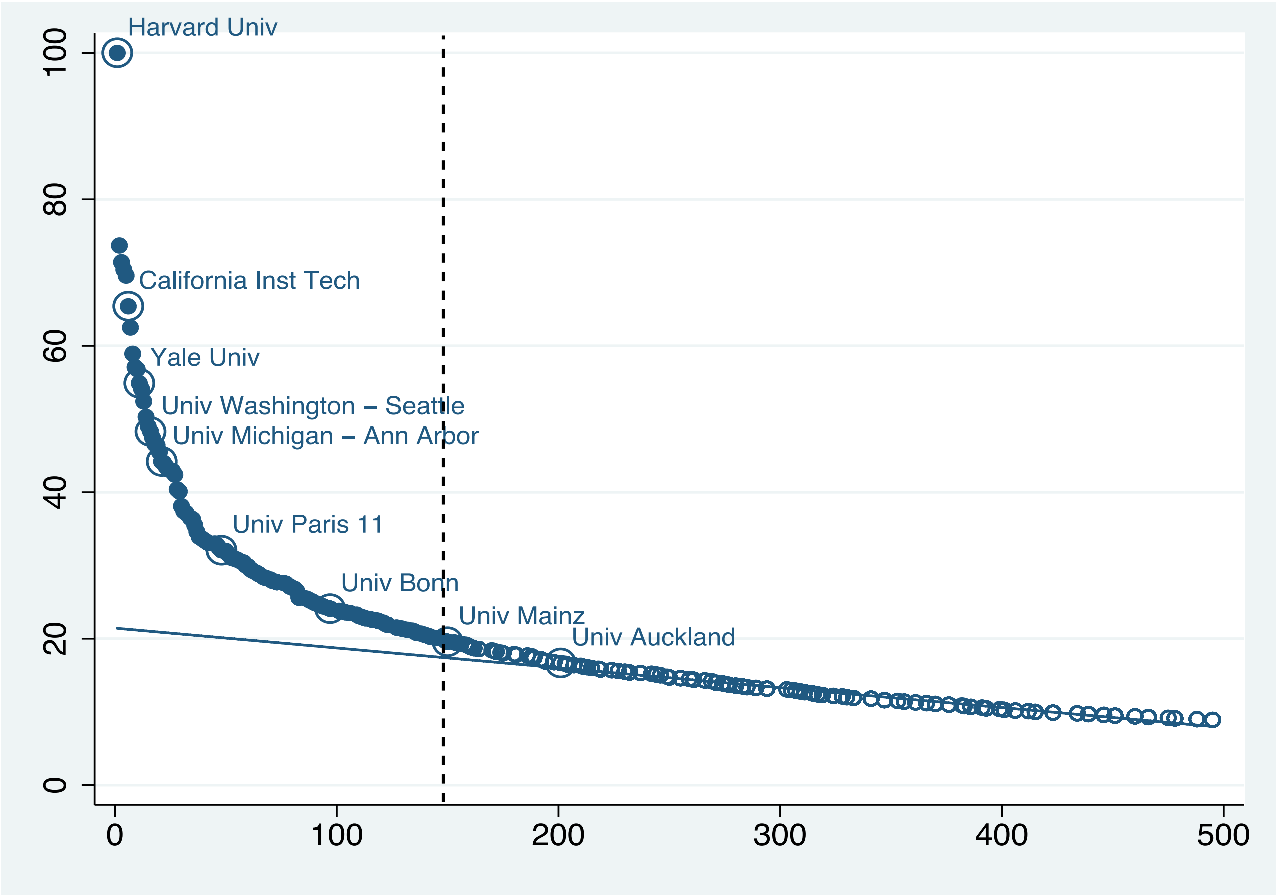


Figure: Overall score relative to rank

CRITICISM OF THE SHANGHAI RANKING:

- Limited scope despite the complexity of a university;
- Favours English-speaking countries;
- Very heavily biased towards science and technology subjects;
- Production versus efficiency: "Bigger is better";
- Input variables not taken in consideration (Aghion et al, 2007);
- Highly sensitive due to the normalization step;
- Confidence intervals needed.

PRINCIPAL COMPONENT ANALYSIS on TOP 150

QUESTION: Can a single "indicator" accurately sum up research excellence ?

GOAL: To determine the underlying factors measured by the variables used in the Shanghai ranking

$\Rightarrow$ Principal component analysis

PRINCIPAL COMPONENT ANALYSIS

The first component accounts for 64% of the inertia and is given by:

$\Phi_1 = 0.42 * Alumni + 0.44 * Awards + 0.48 * HiCi + 0.50 * NS + 0.38 * PUB$

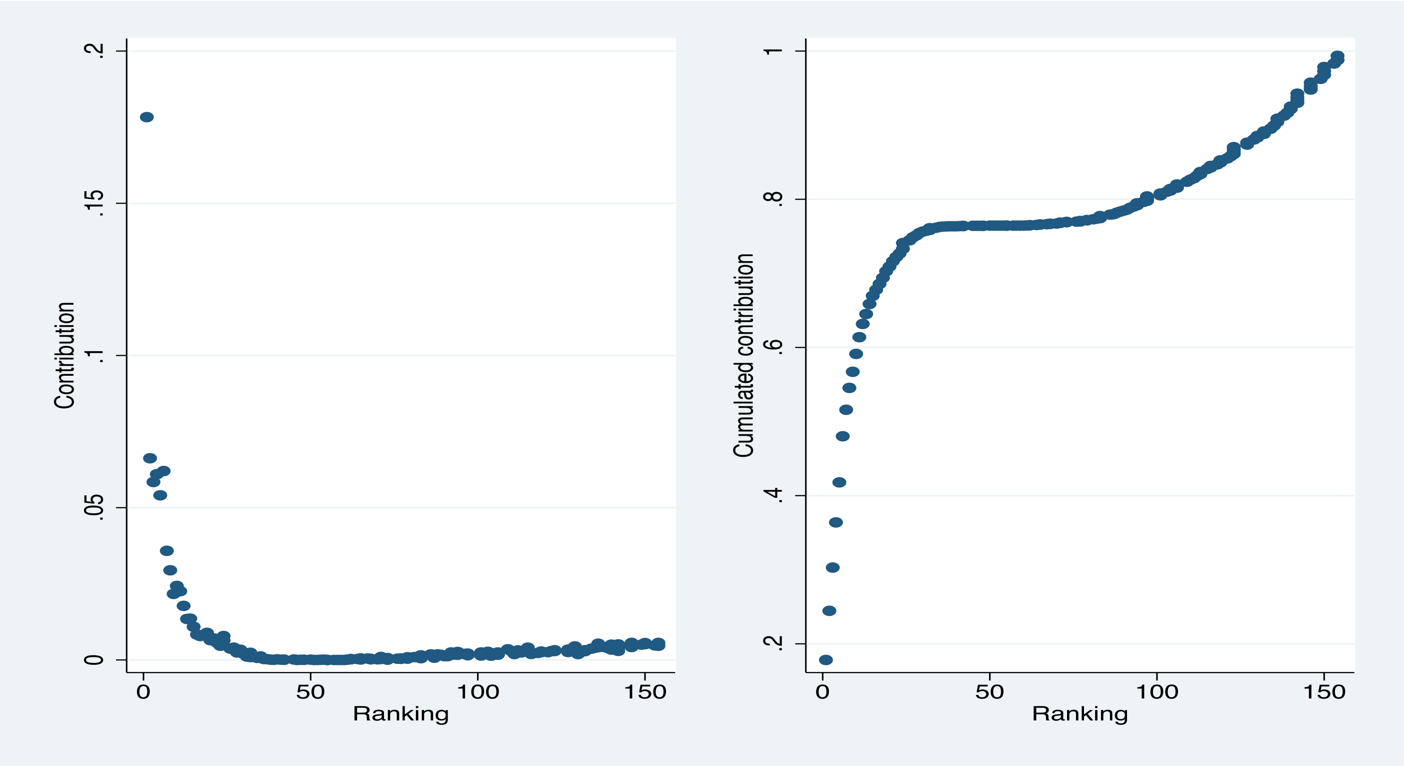
What does this component measure?? The quality of research??

Variable	Corr(ϕ_1, .)
Alumni	78%
Awards	81%
HiCi	89%
N&S	92%
PUB	70%
Total score	99%

BUT ...

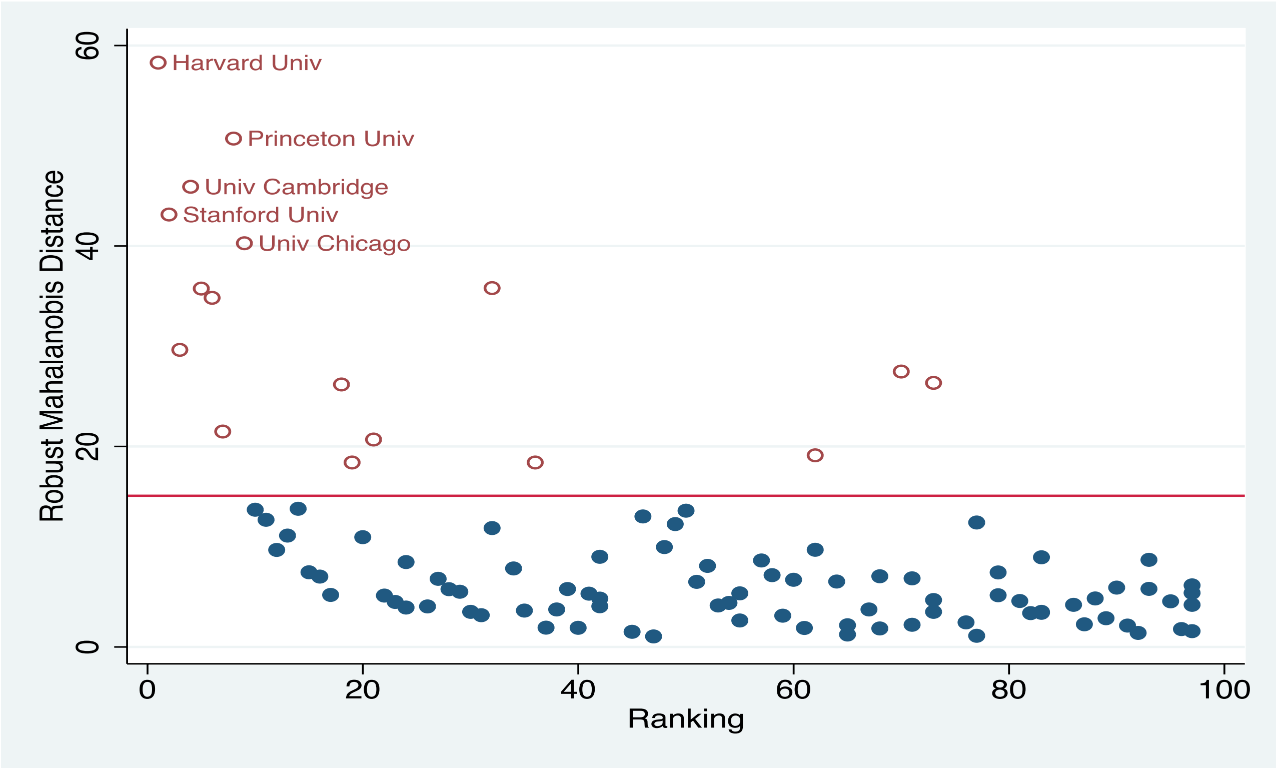
Harvard is an outlier ⇒ 18% of Φ_1 is due solely to Harvard

The Top 10 universities account for over 60% of Φ_1!



DETECTION OF OUTLIERS - Robust distances:

$$RD_i = \sqrt{((x_i - T(X))'C(X)^{-1}(x_i - T(X)))}$$



ROBUST PCA based on RMCD ESTIMATORS
(Croux and Haesbroeck, 2000)

IDEA : Robustify matrix of correlations by working with robust estimators (MCD, RMCD).

Suppose that $p = 2$ for simplicity: $Z = (X, Y) \in I\!R^2$, with

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{bmatrix} \Longrightarrow \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The generalized variance (Wilks, 1932) defined as:

$$\det(\Sigma) = \sigma_X^2 \sigma_Y^2 - \sigma_{YX}^2$$

can be seen as a generalization of the variance.

Minimum Covariance Determinant Estimator (Rousseeuw, 1985):

MCD estimators T_n and C_n: For the sample $\{z_1, \ldots, z_n\}$, select that subsample $\{z_{i_1}, \ldots, z_{i_h}\}$ of size $h \quad (h \leq n)$ with minimum determinant of its covariance matrix. Then compute sample covariance estimator over that subsample. Take $h \approx \frac{n}{2}$.

RMCD estimators are defined by

$$\begin{aligned} T_n^R &= \frac{\sum_{i=1}^n w_i z_i}{\sum_{i=1}^n w_i} \\ C_n^R &= c_2 \frac{\sum_{i=1}^n w_i (z_i - T_n^R)(z_i - T_n^R)^t}{\sum_{i=1}^n w_i} \end{aligned}$$

where c_2 is a consistency constant and the weight are given by

$$w_i = \begin{cases} 1 & \text{si } (z_i - T_n)^t C_n^{-1} (z_i - T_n) \leq q_\delta \\ 0 & \text{otherwise} \end{cases}$$

Two underlying factors are uncovered:
•Φ_1^R explains 38% of inertia
•Φ_2^R explains 28% of inertia

But what do these two factors represent??

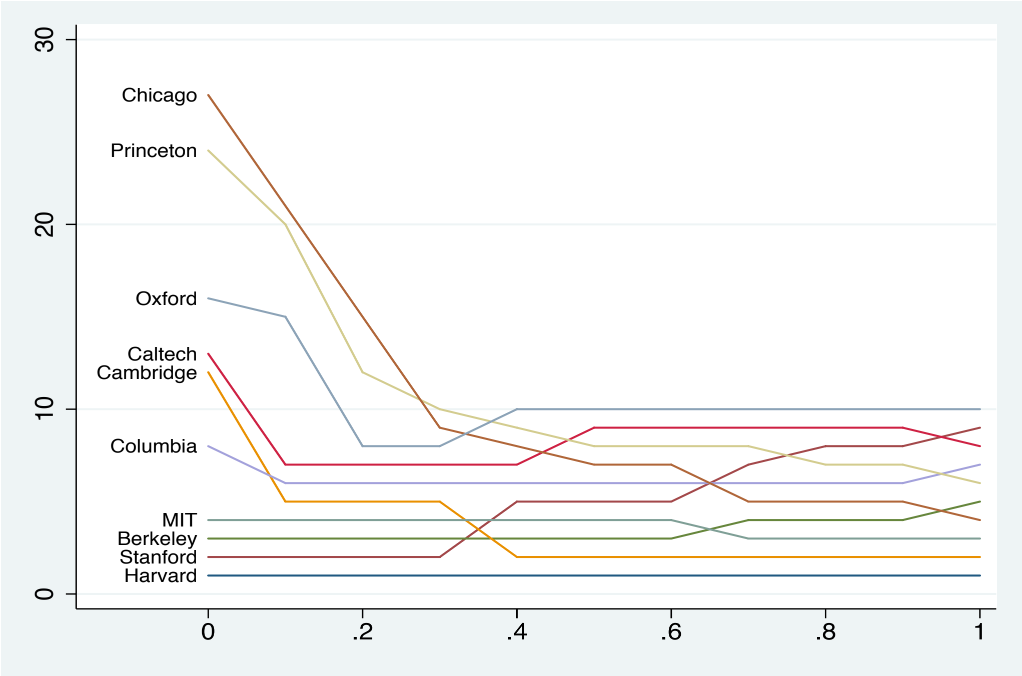
Variable	Corr($\phi_1, .$)	Corr($\phi_2, .$)
Alumni	-20%	80%
Awards	-25%	82%
HiCi	87%	7%
N&S	77%	22%
PUB	68%	-1%
Total score	75%	64%

Highly sensitivity to the weights attributed to the variables $\Rightarrow$

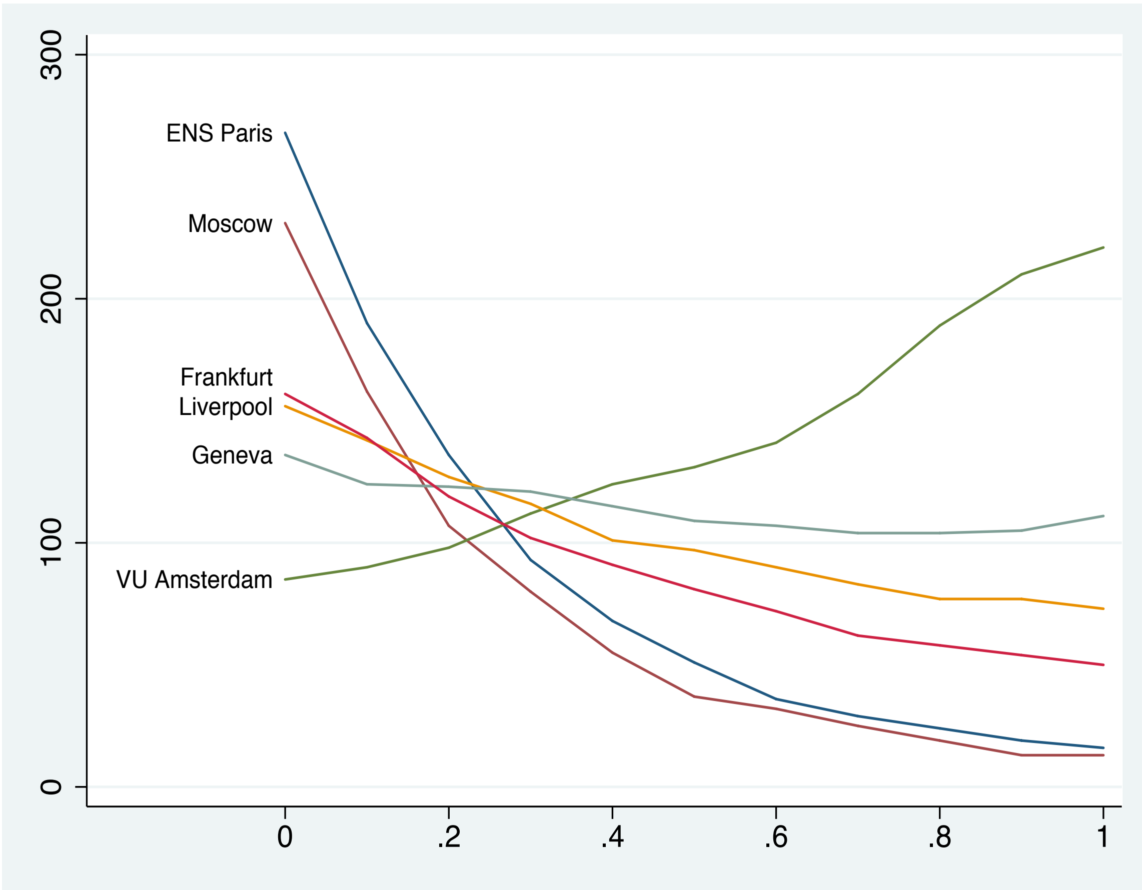
$$SCORE_i = w_i * (Alumni + Award) + (1 - w_i) * (HiCi + N\&S + PUB)$$

with $w_i = 0, 0.1, \ldots, 1$

Example 1: TOP 10



Example 2: Some european universities



USE RANKINGS WITH CAUTION!!

3.5 References

- Dehon, C. , Droesbeke, J-J. et Vermandele C. (2008), *Eléments de statistique*, Bruxelles, Editions de L'Unviversité de Bruxelles
- Joliffe I. T. (1986), *Principal Component Analysis*, 2nd edition, New York Springer.
- Hotelling H. (1933), "Analysis of a complex statistical variable into principal component", *J. Edu. Psy.* , Vol 24, 417-441 and 498-520.
- Pearson K. (1901), "On lines and planes of closest fit to systems of points in space", *Phil. Mag.*,2, 11, 559-572
- Rao C.R. (1964), "The use and interpretation of principal components analysis in applied research", *Sankhya*, serie A, Vol 26, 329-357

Chapitre 4

Correspondence analysis (CA)

4.1 Introduction

- Method that displays and summarizes the information contained in a dataset with qualitative type of variables
- CA is conceptually similar to PCA
- Can be divided into 2 areas:
  - Binary correspondence analysis (BCA): Technique that displays the rows and the columns of a two-way contingency table
  - Multiple correspondence analysis (MCA): Extension of BCA to more than 2 variables

Goals of BCA

Study the associations between the categories of two qualitative variables using the two-way contingency table:

2 qualitative (categorical) variables X and Y:

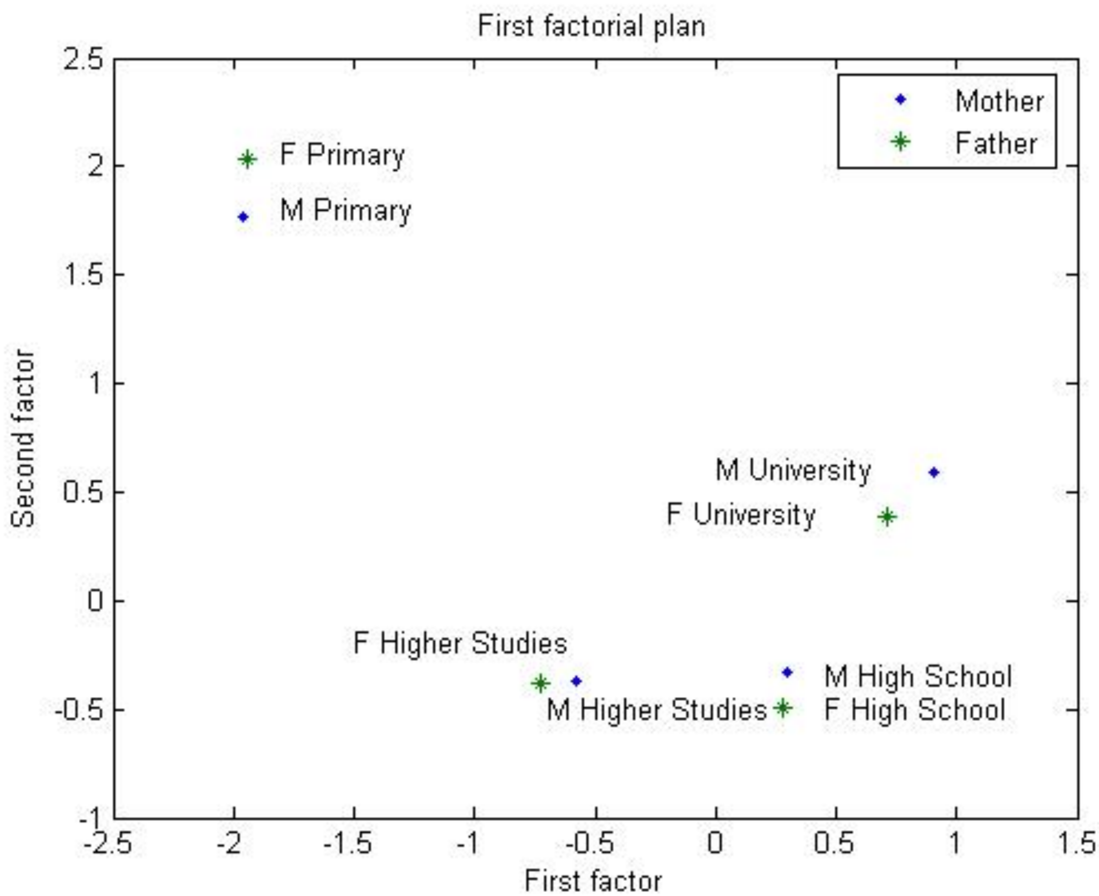
- X has J categories (or modalities): $A_1, \ldots, A_J$
- Y has K categories (or modalities): $B_1, \ldots, B_K$.

Examples

1. In education, can we suppose that the variables concerning work/study habits of students (regularity and work during the exam) are coherent?

2. In a research in education can we suppose that the father's level of education will tend to be very close to the level of education of the mother?

For the students in ULB, the answer is positive



The methodology can be summed up as follows:

- Step 1: Perform PCA on the table of row profiles where the A_j ($j \in 1, \ldots, J$) play the role of individuals and the B_k ($k \in 1, \ldots, K$) the role of variables

- Step 2: Perform PCA on the table of column profiles where the B_k ($k \in 1, \ldots, K$) play the role of individuals and the A_j ($j \in 1, \ldots, J$) the role of variables

- Step 3: Study the links between both PCAs

- Step 4: Plot graphs to show the proximity between row profiles, the proximity between column profiles and put forward the relationship between rows and columns.

Generalization of PCA in two directions :

- The weight associated to each individual (category) depends on the following frequencies:
  - Step 1: the weight allocated to the individual (category) A_j is equal to the frequency of this category ($f_{j.}$)
  - Step 2: the weight assigned to the individual (category) B_k is equal to the frequency of this category ($f_{.k}$)
- In PCA, the distance between observations corresponds to Euclidean distance. In correspondance analysis the distance between modalities corresponds to chi square type of distance

4.2 Example

Survey on 1000 workers:

- Variable X: "Diploma"
  3 categories: A_1, A_2, A_3 (Primary school, High school, University)
- Variable Y: "Salary"
  3 categories: B_1, B_2, B_3 (low, middle, high)

Two-way contingency table:

n_{jk}	B_1	B_2	B_3	$n_{j.}$
A_1	150	40	10	200
A_2	190	350	60	600
A_3	10	110	80	200
$n_{.k}$	350	500	150	1000

Notations

2 qualitative (categorical) variables X and Y:

- X has J categories (or modalities): $A_1, \ldots, A_J$
- Y has K categories (or modalities): $B_1, \ldots, B_K$.

A sample of size is n leads to the following two-way contingency table:

$X\|Y$	B_1	$\ldots$	B_k	$\ldots$	B_K	$\sum_{k=1}^{K}$
A_1	n_{11}	$\ldots$	n_{1k}	$\ldots$	n_{1K}	$n_{1.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	
A_j	n_{j1}	$\ldots$	n_{jk}	$\ldots$	n_{jK}	$n_{j.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	
A_J	n_{J1}	$\ldots$	n_{Jk}	$\ldots$	n_{JK}	$n_{J.}$
$\sum_{j=1}^{J}$	$n_{.1}$	$\ldots$	$n_{.k}$	$\ldots$	$n_{.K}$	n

where n_{jk} counts the number of individuals that are in category A_j for the variable X **and** in category B_k for the variable Y

Remark: $n_{j.} = \sum_{k=1}^{K} n_{jk}$ et $n_{.k} = \sum_{j=1}^{J} n_{jk}$

4.3 Explonatory analysis

Two-way contingency table of relative frequencies F:

Proportion of individuals that belong to category A_j for the variable X **and** into category B_k for the variable Y

$$f_{jk} = \frac{n_{jk}}{n} \qquad (j = 1 \ldots, J; k = 1, \ldots, K).$$

f_{jk}	B_1	B_2	B_3	$f_{j.}$
A_1	0.15	0.04	0.01	0.20
A_2	0.19	0.35	0.06	0.60
A_3	0.01	0.11	0.08	0.20
$f_{.k}$	0.35	0.50	0.15	1

The marginal frequencies are given by:

$$f_{j.} = \frac{n_{j.}}{n} \qquad (j = 1 \ldots, J)$$

and

$$f_{.k} = \frac{n_{.k}}{n} \qquad (k = 1, \ldots, K).$$

To formalize the notion of independence between the two variables X and Y, let us consider that:

f_{jk} is the estimation of

$$\pi_{jk} = P(X \in A_j, Y \in B_k)$$

$f_{j.}$ is the estimation of $\pi_{j.} = P(X \in A_j)$

$f_{.k}$ is the estimation of $\pi_{.k} = P(Y \in B_k)$

Tables of conditional frequencies:

• Table of row profiles:

Proportion of individuals that belong to category B_k for the variable Y among the individuals that have the modality A_j for the variable X:

$$f_{k|j} = \frac{n_{jk}}{n_{j.}} = \frac{n_{jk}/n}{n_{j.}/n} = \frac{f_{jk}}{f_{j.}} \quad (j \text{ fixed}; k = 1, \ldots, K).$$

$f_{k|j}$ is the estimation of $P(Y \in B_k | X \in A_j)$

$\frac{f_{jk}}{f_{j.}}$	B_1	B_2	B_3	
A_1	0.75	0.20	0.05	1
A_2	0.32	0.58	0.10	1
A_3	0.05	0.55	0.40	1
$f_{.k}$	0.35	0.50	0.15	1

• Table of column profiles:

Proportion of individuals that belong to category A_j for the variable X among the individuals that have the modality B_k for the variable Y:

$$f_{j|k} = \frac{n_{jk}}{n_{.k}} = \frac{n_{jk}/n}{n_{.k}/n} = \frac{f_{jk}}{f_{.k}} \quad (j = 1, \ldots, J; k \text{fixed}).$$

$f_{j|k}$ is the estimation of $P(X \in A_j | Y \in B_k)$

$\frac{f_{jk}}{f_{j.}}$	B_1	B_2	B_3	$f_{j.}$
A_1	0.43	0.08	0.07	0.20
A_2	0.54	0.70	0.40	0.40
A_3	0.03	0.22	0.53	0.20
	1	1	1	1

Independence between X and Y

• Two random variables X and Y are independent iff $\forall j \in \{1, \ldots, J\}$ and $\forall k \in \{1, \ldots, K\}$:

$a) P(X \in A_j, Y \in B_k) = P(X \in A_j) P(Y \in B_k)$

$b) P(Y \in B_k | X \in A_j) = P(Y \in B_k)$

$c) P(X \in A_j | Y \in B_k) = P(X \in A_j)$

• At the sample level, these equalities can be estimated by:

$$a\ ) f_{jk} \approx f_{j.} f_{.k} \quad \forall j \in \{1, \ldots, J\}\ \forall k \in \{1, \ldots, K\}$$

$$b\ ) f_{k|j} = \frac{f_{jk}}{f_{j.}} \approx f_{.k} \qquad \forall j,\ \forall k$$

$$c\ ) f_{j|k} = \frac{f_{jk}}{f_{.k}} \approx f_{j.} \qquad \forall j,\ \forall k.$$

We can therefore define the theoretical frequencies and relative frequencies under the assumption of independence as follows:

$$f^*_{jk} = f_{j.}f_{.k} \qquad \text{and} \qquad n^*_{jk} = nf^*_{jk} = \frac{n_{j.}n_{.k}}{n}$$

Observed frequencies

n_{jk}	B_1	B_2	B_3	$n_{j.}$
A_1	150	40	10	200
A_2	190	350	60	600
A_3	10	110	80	200
$n_{.k}$	350	500	150	1000

Theoretical frequencies under independence

n^*_{jk}	B_1	B_2	B_3	$n_{j.}$
A_1	70	100	30	200
A_2	210	300	90	600
A_3	70	100	30	200
$n_{.k}$	350	500	150	1000

Observed relative frequencies

f_{jk}	B_1	B_2	B_3	$f_{j.}$
A_1	0.15	0.04	0.01	0.20
A_2	0.19	0.35	0.06	0.60
A_3	0.01	0.11	0.08	0.20
$f_{.k}$	0.35	0.50	0.15	1

Theoretical relative frequencies under independence

f^*_{jk}	B_1	B_2	B_3	$f_{j.}$
A_1	0.07	0.10	0.03	0.20
A_2	0.21	0.30	0.09	0.60
A_3	0.07	0.10	0.03	0.20
$f_{.k}$	0.35	0.50	0.15	1

Attraction/repulsion matrix D

• The element jk of the Attraction/repulsion matrix D $(J \times K)$ is defined by:

$$d_{jk} = \frac{n_{jk}}{n^*_{jk}} = \frac{f_{jk}}{f^*_{jk}} = \frac{f_{jk}}{f_{j.}f_{.k}}$$

• Interpretations:

$$d_{jk} > 1 \iff f_{jk} > f_{j.}f_{.k}$$

$$f_{jk} > f_{j.}f_{.k} \iff f_{k|j} > f_{.k} \text{ and } f_{j|k} > f_{j.}$$

$\rightarrow$ The modalities (categories) A_j and B_k are attracted to each other

$$d_{jk} < 1 \iff f_{jk} < f_{j.}f_{.k}$$

$$f_{jk} < f_{j.}f_{.k} \iff f_{k|j} < f_{.k} \text{ and } f_{j|k} < f_{j.}$$

$\rightarrow$ The modalities (categories) A_j and B_k are repulse to each other

Example

f_{jk}	B_1	B_2	B_3	f^*_{jk}	B_1	B_2	B_3
A_1	0.15	0.04	0.01	A_1	0.07	0.10	0.03
A_2	0.19	0.35	0.06	A_2	0.21	0.30	0.09
A_3	0.01	0.11	0.08	A_3	0.07	0.10	0.03

d_{jk}	B_1	B_2	B_3
A_1	2.14	0.40	0.33
A_2	0.90	1.16	0.67
A_3	0.14	1.10	2.67

- High salary is more frequent for people with university diploma
- High salary is less frequent for people with at most a primary diploma
- Low salary is less frequent for people with university diploma
- …

Measures of association

• **The χ^2 statistic**:

Conditions for application:

$$\begin{aligned} n &\geq 30 \\ n^*_{jk} &\geq 1 \quad \forall j,k \\ \text{at least } 80\% \text{ of } n^*_{jk} &\geq 5 \end{aligned}$$

If these conditions are not met $\Longrightarrow$ group classes (modalities).

Statistic of test:

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(n_{jk}-n^*_{jk})^2}{n^*_{jk}}$$

Reject the null hypothesis (independence between X and Y) at the level $\alpha\%$ if

$$\chi^2 > \chi^2_{(J-1)(K-1);1-\alpha}$$

- **The statistic** $\phi^2 = \frac{\chi^2}{n}$:

$$\phi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(f_{jk}-f_{jk}^*)^2}{f_{jk}^*} = \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(\frac{n_{jk}}{n}-\frac{n_{jk}^*}{n})^2}{\frac{n_{jk}^*}{n}}$$

Remark: Using weights for the attraction/repulsion indices ($\sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}^* = 1$):

$$\begin{aligned}\bar{d} &= \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}^* d_{jk} = \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}^*\frac{f_{jk}}{f_{jk}^*}\\ &= \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} = 1\\ s_d^2 &= \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}^*(d_{jk}-1)^2 = \frac{\chi^2}{n} = \phi^2\end{aligned}$$

$\Longrightarrow$ The dispersion of the attraction/repulsion indices (around the mean) is given by ϕ^2

4.4 Analysis of row profiles

The point cloud $\aleph_l$ of row profiles

• At each line A_j of the table of row profiles is associated a point L_j in IR^K with coordinates:

$$\underline{l}_j = (f_{1|j}, \ldots, f_{k|j}, \ldots, f_{K|j})'.$$

• A weight $f_{j.}$ (% of individuals that have the modality A_j) is associated with the row profile $\underline{l}_j$ ($j \in \{1, \ldots, J\}$)

$\Longrightarrow$ The point cloud $\aleph_l$ of observations in IR^K contains J weighted row profiles:

$$\aleph_l = \{(L_1; f_{1.}), (L_2; f_{2.}), \ldots, L_J; f_{J.})\}.$$

Center of gravity of $\aleph_l$

The coordinates of the center of gravity are given by a weighted mean of the J row profiles:

$$\underline{g}_l = \sum_{j=1}^{J} f_{j.}\, \underline{l}_j$$

Consequently, the coordinate k of g_l is :

$$\sum_{j=1}^{J} f_{j.} f_{k|j} = \sum_{j=1}^{J} f_{j.} \frac{f_{jk}}{f_{j.}} = \sum_{j=1}^{J} f_{jk} = f_{.k}$$

$$\Downarrow$$

$$\underline{g}_l = (f_{.1}, \ldots, f_{.K})'$$

The center of gravity G_l of the J (weighted) row profiles is equal to the marginal profile (% of individuals having the modality B_k).

The χ^2 distance in IR^K

• Definition: The χ^2 distance in IR^K between two points X and Y with coordinates $(x_1, \ldots, x_K)$ and $(y_1, \ldots, y_K)$ is given by:

$$d^2_{\chi^2}(X,Y) = \sum_{k=1}^{K} \frac{(x_k - y_k)^2}{f_{.k}}$$

The euclidian distance gives the same weight to each column. The χ^2 distance gives the same relative importance to each column proportionally to the frequency B_k

Total inertia of $\aleph_l$

Total inertia based on the χ^2 distance and the weighted row profiles in IR^K:

$$
\begin{aligned}
I_{\chi^2}(\aleph_l, G_l) &= \sum_{j=1}^{J} f_{j.} d^2_{\chi^2}(L_j, G_l) \\
&= \sum_{j=1}^{J} f_{j.} \sum_{k=1}^{K} \frac{1}{f_{.k}} (f_{k|j} - f_{.k})^2 \\
&= \sum_{j=1}^{J} f_{j.} \sum_{k=1}^{K} \frac{1}{f_{.k}} (\frac{f_{jk}}{f_{j.}} - f_{.k})^2 \\
&= \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{f_{j.}}{f_{.k}} (\frac{f_{jk} - f_{j.} f_{.k}}{f_{j.}})^2 \\
&= \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{.k} f_{j.})^2}{f_{j.} f_{.k}} \\
&= \phi^2 = \frac{\chi^2}{n}
\end{aligned}
$$

$\Longrightarrow$ This explains why this distance is called the chi square distance!

Interpretation of the inertia :

• It measures the dependence between the two qualitative variables X and Y

• This measure is independent of the sample size n

• $I_{\chi^2}(\aleph_l, G_l) = 0$ means that all row profiles $L_1, \dots, L_J$ are equal to the center of gravity G_l:

$$\forall k \in \{1, \dots, K\} \quad \text{et} \quad \forall j \in \{1, \dots, J\}$$

$$\begin{aligned} f_{k|j} &= f_{.k} \\ \frac{f_{jk}}{f_{j.}} &= f_{.k} \\ f_{jk} &= f_{j.}f_{.k} \end{aligned}$$

leading to the independence of X and Y.

4.5 Step 1: PCA on the row profiles $\aleph_l$

Same methodology than PCA applied to quantitative variables with two modifications:

• The weights of "individuals (categories)" are not the same: the weight of A_j is equal to $f_{j.}$

• The distance used to measure the proximity between two "individuals" is the χ^2 distance.

$$\Downarrow$$

The PCA is not directly applied to the initial point cloud $\aleph_l$:

$$\aleph_l = \{(L_1, f_{1.}), \ldots, (L_J, f_{J.})\}$$

but on a normalized point cloud $\aleph_l^*$:

$$\aleph^* = \{(L_1^*, f_{1.}), \ldots, (L_J^*, f_{J.})\}$$

where the coordinates of L_j^* are given by:

$$\underline{l}_j^* = (\frac{f_{j1}}{f_{j.}\sqrt{f_{.1}}} - \sqrt{f_{.1}}, \ldots, \frac{f_{jK}}{f_{j.}\sqrt{f_{.K}}} - \sqrt{f_{.K}})'$$

The center of gravity of $\aleph_l^*$ is the origin

First projecting direction Δ_1

The first projecting direction Δ_1 is the direction passing through the origin that "fits in an optimal way" the point cloud $\aleph_l^*$ in terms of inertia:

$$I(\aleph_l^*, \Delta_1) = \min_{\Delta\text{:direction through the origin}} I(\aleph_l^*, \Delta)$$

where $I(\aleph_l^*, \Delta) = \sum_{j=1}^{J} f_{j.} d^2(L_j^*, P_\Delta(L_j^*))$.

Problem: Find the direction given by the vector u_1 such that $I(0, P_{\Delta_1}(L_j^*))$ is maximized:

$$\max \sum_{j=1}^{J} f_{j.} d^2(0, P_{\Delta_1}(L_j^*))$$

under the constriant

$$\|u_1\| = 1$$

It is again a problem of maximization under constraint, and as in PCA, the solution is given by the eigenvalues and eigenvectors of the matrix:

$$V = \sum_{j=1}^{J} f_{j.} \underline{l}_j^* (\underline{l}_j^*)'$$

$\Longrightarrow u_1$ is the eigenvector associated with the largest eigenvalue $\lambda_1 = I(0, P_{\Delta_1}(L_j^*))$.

Note that the element (k, k') of the matrix $V(K \times K)$ is given by :

$$v_{kk'} = \sum_{j=1}^{J} \left(\frac{f_{jk} - f_{j.} f_{.k}}{\sqrt{f_{j.} f_{.k}}} \right) \left(\frac{f_{jk'} - f_{j.} f_{.k'}}{\sqrt{f_{j.} f_{.k'}}} \right)$$

which yields $V = X'X$ with elements of $X(J \times K)$ given as:

$$x_{jk} = \frac{f_{jk} - f_{j.} f_{.k}}{\sqrt{f_{j.} f_{.k}}}$$

First principal component

To create the first principal component Φ_1, the point cloud $\aleph_l^*$ is projected on Δ_1:

$$P_{\Delta_1}(\aleph_l^*) = \{P_{\Delta_1}(L_1^*), \ldots, P_{\Delta_1}(L_J^*)\}.$$

The coordinate for each point associated with modality A_j ($\forall j = 1, \ldots, J$) is given by:

$$\begin{aligned}\phi_{1,j} &= \|OP_{\Delta_1}(L_j^*)\| =< OL_j^*, u_1 >= \sum_{k=1}^{K} u_{1,k}(\underline{l}_j^*)_k \\ &= u_{1,1}(\underline{l}_j^*)_1 + u_{1,2}(\underline{l}_j^*)_2 + \ldots + u_{1,K}(\underline{l}_j^*)_K\end{aligned}$$

Then $\phi_{1,j}$ is the value of the row profile j (associated with A_j) on the first principal component.

It can be proven that

- ϕ_1 is centered: $\sum_{j=1}^{J} f_{j}.\phi_{1,j} = 0$
- the variance of ϕ_1 is equal to λ_1

Global quality of the first principal component

Using the decomposition of total inertia, it can be shown that the percentage of inertia that is kept by projecting on Δ_1 is given by :
$\dfrac{\lambda_1}{\phi^2}$ since $I(\aleph_l^*, 0) = I(\aleph_l^*, \Delta_1) + I(0, P_{\Delta_1}(L_j^*))$

Contribution of modality A_j ($j = 1, \ldots, J$)

Knowing that

$$\lambda_1 = s^2_{\phi_1} = \sum_{j=1}^{J} f_{j.}\phi^2_{1,j} = \sum_{j=1}^{J} f_{j.}d^2(0, P_{\Delta_1}(L_j^*))$$

the contribution of the modality A_j is given by:

$$CTR_{\Delta_1}(A_j) = \frac{f_{j.}\phi^2_{1,j}}{\lambda_1}.$$

$\Longrightarrow$ The interpretation of ϕ_1 is mainly based on modalities A_j that have a high contribution

Quality of representation on the first axis

The quality of representation of the row profile L_j^* on the first axis Δ_1 is measured by the squared cosine of the angle formed by the vector OL_j^* and the axis Δ_1:

$$\cos^2(OL_j^*, \Delta_1) = \left(\frac{< OL_j^*, u_1 >}{\|OL_j^*\| \|u_1\|} \right)^2 = \frac{\phi_{1,j}^2}{\|OL_j^*\|^2}.$$

This formula does not contain the weight $f_{j.}$

$\Longrightarrow$ one modality can be:

- close to the axis Δ_1 and and therefore be well represented (well explained)
- because of a low weight $f_{j.}$, it can have a low contribution to the axis

Extended dimensions

The second projecting axis Δ_2 is defined by the vector u_2:

- through the origin (the center of gravity)
- orthogonal to u_1 ($u_2 \perp u_1$)
- minimizing the residual inertia

$\Longrightarrow u_2$ is the eigenvector of V associated to the second largest eigenvalues λ_2.

In the same way, we can find the other projecting axis $\Delta_3, \Delta_4, \ldots$

How many principal components ?

$\aleph_l^*$ is contained in a space of dimension

$$H \leq min(J-1, K-1)$$

where H is equal to the rank of the matrix V $(K \times K)$

$$\Downarrow$$

at most H orthogonal projecting directions

4.6 Step 2: PCA on the column profiles $\aleph_c$

The previous results and definitions based on the point cloud $\aleph_l$ are directly transposable to the point cloud $\aleph_c$ of column profiles

The point cloud $\aleph_c$ in IR^J of the K column profiles is defined by:

$$\aleph_c = \{(C_1; f_{.1}), (L_2; f_{.2}), \ldots, (C_K; f_{.K})\}$$

where the point C_k in IR^J has coordinates:

$$\underline{\mathrm{c}}_k = (f_{1|k}, \ldots, f_{j|k}, \ldots, f_{J|k})'.$$

Instead of working directly with this point cloud, we prefer to transform it such that the center of gravity is the origin:

$$\aleph_c^* = \{(C_1^*, f_{.1}), \ldots, (C_K^*, f_{.K})\}$$

where C_j^* has the coordinates:

$$\underline{\mathrm{c}}_j^* = (\frac{f_{1|k}}{\sqrt{f_{1.}}} - \sqrt{f_{1.}}, \ldots, \frac{f_{J|k}}{\sqrt{f_{J.}}} - \sqrt{f_{J.}})'$$

Projecting directions

The projecting directions $\Gamma_1, \ldots, \Gamma_H$ of $\aleph_c^*$ are defined by the orthogonal eigenvectors $v_1, \ldots, v_H$ of the matrix

$$W = XX'$$

associated with $H(= \min(J-1, K-1))$ non zero eigenvalues $\lambda_1, \ldots, \lambda_H$. v_1 is associated with the largest eigenvalue, ...

The elements of the matrix $X(J \times K)$ are defined as:

$$x_{jk} = \frac{f_{jk} - f_{j.}f_{.k}}{\sqrt{f_{j.}f_{.k}}}$$

The eigenvalues of W are the same as the eigenvalues of V

Principal components

The principal components $\psi_1, \ldots, \psi_H$ are defined by $\forall k = 1, \ldots, K$::

$$\psi_{h,k} = \|OP_{\Gamma_h}(C_k^*)\| =< OC_k^*, v_h >= \sum_{j=1}^{J} v_{h,j}(\underline{c}_k^*)_j$$
$$= v_{h,1}(\underline{c}_k^*)_1 + v_{h,2}(\underline{c}_k^*)_2 + \ldots + v_{h,J}(\underline{c}_k^*)_J$$

Properties of principal components $\psi_1, \psi_2, \ldots, \psi_H$ $\forall h \in \{1, \ldots, H\}$:

- Principal components are centered:

$$\sum_{k=1}^{J} f_{.k}\psi_{h,k} = 0$$

- The variance of ψ_h is given by λ_h
- Principal components are uncorrelated.

Global quality of Γ_h

The percentage of inertia that is kept when projecting on Γ_h is given by

$$\frac{\lambda_h}{\phi^2}$$

Contribution of modality B_k, $j = 1, \ldots, J$

Knowing that

$$\lambda_h = s^2_{\psi_h} = \sum_{k=1}^{K} f_{.k} \psi^2_{h,k}$$

the contribution of the modality B_k is given by:

$$CTR_{\lambda_h}(B_k) = \frac{f_{.k}\psi^2_{h,k}}{\lambda_h}.$$

Quality of the representation of C_k^* on Γ_h

$$\cos^2(OC_k^*, \Gamma_h) = \left(\frac{< OC_k^*, v_h >}{\|OC_k^*\| \|v_h\|}\right)^2 = \frac{\psi^2_{h,k}}{\|OC_k^*\|^2}.$$

4.7 Step 3: Links between both PCAs

The analysis of point cloud $\aleph_c^*$ could be deduced from the analysis of point cloud $\aleph_l^*$ and vice versa.

$\Longrightarrow$ The possibility to study the associations between the two variables is due to the links between the two analysis.

Row profiles $\aleph_l^*$: IR^K	Column profiles $\aleph_c^*$: IR^J
(λ_h, u_h) where $h = 1, \ldots, H$	(λ_h, v_h) where $h = 1, \ldots, H$
are the eigenvalues and the eigenvectors of	
$V = X'X$	$W = XX'$
leading to the relations	
$Vu_h = \lambda_h u_h$	$Wv_h = \lambda_h v_h$
Hence we have	
$X'Xu_h = \lambda_h u_h$	$XX'v_h = \lambda_h v_h$
$XX'Xu_h = \lambda_h Xu_h$	$X'XX'v_h = \lambda_h X'v_h$
$WXu_h = \lambda_h Xu_h$	$VX'v_h = \lambda_h X'v_h$
$\Longrightarrow$	
Xu_h eigenvector of W	$X'v_h$ eigenvector of V
The norm of these vectors is given by	
$\Vert Xu_h \Vert = \sqrt{\lambda_h}$	$\Vert X'v_h \Vert = \sqrt{\lambda_h}$
the normed eigenvectors associated to λ_h are:	
$\frac{1}{\sqrt{\lambda_h}} Xu_h$	$\frac{1}{\sqrt{\lambda_h}} X'v_h$
To conclude, we have the following relations:	
$v_h = \frac{1}{\sqrt{\lambda_h}} Xu_h$	$u_h = \frac{1}{\sqrt{\lambda_h}} X'v_h$

These relations between both PCA leads (after some developments) to a relation between the attraction/repulsion index and the coordinates of modalities in the two new system.

The distance for the couple (A_j, B_k) to the independence situation is measured by:

$$\Rightarrow \frac{f_{jk}}{f_{j.}f_{.k}} = 1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,j} \psi_{h,k}$$

$$\Rightarrow d_{jk} = 1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,j} \psi_{h,k}$$

$$\Downarrow$$

We can visualize graphically the attraction/repulsion indices using the first principal plan (in a first approximation)

4.8 Graphical representations

4.8.1 Pseudo-barycentric representation

Superposition of both PCAs:

- the point cloud of row profiles $\aleph_l^*$ is projected on the first factorial plan (Δ_1, Δ_2)
- the point cloud of column profiles $\aleph_c^*$ is projected on the first factorial plan (Γ_1, Γ_2)

$\Longrightarrow$ Simultaneous representation of the modalities $\{A_1, \ldots, A_J\}$ and $\{B_1, \ldots, B_K\}$

The modality A_j is associated to A_j^* which has coordinates $(\phi_{1,j}, \phi_{2,j})'$ and the modality B_k is associated to B_k^* which has coordinates $(\psi_{1,k}, \psi_{2,k})'$.

Interpretation of projections on Δ_1, Γ_1

If $\cos^2(OL_j^*, \Delta_1)$ is close to one $\Longrightarrow$ the profil L_j^* is close to its projection $P_{\Delta_1}(L_j^*)$ on Δ_1

$$\Longrightarrow \underline{\mathrm{l}}_j^* = \sum_{h=1}^{H} \phi_{h,j} \underline{\mathrm{u}}_h \Longrightarrow \underline{\mathrm{l}}_j^* \approx \phi_{1,j} \underline{\mathrm{u}}_1$$

This implies that $\forall k \in \{1, \ldots, K\}$:

$$d_{jk} = \frac{f_{jk}}{f_{j.} f_{.k}} \approx 1 + \frac{1}{\sqrt{\lambda_1}} \phi_{1,j} \psi_{1,k}.$$

We can therefore say that:

- The modalities A_j and B_k are attracted to each other ($d_{jk} > 1$)

if $\phi_{1,j} > 0$ and $\psi_{1,k} > 0$

if $\phi_{1,j} < 0$ and $\psi_{1,k} < 0$

- The modalities A_j and B_k are repulse each other ($d_{jk} < 1$)

if $\phi_{1,j} > 0$ and $\psi_{1,k} < 0$

if $\phi_{1,j} < 0$ and $\psi_{1,k} > 0$

Interpretation of the first principal map

If $\cos^2(OL_j^*, (\Delta_1, \Delta_2))$ is close to one $\Longrightarrow$ the profil L_j^* is close to its projection $P_{(\Delta_1,\Delta_2)}(L_j^*)$

$$\Longrightarrow \underline{\mathrm{l}}_j^* = \sum_{h=1}^{H} \phi_{h,j}\underline{\mathrm{u}}_h \Longrightarrow \underline{\mathrm{l}}_j^* \approx \phi_{1,j}\underline{\mathrm{u}}_1 + \phi_{2,j}\underline{\mathrm{u}}_2$$

This implies that $\forall k \in \{1, \ldots, K\}$:

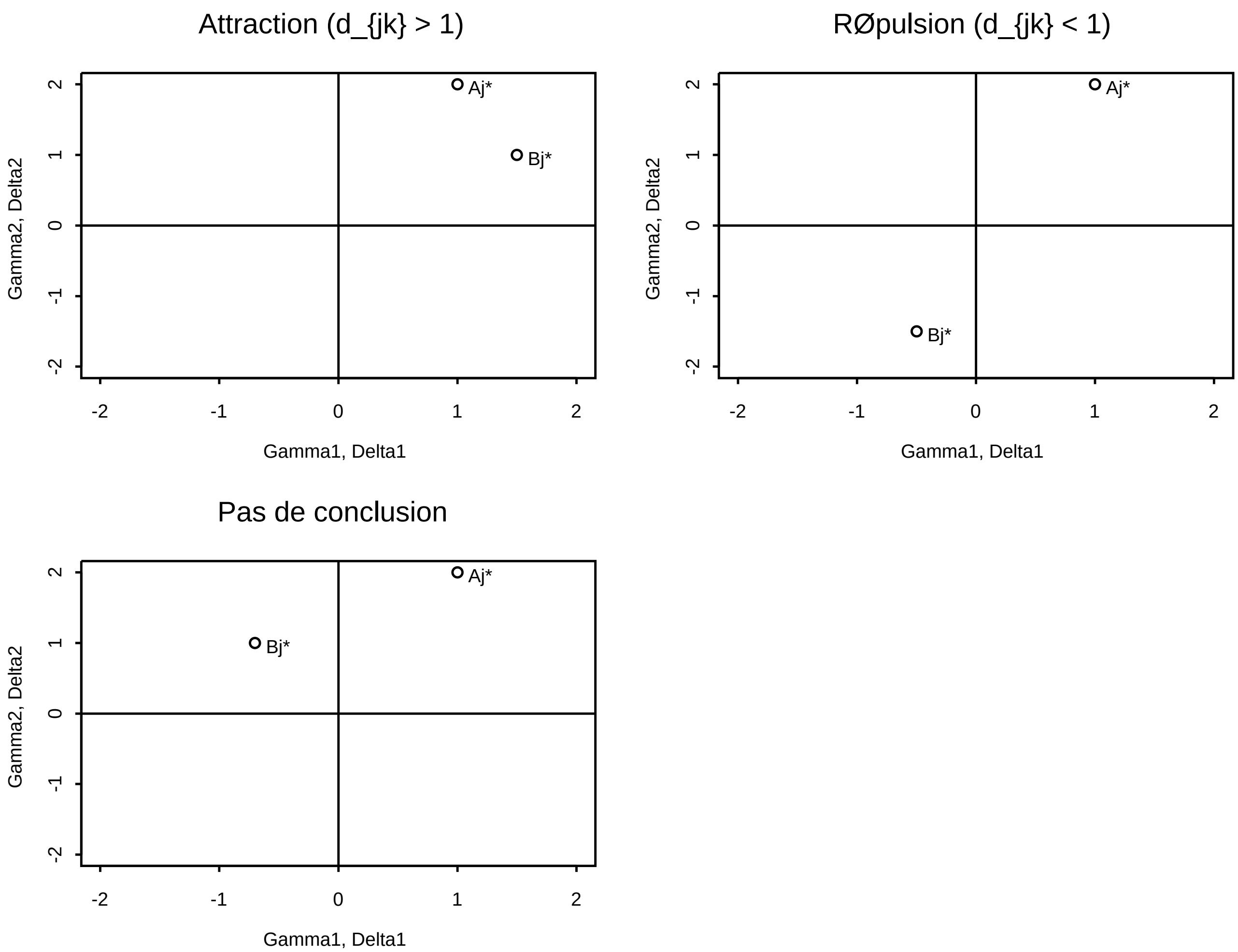
$$d_{jk} = \frac{f_{jk}}{f_{j.}f_{.k}} \approx 1 + \frac{1}{\sqrt{\lambda_1}}\phi_{1,j}\psi_{1,k} + \frac{1}{\sqrt{\lambda_2}}\phi_{2,j}\psi_{2,k}.$$

Therefore:

- The modalities A_j and B_k are attracted to each other ($d_{jk} > 1$) if A_j^* and $B_k^* \in$ are belong to the same quadrant

- The modalities A_j and B_k are repulse each other ($d_{jk} < 1$) if A_j^* and $B_k^* \in$ are in opposite quadrants

- We cannot conclude if A_j^* and $B_k^* \in$ belong to adjacent quadrants.



If a modality A_j^* is **well represented** on the first factorial plan, it is possible to determine graphically whether this modality is attracted or repulsed by some modalities B_k

4.8.2 Barycentric representation

In case of uncertainty about the attraction/repulsion between modalities, this representation can give an answer:

The attraction/repulsion indices are given by:

$$d_{jk} = 1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,j} \psi_{h,k}$$

$\Longrightarrow$ we are going to use the standardized principal components $\tilde{\psi}_h$ instead of ψ_h:

$$\tilde{\psi}_h = \frac{\psi_h}{\sqrt{\lambda_h}}.$$

$\Longrightarrow$ Superposition of both PCAs:

- the row profile A_j is associated to A_j^* which has coordinates $(\phi_{1,j}, \phi_{2,j})'$
- the column profile B_k is associated to $\tilde{B}_k^*$ which has coordinates $(\tilde{\psi}_{1,k}, \tilde{\psi}_{2,k})' = (\frac{\psi_{1,k}}{\sqrt{\lambda_1}}, \frac{\psi_{2,k}}{\sqrt{\lambda_2}})'$

Interpretation for the first factorial plan

If a modality A_j^* is **well represented** on the first principal plan Δ_1, Δ_2:

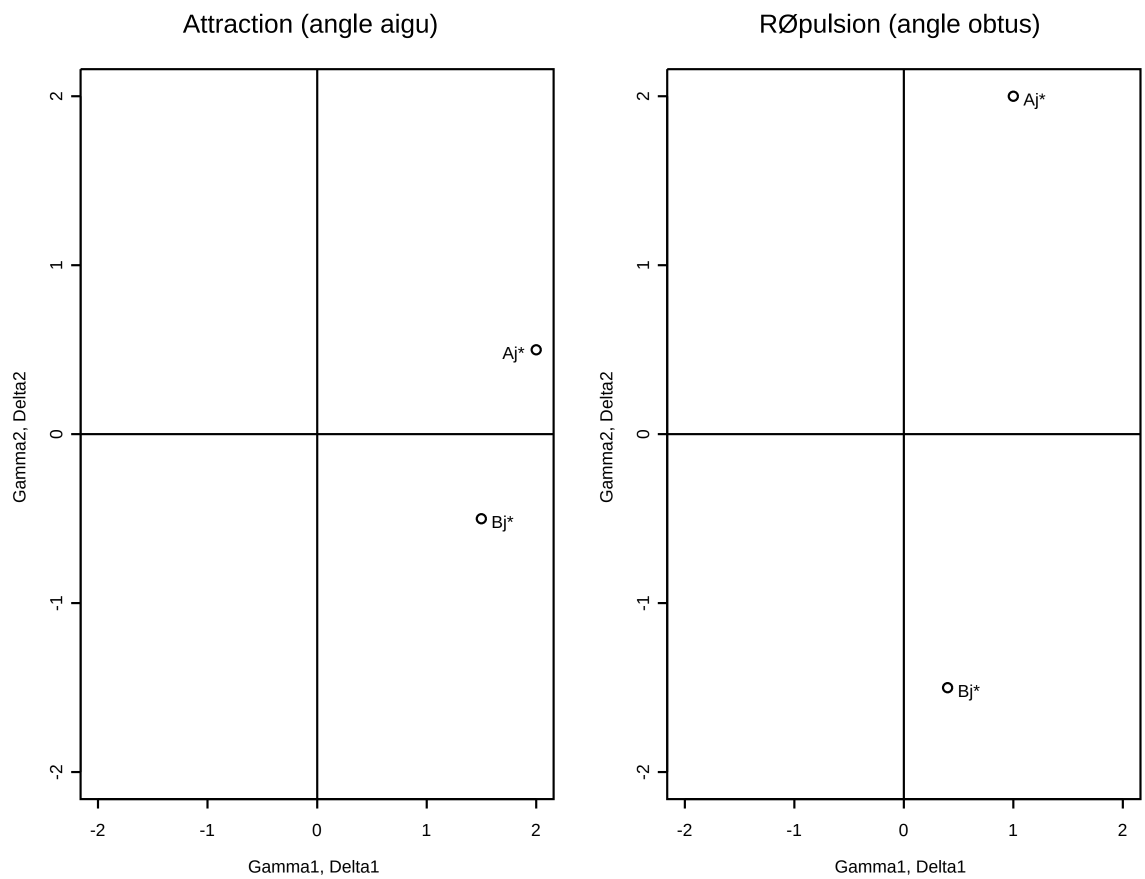
$$\begin{aligned} d_{jk} &\approx 1 + \phi_{1,j}\tilde{\psi}_{1,k} + \phi_{2,j}\tilde{\psi}_{2,k} \\ &\approx 1 + < OA_j^*, O\tilde{B}_k^* > \end{aligned}$$

where $< ., . >$ is the usual scalar product in IR^2

We can therefore say that:

The modalities A_j and B_k are attracted to each other ($d_{jk} > 1$) if the angle between OA_j^* and $O\tilde{B}_k^*$ is acute ($< OA_j^*, O\tilde{B}_k^* >$ is therefore positive)

The modalities A_j and B_k are repulse each other ($d_{jk} < 1$) if the angle between OA_j^* and $O\tilde{B}_k^*$ is obtuse ($< OA_j^*, O\tilde{B}_k^* >$ is therefore negative)



Examples where no conclusion can be drawn with the pseudo-barycentric representation. But with the barycentric representation, the rule is: Draw $A_j^{\perp}$ which passes through the origin and which is orthogonal to OA_j^*. This line separates the space into two parts: the modalities B_k that are on the same side than A_j^* are attracted by it and the modalities on the other side are repulsed by A_j^*.

4.8.3 Biplot

The angles between the modalities and the factors yield most of the information. We therefore introduce a new variable where the coordinates of row profiles are divided by $\sqrt{\lambda_1}$. This leads to a better visibility of the first principal plan.

$\Longrightarrow$ Simultaneous representation of the modalities $\{A_1, \ldots, A_J\}$ and $\{B_1, \ldots, B_K\}$ in the first principal map:

- The modality A_j is associated to $\tilde{A}_j^*$ which has coordinates $(\tilde{\phi}_{1,j}, \tilde{\phi}_{2,j})' = (\frac{\phi_{1,j}}{\sqrt{\lambda_1}}, \frac{\phi_{2,j}}{\sqrt{\lambda_1}})'$.

- The modality B_k is associated to $\tilde{B}_k^*$ which has coordinates $(\tilde{\psi}_{1,k}, \tilde{\psi}_{2,k})' = (\frac{\psi_{1,k}}{\sqrt{\lambda_1}}, \frac{\psi_{2,k}}{\sqrt{\lambda_2}})'$.

This type of standardization is called BIPLOT.

4.9 References

• Benzecri, (1973), *L'analyse des données.* Tome 1: *La taxinomie.* Tome 2: *L'analyse des correspondances* (2^{de}. éd. 1976). Dunod, Paris.

• Escofier and Pages (2008), *Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation.* Dunod, Paris.

• Hirschfeld, (1935), "A connection between correlation and contingency.", *Proc. Camb. Phil. Soc.*, **31**, 520-524.

• Guttman, (1941), *The quantification of a class of attributes: a theory and method of a scale construction.* In: The prediction of personal adjustement (Horst P., Ed.), 251-264, SSCR New York.

Chapitre 5

Multiple correspondence analysis (MCA)

• Extension of BCA to more than 2 variables.

• Goal: Analysis of a table $n \times P$ of "individuals $\times$ qualitative variables".

• Method: apply BCA to a table called "complete disjunctive table".

5.1 Data, tables and distances

5.1.1 The complete disjunctive table

Example

4 individuals: $n = 4$

3 variables: $P = 3$

- Y_1: gender $\longrightarrow$ 2 modalities: $K_1 = 2$ (male=1, female=2)

- Y_2: civil status $\longrightarrow$ 3 modalities: $K_2 = 3$ (single=1, married=2, divorced or widower=3)

- Y_3: level of education $\rightarrow$ 2 modalities: $K_3 = 2$ (primary or secondary school=1, higher or university diploma=2)

$K = K_1 + K_2 + K_3 = 2 + 3 + 2 = 7.$

Logic table (the modalities are coded)

$n \mid P$	Y_1	Y_2	Y_3
1	2	1	1
2	2	1	2
3	1	3	2
4	2	2	1

Complete disjunctive table (CDT)

	X_1		X_2			X_3		
	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	P
1	0	1	1	0	0	1	0	3
2	0	1	1	0	0	0	1	3
3	1	0	0	0	1	0	1	3
4	0	1	0	1	0	1	0	3
n_{pl}	1	3	2	1	1	2	2	12

Notations:

- n individuals, P variables: $Y_1, \ldots, Y_P$
- The variable Y_p has K_p modalities $\Longrightarrow K = \sum_{p=1}^{P} K_p$ total number of modalities in the dataset
- n_{pl} number of individuals having the modality l for the variable Y_p
- $x_{ipl} = 1$ if individual i has modality l of Y_p, 0 otherwise
- X_{pl} is a dummy (binary) variable which is associated with modality l of Y_p
- $X_p = (X_{p1}, \ldots, X_{pK_p})$ vectors of dummy variables of Y_p

The following relations hold:

$$\sum_{l=1}^{K_p} n_{pl} = n \text{ and } \sum_{p=1}^{P} \sum_{l=1}^{K_p} n_{pl} = nP$$

Table of dummy variables X_p associated to Y_p:

	1	$\dots$	l	$\dots$	K_p	$\sum_{l=1}^{K_p}$
1	x_{1p1}	$\dots$	x_{1pl}	$\dots$	x_{1pK_p}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
i	x_{ip1}	$\dots$	x_{ipl}	$\dots$	x_{ipK_p}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	x_{np1}	$\dots$	x_{npl}	$\dots$	x_{npK_p}	1
$\sum_{i=1}^{n}$	n_{p1}	$\dots$	n_{pl}	$\dots$	n_{pK_p}	n

Complete disjunctive table $X = (X_1, \dots, X_P)$:

x	1	$\dots$	p	$\dots$	P	$\sum_{p=1}^{P}\sum_{l=1}^{K_p}$
1		$\dots$		$\dots$		P
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
i	$x_1(n \times K_1)$	$\dots$	$x_p(n \times K_p)$	$\dots$	$x_P(n \times K_P)$	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n		$\dots$		$\dots$		P
$\sum_{i=1}^{n}$						nP

5.1.2 Row and column profiles, attraction/repulsion indices

MCA on $Y_1, \ldots, Y_P$ = BCA on the complete disjunctive table.

Relative frequencies of the complete disjunctive table:

	Y_1	$\ldots$	Y_p	$\ldots$	Y_P	
	$1 \ \ldots \ l \ \ldots \ K_1$	$\ldots$	$1 \ \ldots \ l \ \ldots \ K_p$	$\ldots$	$1 \ \ldots \ l \ \ldots \ K_P$	
1		$\ldots$		$\ldots$		$\frac{1}{n}$
$\vdots$		$\ldots$		$\ldots$		$\frac{1}{n}$
i		$\ldots$	$f_{ipl} = \frac{x_{ipl}}{nP}$	$\ldots$		$\frac{1}{n}$
$\vdots$		$\ldots$		$\ldots$		$\frac{1}{n}$
n		$\ldots$		$\ldots$		$\frac{1}{n}$
		$\ldots$	$f_{.pl} = \frac{n_{pl}}{nP}$	$\ldots$		1

where the marginal relative frequencies are given by:

$$f_{i..} = \frac{1}{n} \text{ and } f_{.pl} = \frac{n_{pl}}{nP}$$

Row profiles L_i of individual i: $l_i(1 \times K)$

$\Rightarrow$ the coordinate pl of the row profile i:

$$(\underline{l}_i)_{pl} = \frac{f_{ipl}}{f_{i..}} = \frac{x_{ipl}/nP}{1/n} = \frac{x_{ipl}}{P}$$

$\forall p = 1, \dots, P; \quad l = 1, \dots, K_p$

Column profile C_{pl} associated to the modality l of Y_p:

$$c_{pl}(n \times 1)$$

$\Rightarrow$ the coordinate i of the column profile pl:

$$(\underline{c}_{pl})_i = \frac{f_{ipl}}{f_{.pl}} = \frac{x_{ipl}/nP}{n_{pl}/nP} = \frac{x_{ipl}}{n_{pl}}$$

$\forall i = 1, \dots, n.$

Notations

$(\underline{l}_i)_{pl}$: coordinate pl of the row profile i

$(\underline{c}_{pl})_i$: coordinate i of the column profile pl

Example

- Row profiles table:

	X_1		X_2			X_3		
	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	
1	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	$\frac{1}{3}$	0	1
2	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$	1
3	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$	0	$\frac{1}{3}$	1
4	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0	1
	$\frac{1}{12}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	1

- Column profiles table:

	X_1		X_2			X_3		
	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	
1	0	$\frac{1}{3}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0	$\frac{1}{4}$
2	0	$\frac{1}{3}$	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	$\frac{1}{4}$
3	1	0	0	0	1	0	$\frac{1}{2}$	$\frac{1}{4}$
4	0	$\frac{1}{3}$	0	1	0	$\frac{1}{2}$	0	$\frac{1}{4}$
	1	1	1	1	1	1	1	1

Attraction/repulsion indices between individual i and modality l of Y_p:

$$d_{i,pl} = \frac{f_{ipl}}{f_{i.}.f_{.pl}} = \frac{\frac{x_{ipl}}{nP}}{\frac{1}{n}\frac{n_{pl}}{nP}} = \frac{x_{ipl}}{n_{pl}/n}$$

As $x_{ipl} = \{0, 1\}$ and $n_{pl}/n \leq 1$, we have that

$$d_{i,pl} = 0 \quad \text{if} \quad x_{ipl} = 0$$
$$d_{i,pl} = \frac{n}{n_{pl}} \geq 1 \quad \text{if} \quad x_{ipl} = 1$$

Interpretation: If one individual i has the modality l of the variable Y_p, then the attraction/repulsion index $d_{i,pl}$ increases as the modality l of the variable Y_p becomes rare (n_{pl} small).

5.1.3 Point cloud and distances between row profiles

Point cloud

- n row profiles $L_1, \ldots, L_n$
- in IR^K where $K = \sum_{p=1}^{P} K_p$
- with weight $1/n$
- and the χ^2 distance.

The center of gravity G_l has coordinate pl $(p = 1, \ldots, P; l = 1, \ldots, K_p)$ given by:

$$\sum_{i=1}^{n} \frac{1}{n} (\underline{l}_i)_{pl} = \frac{1}{nP} \sum_{i=1}^{n} x_{ipl} = \frac{n_{pl}}{nP}$$

$\Longrightarrow G_l$ is the marginal profile (marginal relative profile)

Properties

- Distance between individuals (row profiles)

$$
\begin{aligned}
d^2_{\chi^2}(L_{i_1}, L_{i_2}) &= \sum_{p=1}^{P} \sum_{l=1}^{K_p} \frac{1}{f_{.pl}}((\underline{1}_{i_1})_{pl} - (\underline{1}_{i_2})_{pl})^2 \\
&= \sum_{p=1}^{P} \sum_{l=1}^{K_p} \frac{1}{\frac{n_{pl}}{nP}}(\frac{x_{i_1pl}}{P} - \frac{x_{i_2pl}}{P})^2 \\
&= \frac{n}{P} \sum_{p=1}^{P} \sum_{l=1}^{K_p} \frac{1}{n_{pl}}(x_{i_1pl} - x_{i_2pl})^2
\end{aligned}
$$

Interpretation:
The distance between 2 individuals is small if they have many modalities that are the same.

Example

Distance between individual 1 (female, single with primary or secondary diploma) and 2 (female, single with a higher or university formation):

$$d^2_{\chi^2}(L_1, L_2) = \sum_{p=1}^{3} \sum_{l=1}^{K_p} \frac{1}{f_{.pl}}((\underline{l}_1)_{pl} - (\underline{l}_2)_{pl})^2$$

$$= 12(0-0)^2 + \frac{12}{3}(\frac{1}{3} - \frac{1}{3})^2$$

$$+ \frac{12}{2}(\frac{1}{3} - \frac{1}{3})^2 + \frac{12}{2}(0-0)^2 + 12(0-0)^2$$

$$+ 6(\frac{1}{3} - 0)^2 + 6(0 - \frac{1}{3})^2 = \frac{4}{3} = 1.33$$

Another way to compute it:

$$\begin{aligned} d^2_{\chi^2}(L_1, L_2) &= \frac{n}{P} \sum_{p=1}^{3} \sum_{l=1}^{K_p} \frac{1}{n_{pl}}(x_{i_1pl} - x_{i_2pl})^2 \\ &= \frac{4}{3}(1(0-0)^2 + \frac{1}{3}(1-1)^2 \\ &+ \frac{1}{2}(1-1)^2 + 1(0-0)^2 + 1(0-0)^2 \\ &+ \frac{1}{2}(1-0)^2 + \frac{1}{2}(0-1)^2) = \frac{4}{3} = 1.33 \end{aligned}$$

Matrix of distances and matrix of squared distances between individuals (row profiles)

$d^2_{\chi^2}(L_i, L_j)$	L_1	L_2	L_3	L_4
L_1	-	1.33	5.11	2.00
L_2	1.33	-	3.78	3.33
L_3	5.11	3.78	-	5.78
L_4	2.00	3.33	5.78	-

$d_{\chi^2}(L_i, L_j)$	L_1	L_2	L_3	L_4
L_1	-	1.15	2.26	1.41
L_2	1.15	-	1.94	1.83
L_3	2.26	1.94	-	2.40
L_4	1.41	1.83	2.40	-

Conclusions

- *individuals 1 and 2 are close to each other (both are female and single)*
- *individuals 1 and 3 are very different (all the modalities between those individuals are different).*

- Distance between the row profile L_i and the center of gravity:

$$\begin{aligned}
d^2_{\chi^2}(L_i, G_l) &= \sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{1}{f_{.pl}}((\underline{1}_i)_{pl} - \frac{n_{pl}}{nP})^2 \\
&= \sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{nP}{n_{pl}}(\frac{x_{ipl}}{P} - \frac{n_{pl}}{nP})^2 \\
&= \sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{n}{Pn_{pl}}\left(x_{ipl}^2 + \frac{n_{pl}^2}{n^2} - 2x_{ipl}\frac{n_{pl}}{n}\right) \\
&= \frac{n}{P}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{x_{ipl}}{n_{pl}} + \frac{1}{nP}\sum_{p=1}^{P}\sum_{l=1}^{K_p}n_{pl} - \frac{2}{P}\sum_{p=1}^{P}\sum_{l=1}^{K_p}x_{ipl} \\
&= \frac{n}{P}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{x_{ipl}}{n_{pl}} + \frac{1}{nP}nP - \frac{2}{P}P \\
&= \frac{n}{P}\sum_{p=1}^{P}\sum_{l=1}^{K_p}\frac{x_{ipl}}{n_{pl}} - 1
\end{aligned}$$

$\Longrightarrow$ *The distance between the individual i and the center of gravity G_l increases as the modalities taking by the individual i becomes rare ($x_{ipl} = 1$ and n_{pl} small).*

- Total inertia of point cloud $\aleph_l$ around G_l:

$$\begin{aligned}
I_{\chi^2}(\aleph_l, G_l) &= \sum_{i=1}^{n} f_{i..} d^2_{\chi^2}(L_i, G_l) \\
&= \sum_{i=1}^{n} \frac{1}{n} \left(\frac{n}{P} \sum_{p=1}^{P} \sum_{l=1}^{K_p} \frac{x_{ipl}}{n_{pl}} - 1 \right) \\
&= \frac{1}{P} \sum_{p=1}^{P} \sum_{l=1}^{K_p} \sum_{i=1}^{n} \frac{x_{ipl}}{n_{pl}} - \frac{1}{n} \sum_{i=1}^{n} 1 \\
&= \frac{1}{P} \sum_{p=1}^{P} \sum_{l=1}^{K_p} \frac{n_{pl}}{n_{pl}} - \frac{1}{n} \sum_{i=1}^{n} 1 \\
&= \frac{K}{P} - 1
\end{aligned}$$

where $\frac{K}{P}$ is the average number of modalities by variables $\Downarrow$

The total inertia depends only on the number of variables and on the number of modalities. It does not depend at all on the relations between the variables. From a statistical point of view, this quantity cannot be interpreted (as in PCA).

- $\forall i \in \{1, \ldots, n\}$ the row profile $\underline{l}_i$ satisfies the P linear constraints:

$$\sum_{l=1}^{K_p} (l_i)_{pl} = \sum_{l=1}^{K_p} \frac{x_{ipl}}{P} = \frac{1}{P} \qquad p = 1, \ldots, P$$

$\Longrightarrow$ *the point cloud* $\aleph_l$ *is inside a sub-space of at most* $K - P$ *dimensions.*

5.1.4 Point cloud and distances between column profiles

Point cloud

- $K = \sum_{p=1}^{P} K_p$ column profiles C_{pl}
- in IR^n
- with weight $f_{.pl} = \frac{n_{pl}}{nP}$
- and the χ^2 distance.

The i^{th} coordinate of the center of gravity G_c is given by:

$$\sum_{p=1}^{P}\sum_{l=1}^{K_p} f_{.pl}(\underline{c}_{pl})_i = \sum_{p=1}^{P}\sum_{l=1}^{K_p} \frac{n_{pl}}{nP}\frac{x_{ipl}}{n_{pl}} = \frac{1}{n}$$

$\Longrightarrow G_c$ is the marginal profile (marginal relative profile)

Properties

- Distance between modalities (column profiles)
  The χ^2 distance between modality l_1 of variable Y_{p1} and modality l_2 of variable Y_{p2} is:

$$
\begin{aligned}
d^2_{\chi^2}(c_{p1l1}, c_{p2l2}) &= \sum_{i=1}^{n} \frac{1}{f_{i..}}((\underline{c}_{p1l1})_i - (\underline{c}_{p2l2})_i)^2 \\
&= \sum_{i=1}^{n} \frac{1}{\frac{1}{n}}(\frac{x_{ip1l1}}{n_{p1l1}} - \frac{x_{ip2l2}}{n_{p2l2}})^2 \\
&= n \sum_{p=i}^{n}(\frac{x_{ip1l1}}{n_{p1l1}} - \frac{x_{ip2l2}}{n_{p2l2}})^2
\end{aligned}
$$

Interpretation:

- if the same individuals take these 2 modalities, the distance between the 2 modalities is small

- if a modality is rare, it is far away from the other modalities.

Example

Distance between modality 1 of Y_1 (male) and 2 of Y_2 (married):

$$d^2_{\chi^2}(c_{11}, c_{22}) = \sum_{i=1}^{n} \frac{1}{f_{i..}}((c_{11})_i - (c_{22})_i)^2$$

$$= 4\left((0-0)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2\right)$$

$$= 8$$

$d_{\chi^2}(,)$	11	12	21	22	23	31	32
11	-	2.31	2.45	2.83	0	2.45	1
12		-	0.67	0.94	2.31	0.67	1.37
21			-	2.45	2.45	1.41	1.41
22				-	2.83	1	2.45
23					-	2.45	1
31						-	2
32							-

- “12” and “21” are close to each other (50% of individuals have chosen these two modalities)

- Distance between the column profile C_{pl} and the center of gravity:

$$
\begin{aligned}
d^2_{\chi^2}(C_{pl}, G_c) &= \sum_{i=1}^{n} n((\underline{c}_{pl})_i - \frac{1}{n})^2 \\
&= \sum_{i=1}^{n} n(\frac{x_{ipl}}{n_{pl}} - \frac{1}{n})^2 \\
&= \sum_{i=1}^{n} n\frac{x_{ipl}^2}{n_{pl}^2} + \sum_{i=1}^{n} n\frac{1}{n^2} - 2\sum_{i=1}^{n} \frac{x_{ipl}}{n_{pl}} \\
&= \frac{n}{n_{pl}^2}\sum_{i=1}^{n} x_{ipl} + 1 - \frac{2}{n_{pl}}\sum_{i=1}^{n} x_{ipl} \\
&= \frac{n}{n_{pl}} - 1
\end{aligned}
$$

$\Longrightarrow$ *The distance between the modality l of Y_p and the center of gravity G_c increases as the modality becomes more rare (n_{pl} small).*

- Total inertia of point cloud $\aleph_c$ around G_c:

$$
\begin{aligned}
I_{\chi^2}(\aleph_c, G_c) &= \sum_{p=1}^{P}\sum_{l=1}^{K_p} f_{.pl} d^2_{\chi^2}(C_{pl}, G_c) \\
&= \sum_{p=1}^{P}\sum_{l=1}^{K_p} \frac{n_{pl}}{nP}(\frac{n}{n_{pl}} - 1) \\
&= \sum_{p=1}^{P}\sum_{l=1}^{K_p} \frac{1}{P}(1 - \frac{n_{pl}}{n}) \\
&= \sum_{p=1}^{P} \frac{1}{P}(K_p - 1) = \frac{1}{P}(K - P) \\
&= \frac{K}{P} - 1
\end{aligned}
$$

Notice that $I_{\chi^2}(\aleph_c, G_c) = 1$ if all the variables have exactly two modalities.

- Contribution of the modality l of the variable Y_p to the total inertia of the point cloud $\aleph_c$:

$$\begin{aligned} f_{.pl} d^2_{\chi^2}(C_{pl}, G_c) &= \frac{n_{pl}}{nP}(\frac{n}{n_{pl}} - 1) \\ &= \frac{1}{P} - \frac{n_{pl}}{nP} = \frac{1}{P}(1 - \frac{n_{pl}}{n}) \end{aligned}$$

$\Longrightarrow$ *The contribution of the modality l of the variable Y_p increases when n_{pl} decreases. A rare modality has therefore a larger impact than a common modality.*

- The contribution of the variable Y_p (sum of the contributions of the modalities) is given by:

$$\sum_{l=1}^{K_p} \frac{1}{P}(1 - \frac{n_{pl}}{n}) = \frac{1}{P}(K_p - 1)$$

$\Longrightarrow$ *The contribution of a variable increases with the number of modalities.*

$$\Downarrow$$

When doing a survey, it is better to take into account variables that have more or less the same number of modalities.

It is also adviced to avoid having rare modalities.

5.2 MCA

5.2.1 Projecting directions (similar results than BCA)

Row profiles

$\aleph_l = \{(L_1; \frac{1}{n}), \ldots, (L_n; \frac{1}{n})\}$ with χ^2 distances in IR^K where L_i has coordinates:

$$\underline{l}_i = \frac{x_{ipl}}{P} \qquad p = 1, \ldots, P; l = 1, \ldots, K_p$$

Column profiles

$\aleph_c = \{(C_{pl}; f_{.pl} = \frac{n_{pl}}{n})$ where $p = 1, \ldots, P$ and $l = 1, \ldots, K_p\}$ with χ^2 distances in IR^n where C_{pl} has coordinates:

$$\underline{c}_{pl} = \frac{x_{ipl}}{n_{pl}} \qquad i = 1, \ldots, n$$

Row profiles $\aleph_l^*$: IR^K	Columb profiles $\aleph_c^*$: IR^n
(λ_h, u_h) where $h = 1, \ldots, H$	(λ_h, v_h) where $h = 1, \ldots, H$
are the eigenvalues and the eigenvectors of	
$V = T'T$	$W = TT'$
Hence we have	
$Vu_h = \lambda_h u_h$	$Wv_h = \lambda_h v_h$

where T is a matrix $n \times K$ with coordinates:

$$t_{i,pl} = \frac{f_{ipl} - f_{i..}f_{.pl}}{\sqrt{f_{i..}f_{.pl}}} = \frac{x_{ipl} - \frac{n_{pl}}{n}}{\sqrt{Pn_{pl}}}$$

Construction of the principal components (projection of the row and column profiles):

$$\phi_{h,j} = \|OP_{\Delta_h}(L_j^*)\| =< OL_j^*, u_h >= \sum_{k=1}^{K} u_{h,k}(\underline{\mathrm{l}}_j^*)_k$$

$$\psi_{h,pl} = \|OP_{\Gamma_h}(C_{pl}^*)\| =< OC_{pl}^*, v_h >= \sum_{i=1}^{n} v_{h,j}(\underline{\mathrm{c}}_{pl}^*)_i$$

How many principal components ?

Stopping rule in PCA:

Keep principal component iff the associated eigenvalue is larger than 1 (mean of eigenvalues).

This rule is adapted to MCA as follows:

Keep principal component iff the associated eigenvalue is larger than $\frac{1}{P}$.

Indeed, suppose that $H = K - P$ (usual situation), then the mean of all non-zero eigenvalues is given by:

$$\frac{1}{K-P} \sum \text{non zero eigenvalues}$$
$$= \frac{1}{K-P} \text{ total inertia of point cloud } \aleph_l \text{ around } G_l$$
$$= \frac{1}{K-P}\left(\frac{K}{P} - 1\right) = \frac{1}{P}.$$

This results explains the criteria given above.

5.2.2 Quality of the representation of each modality

• Quality of representation of each modality l of the variable Y_p on the axis Γ_h is given by:

$$\cos^2 \text{(angle between } OC^*_{pl} \text{ and the axis } \Gamma_h)$$

$$\cos^2 (\beta_{h,pl}) = \frac{\psi^2_{h,pl}}{\|OC^*_{pl}\|^2}$$

It can be proven that:

$$\cos(\beta_{h,pl}) = r_{X_{pl},\phi_h}$$

As for PCA, it is possible to construct a correlation circle with the modalities.

5.2.3 Contribution of each modality

• Contribution of the modality l of Y_p on the variance of the new variable ψ_h:

$$CTR_{\Gamma_h}(X_{pl}) = \frac{f_{.pl}\psi^2_{h,pl}}{\lambda_h} = \frac{n_{pl}}{nP\lambda_h}\psi^2_{h,pl}$$

• Global contribution of the variable Y_p (sum on all modalities) on the variance of ψ_h:

$$CTR_{\Gamma_h}(Y_p) = \sum_{l=1}^{K_p} CTR_{\Gamma_h}(X_{pl})$$

5.2.4 Reconstitution formula

The formula introduced for BCA becomes:

$$f_{ipl} = f_{i..}f_{.pl}(1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i}\psi_{h,pl})$$

$$\Longrightarrow \frac{x_{ipl}}{nP} = \frac{1}{n}\frac{n_{pl}}{nP}(1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i}\psi_{h,pl})$$

$$\Longrightarrow x_{ipl} = \frac{n_{pl}}{n}(1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i}\psi_{h,pl})$$

The distance between the "observed probability" that individual i has modality l on variable Y_p (x_{ipl}) and the "mean probability" to have this modality ($\frac{n_{pl}}{n}$) is given as a function of principal components

$$\Downarrow$$

This leads to the link between individual i and the modality l associated to the variable Y_p

Two other formulas can be introduced :

• The number of individuals with modality l on Y_p **and** modality l' on $Y_{p'} = n_{pl,p'l'}$ is given by:

$$
\begin{aligned}
n_{pl,p'l'} &= \sum_{i=1}^{n} x_{ipl} x_{ip'l'} \\
&= \sum_{i=1}^{n} \frac{n_{pl}}{n}\Big(1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i} \psi_{h,pl}\Big) \\
&\times \frac{n_{p'l'}}{n}\Big(1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i} \psi_{h,p'l'}\Big) \\
&= \dots \\
&= \frac{n_{pl} n_{p'l'}}{n}\Big(1 + \sum_{h=1}^{H} \psi_{h,pl} \psi_{h,p'l'}\Big)
\end{aligned}
$$

$\Longrightarrow$ *Comparison between modalities*

But the attraction/repulsion index $d_{pl,p'l'}$ between the modality l of Y_p and the modality l' de Y_p' is given by:

$$d_{pl,p'l'} = \frac{n_{pl,p'l'}/n}{\frac{n_{pl}}{n}\frac{n_{p'l'}}{n}} = \frac{n_{pl,p'l'}}{\frac{n_{pl}n_{p'l'}}{n}}$$

$$\Longrightarrow d_{pl,p'l'} = 1 + \sum_{h=1}^{H} \psi_{h,pl}\psi_{h,p'l'}$$

• The proximity between two individuals i and i' is defined by :

$$p_{i,i'} = 1 + \sum_{h=1}^{H} \phi_{h,i}\phi_{h,i'}$$

Two individuals are close (same behaviour) if they have in general the same modalities.

5.3 Graphical representations

Two types of graphical representations:

- Pseudo-barycentric representation (standard)
- Biplot representation (barycentric)

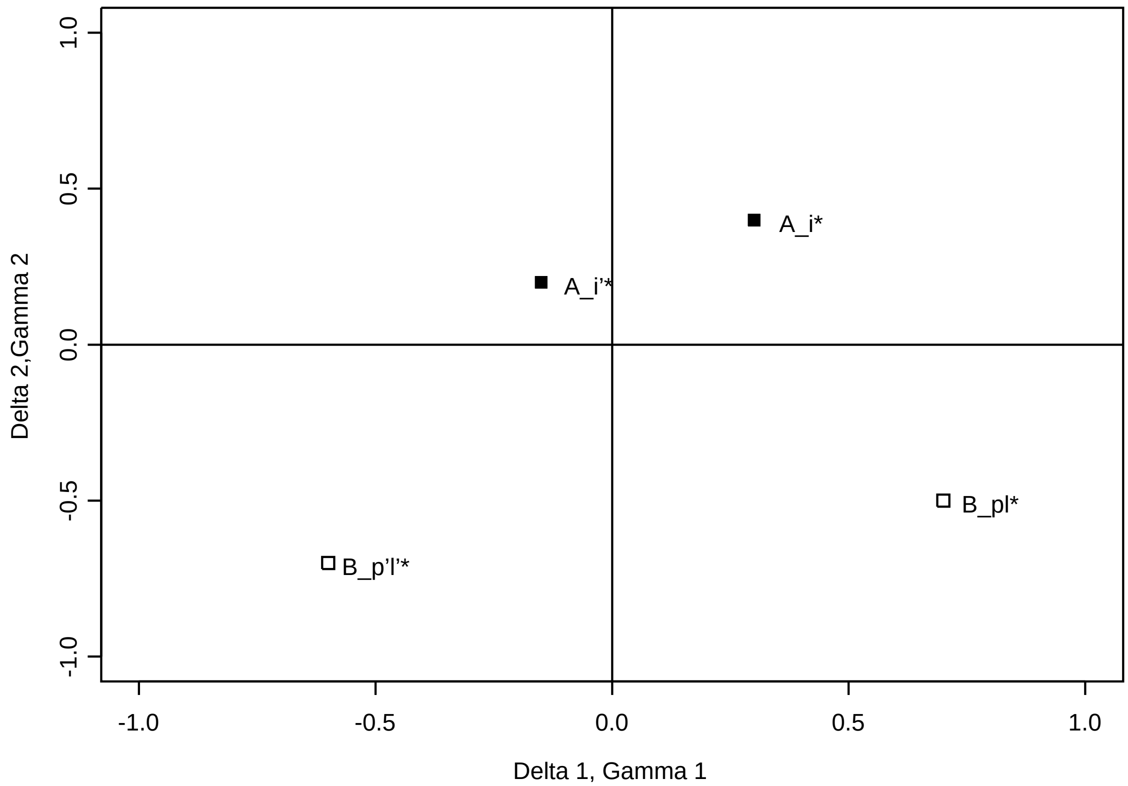
5.3.1 Standard representation (Pseudo-barycentric)

We focus on the first principal plan but more dimensions can be analyzed with the same methodology

The first principal plan is constructed using both PCAs:

- individual A_i^* $(i = 1, \ldots, n)$ is projected on the first factorial plan leading to coordinate $(\phi_{1,i}, \phi_{2,i})$

- modality B_{pl}^* $(p = 1, \ldots, P; l = 1, \ldots, K_p)$ is projected on the first factorial plan leading to coordinate $(\psi_{1,pl}, \psi_{2,pl})$



This representation is the closest representation of the simultaneous information inside point clouds $\aleph_l^*$ and $\aleph_c^*$

Interpretation:

• The well represented modalities on the first principal plan are compared using the following approximated formula:

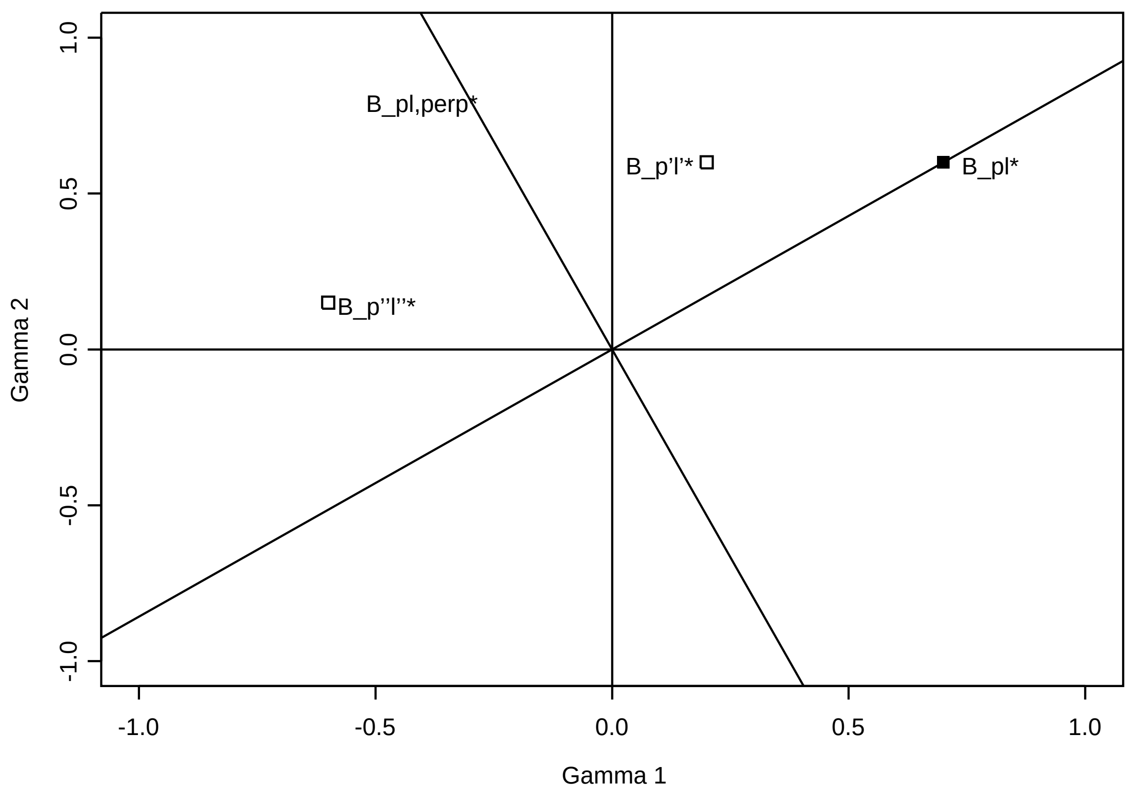
$$\begin{aligned} d_{pl,p'l'} &\approx 1 + \sum_{h=1}^{2} \psi_{h,pl}\psi_{h,pl} \\ &= 1 + < 0B^*_{pl}, 0B^*_{p'l'} > \\ &= 1 + \|0B^*_{pl}\| \|0B^*_{p'l'}\| \cos(0B^*_{pl}, 0B^*_{p'l'}) \end{aligned}$$

Draw $B^{\perp}_{pl}$ which passes through the origin and which is orthogonal to $0B^*_{pl}$. This line separates the space into two parts:

- the modalities that are on the same side than B^*_{pl} are attracted by it

- the modalities on the other side are repulsed by B^*_{pl}

The attraction/repulsion index increases with $| < 0B^*_{pl}, 0B^*_{p'l'} > |$.



If the modalities pl, $p'l'$ *and* $p''l''$ *are well represented on the first principal plan, therefore we can conclude that* pl *and* $p'l'$ *are attracted by each other, and modalities* pl *and* $p''l''$ *are repulse by each other.*

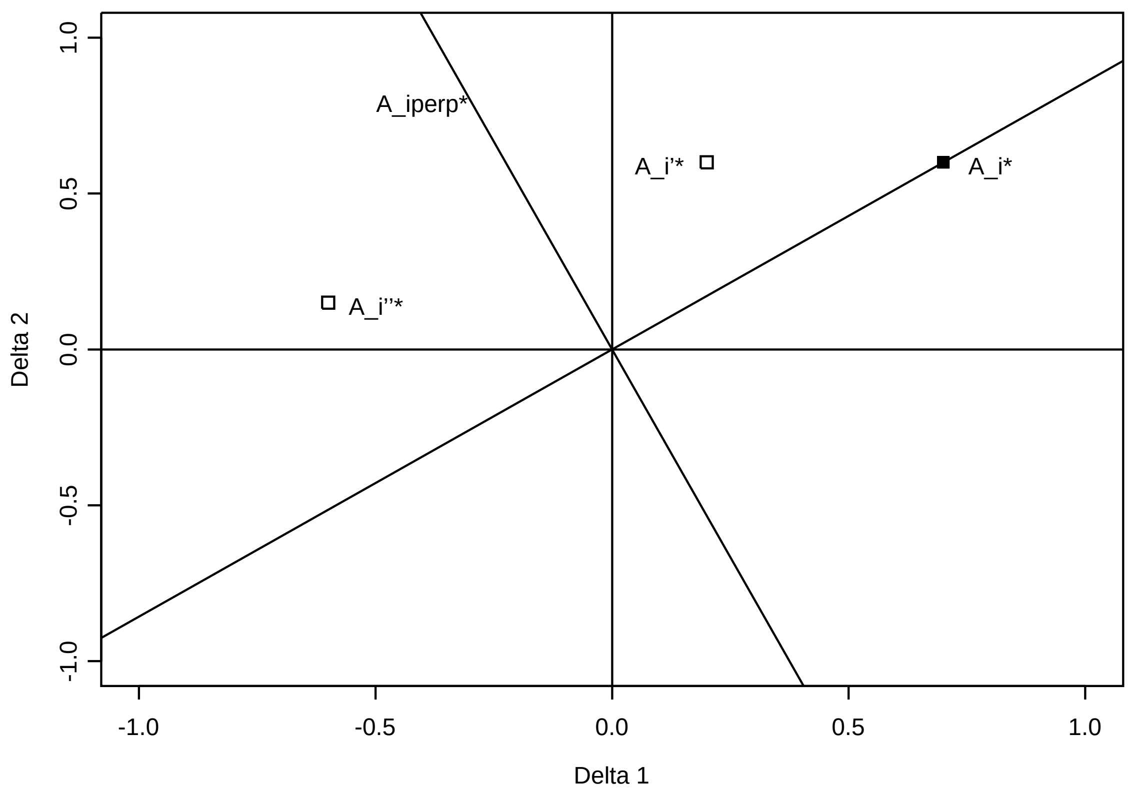
• The well represented individuals on the first principal plan are compared using the following approximated formula:

$$\begin{aligned} p_{i,i'} &\approx 1+\sum_{h=1}^{2}\phi_{h,i}\phi_{h,i'} \\ &= 1+ < 0A_i^*, 0A_{i'}^* > \\ &= 1+\|0A_i^*\|\|0A_{i'}^*\|\cos(0A_i^*, 0A_{i'}^*) \end{aligned}$$

Draw $A_i^{\perp}$ which passes through the origin and which is orthogonal to $0A_i^*$. This line separates the space into two parts:

- the modalities that are on the same side than A_i^* are individuals who share a set of modalities with individual i. And the common set increases with $< 0A_i^*, 0A_{i'}^* >$.

- the modalities on the other side than A_i^* are individuals who have few characteristic in common with individual i.



If the individuals i, i' and i'' are well represented on the first principal plan, therefore we can conclude that individual i is close to individual i' and has few characteristic in common with individual i''

• The well represented modalities and individuals on the first principal plan are compared using the following approximated formula:

$$x_{ipl} \approx \frac{n_{pl}}{n}(1 + \sum_{h=1}^{2} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i}\psi_{h,pl})$$

The coefficient $\frac{1}{\sqrt{\lambda_h}}$ implies some difficulties in the interpretation.

If A_i^* and B_{pl}^* are well represented on the first principal plan:

- *The probability that the individual A_i^* has modality l on variable Y_p is high if they are belong to the same quadrant*
- *The probability that the individual A_i^* has modality l on variable Y_p is low if they are in opposite quadrants*
- *We cannot conclude if they belong to adjacent quadrants.*

5.3.2 Biplot

The Biplot representation leads to a better visibility of the first principal plan to compare the individuals with the modalities.

• The individual i is associated to $\tilde{A}_i^*$ which has coordinates:

$$(\tilde{\phi}_{1,i}, \tilde{\phi}_{2,i})' = (\frac{\phi_{1,i}}{\sqrt{\lambda_1}}, \frac{\phi_{2,i}}{\sqrt{\lambda_2}})'$$

• The modality l on variable Y_p ($p = 1, \ldots, P; l = 1, \ldots, K_p$) is associated with B_{pl}^* which has coordinates:

$$\psi_{1,pl}, \psi_{2,pl}.$$

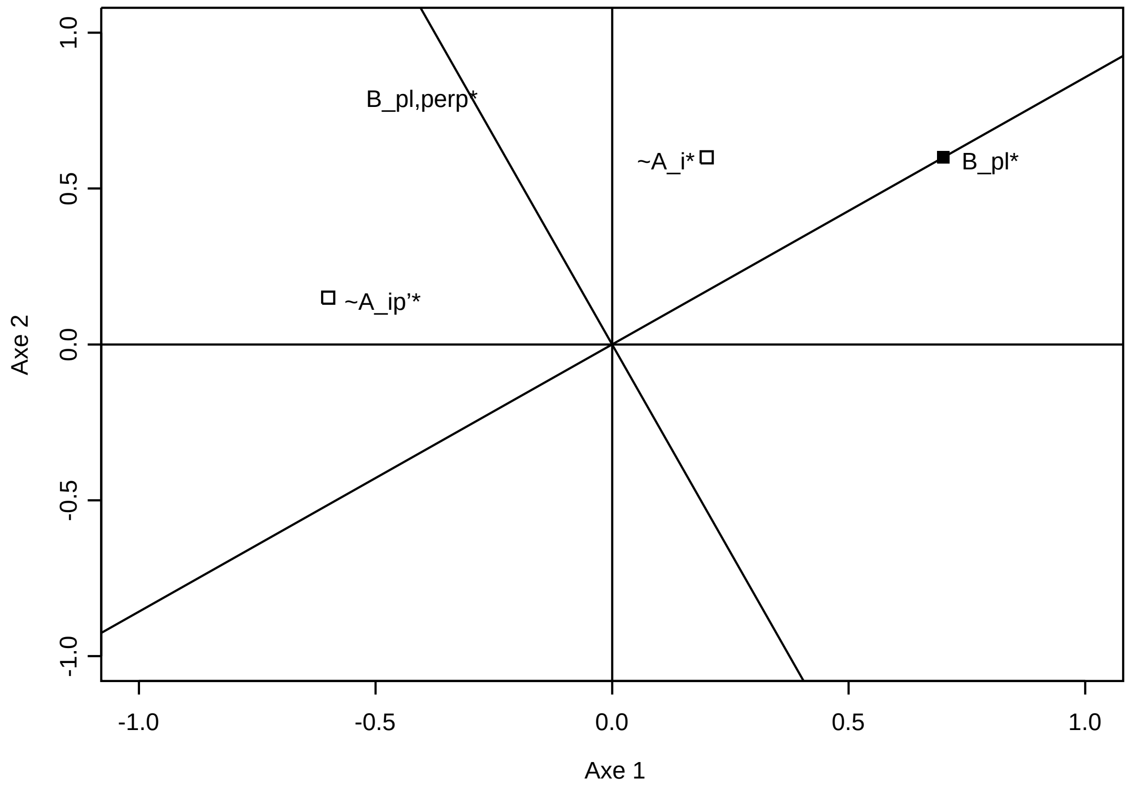
Reconstitution formula to compare the individuals with the modalities:

$$
\begin{aligned}
x_{ipl} &\approx \frac{n_{pl}}{n}(1+\sum_{h=1}^{2}\tilde{\phi}_{h,i}\psi_{h,pl}) \\
&= \frac{n_{pl}}{n}(1+ < 0\tilde{A}_i^*, 0B_{pl}^* >) \\
&= \frac{n_{pl}}{n}(1+\|0\tilde{A}_i^*\|\|0B_{pl}^*\|\cos(0\tilde{A}_i^*, 0B_{pl}^*))
\end{aligned}
$$

Draw $B_{pl}^{\perp}$ *which passes through the origin and which is orthogonal to* $0B_{pl}$*. This line separates the space into two parts:*

- the individuals that are on the same side than B_{pl} *have, with high probability, the modality* l *on variable* Y_p

- the individuals on the other side have, with low probability, the modality l *on variable* Y_p*.*



If the modality l on variable Y_p is well represented on the first principal plan, therefore the probability that individual i has modality l on variable Y_p is high and the probability that individual i' has modality l on variable Y_p is low.

5.4 The Burt table (BT)

When the use of BT is more appropriate than the use of CDT?

• If n is large, the simultaneous representation of individuals and modalities is unreadable.

• If the individuals are anonymous, the interest is only based on the modalities.

$$\Downarrow$$

Contingency table (symmetric) with $K = K_1 + \ldots + K_P$ modalities on P variables.

		Y_1			$\ldots$		Y_p		$\ldots$		Y_P		
		1	$\ldots$	K_1	$\ldots$	1	$\ldots$	K_p	$\ldots$	1	$\ldots$	K_P	
	1	n_{11}		0	$\ldots$				$\ldots$				Pn_{11}
Y_1	$\vdots$		$\ddots$		$\vdots$		$n_{1l,pl'}$		$\vdots$		$n_{1l,Pl'}$		$\vdots$
	K_1	0		n_{1K_1}	$\ldots$				$\ldots$				Pn_{1K_1}
	$\vdots$				$\vdots$				$\vdots$				$\vdots$
	1				$\ldots$	n_{p1}		0	$\ldots$				Pn_{p1}
Y_p	$\vdots$		$n_{pl,1l'}$		$\vdots$		$\ddots$		$\vdots$		$n_{pl,Pl'}$		$\vdots$
	K_p				$\ldots$	0		n_{pK_p}	$\ldots$				Pn_{pK_p}
	$\vdots$				$\vdots$				$\vdots$				$\vdots$
	1				$\ldots$		$n_{Pl,pl'}$		$\ldots$	n_{P1}		0	Pn_{P1}
Y_p	$\vdots$		$n_{Pl,1l'}$		$\vdots$				$\vdots$		$\ddots$		$\vdots$
	K_p				$\ldots$				$\ldots$	0		n_{PK_P}	Pn_{PK_P}
		Pn_{11}	$\ldots$	Pn_{1K_1}	$\ldots$	Pn_{p1}	$\ldots$	Pn_{pK_p}	$\ldots$	Pn_{P1}	$\ldots$	Pn_{PK_P}	nP^2

We use the BCA on the Burt table, instead of the application of the BCA on the complete disjunctive table (CDT).

Remark: The row profiles and the column profiles are identical since the Burt table is symmetric.

5.4.1 Links between MCA on CDT and MCA on BT

• The inertia obtained by MCA on BT are given by the squared inertia obtained by MCA on CDT:

$$\lambda_{BT,h} = \lambda_h^2 \qquad h = 1, \ldots, H$$

• The variances of the principal component $\psi_{BT,h}$ obtained by MCA on BT are given by the squared variances of the principal component obtained by MCA on CDT:

$$s^2_{\psi_h} = \lambda_h \text{ and } s^2_{\Psi_{BT,h}} = \lambda_{BT,h} = \lambda_h^2$$

• It holds also that $\forall h = 1, \ldots, H$:

$$\psi_{BT,h} = \sqrt{\lambda_h}\psi_h$$

5.5 Practical example

Research question:
Determining if, *inside the PS electorate*, Muslims behave differently from non-believers and Catholics.

Database:
Votes for the PS in the regional elections of June 2004 in the Brussels Region

Method:
To this end, we will look into the answers given to society-oriented questions using multiple correspondence analysis.

5.5.1 Society-oriented questions:

• Mail services should be privatized;

• Trade Unions should weigh heavily in major economic decisions;

• Homosexual couples should be allowed to adopt children;

• Consumption of cannabis should be forbidden;

• People don't feel at home in Belgium anymore;

• Abolishing the death penalty was the right decision.

The answers proposed to these questions are:
Total agreement (1),
Rather in agreement (2),
Rather opposed (3),
Totally opposed (4),
No opinion (5).

The questionnaire also includes a question concerning a subjective judgment of the individual about his general behavior on a left-right scale:

"Here is a political left-right scale. 0 is the most left-wing position 9 the most right-wing. Where would you locate yourself?"

The variable "Belief" with three categories (Muslims, non-believers and Catholics) is also available

5.5.2 χ^2 independence test

First, we analyze each society-oriented question separately by testing its dependency with respect to the *belief* variable using a χ^2 independence test.

χ^2	Mail	Trade Union	Homosexual
Test	26.78	27.13	144.82
p-value	(0.00)	(0.00)	(0.00)

χ^2	Cannabis	Home	D. Penalty
Test	86.98	27.94	11.75
p-value	(0.00)	(0.00)	(0.16)

The assumption of independence between the society-oriented questions and belief-oriented question is rejected for all of the questions (at the 5% level) except for the question on the death penalty (very small variation inside the question).

5.5.3 Attraction-repulsion indexes

Links between each pair of modalities of two variables with the attraction-repulsion indexes d_{jk} defined as

$$d_{jk} = \frac{f_{jk}}{f_{j.}f_{.k}}$$

where f_{jk} is the observed frequency and $f_{j.}f_{.k}$ is the theoretical frequency under the independence hypothesis.

Interpretation:

$d_{jk} > 1 \Longleftrightarrow$ the two modalities attract each others

$d_{jk} < 1 \Longleftrightarrow$ the two modalities push each other away

$d_{jk} \approx 1 \Longleftrightarrow$ the two modalities are close to being.
independent

Mail services should be privatized

Attraction Index	Non-believer	Catholic	Muslim
Total agreement	0.712	1.411	1.196
Rather in agreement	1.055	0.707	1.113
Rather opposed	1.080	1.001	0.866
Totally opposed	1.119	1.062	0.757
No opinion	0.779	0.857	1.472

- Proportion of Muslim PS-voters who declare having no opinion on the subject is much higher than the corresponding proportions of Catholic and Non-believer PS-voters.
- Proportion of Catholics who are in total agreement to a privatization of mail services is much higher.

Trade Unions should weigh heavily in major economic decisions

Attraction Index	Non-believer	Catholic	Muslim
Total agreement	0.878	0.920	1.261
Rather in agreement	1.117	0.930	0.853
Rather opposed	1.203	1.102	0.588
Totally opposed	0.953	1.779	0.534
No opinion	0.847	0.953	1.290

- As for the influence of Trade Unions in major political decisions, Muslim PS-voters are more prone to agree with the necessity of more influence than the others, while Catholics seem to be very opposed to the latter.

Homosexual couples should be allowed to adopt children

Attraction Index	Non-believer	Catholic	Muslim
Total agreement	1.311	0.886	0.558
Rather in agreement	1.470	0.959	0.240
Rather opposed	1.101	1.220	0.676
Totally opposed	0.468	1.104	1.821
No opinion	1.240	0.674	0.825

- The answers to the question of allowing adoption by homosexual couples is very clear-cut.
- Non-believers are proportionally much more in agreement with the assertion than others
- Catholics generally seem to oppose or totally oppose it.
- A vast majority of Muslims declare themselves totally opposed to the proposition.

Consumption of cannabis should be forbidden

Attraction Index	Non-believer	Catholic	Muslim
Total agreement	0.626	1.116	1.548
Rather in agreement	0.748	1.176	1.300
Rather opposed	1.341	0.948	0.463
Totally opposed	1.371	0.680	0.601
No opinion	1.024	1.186	0.830

- Majority of Muslims agree with the proposal
- Majority of Non-believers declare themselves opposed to it.

People don't feel at home in Belgium anymore

Attraction Index	Non-believer	Catholic	Muslim
Total agreement	0.786	1.433	1.056
Rather in agreement	0.677	1.330	1.311
Rather opposed	0.937	1.207	0.962
Totally opposed	1.178	0.738	0.885
No opinion	0.867	1.082	1.166

- Strong opposition between Non-believers and Catholics. The Catholic are proportionally more prone to agree with the assertion than Non-believers.
- Muslims also seem to agree on the fact that they "don't feel at home in Belgium anymore".

Abolishing the death penalty was the right decision

Attraction Index	Non-believer	Catholic	Muslim
Total agreement	1.069	0.881	0.967
Rather in agreement	1.020	0.926	1.019
Rather opposed	0.735	1.486	1.105
Totally opposed	0.762	1.390	1.127
No opinion	0.932	1.178	0.989

- High number of "totally in agreement" with abolishing it
- Muslims don't really show a tendency one way or another with respect to the others.
- Catholics seem to be more prone than Non-believers to be against the abolishment of the death penalty.

5.5.4 Multiple correspondance analysis (AFCM)

Multivariate vision of the set of society-oriented questions (active variables)

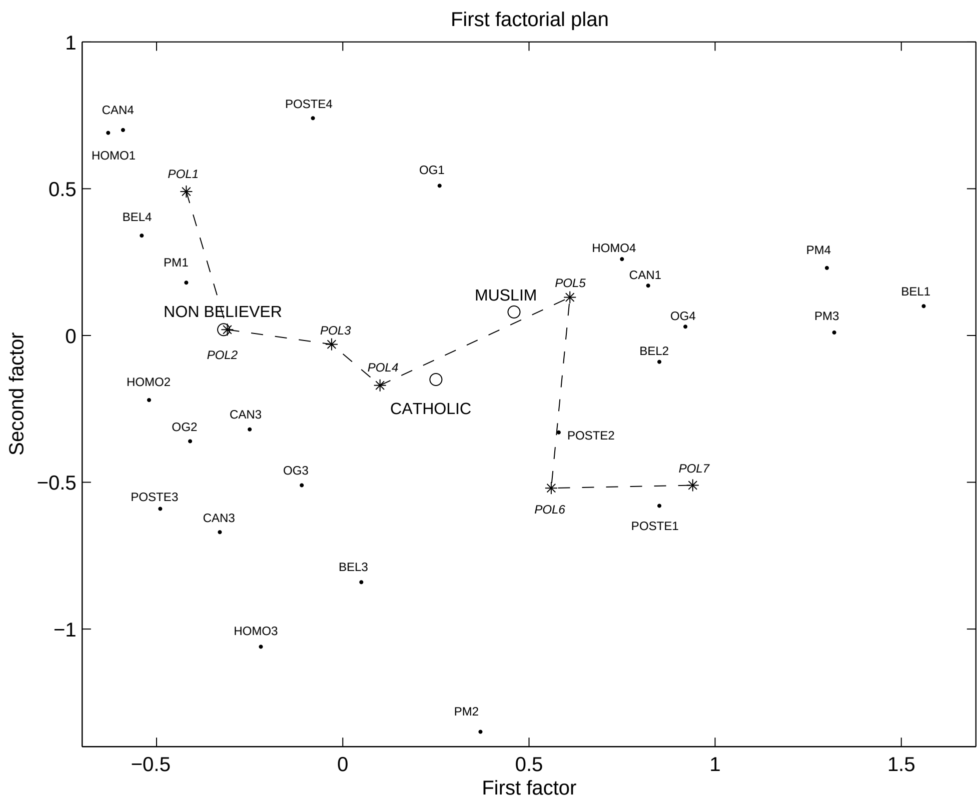


Figure 5.1: Multiple Correspondence Analysis on society-oriented questions. *Belief* and the *political scale* are added as illustrative variables.

Two illustrative variables: *belief* and the *political scale*

The first axis represents a left-right dimension.

To visualize better, we deleted modality "no opinion" for the society-oriented questions.

- Inertia explained by the first plane: 20%

- Contributors on first factorial axis:

  24.8% feeling at home in Belgium

  22.7% the death penalty

  17.9% adoption by homosexual couples

  17% prohibition of cannabis consumption

  10.4% privitization of mail services

  7.2% Trade Unions in political decisions

- Contributors on second factorial axis:

  24.2% privitization of mail services

  19.3% adoption by homosexual couples

  16.5% prohibition of cannabis consumption

  14.7% the death penalty

  13.6% feeling at home in Belgium

  11.8% Trade Unions in political decisions

5.5.5 Econometric Model

Multivariate data analysis doesn't take into account the influence of other variables which may strongly influence the results

Dependent variable: the left-right indicator built on the basis of the six society-oriented questions

	Regression 1		Regression 2	
Variable	Coefficient	Std. Error	Coefficient	Std. Error
C	-0.166***	(0.027)	-0.457***	(0.078)
NONCROYANT	-0.319***	(0.050)	-0.225***	(0.048)
MUSULMAN	0.089	(0.055)	0.152***	(0.055)
AGE			0.008***	(0.001)
AUCUN			0.371***	(0.112)
PRIMAIRE			0.421***	(0.094)
PROFESSIONNEL			0.310***	(0.083)
SECINF			0.416***	(0.068)
SECSUP			0.274***	(0.053)
SUPNONUNIV			0.163***	(0.054)
TECHNIQUE			0.151	(0.096)
	R-squared:	12.6 %	R-squared:	24.4 %

Sample size: 676, *Statistically different from zero at 10%,

Chapitre 6

Canonical correlation analysis

6.1 Introduction

Objective: Characterize the linear relation between 2 sets of quantitative variables

Canonical correlation analysis seeks to identify and quantify the associations between two sets of variables

Key reference:
Hotelling, H. (1936), "Relations between two Sets of Variables", Biometrika, 28, 321-377

EXAMPLES:

• Relationships between job evaluation ratings and self-ratings of job characteristics (Dunham, 1977)

Measures of job characteristics

X_1: Task Feedback

X_2: Task significance

X_3: Task variety

X_4: Task identity

X_5: Autonomy

Self-ratings of job characteristics

Y_1: Supervision satisfaction

Y_2: Career future satisfaction

Y_3: Financial satisfaction

Y_4: Amount of work satisfaction

Y_5: Company identification

Y_6: Kind of work satisfaction

Y_7: General satisfaction

• Determine associations between socio-economic variables and consumption behaviors

Socio-economic variables

X_1: Household income

X_2: Number of school years of the husband

X_3: Number of school years of the wife

X_4: Age of the husband

X_5: Age of the wife

X_6: Number of children

Consumption behaviors

Y_1: Number of times that the family goes to a restaurant (per year)

Y_2: Number of times that the family goes to the cinema (per year)

6.2 Canonical variates and canonical correlations

Let $X = (X_1, X_2, \ldots, X_p)'$

and $Y = (Y_1, Y_2, \ldots, Y_q)'$.

IDEA: Find linear combinations (Canonical variates)

$$U_k = \alpha_k' X \quad \text{and} \quad V_k = \beta_k' Y$$

with maximal

$$|\text{corr}(U_k, V_k)|$$

subject to the following constraints: :

-Var(U_k) = Var$(V_k) = 1$

-uncorrelated with previously found canonical variates.

Canonical vectors: α_k and β_k $(k \leq min\{p, q\})$

Canonical correlations: $\rho_k = |\text{corr}(U_k, V_k)|$.

To solve this maximization problem under constraint, denote: $Z = (X, Y) \in IR^{p+q}$, where

$$\mathrm{Cov}(Z) = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} := \Sigma.$$

Solution of canonical analysis problem at the population level
(proof page 546, Johnson and Wichern):

- α_k are the eigenvectors of

$$\mathcal{M}_X = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$$

- β_k are the eigenvectors of

$$\mathcal{M}_Y = \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

(we get also the following link: $\beta_k = \frac{1}{\rho_k}\Sigma_{YY}^{-1}\Sigma_{YX}\alpha_k$)

- ρ_k^2 are the eigenvalues of $\mathcal{M}_X$ or $\mathcal{M}_Y$.

The first couple (α_1, β_1) is associated with the largest eigenvalue, etc.

Remark: In practice, it is sometimes more relevant to apply canonical correlation analysis to the correlation matrix instead of the covariance matrix (use standardized variables)

$$R(Z) = \begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{bmatrix}$$

Using the correlation matrix instead of the covariance matrix, the canonical correlations are the same but the canonical vectors are modified. Nevertheless, a simple relation exists between both formulations:

$$\tilde{\alpha}_k = D_X^{1/2} \alpha_k$$
$$\tilde{\beta}_k = D_Y^{1/2} \beta_k$$

where D_X is the diagonal matrix with variances of X on the diagonal and D_Y the matrix with the variances of Y on the diagonal

6.3 Estimation

QUESTION: How to estimate canonical variates $U_k = \alpha_k' X$ and $V_k = \beta_k' Y$?

ANSWER: Estimation of the covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

by the sample covariance matrix

$$\mathrm{S} = \begin{bmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{bmatrix}$$

Solution to the problem at the sample level:

- $\hat{\alpha}_k$ are the eigenvectors of

$$M_X = S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX}$$

- $\hat{\beta}_k$ are the eigenvectors of

$$M_Y = S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY}$$

- $\hat{\rho}_k^2$ are the eigenvalues of M_X or M_Y.

6.4 Interpreting the sample canonical variables

The canonical variables are artificial and based on X et $Y \Longrightarrow$ Try to identify the meaning of these new variables.

Two schools of thought are opposed in this field

• Contribution in the construction of U_k and V_k
Rencher (1998) proposed to use the coordinates of canonical vectors which measure the marginal impact of each variables in the construction of canonical variables $\Longrightarrow$ Multivariate approach
• Correlations with initial variables (as in PCA)
Tenenhaus (page 18, 1998) preferred to use the correlations between initial variables and canonical variates $\Longrightarrow$ easy but bivariate

$$\Downarrow$$

Use the two directions to have an idea

6.5 Some descriptive measures of the quality of the reduction

6.5.1 Error matrices of approximations

Since: $\hat{U} = \hat{A}X$ and $\hat{V} = \hat{B}Y$ with

$\hat{A} = [\hat{\alpha}_1', \hat{\alpha}_2', \ldots, \hat{\alpha}_p']$ and $\hat{B} = [\hat{\beta}_1', \hat{\beta}_2', \ldots, \hat{\beta}_q']$,

it follows that

$$X = \hat{A}^{-1}\hat{U} \text{ and } Y = \hat{B}^{-1}\hat{V}$$

Hence the covariance matrices can be written on the basis of canonical variates:

$$\begin{aligned}
S_{XY} &= (\hat{A}^{-1})\mathrm{cov}(\hat{U}, \hat{V})(\hat{B}^{-1})' = \hat{\rho}_1\hat{\alpha}^{(1)}\hat{\beta}^{(1)\prime} + \ldots + \hat{\rho}_p\hat{\alpha}^{(p)}\hat{\beta}^{(p)\prime} \\
S_{XX} &= (\hat{A}^{-1})(\hat{A}^{-1})' = \hat{\alpha}^{(1)}\hat{\alpha}^{(1)\prime} + \ldots + \hat{\alpha}^{(p)}\hat{\alpha}^{(p)\prime} \\
S_{YY} &= (\hat{B}^{-1})(\hat{B}^{-1})' = \hat{\beta}^{(1)}\hat{\beta}^{(1)\prime} + \ldots + \hat{\beta}^{(q)}\hat{\beta}^{(q)\prime}
\end{aligned}$$

where $\hat{\alpha}^{(i)}$ and $\hat{\beta}^{(i)}$ are the i^{th} columns respectively of the inverse matrices $\hat{A}^{-1}$ and $\hat{B}^{-1}$.

QUESTION:

Which proportion of the information on S_{XX}, S_{YY} and S_{XY} is lost when only $r(< p)$ canonical variates are used?

$$S_{XY}-\hat{\rho}_1\hat{\alpha}^{(1)}\hat{\beta}^{(1)\prime}+\ldots+\hat{\rho}_r\hat{\alpha}^{(r)}\hat{\beta}^{(r)\prime} = \hat{\rho}_{r+1}\hat{\alpha}^{(r+1)}\hat{\beta}^{(r+1)\prime}+\ldots+\hat{\rho}_p\hat{\alpha}^{(p)}\hat{\beta}^{(}$$

$$S_{XX}-\hat{\alpha}^{(1)}\hat{\alpha}^{(1)\prime}+\ldots+\hat{\alpha}^{(r)}\hat{\alpha}^{(r)\prime} = \hat{\alpha}^{(r+1)}\hat{\alpha}^{(r+1)\prime}+\ldots+\hat{\alpha}^{(p)}\hat{\alpha}^{(p)\prime}$$

$$S_{YY}-\hat{\beta}^{(1)}\hat{\beta}^{(1)\prime}+\ldots+\hat{\beta}^{(r)}\hat{\beta}^{(r)\prime} = \hat{\beta}^{(r+1)}\hat{\beta}^{(r+1)\prime}+\ldots+\hat{\beta}^{(q)}\hat{\beta}^{(q)\prime}$$

It is straightforward to note that most of the time S_{XY} is better explained than S_{XX} and S$_{YY}$

6.5.2 Proportions of explained sample variances

When the observations are standardized, the sample covariance matrices are correlation matrices.

Proportions of total sample variances explained by the first r canonical variates:

$$R^2_{\tilde{X}|\hat{U}_1,\ldots,\hat{U}_r} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{\hat{U}_i,\tilde{X}_k}}{p}$$

$$R^2_{\tilde{Y}|\hat{V}_1,\ldots,\hat{V}_r} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{q} r^2_{\hat{V}_i,\tilde{Y}_k}}{q}$$

6.6 Large sample inferences

Suppose that $Z = (X, Y) \in IR^{p+q} \sim N_{p+q}(\mu, \Sigma)$

6.6.1 Testing procedure on Σ_{XY}

Idea: Perform a testing procedure looking at the association between the two groups of variables (proof in Kshirsagar, 1972)

H_0: $\Sigma_{XY} = 0$ $(\rho_1 = \ldots = \rho_p = 0)$
H_1: $\Sigma_{XY} \neq 0$

Test statistic: $MV = -n \ln \prod_{i=1}^{p} (1 - \hat{\rho}_i^2)$

$$(MV = n \ln(\frac{\det(S_{XX}) \det(S_{YY})}{\det(S)}))$$

Distribution under H_0: $MV \sim \chi^2_{pq}$

Reject H_0 at significance level $\alpha = 5\%$ if

$$MV > \chi^2_{pq;0.95}$$

6.6.2 Individual tests on canonical correlations

If $H_0 : \Sigma_{XY} = 0$ is rejected, it is natural to examine the "significance" of the individual canonical correlations. First step: $\rho_1 \neq 0$:
H_0^1: $\rho_1 \neq 0, \rho_2 = \rho_3 = \ldots = \rho_p = 0$
$H_1^1 : \rho_i \neq 0$ pour $i \geq 2$

If H_0^1 is rejected, the next step is:
H_0^2:$\rho_1 \neq 0, \rho_2 \neq 0, \rho_3 = \rho_4 = \ldots = \rho_p = 0.$
$H_1^2 : \rho_i \neq 0$ pour $i \geq 3$

and so on $\forall k \in \{2, \ldots, p-1\}$:
H_0^k:$\rho_1 \neq 0, \rho_k \neq 0, \rho_{k+1} = \ldots = \rho_p = 0.$
$H_1^k : \rho_i \neq 0$ pour $i \geq k+1$

Decision rule: Reject H_0 at significance level α if

$$-(n-1-\frac{1}{2}(p+q+1)) \ln \prod_{i=k+1}^{p} (1-\hat{\rho}_i^2)$$
$$> \chi^2_{(p-k)(q-k);1-\alpha}$$

6.7 Example: Relationships between job evaluation ratings and self-ratings of job characteristics (Dunham, 1977; see Johnson & Wichern (2002))

Measures of job characteristics

X_1: Task Feedback

X_2: Task significance

X_3: Task variety

X_4: Task identity

X_5: Autonomy

Self-ratings of job characteristics

Y_1: Supervision satisfaction

Y_2: Career future satisfaction

Y_3: Financial satisfaction

Y_4: Amount of work satisfaction

Y_5: Company identification

Y_6: Kind of work satisfaction

Y_7: General satisfaction

Chapitre 7

Discriminant and classification

7.1 Introduction

OBJECTIVES:

1. Discrimination or separation: Separate two (or more) classes of objects. Describe the different caracteristics of observations arising from different known populations.

2. Classification or allocation: Define rules that assign an individual to a certain class.

Overlap between the two approaches since the variables that discriminate can also be used to allocate new observation to one group and vice-versa.

EXAMPLES

Populations π_1 and π_2	Measured variables
Good and poor credit risks	Income, age, number of credit cards, family size
Successful and unsuccessful students	Socio-economic variables, secondary path, gender
Males and females	Anthropological measurements
Purchasers of a new product and laggards	Income, education, family size amount of previous brand switching
Papers written by two authors	Frequencies of different words and lengths of sentences
Two species of flowers	Sepal and petal length, pollen diameter

Remark: In the sequel we present the problem using two populations but the generalization to more than two populations is straightforward.

THEORITICAL CONTEXT:

Let denote the 2 populations by : π_1 and π_2.

The information on observations can be summarized in p variables:

$$X' = [X_1, \ldots, X_p]$$

The behavior of the variables is different in the two populations

$$\Downarrow$$

The joint density functions on X are respectively given by : $f_1(x)$ et $f_2(x)$

IDEA: Separate the space IR^p into 2 parts R_1 and R_2 using the sample.

RULE: If a new observation $\in R_1$ ($\in R_2$) then we suppose that it belongs to π_1 (π_2).

For the sample, we known the values of X **and also to which population it belongs to**. But for new observation, the population is unknown : WHY ?

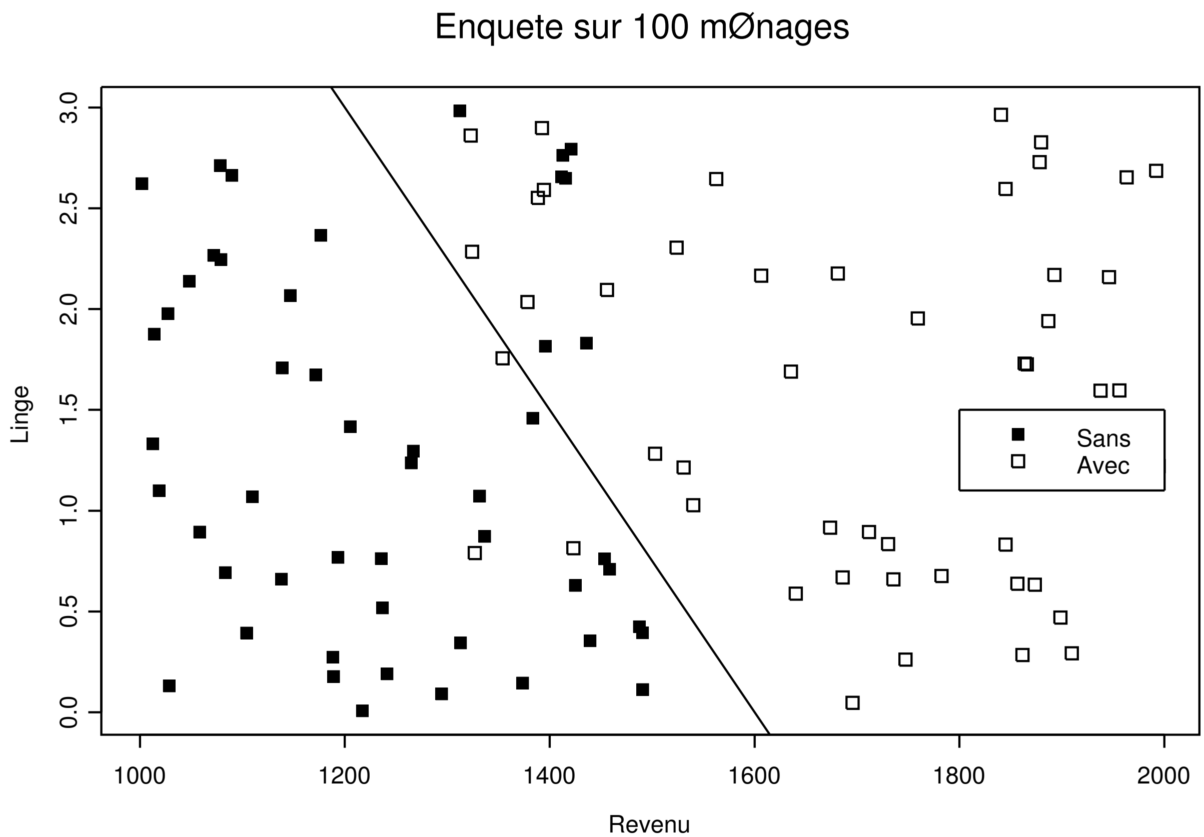
1. Incomplete knowledge of future performance (example: future firm's bankruptcy)
2. Information on the memberships of π_1 or π_2 requires the destruction (example: lifetime of a battery)
3. Unavailable or expensive information (example: medical problems)

$$\Downarrow$$

Find optimal rules based on the sample to classify observations to reduce misclassification as much as possible.

Example: Separate the space (by a segment in this case) to target the population that could be interested in buying a new washing machine (fictive data).

Variables: X_1: income of the family in euros, X_2: quantity (in kilo) dirty laundry per week.



The way the variables X are distributed in the space IR^2 does not allow to obtain a complete separation of the two populations.

7.2 Rules of classification based on the expected cost of misclassification

Let denote Ω the support of vector X. Let R_1 and $R_2 = \Omega - R_1$ be mutually exclusive and exhaustive:

$$R_1 \cup R_2 = \Omega$$
$$R_1 \cap R_2 = \emptyset$$

RULE: If a new observation $\in R_1$ ($\in R_2$) then we suppose that it belongs to π_1 (π_2). It is then possible to measure the conditional probability of misclassification.

The conditional probability of classifying an object as π_2 when in fact it is from π_1 is:

$$P(2|1) = P(X \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(x)dx$$

and similarly the conditional probability is:

$$P(1|2) = P(X \in R_1|\pi_2) = \int_{R_1} f_2(x)dx$$

But we have also to take into account **prior probabilities**:

$$
\begin{aligned}
p_1 &= P(\text{belong to } \pi_1) \\
p_2 &= P(\text{belong to } \pi_2)
\end{aligned}
$$

Hence probabilities of correctly or incorrectly classifying an observation can be derived:

$$
\begin{aligned}
P(\text{obs. from } \pi_1 \quad & \text{is correctly classified as } \pi_1) \\
&= P(\pi_1)P(X \in R_1|\pi_1) \\
&= p_1 P(1|1) \\
P(\text{obs. from } \pi_1 \quad & \text{is uncorrectly classified}) \\
&= P(\pi_1)P(X \in R_2|\pi_1) = p_1 P(1|2) \\
P(\text{obs. from } \pi_2 \quad & \text{is correctly classified as } \pi_2) \\
&= P(\pi_2)P(X \in R_2|\pi_2) = p_2 P(2|2) \\
P(\text{obs. from } \pi_2 \quad & \text{is uncorrectly classified}) \\
&= P(\pi_2)P(X \in R_1|\pi_2) = p_2 P(2|1)
\end{aligned}
$$

The cost of misclassification

Example: Not detecting a disease for a sick person is more important than detecting a disease for a healthy person

The cost of misclassification can be defined by a cost matrix:

	R_1	R_2
π_1	0	$c(2\|1)$
π_2	$c(1\|2)$	0

Expected cost of misclassification (ECM)

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

RESULT: The regions R_1 and R_2 that minimize ECM are defined by the values of x for which the following inequalities hold:

$$R_1 \; : \; \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$
$$R_2 \; : \; \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

Proof: Johnson & Wichern (2002) page 647.

Particular cases:

- Equal prior probabilities:

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \text{ et } R_2 : \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}$$

- Equal misclassification costs:

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \text{ et } R_2 : \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$$

- Equal prior probabilities and misclassification costs

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1 \text{ et } R_2 : \frac{f_1(x)}{f_2(x)} < 1.$$

Other criteria to derive optimal classification procedure

• Minimize the total probability of misclassification (TPM):

$$TPM = p_1 P(2|1) + p_2 P(1|2)$$

$\Rightarrow$ Mathematically, this problem is equivalent to minimizing ECM when the costs of misclassification are equal.

• Allocate a new observation x_0 to the population with the largest "posterior" probability $P(\pi_i|x_0)$. By Bayes 's rule, we obtain:

$$P(\pi_1|x_0) = \frac{p_1 f_1(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)}$$
$$P(\pi_2|x_0) = \frac{p_2 f_2(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)}$$

7.3 Classification with two multivariate normal populations

Often used in theory and practice because of their simplicity and reasonably high efficiency across a wide variety of population models.

HYPOTHESES:

$$f_1(x) = N_p(\mu_1, \Sigma_1) \text{ et } f_2(x) = N_p(\mu_2, \Sigma_2)$$

If $X \sim N_p(\mu, \Sigma)$ then:

$$f(x) = \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{\frac{1}{2}}} \exp[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)]$$

Before using these rules, it is necessary to test the normality hypothesis (e.g. QQ-plot). If the data reject the gaussianity assumption, we can try to obtain this assumption by a transformation of the data(e.g. by logarithm transformation).

Linear classification: $\Sigma_1 = \Sigma_2 = \Sigma$

RESULT: The regions R_1 and R_2 that minimize ECM are defined by the values of x for which the following inequalities hold:

$$R_1 \ : \ \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$
$$R_2 \ : \ \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

which is after simplification:

$$R_1 \ : \ (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}]$$
$$R_2 \ : \ (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) < \ln[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}]$$

But in practice μ_1, μ_2 and Σ are unknwon

$$\Downarrow$$

Estimate these parameters with unbiased estimators.

Estimate μ_1 and Σ_1 using the sample from π_1 of size n_1:

$$\hat{\mu}_1 = \begin{bmatrix} \bar{x}_1^{(1)} \\ \bar{x}_2^{(1)} \\ \cdots \\ \bar{x}_p^{(1)} \end{bmatrix} \text{ et } \hat{\Sigma}_1 = S_1 \begin{bmatrix} S_{11}^{(1)} & S_{12}^{(1)} & \dots & S_{1p}^{(1)} \\ S_{21}^{(1)} & S_{22}^{(1)} & \dots & S_{2p}^{(1)} \\ \cdots & \dots & \dots & \cdots \\ S_{p1}^{(1)} & S_{p2}^{(1)} & \dots & S_{pp}^{(1)} \end{bmatrix}$$

Estimate μ_2 and Σ_2 using the sample from π_2 of size n_2:

$$\hat{\mu}_2 = \begin{bmatrix} \bar{x}_1^{(2)} \\ \bar{x}_2^{(2)} \\ \cdots \\ \bar{x}_p^{(2)} \end{bmatrix} \text{ et } \hat{\Sigma}_1 = S_1 \begin{bmatrix} S_{11}^{(2)} & S_{12}^{(2)} & \dots & S_{1p}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} & \dots & S_{2p}^{(2)} \\ \cdots & \dots & \dots & \cdots \\ S_{p1}^{(2)} & S_{p2}^{(2)} & \dots & S_{pp}^{(2)} \end{bmatrix}$$

Under the hypothesis $\Sigma_1 = \Sigma_2$, we can use an unbiased pooled estimator of Σ:

$$\hat{\Sigma} = S_{pooled} = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} S_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} S_2$$

The estimated rule minimizing ECM is then:

$$R_1 \; : \; (\hat{\mu}_1 - \hat{\mu}_2)' S_{pooled}^{-1} x - \frac{1}{2} (\hat{\mu}_1 - \hat{\mu}_2)' S_{pooled}^{-1} (\hat{\mu}_1 + \hat{\mu}_2) \geq \ln[\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}]$$

$$R_2 \; : \; (\hat{\mu}_1 - \hat{\mu}_2)' S_{pooled}^{-1} x - \frac{1}{2} (\hat{\mu}_1 - \hat{\mu}_2)' S_{pooled}^{-1} (\hat{\mu}_1 + \hat{\mu}_2) < \ln[\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}]$$

Quadratic classification: $\Sigma_1 \neq \Sigma_2$

RESULT: The regions R_1 and R_2 that minimize ECM are defined by the values of x for which the following inequalities hold:

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1} \text{ and } R_2 : \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

which is after simplification:

$$\begin{aligned} R_1 &: -\frac{1}{2}x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x - k \geq \ln[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}] \\ R_2 &: -\frac{1}{2}x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x - k < \ln[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}] \end{aligned}$$

where

$$k = \frac{1}{2}\ln(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}) + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)$$

The estimated rule minimizing ECM is then:

$$\begin{aligned} R_1 &: -\frac{1}{2}x'(S_1^{-1} - S_2^{-1})x + (\hat{\mu}_1'S_1^{-1} - \hat{\mu}_2'S_2^{-1})x - k \geq \ln[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}] \\ R_2 &: -\frac{1}{2}x'(S_1^{-1} - S_2^{-1})x + (\hat{\mu}_1'S_1^{-1} - \hat{\mu}_2'S_2^{-1})x - k < \ln[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}] \end{aligned}$$

7.4 Evaluation of classification rules

Total probability of misclassification (TPM):

$$TPM = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

The lowest value of this quantity is called the optimum error rate (OER).

Suppose that $p_1 = p_2$, $C(2|1) = C(1|2)$ and $f_1(x) = N(\mu_1, \Sigma)$ and $f_2(x) = N(\mu_2, \Sigma)$, then the regions minimizing TPM are:

$$\begin{aligned} R_1 &: (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq 0 \\ R_2 &: (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) < 0 \end{aligned}$$

RESULT: The optimum Error Rate is:

$$OER = \Phi(\frac{-\Delta}{2}) \text{ where } \Delta^2 = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$

Example: if $\Delta^2 = 2.56$ then $OER = 0.2119$, hence then optimal rule of classification fails in 21% of cases.

But the rule is generally based on estimators

$$\Downarrow$$

We need to calculate the actual error rate (AER):

$$AER = p_1 \int_{\hat{R}_2} f_1(x)dx + p_2 \int_{\hat{R}_1} f_2(x)dx$$

where

$$\begin{aligned}
\hat{R}_1 &: (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) \geq 0 \\
\hat{R}_2 &: (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) < 0
\end{aligned}$$

But calculus to obtain AER are difficult and depend on $f_1(x)$ and $f_2(x)$.

Apparent Error rate (APER):

APER = % of obs. in the sample misclassified

$\Longrightarrow$ Easy to calculate and does not require knowledge on density functions

But underestimates AER even if n_i are large.

Solution: the problem comes from the fact that the same sample is used to construct the rule and also to test the quality of the classification

$$\Downarrow$$

Divide the sample in two parts : the training sample to construct the rule ($\pm 80\%$) and the validation sample to calculate APER.

But: • It requires large sample size

• The evaluated classification rule is not the one that is used (with all observations) (using all observations).

7.5 Extensions and remarks

• The generalization to the case where $p > 2$ is straighforward

• If some variables in the database are binary, it is better to use the logistic regression instead of classification rules which are usually based on normality assumption

• If the dataset is too large (too many variables), you can perform a stepwise discriminant analysis

• Others methods: Classification trees (CART), Neural Networks (NN), ...

Chapitre 8

Clustering

8.1 Introduction

Aim: Grouping "objects" based on measures of distances using stepwise algorithm

No assumptions are made concerning the number of groups or the group structure (which is different from classification)

Method: Grouping is based on similarities or distances calculated from

- data matrix X $(n \times p)$
- contingency table
- measures of association
- correlation coefficients

Similarity measures for individuals

Quantitatives variables

Let x and y be two p-dimensional observations $\in IR^P$. Different distances can be computed:

- Euclidean distance:

$$\begin{aligned} d(x,y) &= \sqrt{(x_1-y_1)^2+\ldots+(x_p-y_p)^2)} \\ &= \sqrt{((\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})} \end{aligned}$$

- Statistical distances:

$$d(x,y) = \sqrt{((\mathbf{x}-\mathbf{y})'\mathbf{A}(\mathbf{x}-\mathbf{y})}$$

where A is often given by $A = S^{-1}$ with S the sample covariance matrix.

- Minkowski distance:

$$d(x,y) = \left(\sum_{i=1}^{P}|x_i-y_i|^m\right)^{1/m}$$

For $m = 1$, $d(x,y)$ is the“city-block” distance and for $m = 2$ we recover the euclidian distance

Similarity measures for variables

Quantitatives variables

- Sample correlation coefficients
- Absolute values of correlation coefficients
- ...

Binary variables

- $\phi^2 = r^2 = \chi^2/n$
- Frequencies
- ...

Qualitative variables

- χ^2 statistics
- $\phi^2 = \chi^2/n$
- ...

There are many ways to measure similarity between individuals or variables

Stepwise algorithms:

Two families of algorithms:

- Nonhierarchical clustering methods: Direct partition into a fixed number of groups (clusters)
  - Moving centers method
  - "K-means" method
- Hierarchical clustering methods
  - Agglomerative hierarchical methods: start with individual objects, then the most similar objects are first grouped, and so on.
  - Divisive hierarchical methods: work in the opposite direction

A large literature exist on this subject

8.2 Nonhierarchical clustering methods

Mainly used for large database

Goal: Find q (fixed) groups of n individuals with

- homogeneity in the group
- heterogeneity between groups

$\Longrightarrow$ Find a criteria to measure the proximity among individuals of the same group and compare this measure for all possible partitions BUT .....

Example: 4 groups for 14 individuals : more than 10 millions of partitions

It is then impossible to find the "best partition"
$\Longrightarrow$ Used algorithm to find a partition "close"

to the "best" partition

8.2.1 Algorithm: Moving centers method

Let a set of n individuals with P characteristics

Let d be a distance in IR^P (euclidean, $\chi^2, \ldots$)

The number of groups is fixed to q

Step 0: Chose q starting centers (random selection of q individuals):

$$\{C_1^0, \ldots, C_k^0, \ldots, C_q^0\}$$

$$\Downarrow$$

Creation of a partition P^0: $\{I_1^0, \ldots, I_k^0, \ldots, I_q^0\}$ in q groups of the n individuals such that

$$i \in I_k^0 \quad \text{if } d(i, C_k^0) < d(i, C_j^0) \quad \forall j \in \{1, \ldots, q\} \neq k$$

Step 1: Let the new centers of the q groups be:

$$\{C_1^1, \ldots, C_k^1, \ldots, C_q^1\}$$

calculated as the centers of gravity of the q groups obtained in step 0: $\{I_1^0, \ldots, I_k^0, \ldots, I_q^0\}$

$$\Downarrow$$

Creation of a partition P^1 in q groups, using the same distance rule, of n individuals:

$$\{I_1^1, \ldots, I_k^1, \ldots, I_q^1\}$$

$$\vdots$$

Step m: Let the new centers of the q groups be:

$$\{C_1^m, \ldots, C_k^m, \ldots, C_q^m\}$$

calculated as centers of gravity of the q groups obtained in step $m-1$:

$$\{I_1^{m-1}, \ldots, I_k^{m-1}, \ldots, I_q^{m-1}\}$$

$\Longrightarrow$ Creation of a new partition P^m using the same methodology:

$$\{I_1^m, \ldots, I_k^m, \ldots, I_q^m\}$$

$$\vdots$$

Final Step: Stop the iterations

• if the number of iterations exceeds a given number of iterations which is chosen à priori (security)

• if two consecutive steps give the same partition

• if a statistical criteria (intra-class variance) doesn't decrease sufficiently

:-) This algorithm converges since we can prove that the intra-class variance never increases from step m to step $m+1$

:-(The final partition depends of the initial centers chosen randomly in step 0

8.2.2 Stable groups

The algorithm of moving centers method converges to local optimum since the final partition depends of the initial centers chosen randomly in step 0

$$\Downarrow$$

Find stable groups using several initial centers in step 0

Definition of stable groups: Set of individuals being always affected to the same cluster regardless of the initial conditions

Let $\{P_1, \ldots, P_s\}$ be s partitions in q groups

Product- Partition:

The group noted by $\{k_1, k_2, \ldots, k_s\}$ denotes the individuals $\in$ group k_1 of partition P_1

$\in$ group k_2 of partition P_2

$\vdots$

$\in$ group k_s of partition P_s

The groups of the product-partition containing a large number of individuals are called stable clusters (groups)

Example: 2 partitions of 113 individuals in 3 groups:

113	38	35	40
30	5	25	0
43	30	8	5
40	3	2	35

With 2 partitions: 9 possibilities

With 3 partitions: $3^3 = 27$ possibilities

Remarks

- Allow to explore high density areas
- The number of possibilities grows very fast

8.2.3 Algorithm: k-means method

The k-means method recomputes the new centers of gravity after each individual modification of clusters

Algorithm

- **Step 0**: Chose q starting centers

- **Step 1**: For the first individual in the database: chose the nearest center/cluster in terms of distance. And then recompute directly the gravity centers of the cluster "out" and the cluster "in". Perform the same procedure with the next individual

- **Step 2**: Repeat step 2 until convergence

Example: 4 individuals (A,B,C,D), 2 variables X_1 and X_2, 2 groups and euclidean distance.

Item	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

• Step 0: Determine randomly two groups: (AB) and (CD) and compute the gravity centers of these two groups

Groups	$\bar{x}_1$	$\bar{x}_2$
(AB)	$\frac{5+(-1)}{2} = 2$	$\frac{3+1}{2} = 2$
(CD)	$\frac{1+(-3)}{2} = -1$	$\frac{-2+(-2)}{2} = -2$

• Step 1: Determine the euclidean distance between A and the two centers of gravity:

$$\begin{aligned} d^2(A,(AB)) &= (5-2)^2 + (3-2)^2 = 10 \\ d^2(A,(CD)) &= (5+1)^2 + (3+2)^2 = 61 \end{aligned}$$

Therefore A remains in the same group (AB)

Determine the euclidean distance between B and the two centers of gravity:

$$\begin{aligned} d^2(B,(AB)) &= (-1-2)^2 + (1-2)^2 = 10 \\ d^2(B,(CD)) &= (-1+1)^2 + (1+2)^2 = 9 \end{aligned}$$

Therefore B is moved from cluster (AB) to cluster (CD). Recompute the centers of gravity:

Groups	$\bar{x}_1$	$\bar{x}_2$
A	5	3
(BCD)	-1	-1

Determine the euclidean distance between all individuals and the two centers of gravity:

Groups	A	B	C	D
A	0	40	41	89
(BCD)	52	4	5	5

C remains in group (BCD), D remains in group (BCD).

• Step 2: Repeat step 1. No modification of the clusters then the algorithm can STOP.

Conclusions:

:-) 1 iteration gives already a good idea of the final partition

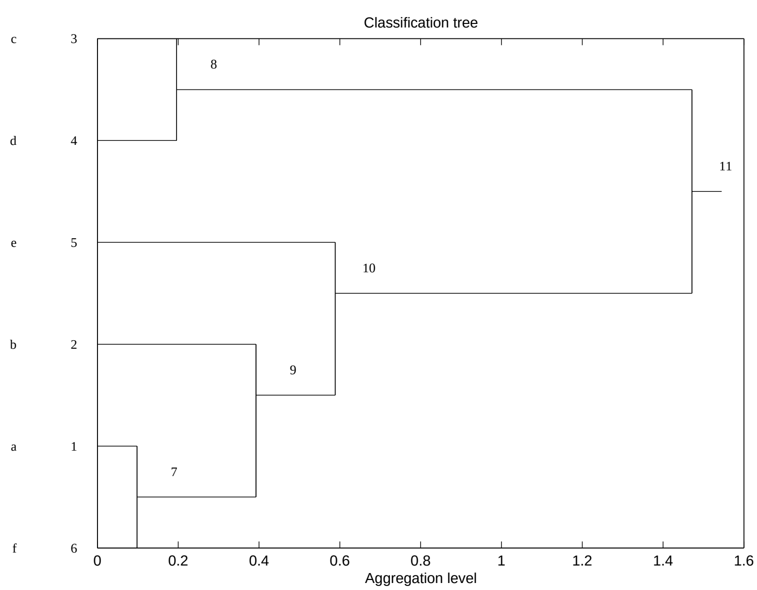
:-(the final partition depends of the ordering of individuals in the database

8.3 Agglomerative hierarchical clustering methods

Start with n clusters and aggregate two by two the nearest clusters

$$\Downarrow$$

Classification Tree or Dendrogram: set the $n-1$ partitions



- Cut the tree at a desired level: the cut branches on the left describe the corresponding clusters.
- gives a good idea of the number of groups but where to cut he tree ?: No real answer.

Principal problem: define the criteria to aggregate two clusters

$\Longrightarrow$ use an inertia criteria or define a distance between clusters:

Let d be the distance used between individuals

Question: How to measure the distance between the cluster $\{z\}$ and the cluster $h = \{x, y\}$?

- Single linkage:

$$d(h, z) = \min\{d(x, z), d(y, z)\}$$

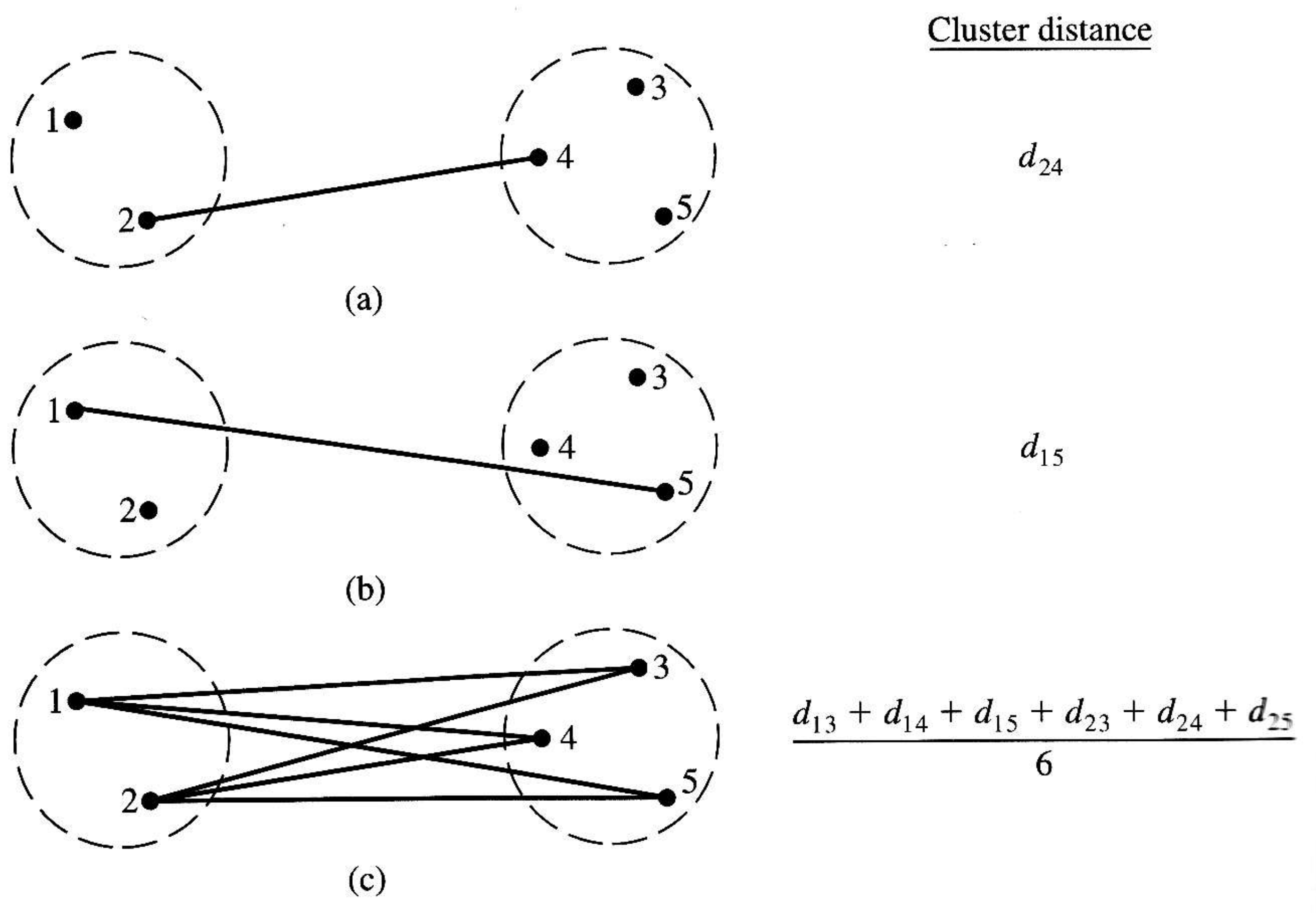
- Complete linkage:

$$d(h, z) = \max\{d(x, z), d(y, z)\}$$

- Average linkage:

$$d(h, z) = \frac{d(x, z) + d(y, z)}{2}$$

Schema of the three proposed linkages:



Algorithm for n individuals

Step 1: Compute the matrix of distances and form a cluster with the two nearest individuals

$\Longrightarrow$ Partition of $n-1$ clusters.

Step 2: Compute the matrix of distances between the $n-1$ clusters/objects and aggregate the two nearest clusters

$\Longrightarrow$ Partition of $n-2$ clusters.

$\vdots$

$\vdots$

Final step (n): Calculate the matrix of distances between the 2 last clusters

$\Longrightarrow$ Final partition containing all objects

Example: Single linkage.

The matrix of distances between 5 individuals is given by

$d(.,.)$	(1)	(2)	(3)	(4)	(5)
(1)	0				
(2)	9	0			
(3)	3	7	0		
(4)	6	5	9	0	
(5)	11	10	2	8	0

Step 1: $\min(d_{ij}) = d_{53} = 2 \Longrightarrow$ aggregate objets 5 and 3 in the new cluster (35). Partition in 4 clusters.

Step 2: Compute the new matrix of distances:

$$\begin{aligned} d_{(35)1} &= \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3 \\ d_{(35)2} &= \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7 \\ d_{(35)4} &= \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8 \end{aligned}$$

$d(.,.)$	(35)	(1)	(2)	(4)
(35)	0			
(1)	3	0		
(2)	7	9	0	
(4)	8	6	5	0

$\Longrightarrow$ aggregate cluster (35) with object 1. Partition in 3 clusters.

Step 3: Compute the new matrix of distances:

$d(.,.)$	(135)	(2)	(4)
(135)	0		
(2)	7	0	
(4)	6	5	0

$\Longrightarrow$ aggregate objects 4 and 2 to form cluster (24). Partition in 2 clusters.

Step 4: Compute the new matrix of distances:

$d(.,.)$	(135)	(24)
(135)	0	
(24)	6	0

$\Longrightarrow$ Final partition containing all objects

Dendrogram

Example: Complete linkage.

The matrix of distances between 5 individuals is given by

$d(.,.)$	(1)	(2)	(3)	(4)	(5)
(1)	0				
(2)	9	0			
(3)	3	7	0		
(4)	6	5	9	0	
(5)	11	10	2	8	0

Step 1: $\min(d_{ij}) = d_{53} = 2 \Longrightarrow$ aggregate objects 5 and 3 to form cluster (35). Partition in 4 clusters.

Step 2: Compute the new matrix of distances:

$$
\begin{aligned}
d_{(35)1} &= \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11 \\
d_{(35)2} &= \max\{d_{32}, d_{52}\} = \max\{7, 10\} = 10 \\
d_{(35)4} &= \max\{d_{34}, d_{54}\} = \max\{9, 8\} = 9
\end{aligned}
$$

$d(.,.)$	(35)	(1)	(2)	(4)
(35)	0			
(1)	11	0		
(2)	10	9	0	
(4)	9	6	5	0

$\Longrightarrow$ aggregate objects 2 and 4 to form cluster (24). Partition in 3 clusters.

Step 3: Calculate the new matrix of distances:

$d(.,.)$	(35)	(24)	(1)
(35)	0		
(24)	10	0	
(1)	11	9	0

$\Longrightarrow$ aggregate cluster (24) and object 1. Partition in 2 clusters.

Step 4: Compute the new matrix of distances:

$d(.,.)$	(35)	(124)
(35)	0	
(124)	11	0

$\Longrightarrow$ Final partition containing all objects

Dendrogram