Statistique des données d'expression
Chapitre 1 : Introduction

Alain Baccini & Philippe Besse

Laboratoire de Statistique et Probabilités
Université de Toulouse

Institut de Mathématiques
math.univ-toulouse.fr/biostat

Motivations

- Moyens informatiques de calcul et stockage
- analyse statistique des transcrits
- interface graphique, logiciels boîte noire et R
- contrôle et liberté des options

Motivations

- Moyens informatiques de calcul et stockage
- analyse statistique des transcrits
- interface graphique, logiciels boîte noire et R
- contrôle et liberté des options

Motivations

- Moyens informatiques de calcul et stockage
- analyse statistique des transcrits
- interface graphique, logiciels boîte noire et R
- contrôle et liberté des options

Motivations

- Moyens informatiques de calcul et stockage
- analyse statistique des transcrits
- interface graphique, logiciels boîte noire et R
- contrôle et liberté des options

Objectifs généraux

- Techniques statistiques exploratoires dites multidimensionnelles
- Méthodes factorielles et de classification
- Document complet mais présentation intuitive
- Illustration par des exemples simples et des données d'expression
- après normalisation, avant modélisation
- Fixer les choix pour des représentations pertinentes
- Utilisation réfléchie de R

Objectifs généraux

- Techniques statistiques exploratoires dites multidimensionnelles
- Méthodes factorielles et de classification
- Document complet mais présentation intuitive
- Illustration par des exemples simples et des données d'expression
- après normalisation, avant modélisation
- Fixer les choix pour des représentations pertinentes
- Utilisation réfléchie de R

Objectifs généraux

- Techniques statistiques exploratoires dites multidimensionnelles
- Méthodes factorielles et de classification
- Document complet mais présentation intuitive
- Illustration par des exemples simples et des données d'expression
- après normalisation, avant modélisation
- Fixer les choix pour des représentations pertinentes
- Utilisation réfléchie de R

Objectifs généraux

- Techniques statistiques exploratoires dites multidimensionnelles
- Méthodes factorielles et de classification
- Document complet mais présentation intuitive
- Illustration par des exemples simples et des données d'expression
- après normalisation, avant modélisation
- Fixer les choix pour des représentations pertinentes
- Utilisation réfléchie de R

Objectifs généraux

- Techniques statistiques exploratoires dites multidimensionnelles
- Méthodes factorielles et de classification
- Document complet mais présentation intuitive
- Illustration par des exemples simples et des données d'expression
- après normalisation, avant modélisation
- Fixer les choix pour des représentations pertinentes
- Utilisation réfléchie de R

Objectifs généraux

- Techniques statistiques exploratoires dites multidimensionnelles
- Méthodes factorielles et de classification
- Document complet mais présentation intuitive
- Illustration par des exemples simples et des données d'expression
- après normalisation, avant modélisation
- Fixer les choix pour des représentations pertinentes
- Utilisation réfléchie de R

Objectifs généraux

- Techniques statistiques exploratoires dites multidimensionnelles
- Méthodes factorielles et de classification
- Document complet mais présentation intuitive
- Illustration par des exemples simples et des données d'expression
- après normalisation, avant modélisation
- Fixer les choix pour des représentations pertinentes
- Utilisation réfléchie de R

Objectif exploratoire pour dépister les problèmes

- valeurs manquantes, erronées ou atypiques
- modalités ou classes trop rares
- distributions “anormales” (dissymétrie, multimodalité...)
- incohérences, liaisons non linéaires

Pré-traitement des données

- transformations : logarithme, puissance, centrage, réduction, rangs. . .
- imputations, suppression des données manquantes
- Attention aux artefacts

Objectif exploratoire pour dépister les problèmes

- valeurs manquantes, erronées ou atypiques
- modalités ou classes trop rares
- distributions “anormales” (dissymétrie, multimodalité...)
- incohérences, liaisons non linéaires

Pré-traitement des données

- transformations : logarithme, puissance, centrage, réduction, rangs. . .
- imputations, suppression des données manquantes
- Attention aux artefacts

Méthodes

- Description uni puis bi-dimensionnelle
- Texhniques factorielles : ACP, AFD, AFC, AC, MDS
- Classification (non supervisée) : CAH, réallocation dynamique
- Puis : tests et modèle linéaire, étude de cas

Application aux données d'expression

- Nutrition chez la souris (P. Martin)
- Obésité humaine (N. Vignerie)
- Cancer pancréatique humain (H. Laurell)

Méthodes

- Description uni puis bi-dimensionnelle
- Texhniques factorielles : ACP, AFD, AFC, AC, MDS
- Classification (non supervisée) : CAH, réallocation dynamique
- Puis : tests et modèle linéaire, étude de cas

Application aux données d'expression

- Nutrition chez la souris (P. Martin)
- Obésité humaine (N. Vignerie)
- Cancer pancréatique humain (H. Laurell)

Méthodes

- Description uni puis bi-dimensionnelle
- Texhniques factorielles : ACP, AFD, AFC, AC, MDS
- Classification (non supervisée) : CAH, réallocation dynamique
- Puis : tests et modèle linéaire, étude de cas

Application aux données d'expression

- Nutrition chez la souris (P. Martin)
- Obésité humaine (N. Vignerie)
- Cancer pancréatique humain (H. Laurell)

Méthodes

- Description uni puis bi-dimensionnelle
- Texhniques factorielles : ACP, AFD, AFC, AC, MDS
- Classification (non supervisée) : CAH, réallocation dynamique
- Puis : tests et modèle linéaire, étude de cas

Application aux données d'expression

- Nutrition chez la souris (P. Martin)
- Obésité humaine (N. Vignerie)
- Cancer pancréatique humain (H. Laurell)

Méthodes

- Description uni puis bi-dimensionnelle
- Texhniques factorielles : ACP, AFD, AFC, AC, MDS
- Classification (non supervisée) : CAH, réallocation dynamique
- Puis : tests et modèle linéaire, étude de cas

Application aux données d'expression

- Nutrition chez la souris (P. Martin)
- Obésité humaine (N. Vignerie)
- Cancer pancréatique humain (H. Laurell)

Méthodes

- Description uni puis bi-dimensionnelle
- Texhniques factorielles : ACP, AFD, AFC, AC, MDS
- Classification (non supervisée) : CAH, réallocation dynamique
- Puis : tests et modèle linéaire, étude de cas

Application aux données d'expression

- Nutrition chez la souris (P. Martin)
- Obésité humaine (N. Vignerie)
- Cancer pancréatique humain (H. Laurell)

Méthodes

- Description uni puis bi-dimensionnelle
- Texhniques factorielles : ACP, AFD, AFC, AC, MDS
- Classification (non supervisée) : CAH, réallocation dynamique
- Puis : tests et modèle linéaire, étude de cas

Application aux données d'expression

- Nutrition chez la souris (P. Martin)
- Obésité humaine (N. Vignerie)
- Cancer pancréatique humain (H. Laurell)

Spécificités des données d'expression

- Nb de gènes *v.s.* taille de l'échantillon
- gène : "variable" ou "individu" ?
- Distances associées ?
- Variables biologiques complémentaires ?

Spécificités des données d'expression

- Nb de gènes *v.s.* taille de l'échantillon
- gène : "variable" ou "individu" ?
- Distances associées ?
- Variables biologiques complémentaires ?

Spécificités des données d'expression

- Nb de gènes *v.s.* taille de l'échantillon
- gène : "variable" ou "individu" ?
- Distances associées ?
- Variables biologiques complémentaires ?

Spécificités des données d'expression

- Nb de gènes *v.s.* taille de l'échantillon
- gène : "variable" ou "individu" ?
- Distances associées ?
- Variables biologiques complémentaires ?

Transformations après normalisation

- Logarithme : facteurs multiplicatifs plutôt qu'additifs
- Centrage : retrancher la moyenne empirique (lignes et/ou colonnes)
- Réduction : éliminer l'unité de mesure
- Marges unitaires et AFC
- Rangs et corrélation sur les rangs (Spearman) plus robuste

Choix à rendre explicite

- Poids des individus et distance entre variables
- Distances entre individus
- Nombre de facteurs ou de composantes
- Nombre de classes

Transformations après normalisation

- Logarithme : facteurs multiplicatifs plutôt qu'additifs
- Centrage : retrancher la moyenne empirique (lignes et/ou colonnes)
- Réduction : éliminer l'unité de mesure
- Marges unitaires et AFC
- Rangs et corrélation sur les rangs (Spearman) plus robuste

Choix à rendre explicite

- Poids des individus et distance entre variables
- Distances entre individus
- Nombre de facteurs ou de composantes
- Nombre de classes

Statistique des données d'expression
Chapitre 2 : Description statistique élémentaire

Alain Baccini & Philippe Besse

Laboratoire de Statistique et Probabilités
Université de Toulouse

Institut de Mathématiques
math.univ-toulouse.fr/biostat

Indicateurs

- X, variable réelle observée sur
- n individus de poids w_i.
- *tendance centrale* : médiane, moyenne

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

- *dispersion* : écart-type σ, intervalle inter-quartiles
- dissymétrie (skeeness), applatissement (kurtosis)
- test de normalité (Kolmogorov)

Indicateurs

- X, variable réelle observée sur
- n individus de poids w_i.
- *tendance centrale* : médiane, moyenne

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

- *dispersion* : écart-type σ, intervalle inter-quartiles
- dissymétrie (skeeness), applatissement (kurtosis)
- test de normalité (Kolmogorov)

Indicateurs

- X, variable réelle observée sur
- n individus de poids w_i.
- *tendance centrale* : médiane, moyenne

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

- *dispersion* : écart-type σ, intervalle inter-quartiles
- dissymétrie (skeeness), applatissement (kurtosis)
- test de normalité (Kolmogorov)

Indicateurs

- X, variable réelle observée sur
- n individus de poids w_i.
- *tendance centrale* : médiane, moyenne

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

- *dispersion* : écart-type σ, intervalle inter-quartiles
- dissymétrie (skeeness), applatissement (kurtosis)
- test de normalité (Kolmogorov)

Indicateurs

- X, variable réelle observée sur
- n individus de poids w_i.
- *tendance centrale* : médiane, moyenne

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

- *dispersion* : écart-type σ, intervalle inter-quartiles
- dissymétrie (skeeness), applatissement (kurtosis)
- test de normalité (Kolmogorov)

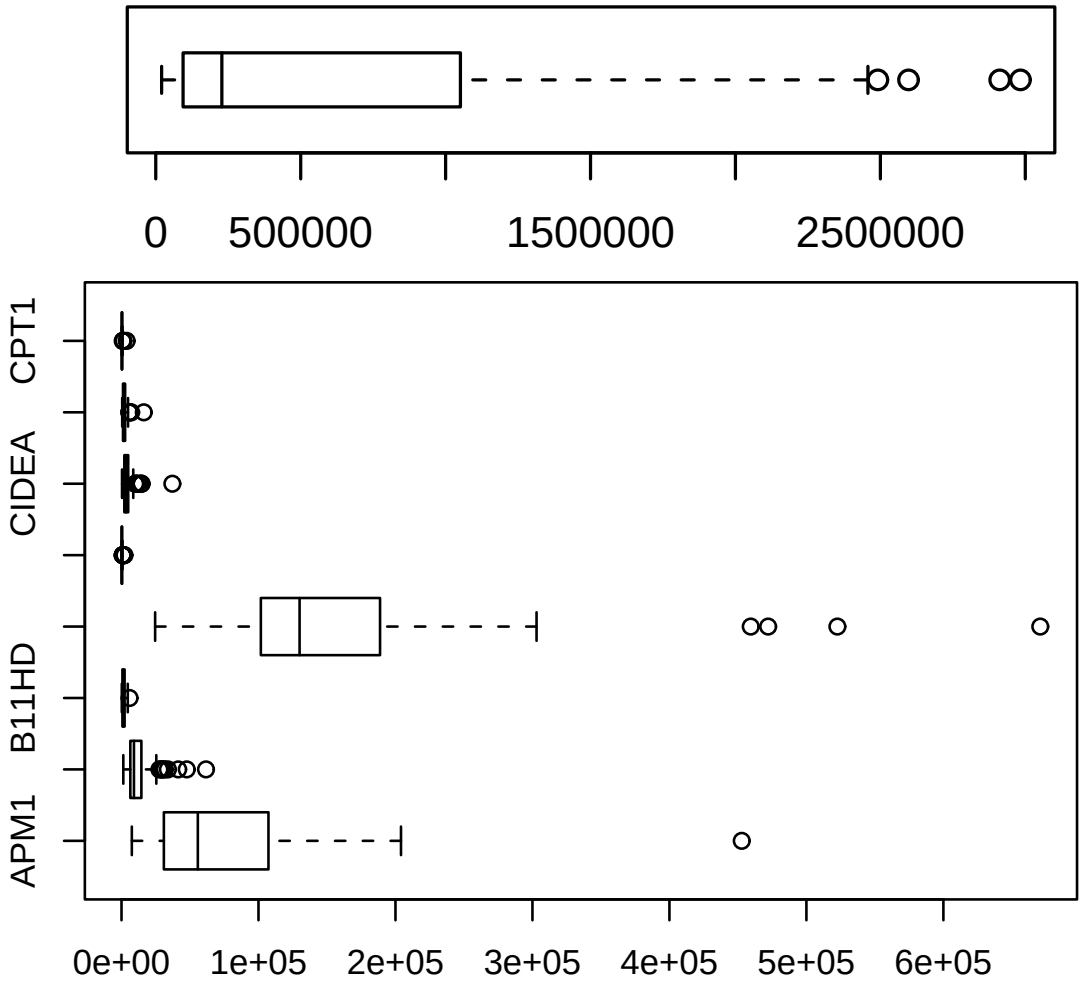
Indicateurs

- X, variable réelle observée sur
- n individus de poids w_i.
- *tendance centrale* : médiane, moyenne

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

- *dispersion* : écart-type σ, intervalle inter-quartiles
- dissymétrie (skeeness), applatissement (kurtosis)
- test de normalité (Kolmogorov)

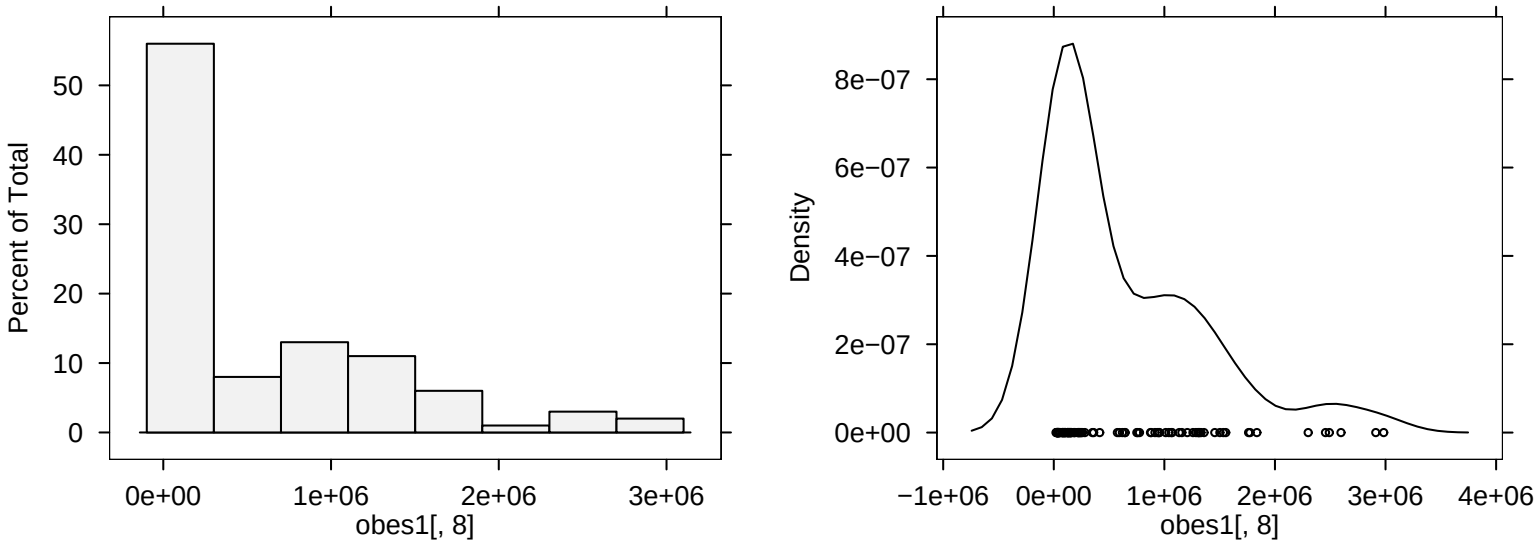
Obésité : Diagrammes boîte et distributions des gènes



Histogramme et estimation fonctionnelle

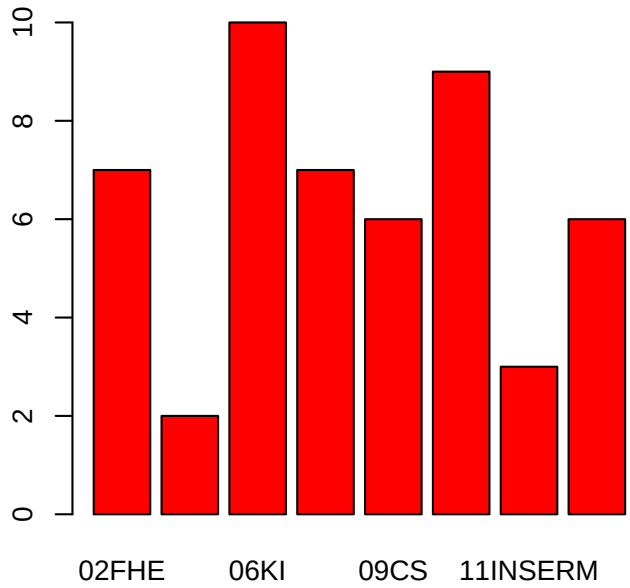
- Histogramme avec découpage en classes
- Estimation fonctionnelle avec λ : paramètre de lissage

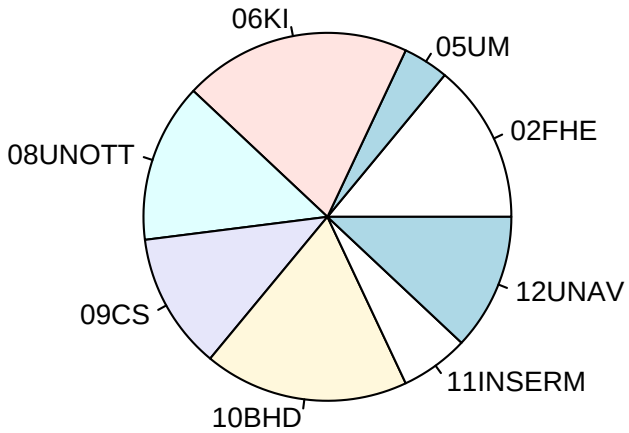
$$\widehat{g}_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) \quad \text{où} \quad K(t) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$$



Obésité : Estimations de la distribution d'un gène

Une variable qualitative : diagramme en barres





Obésité : répartition des centres

Deux variables quantitatives

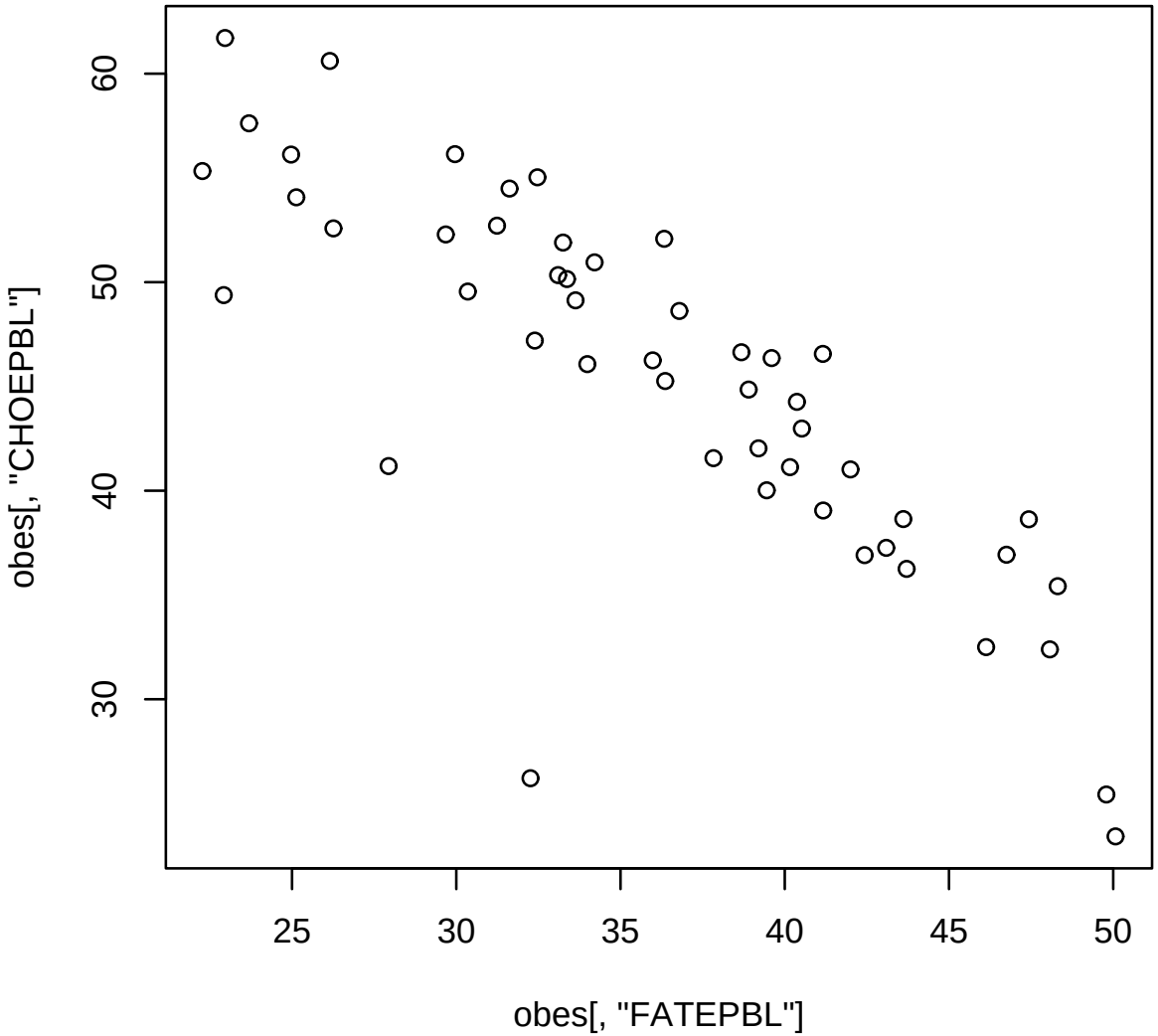
- Covariance

$$\begin{aligned} \text{cov}(X,Y) &= \sum_{i=1}^{n} w_i [x_i - \bar{x}][y_i - \bar{y}] \\ &= [\sum_{i=1}^{n} w_i x_i y_i] - \bar{x}\,\bar{y}. \end{aligned}$$

- Corrélation linéaire

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Obésité : liaison entre deux variables biologiques



Notations

- X variable qualitative à r modalités :

$$x_1, \ldots, x_\ell, \ldots, x_r$$

- Y quantitative de moyenne $\bar{y}$ et de variance σ_Y^2
- Les modalités x_ℓ de X définissent une *partition* de Ω en m classes de cardinaux $n_1, \ldots, n_m$:

$$\sum_{\ell=1}^{m} n_\ell = n \text{ où } n = card(\Omega).$$

$$\bar{y}_\ell = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} Y(\omega_i)\,; \quad \sigma_\ell^2 = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} [Y(\omega_i) - \overline{y_\ell}]^2.$$

Notations

- X variable qualitative à r modalités :

$$x_1, \ldots, x_\ell, \ldots, x_r$$

- Y quantitative de moyenne $\bar{y}$ et de variance σ_Y^2
- Les modalités x_ℓ de X définissent une *partition* de Ω en m classes de cardinaux $n_1, \ldots, n_m$:

$$\sum_{\ell=1}^{m} n_\ell = n \text{ où } n = card(\Omega).$$

$$\bar{y}_\ell = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} Y(\omega_i)\,; \quad \sigma_\ell^2 = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} [Y(\omega_i) - \overline{y_\ell}]^2.$$

Notations

- X variable qualitative à r modalités :

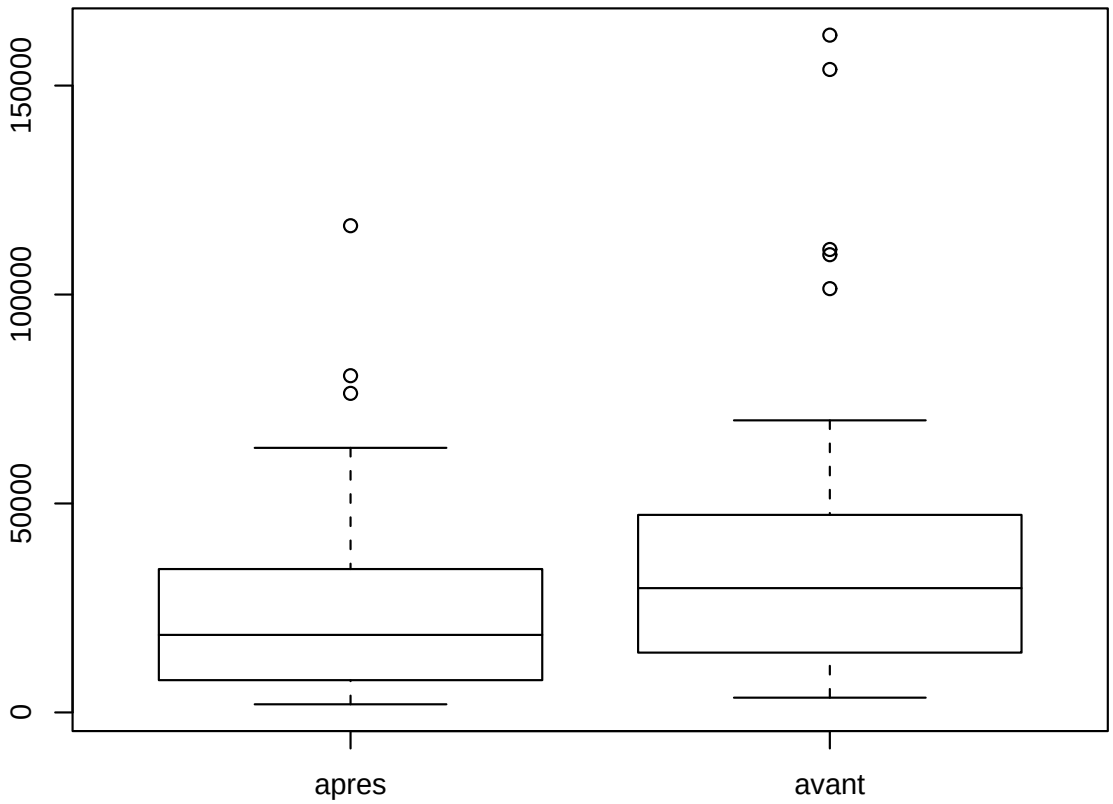
$$x_1, \ldots, x_\ell, \ldots, x_r$$

- Y quantitative de moyenne $\bar{y}$ et de variance σ_Y^2
- Les modalités x_ℓ de X définissent une *partition* de Ω en m classes de cardinaux $n_1, \ldots, n_m$:

$$\sum_{\ell=1}^{m} n_\ell = n \text{ où } n = card(\Omega).$$

$$\bar{y}_\ell = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} Y(\omega_i)\,; \quad \sigma_\ell^2 = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} [Y(\omega_i) - \overline{y_\ell}]^2.$$

Obésité : Distributions d'un gène avant / après régime



Décomposition de la variance

$$\bar{y} = \frac{1}{n}\sum_{\ell=1}^{r} n_\ell \overline{y_\ell}\,;$$

$$\sigma_Y^2 = \frac{1}{n}\sum_{\ell=1}^{r} n_\ell (\overline{y_\ell} - \bar{y})^2 + \frac{1}{n}\sum_{\ell=1}^{r} n_\ell \sigma_\ell^2 = \sigma_E^2 + \sigma_R^2\,.$$

Rapport de corrélation

$$s_{Y/X} = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}}\,;$$

Décomposition de la variance

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{\ell=1}^{r} n_\ell \overline{y_\ell}\,; \\ \sigma_Y^2 &= \frac{1}{n}\sum_{\ell=1}^{r} n_\ell (\overline{y_\ell} - \bar{y})^2 + \frac{1}{n}\sum_{\ell=1}^{r} n_\ell \sigma_\ell^2 = \sigma_E^2 + \sigma_R^2\,. \end{aligned}$$

Rapport de corrélation

$$s_{Y/X} = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}}\,;$$

Table de contingence entre 2 variables qualitatives

- X et Y qualitatives à r et c modalités :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

- Effectifs conjoints : $n_{\ell h}$ et marginaux : $n_{\ell +}$ et n_{+h}
- Profils : $\left\{\frac{n_{\ell 1}}{n_{\ell +}}, \cdots, \frac{n_{\ell h}}{n_{\ell +}}, \cdots, \frac{n_{\ell c}}{n_{\ell +}}\right\}$ $\left\{\frac{n_{1h}}{n_{+h}}, \cdots, \frac{n_{\ell h}}{n_{+h}}, \cdots, \frac{n_{rh}}{n_{+h}}\right\}$.

Table de contingence entre 2 variables qualitatives

- X et Y qualitatives à r et c modalités :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

- Effectifs conjoints : $n_{\ell h}$ et marginaux : $n_{\ell +}$ et n_{+h}
- Profils : $\left\{\frac{n_{\ell 1}}{n_{\ell +}}, \cdots, \frac{n_{\ell h}}{n_{\ell +}}, \cdots, \frac{n_{\ell c}}{n_{\ell +}}\right\}$ $\left\{\frac{n_{1h}}{n_{+h}}, \cdots, \frac{n_{\ell h}}{n_{+h}}, \cdots, \frac{n_{rh}}{n_{+h}}\right\}$.

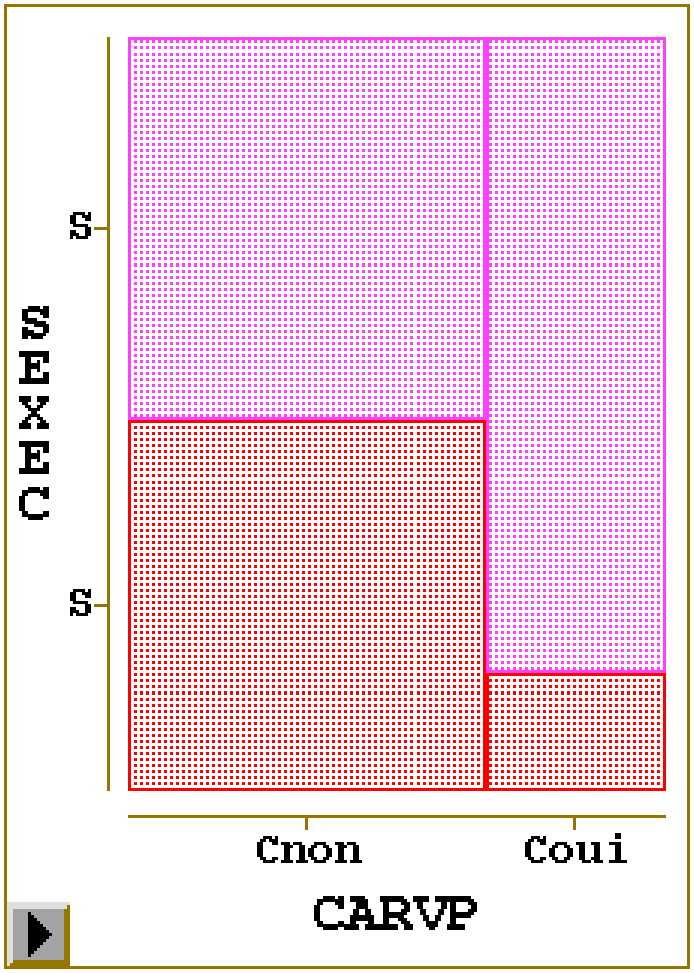
Table de contingence entre 2 variables qualitatives

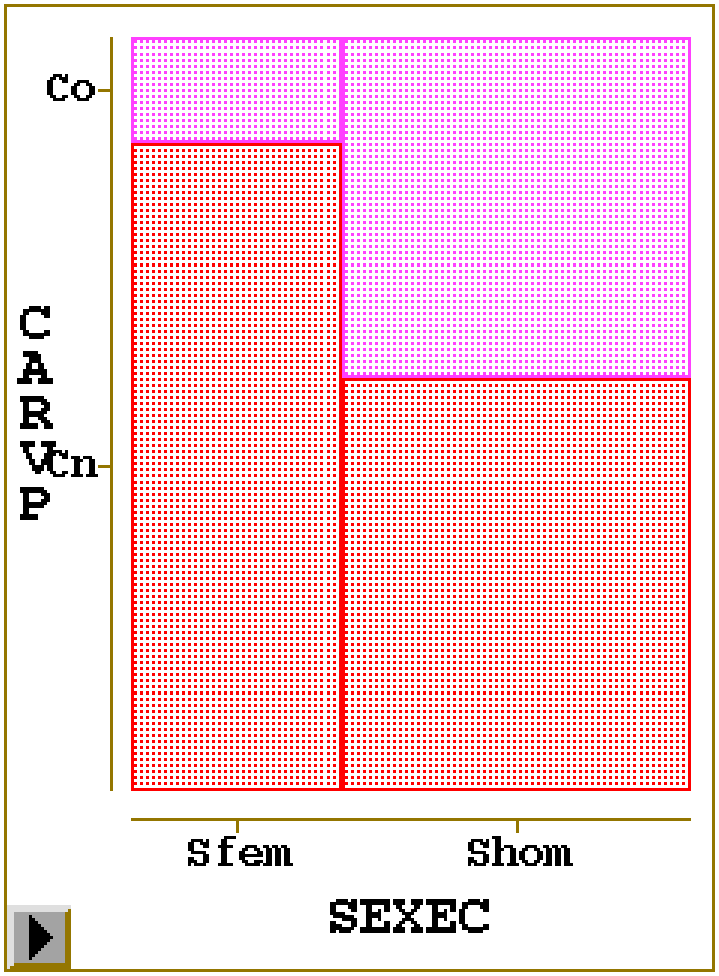
- X et Y qualitatives à r et c modalités :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

- Effectifs conjoints : $n_{\ell h}$ et marginaux : $n_{\ell +}$ et n_{+h}
- Profils : $\left\{\frac{n_{\ell 1}}{n_{\ell +}}, \ldots, \frac{n_{\ell h}}{n_{\ell +}}, \ldots, \frac{n_{\ell c}}{n_{\ell +}}\right\} \left\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\right\}$.

Graphe des profils





Indices de liaison

- Aucune liaison entre X et Y ssi :

$$n_{\ell h} = \frac{n_{\ell +} n_{+h}}{n} \; \forall (\ell, h) \in \{1, \ldots, r\} \times \{1, \ldots, c\}$$

Khi-deux :
$$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n})^2}{\frac{n_{\ell +} n_{+h}}{n}}$$

Phi-deux :
$$\Phi^2 = \frac{\chi^2}{n}$$

T de Tschuprow :
$$T = \sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}$$

- On vérifie : $0 \leq T \leq 1$

Indices de liaison

- Aucune liaison entre X et Y ssi :

$$n_{\ell h} = \frac{n_{\ell +} n_{+h}}{n} \; \forall (\ell, h) \in \{1, \ldots, r\} \times \{1, \ldots, c\}$$

Khi-deux :
$$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n})^2}{\frac{n_{\ell +} n_{+h}}{n}}$$

Phi-deux :
$$\Phi^2 = \frac{\chi^2}{n}$$

T de Tschuprow :
$$T = \sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}$$

- On vérifie : $0 \leq T \leq 1$

Plusieurs variables

- p variables de même type sont observées.
- Matrices $(p \times p)$ des indices précédents :
  - covariances : $\mathbf{S}_i^j = \text{cov}(X^i, X^j)$;
  - corrélations : $\mathbf{R}_i^j = \text{cor}(X^i, X^j)$;
  - coefficients de Tschuprow ,
  - tableau des nuages de points (SPloM).

Plusieurs variables

- p variables de même type sont observées.
- Matrices $(p \times p)$ des indices précédents :
  - covariances : $\mathbf{S}_i^j = \text{cov}(X^i, X^j)$;
  - corrélations : $\mathbf{R}_i^j = \text{cor}(X^i, X^j)$;
  - coefficients de Tschuprow ,
  - tableau des nuages de points (SPloM).

Plusieurs variables

- p variables de même type sont observées.
- Matrices $(p \times p)$ des indices précédents :
  - covariances : $\mathbf{S}_i^j = \text{cov}(X^i, X^j)$;
  - corrélations : $\mathbf{R}_i^j = \text{cor}(X^i, X^j)$;
  - coefficients de Tschuprow ,
  - tableau des nuages de points (SPloM).

Plusieurs variables

- p variables de même type sont observées.
- Matrices $(p \times p)$ des indices précédents :
  - covariances : $\mathbf{S}_i^j = \text{cov}(X^i, X^j)$;
  - corrélations : $\mathbf{R}_i^j = \text{cor}(X^i, X^j)$;
  - coefficients de Tschuprow ,
  - tableau des nuages de points (SPloM).

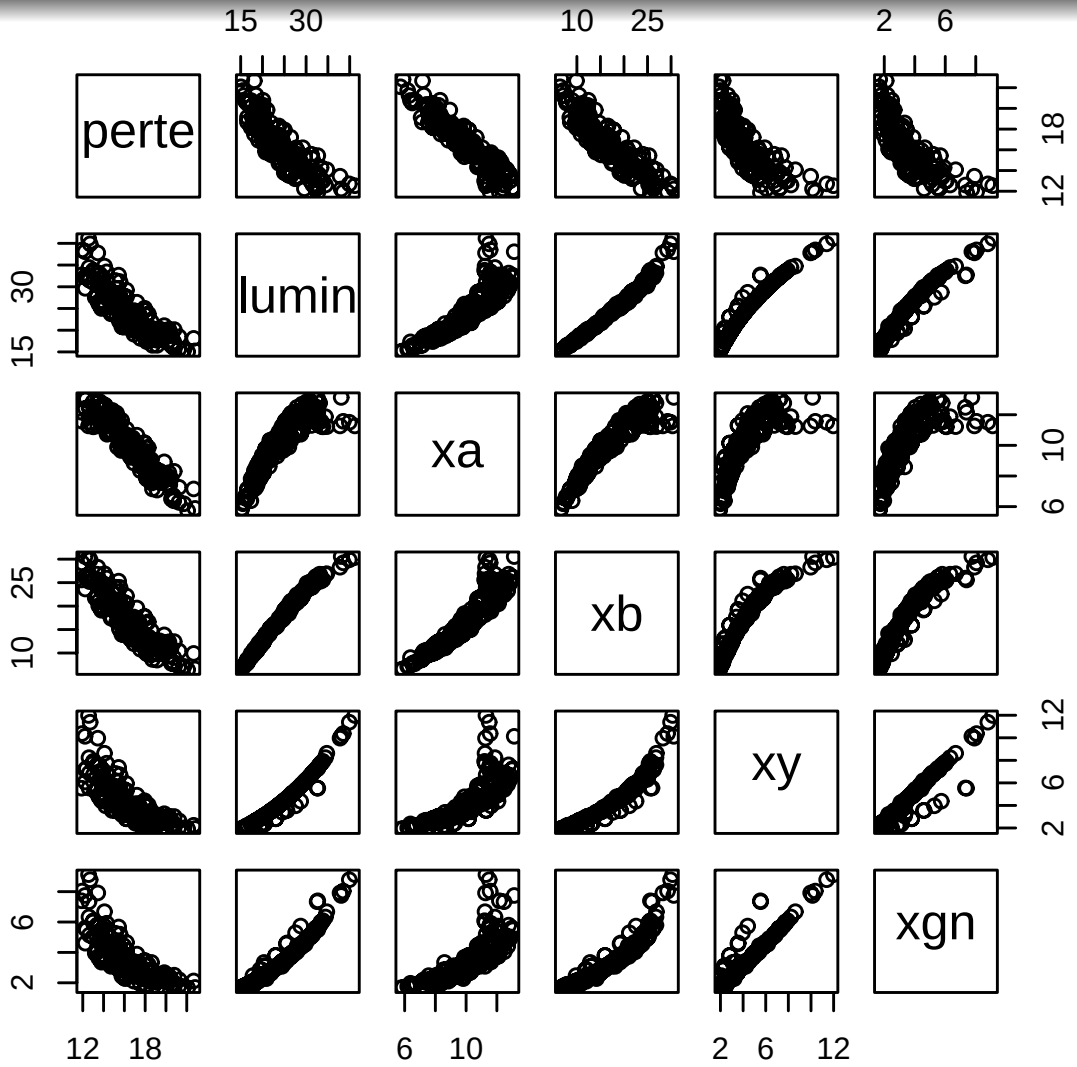
Plusieurs variables

- p variables de même type sont observées.
- Matrices $(p \times p)$ des indices précédents :
  - covariances : $\mathbf{S}_i^j = \text{cov}(X^i, X^j)$;
  - corrélations : $\mathbf{R}_i^j = \text{cor}(X^i, X^j)$;
  - coefficients de Tschuprow ,
  - tableau des nuages de points (SPloM).

Plusieurs variables

- p variables de même type sont observées.
- Matrices $(p \times p)$ des indices précédents :
  - covariances : $\mathbf{S}_i^j = \text{cov}(X^i, X^j)$;
  - corrélations : $\mathbf{R}_i^j = \text{cor}(X^i, X^j)$;
  - coefficients de Tschuprow ,
  - tableau des nuages de points (SPloM).

Matrice de nuages de points



Diagnostics

- détection d'erreurs, d'incohérences
- mitage de l'ensemble des données
- présence de valeurs atypiques
- "normalité" des distributions

Diagnostics

- détection d'erreurs, d'incohérences
- mitage de l'ensemble des données
- présence de valeurs atypiques
- "normalité" des distributions

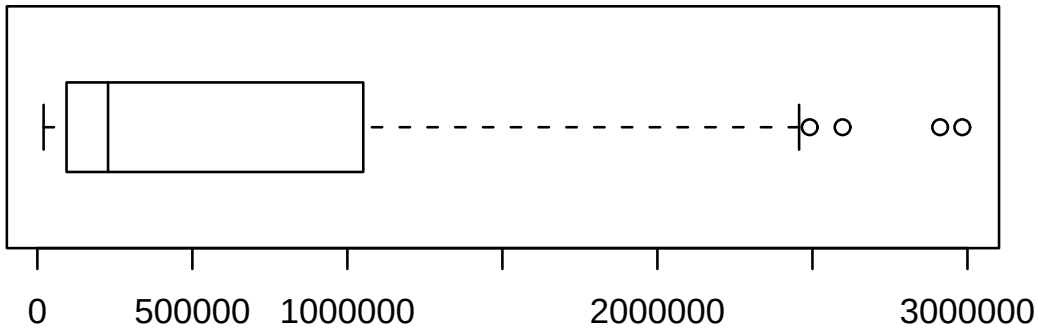
Diagnostics

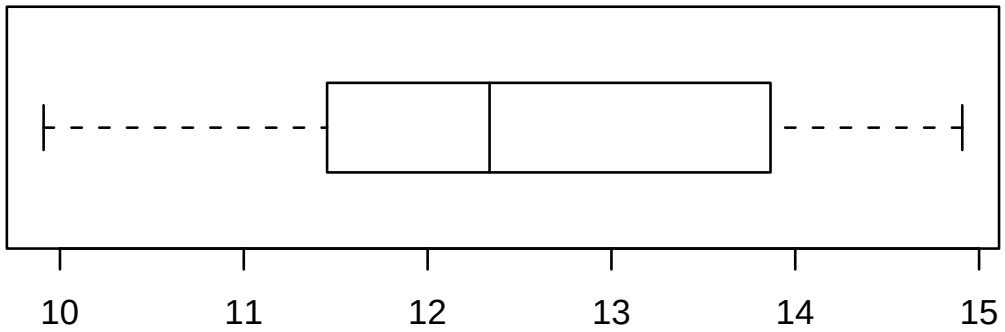
- détection d'erreurs, d'incohérences
- mitage de l'ensemble des données
- présence de valeurs atypiques
- "normalité" des distributions

Diagnostics

- détection d'erreurs, d'incohérences
- mitage de l'ensemble des données
- présence de valeurs atypiques
- "normalité" des distributions

Obésité : transformation en log de l'expression





Données transcriptomiques

- 40 souris, 2 génotypes (sauvages, PParα), 5 régimes
- 120 gènes
- 21 concentrations d'acides gras

Données transcriptomiques

- 40 souris, 2 génotypes (sauvages, PParα), 5 régimes
- 120 gènes
- 21 concentrations d'acides gras

Données transcriptomiques

- 40 souris, 2 génotypes (sauvages, PParα), 5 régimes
- 120 gènes
- 21 concentrations d'acides gras

Diagramme boîte des 40 souris

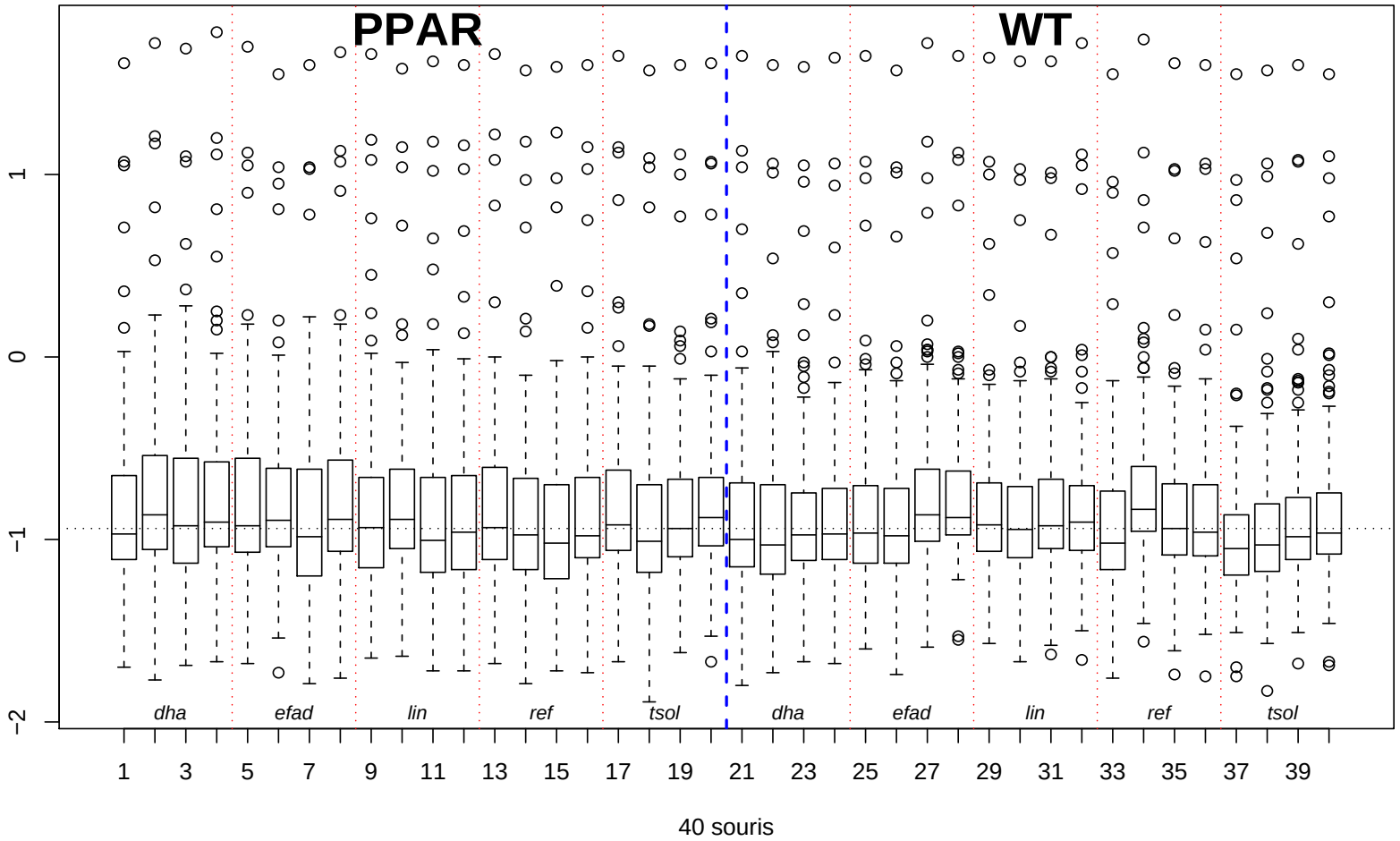
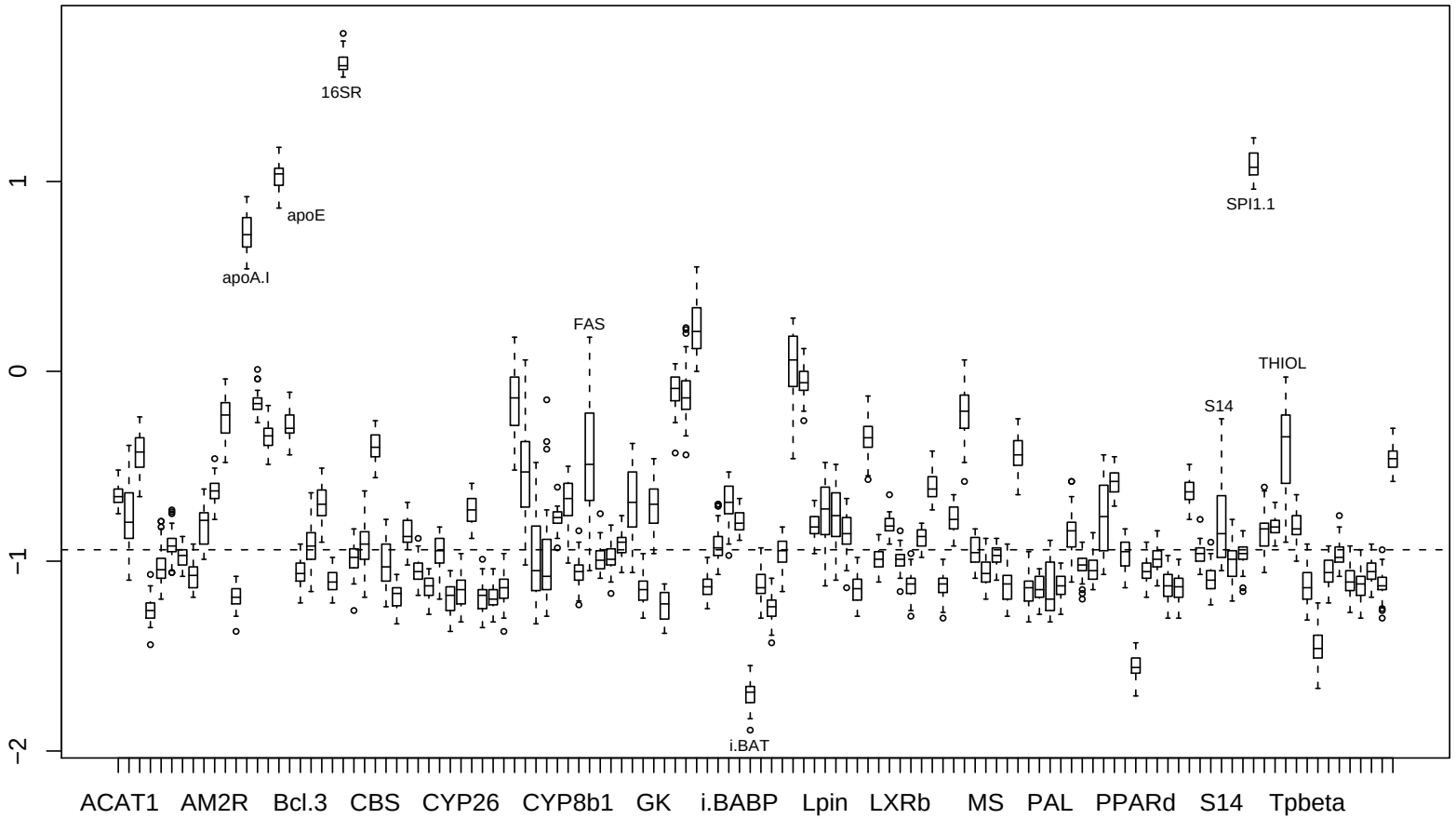


Diagramme boîte des 120 gènes



Statistique des données d'expression Chapitre 3 : Analyse en Composantes Principales

Alain Baccini & Philippe Besse

Laboratoire de Statistique et Probabilités
Université de Toulouse

Institut de Mathématiques
math.univ-toulouse.fr/biostat

Objectifs

- Représenter graphiquement l'observation de $p > 3$ variables
- Recherche de résumés pertinents (nuages de point) dans le plan
- respectant les distances entre individus
- la structure des corrélations entre variables

Objectifs

- Représenter graphiquement l'observation de $p > 3$ variables
- Recherche de résumés pertinents (nuages de point) dans le plan
- respectant les distances entre individus
- la structure des corrélations entre variables

Objectifs

- Représenter graphiquement l'observation de $p > 3$ variables
- Recherche de résumés pertinents (nuages de point) dans le plan
- respectant les distances entre individus
- la structure des corrélations entre variables

Objectifs

- Représenter graphiquement l'observation de $p > 3$ variables
- Recherche de résumés pertinents (nuages de point) dans le plan
- respectant les distances entre individus
- la structure des corrélations entre variables

Notes de $n = 9$ élèves dans $p = 4$ disciplines

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Notes de $n = 9$ élèves dans $p = 4$ disciplines

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Matrice des variances-covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	+ 8.94	4.12	5.48
FRAN	2.66	4.12	+ 12.06	9.29
ANGL	4.82	5.48	9.29	+ 7.91 = 40.30

Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00

Matrice des variances-covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	+ 8.94	4.12	5.48
FRAN	2.66	4.12	+ 12.06	9.29
ANGL	4.82	5.48	9.29	+ 7.91 = 40.30

Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00

Matrice des variances-covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	+ 8.94	4.12	5.48
FRAN	2.66	4.12	+ 12.06	9.29
ANGL	4.82	5.48	9.29	+ 7.91 = 40.30

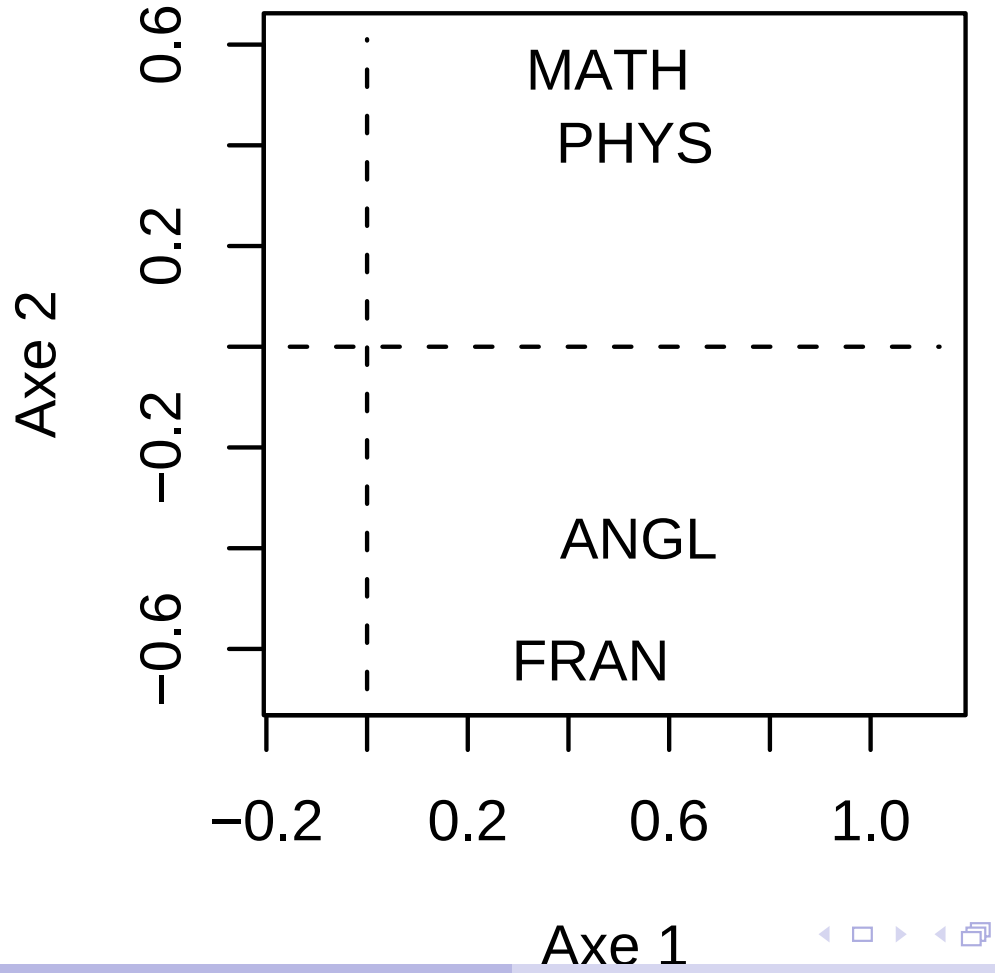
Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00

Espace des variables

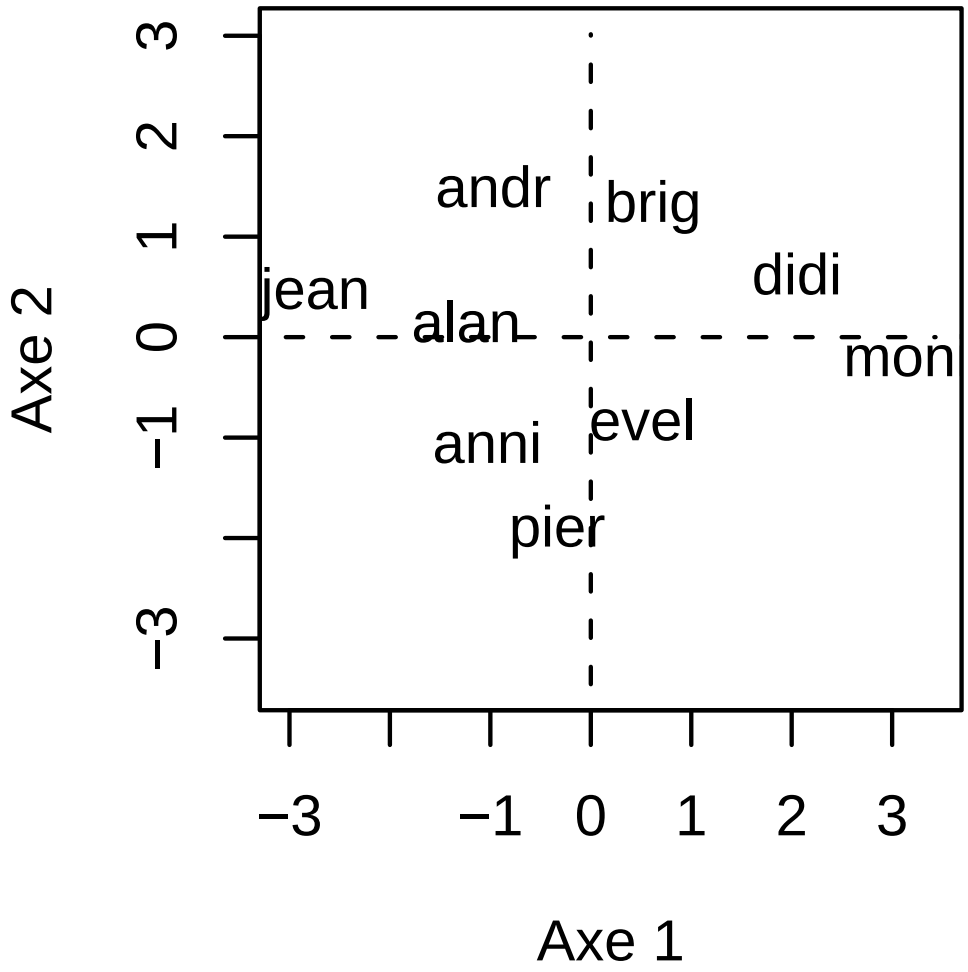


Espace des individus

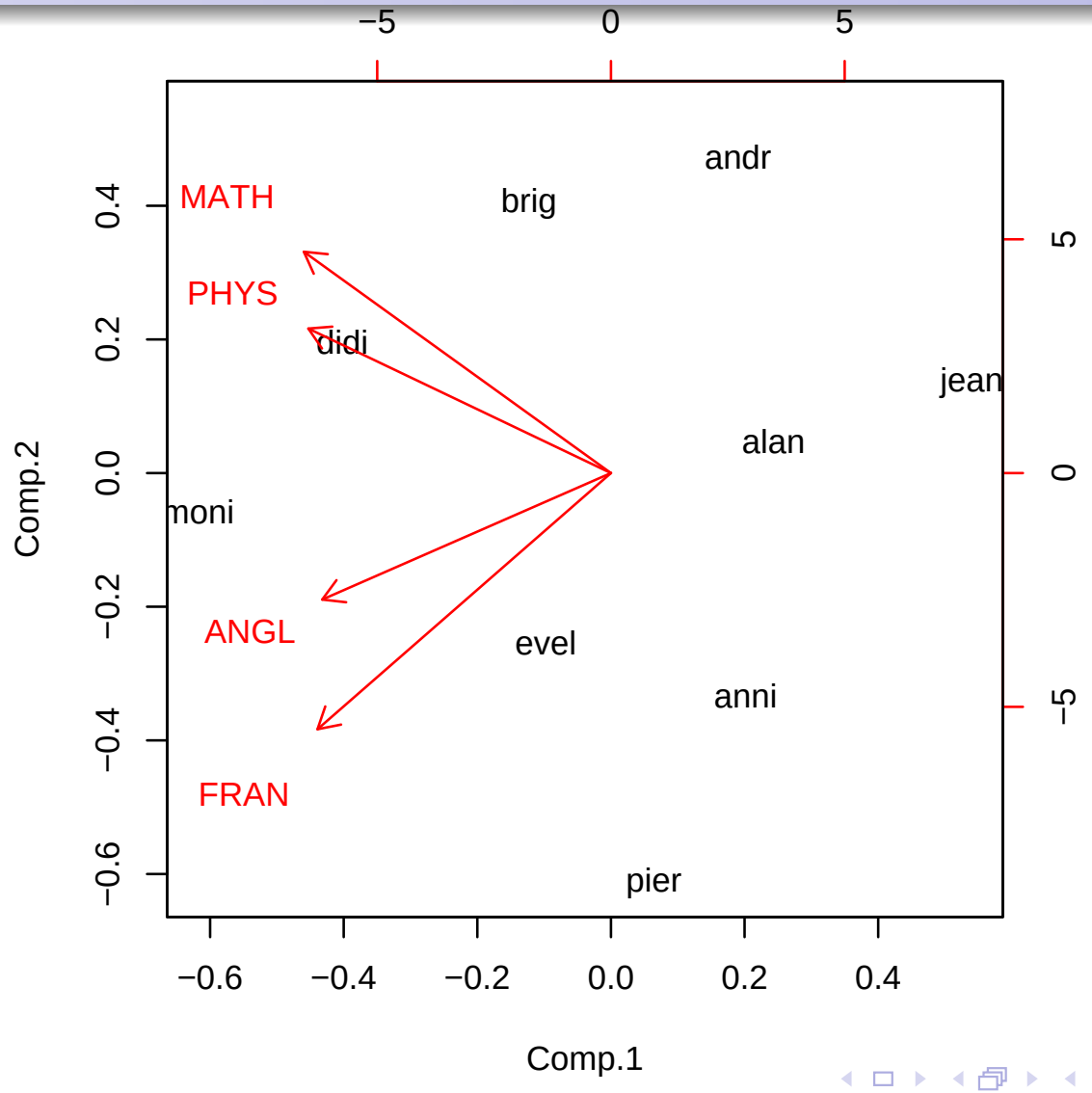
Coordonnées des individus ; contributions ; cosinus carrés

	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
jean	-8.61	-1.41	20.99	29.19	1.83	0.97	0.03
alan	-3.88	-0.50	4.22	5.92	0.23	0.98	0.02
anni	-3.21	3.47	6.17	4.06	11.11	0.46	0.54
moni	9.85	0.60	26.86	38.19	0.33	1.00	0.00
didi	6.41	-2.05	12.48	16.15	3.87	0.91	0.09
andr	-3.03	-4.92	9.22	3.62	22.37	0.28	0.72
pier	-1.03	6.38	11.51	0.41	37.56	0.03	0.97
brig	1.95	-4.20	5.93	1.50	16.29	0.18	0.82
evel	1.55	2.63	2.63	0.95	6.41	0.25	0.73

Espace des individus



Représentation simultanée



Notations

- p variables statistiques réelles X^j observées sur
- n individus de poids w_i

Représentation vectorielle

- Individu i : $\mathbf{x}_i$, i-ème ligne de $\mathbf{X}$ mise en colonne

$$\mathbf{x}_i \in E \text{ isomorphe à } (\mathbb{R}^p, \mathcal{E}, \mathbf{M})$$

- Variable X^j : $\mathbf{x}^j$, j-ème colonne centrée de $\mathbf{X}$

$\mathbf{x}^j \in F$ est isomorphe à $(\mathbb{R}^n, \mathcal{F}, \mathbf{D})$ avec $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$

Notations

- p variables statistiques réelles X^j observées sur
- n individus de poids w_i

Représentation vectorielle

- Individu i : $\mathbf{x}_i$, i-ème ligne de $\mathbf{X}$ mise en colonne

$$\mathbf{x}_i \in E \text{ isomorphe à } (\mathbb{R}^p, \mathcal{E}, \mathbf{M})$$

- Variable X^j : $\mathbf{x}^j$, j-ème colonne centrée de $\mathbf{X}$

$\mathbf{x}^j \in F$ est isomorphe à $(\mathbb{R}^n, \mathcal{F}, \mathbf{D})$ avec $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$

Notations

- p variables statistiques réelles X^j observées sur
- n individus de poids w_i

Représentation vectorielle

- Individu i : $\mathbf{x}_i$, i-ème ligne de $\mathbf{X}$ mise en colonne

$$\mathbf{x}_i \in E \text{ isomorphe à } (\mathbb{R}^p, \mathcal{E}, \mathbf{M})$$

- Variable X^j : $\mathbf{x}^j$, j-ème colonne centrée de $\mathbf{X}$

$$\mathbf{x}^j \in F \text{ est isomorphe à } (\mathbb{R}^n, \mathcal{F}, \mathbf{D}) \text{ avec } \mathbf{D} = \text{diag}(w_1, \ldots, w_n)$$

Notations

- p variables statistiques réelles X^j observées sur
- n individus de poids w_i

Représentation vectorielle

- Individu i : $\mathbf{x}_i$, i-ème ligne de $\mathbf{X}$ mise en colonne

$$\mathbf{x}_i \in E \text{ isomorphe à } (\mathbb{R}^p, \mathcal{E}, \mathbf{M})$$

- Variable X^j : $\mathbf{x}^j$, j-ème colonne centrée de $\mathbf{X}$

$\mathbf{x}^j \in F$ est isomorphe à $(\mathbb{R}^n, \mathcal{F}, \mathbf{D})$ avec $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\dfrac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Interprétation statistique de la métrique des poids

- Moyenne empirique de X^j : $\overline{x^j} = \left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Barycentre des individus : $\overline{\mathbf{x}} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$
- Matrice centrée des données : $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$
- Ecart-type de X^j : $\sigma_j = (\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$
- Covariance : $\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k = \left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$
- Matrice : $\mathbf{S} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$
- Corrélation : $\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}\|\mathbf{x}^k\|_{\mathbf{D}}} = \cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$

Résumé

Avec des variables centrées vecteurs de $(F, \mathbf{D})$:

- la longueur d'un vecteur est un écart-type
- le cosinus d'un angle est une corrélation

Objectifs

- Représentation graphique "optimale" des individus
- Représentation graphique des variables
- Réduction de dimension (compression)
- Une approche parmis d'autres...

Notations

- $\mathbf{X}$ matrice des données issues de l'observation de p variables quantitatives X^j sur n individus de poids w_i,
- E est l'espace des individus, base canonique, métrique de matrice $\mathbf{M}$,
- F est l'espace des variables, base canonique, métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

Objectifs

- Représentation graphique "optimale" des individus
- Représentation graphique des variables
- Réduction de dimension (compression)
- Une approche parmis d'autres...

Notations

- $\mathbf{X}$ matrice des données issues de l'observation de p variables quantitatives X^j sur n individus de poids w_i,
- E est l'espace des individus, base canonique, métrique de matrice $\mathbf{M}$,
- F est l'espace des variables, base canonique, métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

Objectifs

- Représentation graphique "optimale" des individus
- Représentation graphique des variables
- Réduction de dimension (compression)
- Une approche parmis d'autres...

Notations

- $\mathbf{X}$ matrice des données issues de l'observation de p variables quantitatives X^j sur n individus de poids w_i,
- E est l'espace des individus, base canonique, métrique de matrice $\mathbf{M}$,
- F est l'espace des variables, base canonique, métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

Objectifs

- Représentation graphique "optimale" des individus
- Représentation graphique des variables
- Réduction de dimension (compression)
- Une approche parmis d'autres...

Notations

- $\mathbf{X}$ matrice des données issues de l'observation de p variables quantitatives X^j sur n individus de poids w_i,
- E est l'espace des individus, base canonique, métrique de matrice $\mathbf{M}$,
- F est l'espace des variables, base canonique, métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

Objectifs

- Représentation graphique "optimale" des individus
- Représentation graphique des variables
- Réduction de dimension (compression)
- Une approche parmis d'autres...

Notations

- $\mathbf{X}$ matrice des données issues de l'observation de p variables quantitatives X^j sur n individus de poids w_i,
- E est l'espace des individus, base canonique, métrique de matrice $\mathbf{M}$,
- F est l'espace des variables, base canonique, métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

Objectifs

- Représentation graphique "optimale" des individus
- Représentation graphique des variables
- Réduction de dimension (compression)
- Une approche parmis d'autres...

Notations

- $\mathbf{X}$ matrice des données issues de l'observation de p variables quantitatives X^j sur n individus de poids w_i,
- E est l'espace des individus, base canonique, métrique de matrice $\mathbf{M}$,
- F est l'espace des variables, base canonique, métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

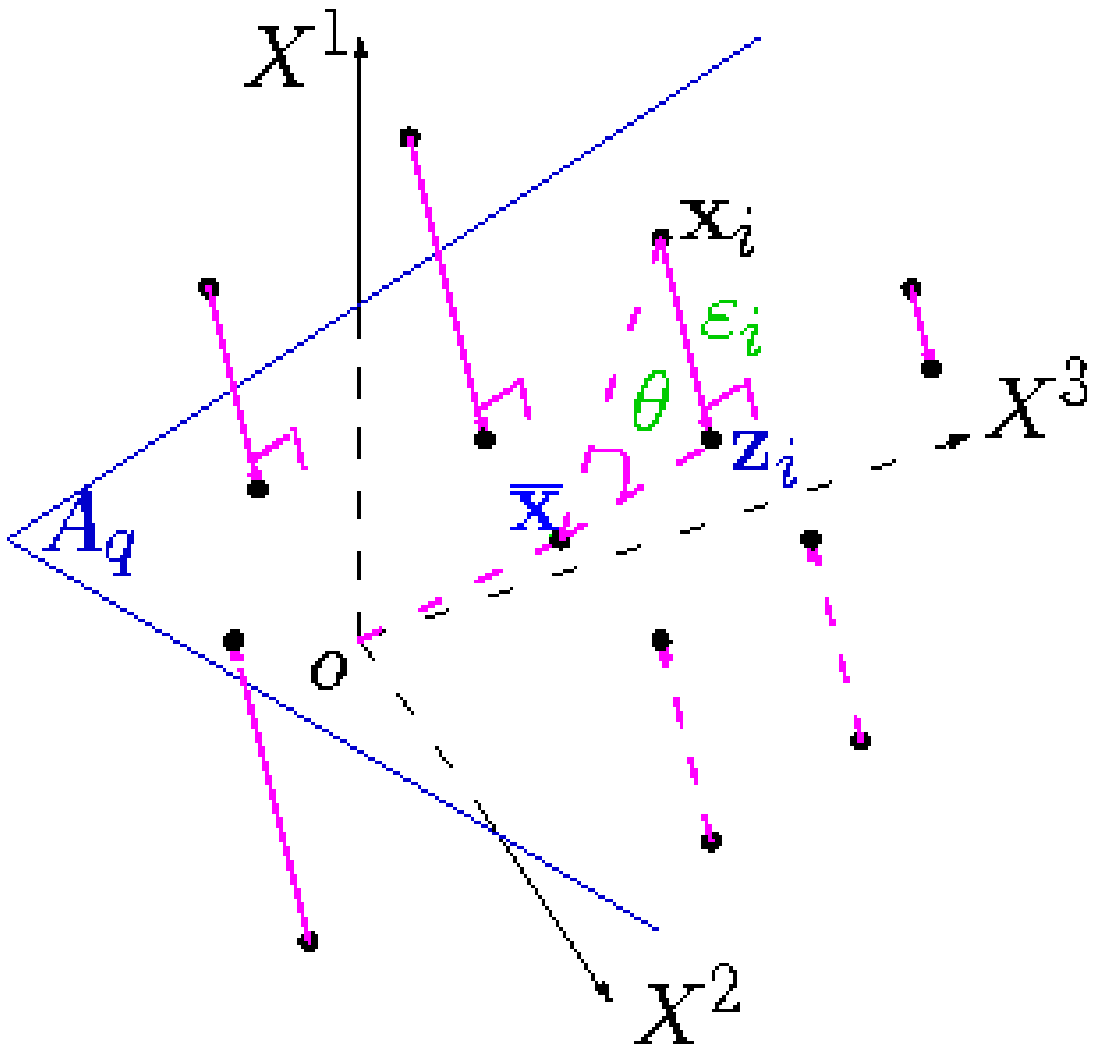
Objectifs

- Représentation graphique "optimale" des individus
- Représentation graphique des variables
- Réduction de dimension (compression)
- Une approche parmis d'autres...

Notations

- $\mathbf{X}$ matrice des données issues de l'observation de p variables quantitatives X^j sur n individus de poids w_i,
- E est l'espace des individus, base canonique, métrique de matrice $\mathbf{M}$,
- F est l'espace des variables, base canonique, métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

ACP dans l'espace des individus avec $p = 3$



Modèle de l'ACP

$$\text{Observation} = \text{Modèle} + \text{Bruit}$$

$$\mathbf{x}_i = \mathbf{z}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n$$

$$\text{avec} \begin{cases} \{\mathbf{x}_i\ ; i = 1, \ldots, n\},\ n \text{ vect. aléat. indép. de } E, \\ E(\boldsymbol{\varepsilon}_i) = 0,\ \text{var}(\boldsymbol{\varepsilon}_i) = \sigma^2 \boldsymbol{\Gamma}, \\ \sigma > 0 \text{ inconnu, } \boldsymbol{\Gamma} \text{ régulière et connue,} \end{cases}$$

$\exists A_q$, sous-espace affine de dimension q de E tel que

$$\forall i, \mathbf{z}_i \in A_q \quad (q < p).$$

Si $\bar{\mathbf{z}} = \sum_{i=1}^{n} w_i \mathbf{z}_i$ alors $\bar{\mathbf{z}}$ appartient à A_q.
Soit E_q tel que : $A_q = \bar{\mathbf{z}} + E_q$.

Moindres carrés

$$\min_{E_q, \mathbf{z}_i} \left\{ \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 \; ; \; \dim(E_q) = q, \mathbf{z}_i - \bar{\mathbf{z}} \in E_q \right\}.$$

Théorème

L'estimation des paramètres : E_q et $\mathbf{z}_i (i = 1, \ldots, n)$, σ est fournie par la décomposition en valeurs singulières de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$:

$$\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda_k^{1/2} \mathbf{u}^k \mathbf{v}^{k\prime} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}_q'.$$

Moindres carrés

$$\min_{E_q, \mathbf{z}_i} \left\{ \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 \; ; \; \dim(E_q) = q, \mathbf{z}_i - \bar{\mathbf{z}} \in E_q \right\}.$$

Théorème

L'estimation des paramètres : E_q et $\mathbf{z}_i (i = 1, \ldots, n)$, σ est fournie par la décomposition en valeurs singulières de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$:

$$\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda_k^{1/2} \mathbf{u}^k \mathbf{v}^{k\prime} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}_q'.$$

Preuve

- Soit $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n \overline{\mathbf{x}}'$ la matrice centrée et $\mathbf{Z}$ la matrice $(n \times p)$ dont les lignes sont les vecteurs $(\mathbf{z}_i - \overline{\mathbf{z}})'$

$$\sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 = \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{z}} - \mathbf{z}_i \right\|_{\mathbf{M}}^2 + \left\| \overline{\mathbf{x}} - \overline{\mathbf{z}} \right\|_{\mathbf{M}}^2 \ ;$$

- Ainsi, $\overline{\mathbf{z}}$ est estimé par $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$ et il reste à résoudre l'approximation matricielle :

$$\min_{\mathbf{Z}} \left\{ \left\| \overline{\mathbf{X}} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}} \ ; \mathbf{Z} \in \mathcal{M}_{n,p}, \ \text{rang}(\mathbf{Z}) = q \right\}$$

Résultats

- Les $\mathbf{u}^k$ sont les vecteurs propres $\mathbf{D}$-orthonormés de la matrice $\overline{\mathbf{X}}\mathbf{M}\overline{\mathbf{X}}'\mathbf{D}$ associés aux valeurs propres λ_k rangées par ordre décroissant
- Les $\mathbf{v}_k$, appelés vecteurs principaux, sont les vecteurs propres $\mathbf{M}$-orthonormés de la matrice $\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathbf{S}\mathbf{M}$ associés aux mêmes valeurs propres

Preuve

- Soit $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n \overline{\mathbf{x}}'$ la matrice centrée et $\mathbf{Z}$ la matrice $(n \times p)$ dont les lignes sont les vecteurs $(\mathbf{z}_i - \overline{\mathbf{z}})'$

$$\sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 = \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{z}} - \mathbf{z}_i \right\|_{\mathbf{M}}^2 + \left\| \overline{\mathbf{x}} - \overline{\mathbf{z}} \right\|_{\mathbf{M}}^2 \ ;$$

- Ainsi, $\overline{\mathbf{z}}$ est estimé par $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$ et il reste à résoudre l'approximation matricielle :

$$\min_{\mathbf{Z}} \left\{ \left\| \overline{\mathbf{X}} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}} \ ; \mathbf{Z} \in \mathcal{M}_{n,p}, \ \text{rang}(\mathbf{Z}) = q \right\}$$

Résultats

- Les $\mathbf{u}^k$ sont les vecteurs propres $\mathbf{D}$-orthonormés de la matrice $\overline{\mathbf{X}}\mathbf{M}\overline{\mathbf{X}}'\mathbf{D}$ associés aux valeurs propres λ_k rangées par ordre décroissant
- Les $\mathbf{v}_k$, appelés vecteurs principaux, sont les vecteurs propres $\mathbf{M}$-orthonormés de la matrice $\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathbf{S}\mathbf{M}$ associés aux mêmes valeurs propres

Preuve

- Soit $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n \overline{\mathbf{x}}'$ la matrice centrée et $\mathbf{Z}$ la matrice $(n \times p)$ dont les lignes sont les vecteurs $(\mathbf{z}_i - \overline{\mathbf{z}})'$

$$\sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 = \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{z}} - \mathbf{z}_i \right\|_{\mathbf{M}}^2 + \left\| \overline{\mathbf{x}} - \overline{\mathbf{z}} \right\|_{\mathbf{M}}^2 \ ;$$

- Ainsi, $\overline{\mathbf{z}}$ est estimé par $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$ et il reste à résoudre l'approximation matricielle :

$$\min_{\mathbf{Z}} \left\{ \left\| \overline{\mathbf{X}} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}} \ ; \mathbf{Z} \in \mathcal{M}_{n,p}, \ \text{rang}(\mathbf{Z}) = q \right\}$$

Résultats

- Les $\mathbf{u}^k$ sont les vecteurs propres $\mathbf{D}$-orthonormés de la matrice $\overline{\mathbf{X}}\mathbf{M}\overline{\mathbf{X}}'\mathbf{D}$ associés aux valeurs propres λ_k rangées par ordre décroissant
- Les $\mathbf{v}_k$, appelés vecteurs principaux, sont les vecteurs propres $\mathbf{M}$-orthonormés de la matrice $\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathbf{S}\mathbf{M}$ associés aux mêmes valeurs propres

Preuve

- Soit $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n \overline{\mathbf{x}}'$ la matrice centrée et $\mathbf{Z}$ la matrice $(n \times p)$ dont les lignes sont les vecteurs $(\mathbf{z}_i - \overline{\mathbf{z}})'$

$$\sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 = \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{z}} - \mathbf{z}_i \right\|_{\mathbf{M}}^2 + \left\| \overline{\mathbf{x}} - \overline{\mathbf{z}} \right\|_{\mathbf{M}}^2 \ ;$$

- Ainsi, $\overline{\mathbf{z}}$ est estimé par $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$ et il reste à résoudre l'approximation matricielle :

$$\min_{\mathbf{Z}} \left\{ \left\| \overline{\mathbf{X}} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}} \ ; \mathbf{Z} \in \mathcal{M}_{n,p}, \ \text{rang}(\mathbf{Z}) = q \right\}$$

Résultats

- Les $\mathbf{u}^k$ sont les vecteurs propres $\mathbf{D}$-orthonormés de la matrice $\overline{\mathbf{X}}\mathbf{M}\overline{\mathbf{X}}'\mathbf{D}$ associés aux valeurs propres λ_k rangées par ordre décroissant
- Les $\mathbf{v}_k$, appelés vecteurs principaux, sont les vecteurs propres $\mathbf{M}$-orthonormés de la matrice $\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathbf{SM}$ associés aux mêmes valeurs propres

Estimation des paramètres

- $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$
- $\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda^{1/2} \mathbf{u}^k \mathbf{v}^{k\prime} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}_q' = \widehat{\mathbf{X}\mathbf{P}_q}'$ où $\widehat{\mathbf{P}_q} = \mathbf{V}_q \mathbf{V}_q' \mathbf{M}$ est la matrice de projection $\mathbf{M}$-orthogonale sur $\widehat{E_q}$
- $\widehat{E_q} = \text{ Vect}\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}$
- $\widehat{E_2}$ est appelé plan principal
- $\widehat{\mathbf{z}_i} = \widehat{\mathbf{P}_q \mathbf{x}_i} + \overline{\mathbf{x}}$

Estimation des paramètres

- $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$
- $\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda^{1/2} \mathbf{u}^k \mathbf{v}^{k'} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}'_q = \widehat{\mathbf{X}\mathbf{P}_q}'$ où $\widehat{\mathbf{P}_q} = \mathbf{V}_q \mathbf{V}'_q \mathbf{M}$ est la matrice de projection $\mathbf{M}$-orthogonale sur $\widehat{E_q}$
- $\widehat{E_q} = \text{Vect}\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}$
- $\widehat{E_2}$ est appelé plan principal
- $\widehat{\mathbf{z}_i} = \widehat{\mathbf{P}_q} \mathbf{x}_i + \overline{\mathbf{x}}$

Estimation des paramètres

- $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$
- $\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda^{1/2} \mathbf{u}^k \mathbf{v}^{k'} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}'_q = \widehat{\mathbf{X}\mathbf{P}_q}'$ où $\widehat{\mathbf{P}_q} = \mathbf{V}_q \mathbf{V}'_q \mathbf{M}$ est la matrice de projection $\mathbf{M}$-orthogonale sur $\widehat{E_q}$
- $\widehat{E_q} =$ Vect$\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}$
- $\widehat{E_2}$ est appelé plan principal
- $\widehat{\mathbf{z}_i} = \widehat{\mathbf{P}_q \mathbf{x}_i} + \overline{\mathbf{x}}$

Estimation des paramètres

- $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$
- $\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda^{1/2} \mathbf{u}^k \mathbf{v}^{k'} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}'_q = \widehat{\mathbf{X}\mathbf{P}_q}'$ où $\widehat{\mathbf{P}_q} = \mathbf{V}_q \mathbf{V}'_q \mathbf{M}$ est la matrice de projection $\mathbf{M}$-orthogonale sur $\widehat{E_q}$
- $\widehat{E_q} =$ Vect$\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}$
- $\widehat{E_2}$ est appelé plan principal
- $\widehat{\mathbf{z}_i} = \widehat{\mathbf{P}_q \mathbf{x}_i} + \overline{\mathbf{x}}$

Estimation des paramètres

- $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$
- $\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda^{1/2} \mathbf{u}^k \mathbf{v}^{k'} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}'_q = \widehat{\mathbf{X}\mathbf{P}_q}'$ où $\widehat{\mathbf{P}_q} = \mathbf{V}_q \mathbf{V}'_q \mathbf{M}$ est la matrice de projection $\mathbf{M}$-orthogonale sur $\widehat{E_q}$
- $\widehat{E_q} = \text{ Vect}\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}$
- $\widehat{E_2}$ est appelé plan principal
- $\widehat{\mathbf{z}_i} = \widehat{\mathbf{P}_q} \mathbf{x}_i + \overline{\mathbf{x}}$

Remarques

1. Solutions emboîtées pour $q = 1, \ldots, p$
2. Les espaces principaux sont uniques sauf dans le cas de valeurs propres multiples
3. Si les variables ne sont pas homogènes elles sont réduites :

$$\widetilde{\overline{\mathbf{X}}} = \overline{\mathbf{X}}\boldsymbol{\Sigma}^{-1/2} \text{ où } \boldsymbol{\Sigma} = \text{ diag } (\sigma_1^2, \ldots, \sigma_p^2);$$

4. $\widetilde{\mathbf{S}} = \mathbf{R} = \boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2}$ matrice des corrélations

Remarques

1. Solutions emboîtées pour $q = 1, \ldots, p$
2. Les espaces principaux sont uniques sauf dans le cas de valeurs propres multiples
3. Si les variables ne sont pas homogènes elles sont réduites :

$$\widetilde{\overline{\mathbf{X}}} = \overline{\mathbf{X}}\boldsymbol{\Sigma}^{-1/2} \text{ où } \boldsymbol{\Sigma} = \text{ diag } (\sigma_1^2, \ldots, \sigma_p^2);$$

4. $\widetilde{\mathbf{S}} = \mathbf{R} = \boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2}$ matrice des corrélations

Remarques

1. Solutions emboîtées pour $q = 1, \ldots, p$
2. Les espaces principaux sont uniques sauf dans le cas de valeurs propres multiples
3. Si les variables ne sont pas homogènes elles sont réduites :

$$\widetilde{\overline{\mathbf{X}}} = \overline{\mathbf{X}}\boldsymbol{\Sigma}^{-1/2} \text{ où } \boldsymbol{\Sigma} = \text{ diag } (\sigma_1^2, \ldots, \sigma_p^2);$$

4. $\widetilde{\mathbf{S}} = \mathbf{R} = \boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2}$ matrice des corrélations

Remarques

1. Solutions emboîtées pour $q = 1, \ldots, p$
2. Les espaces principaux sont uniques sauf dans le cas de valeurs propres multiples
3. Si les variables ne sont pas homogènes elles sont réduites :

$$\widetilde{\overline{\mathbf{X}}} = \overline{\mathbf{X}}\boldsymbol{\Sigma}^{-1/2} \text{ où } \boldsymbol{\Sigma} = \text{ diag } (\sigma_1^2, \ldots, \sigma_p^2);$$

4. $\widetilde{\mathbf{S}} = \mathbf{R} = \boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2}$ matrice des corrélations

Combinaison de variables

- La combinaison linéaire des variable centrées $X^1, \ldots, X^p$

$$\mathbf{c} = \sum_{j=1}^{p} f_j \mathbf{x}^j = \overline{\mathbf{X}}\mathbf{f},$$

- définit une variable centrée C avec $C(\omega_i) = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{f}$

Théorème

Soient p variables quantitatives centrées $X^1, \ldots, X^p$ observées sur n individus de poids w_i ; l'ACP de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ est la recherche des q combinaisons linéaires normées des X^j, non corrélées et dont la somme des variances soit maximale.

Combinaison de variables

- La combinaison linéaire des variable centrées $X^1, \ldots, X^p$

$$\mathbf{c} = \sum_{j=1}^{p} f_j \mathbf{x}^j = \overline{\mathbf{X}}\mathbf{f},$$

- définit une variable centrée C avec $C(\omega_i) = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{f}$

Théorème

Soient p variables quantitatives centrées $X^1, \ldots, X^p$ observées sur n individus de poids w_i ; l'ACP de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ est la recherche des q combinaisons linéaires normées des X^j, non corrélées et dont la somme des variances soit maximale.

Combinaison de variables

- La combinaison linéaire des variable centrées $X^1, \ldots, X^p$

$$\mathbf{c} = \sum_{j=1}^{p} f_j \mathbf{x}^j = \overline{\mathbf{X}}\mathbf{f},$$

- définit une variable centrée C avec $C(\omega_i) = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{f}$

Théorème

Soient p variables quantitatives centrées $X^1, \ldots, X^p$ observées sur n individus de poids w_i ; l'ACP de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ est la recherche des q combinaisons linéaires normées des X^j, non corrélées et dont la somme des variances soit maximale.

Résultats

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les facteurs principaux
- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les composantes principales
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les variables principales

$$\begin{aligned} \text{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)'\mathbf{D}\overline{\mathbf{X}}\mathbf{f}^\ell = \mathbf{f}^{k\prime}\mathbf{S}\mathbf{f}^\ell \\ &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \delta_k^\ell \end{aligned}$$

- $\mathbf{f}^k$: vecteurs propres $\mathbf{M}^{-1}$-orthonormés de $\mathbf{MS}$
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$: matrice des composantes principales
- Les vecteurs $\mathbf{D}$-orthonormés u^k définissent les axes factoriels.

Résultats

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les facteurs principaux
- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les composantes principales
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les variables principales

$$\begin{aligned} \text{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)'\mathbf{D}\overline{\mathbf{X}}\mathbf{f}^\ell = \mathbf{f}^{k\prime}\mathbf{S}\mathbf{f}^\ell \\ &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \delta_k^\ell \end{aligned}$$

- $\mathbf{f}^k$: vecteurs propres $\mathbf{M}^{-1}$-orthonormés de $\mathbf{MS}$
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$: matrice des composantes principales
- Les vecteurs $\mathbf{D}$-orthonormés u^k définissent les axes factoriels.

Résultats

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les facteurs principaux
- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les composantes principales
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les variables principales

$$\begin{aligned} \text{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)'\mathbf{D}\overline{\mathbf{X}}\mathbf{f}^\ell = \mathbf{f}^{k\prime}\mathbf{S}\mathbf{f}^\ell \\ &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \delta_k^\ell \end{aligned}$$

- $\mathbf{f}^k$: vecteurs propres $\mathbf{M}^{-1}$-orthonormés de $\mathbf{MS}$
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\Lambda^{1/2}$: matrice des composantes principales
- Les vecteurs $\mathbf{D}$-orthonormés u^k définissent les axes factoriels.

Résultats

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les facteurs principaux
- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les composantes principales
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les variables principales

$$\begin{aligned} \text{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)'\mathbf{D}\overline{\mathbf{X}}\mathbf{f}^\ell = \mathbf{f}^{k\prime}\mathbf{S}\mathbf{f}^\ell \\ &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \delta_k^\ell \end{aligned}$$

- $\mathbf{f}^k$: vecteurs propres $\mathbf{M}^{-1}$-orthonormés de $\mathbf{MS}$
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\Lambda^{1/2}$: matrice des composantes principales
- Les vecteurs $\mathbf{D}$-orthonormés u^k définissent les axes factoriels.

Résultats

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les facteurs principaux
- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les composantes principales
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les variables principales

$$\begin{aligned} \text{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)'\mathbf{D}\overline{\mathbf{X}}\mathbf{f}^\ell = \mathbf{f}^{k\prime}\mathbf{S}\mathbf{f}^\ell \\ &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \delta_k^\ell \end{aligned}$$

- $\mathbf{f}^k$: vecteurs propres $\mathbf{M}^{-1}$-orthonormés de $\mathbf{MS}$
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$: matrice des composantes principales
- Les vecteurs $\mathbf{D}$-orthonormés u^k définissent les axes factoriels.

Résultats

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les facteurs principaux
- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les composantes principales
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les variables principales

$$\begin{aligned} \text{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)'\mathbf{D}\overline{\mathbf{X}}\mathbf{f}^\ell = \mathbf{f}^{k\prime}\mathbf{S}\mathbf{f}^\ell \\ &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \delta_k^\ell \end{aligned}$$

- $\mathbf{f}^k$: vecteurs propres $\mathbf{M}^{-1}$-orthonormés de $\mathbf{MS}$
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$: matrice des composantes principales
- Les vecteurs $\mathbf{D}$-orthonormés u^k définissent les axes factoriels.

Projection des individus

- Représentation optimale des distances individuelles au sens de $\mathbf{M}$
- L'individu $\mathbf{x}_i$ est approché par la projection $\mathbf{M}$-orthogonale $\widehat{\mathbf{z}_i}^q$ sur le sous-espace $\widehat{E_q} = \text{Vect}(\mathbf{v}^1, \ldots, \mathbf{v}^q)$
- $k$ème coordonnée :
$\left\langle \mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{v}^k \right\rangle_{\mathbf{M}} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{M}\mathbf{v}^k = \mathbf{e}_i'\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = c_i^k$

Proposition

Les coordonnées de la projection $\mathbf{M}$-orthogonale de $\mathbf{x}_i - \overline{\mathbf{x}}$ sur $\widehat{E_q}$ sont les q premiers élément de la i-ème ligne de la matrice $\mathbf{C}$ des composantes principales.

Projection des individus

- Représentation optimale des distances individuelles au sens de $\mathbf{M}$
- L'individu $\mathbf{x}_i$ est approché par la projection $\mathbf{M}$-orthogonale $\widehat{\mathbf{z}_i^q}$ sur le sous-espace $\widehat{E_q} = \text{Vect}(\mathbf{v}^1, \ldots, \mathbf{v}^q)$
- $k$ème coordonnée :
$\left\langle \mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{v}^k \right\rangle_{\mathbf{M}} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{M}\mathbf{v}^k = \mathbf{e}_i'\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = c_i^k$

Proposition

Les coordonnées de la projection $\mathbf{M}$-orthogonale de $\mathbf{x}_i - \overline{\mathbf{x}}$ sur $\widehat{E_q}$ sont les q premiers élément de la i-ème ligne de la matrice $\mathbf{C}$ des composantes principales.

Projection des individus

- Représentation optimale des distances individuelles au sens de $\mathbf{M}$
- L'individu $\mathbf{x}_i$ est approché par la projection $\mathbf{M}$-orthogonale $\widehat{\mathbf{z}_i^q}$ sur le sous-espace $\widehat{E_q} = \text{Vect}(\mathbf{v}^1, \ldots, \mathbf{v}^q)$
- $k$ème coordonnée :
  $\left\langle \mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{v}^k \right\rangle_{\mathbf{M}} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{M}\mathbf{v}^k = \mathbf{e}_i'\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = c_i^k$

Proposition

Les coordonnées de la projection $\mathbf{M}$-orthogonale de $\mathbf{x}_i - \overline{\mathbf{x}}$ sur $\widehat{E_q}$ sont les q premiers élément de la i-ème ligne de la matrice $\mathbf{C}$ des composantes principales.

Projection des individus

- Représentation optimale des distances individuelles au sens de $\mathbf{M}$
- L'individu $\mathbf{x}_i$ est approché par la projection $\mathbf{M}$-orthogonale $\widehat{\mathbf{z}_i}^q$ sur le sous-espace $\widehat{E_q} = \text{Vect}(\mathbf{v}^1, \ldots, \mathbf{v}^q)$
- $k$ème coordonnée :
  $\left\langle \mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{v}^k \right\rangle_{\mathbf{M}} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{M}\mathbf{v}^k = \mathbf{e}_i'\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = c_i^k$

Proposition

Les coordonnées de la projection $\mathbf{M}$-orthogonale de $\mathbf{x}_i - \overline{\mathbf{x}}$ sur $\widehat{E_q}$ sont les q premiers élément de la i-ème ligne de la matrice $\mathbf{C}$ des composantes principales.

Mesures de "qualité"

- Dispersion des points $\overline{\mathbf{x}}$ ou *inertie* :

$$I_g = \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2 = \left\| \overline{\mathbf{X}} \right\|_{\mathbf{M,D}}^2 = \mathrm{tr}\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathrm{tr}\mathbf{SM}.$$

- Qualité globale mesurée par la part de dispersion ou d'inertie expliquée :

$$r_q = \frac{\mathrm{tr}\mathbf{SM}\widehat{\mathbf{P}}_q}{\mathrm{tr}\mathbf{SM}} = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

- Qualité de représentation de x_i

$$\left[\cos\theta(\mathbf{x}_i - \overline{\mathbf{x}}, \widehat{\mathbf{z}}_i^q)\right]^2 = \frac{\left\| \widehat{\mathbf{P}}_q(\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_{\mathbf{M}}^2}{\left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2} = \frac{\sum_{k=1}^{q} (c_i^k)^2}{\sum_{k=1}^{p} (c_i^k)^2}$$

Mesures de "qualité"

- Dispersion des points $\overline{\mathbf{x}}$ ou *inertie* :

$$I_g = \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2 = \left\| \overline{\mathbf{X}} \right\|_{\mathbf{M},\mathbf{D}}^2 = \mathrm{tr}\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathrm{tr}\mathbf{SM}.$$

- Qualité globale mesurée par la part de dispersion ou d'inertie expliquée :

$$r_q = \frac{\mathrm{tr}\mathbf{SM}\widehat{\mathbf{P}_q}}{\mathrm{tr}\mathbf{SM}} = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

- Qualité de représentation de x_i

$$\left[\cos\theta(\mathbf{x}_i - \overline{\mathbf{x}}, \widehat{\mathbf{z}_i^q})\right]^2 = \frac{\left\| \widehat{\mathbf{P}_q}(\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_{\mathbf{M}}^2}{\left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2} = \frac{\sum_{k=1}^{q} (c_i^k)^2}{\sum_{k=1}^{p} (c_i^k)^2}$$

Mesures de "qualité"

- Dispersion des points $\overline{\mathbf{x}}$ ou *inertie* :

$$I_g = \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2 = \left\| \overline{\mathbf{X}} \right\|_{\mathbf{M},\mathbf{D}}^2 = \mathrm{tr}\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathrm{tr}\mathbf{SM}.$$

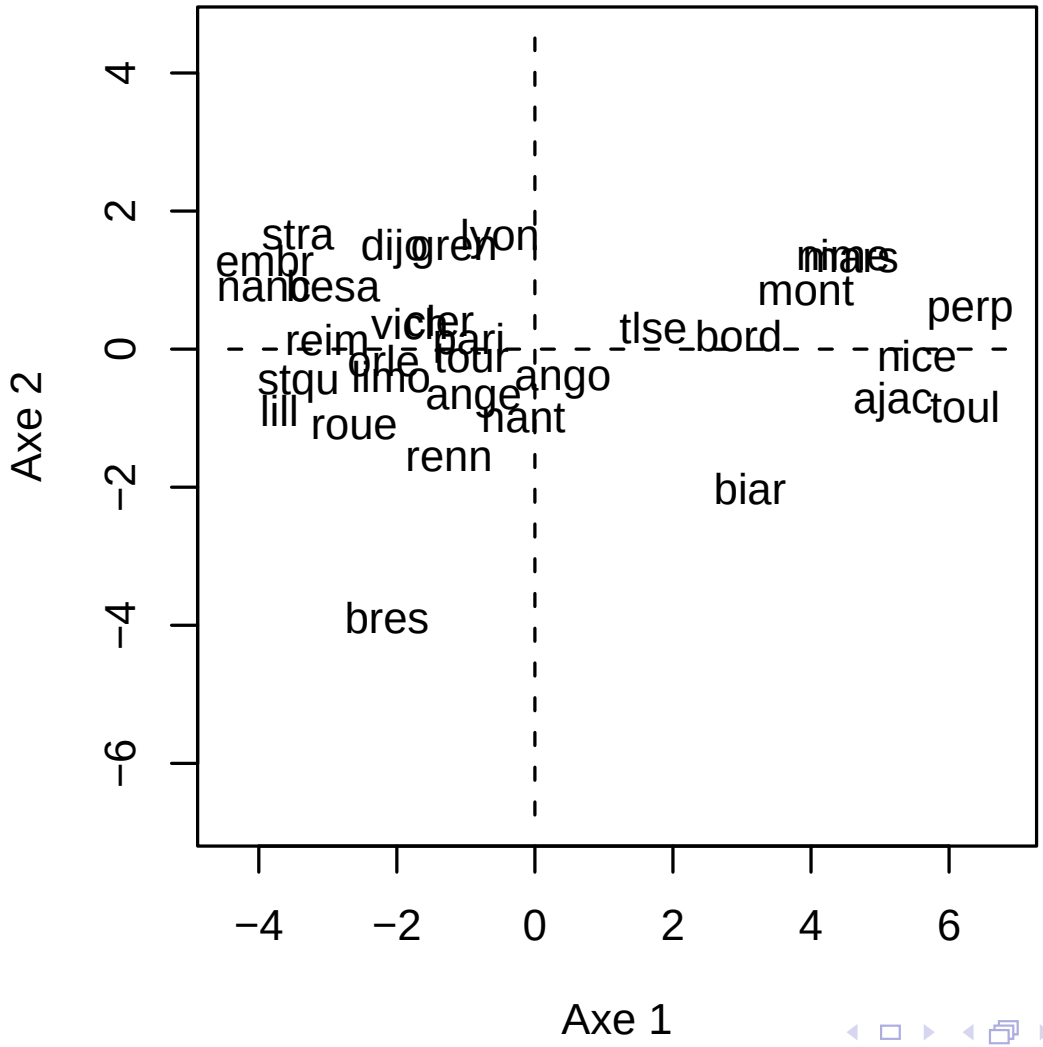
- Qualité globale mesurée par la part de dispersion ou d'inertie expliquée :

$$r_q = \frac{\mathrm{tr}\mathbf{SM}\widehat{\mathbf{P}_q}}{\mathrm{tr}\mathbf{SM}} = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

- Qualité de représentation de x_i

$$\left[\cos\theta(\mathbf{x}_i - \overline{\mathbf{x}}, \widehat{\mathbf{z}_i^q})\right]^2 = \frac{\left\| \widehat{\mathbf{P}_q}(\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_{\mathbf{M}}^2}{\left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2} = \frac{\sum_{k=1}^{q} (c_i^k)^2}{\sum_{k=1}^{p} (c_i^k)^2}$$

Températures : graphe des individus



Diagnostics

- Contributions des individus à l'inertie de leur nuage

$$\gamma_i = \frac{w_i \left\| \mathbf{x}_i - \bar{\mathbf{x}} \right\|^2_{\mathbf{M}}}{\mathbf{trSM}} = \frac{w_i \sum_{k=1}^{p} (c_i^k)^2}{\sum_{k=1}^{p} \lambda_k}$$

- ainsi qu'à la variance d'une variable principale

$$\gamma_i^k = \frac{w_i (c_i^k)^2}{\lambda_k}$$

- Individu supplémentaire s de coordonnées : $\mathbf{V}_q' \mathbf{M} (\mathbf{s} - \bar{\mathbf{x}})$

$$\begin{aligned} \left\langle \mathbf{v}^k, \mathbf{V}_q \mathbf{V}_q' \mathbf{M} (\mathbf{s} - \bar{\mathbf{x}}) \right\rangle_{\mathbf{M}} &= \mathbf{v}^{k\prime} \mathbf{M} \mathbf{V}_q \mathbf{V}_q' \mathbf{M} (\mathbf{s} - \bar{\mathbf{x}}) \\ &= \mathbf{e}^{k\prime} \mathbf{V}_q' \mathbf{M} (\mathbf{s} - \bar{\mathbf{x}}) \end{aligned}$$

Diagnostics

- Contributions des individus à l'inertie de leur nuage

$$\gamma_i = \frac{w_i \left\| \mathbf{x}_i - \bar{\mathbf{x}} \right\|^2_{\mathbf{M}}}{\mathrm{tr}\mathbf{SM}} = \frac{w_i \sum_{k=1}^{p} (c_i^k)^2}{\sum_{k=1}^{p} \lambda_k}$$

- ainsi qu'à la variance d'une variable principale

$$\gamma_i^k = \frac{w_i (c_i^k)^2}{\lambda_k}$$

- Individu supplémentaire s de coordonnées : $\mathbf{V}_q'\mathbf{M}(\mathbf{s} - \bar{\mathbf{x}})$

$$\begin{aligned} \left\langle \mathbf{v}^k, \mathbf{V}_q\mathbf{V}_q'\mathbf{M}(\mathbf{s} - \bar{\mathbf{x}}) \right\rangle_{\mathbf{M}} &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{V}_q\mathbf{V}_q'\mathbf{M}(\mathbf{s} - \bar{\mathbf{x}}) \\ &= \mathbf{e}^{k\prime}\mathbf{V}_q'\mathbf{M}(\mathbf{s} - \bar{\mathbf{x}}) \end{aligned}$$

Diagnostics

- Contributions des individus à l'inertie de leur nuage

$$\gamma_i = \frac{w_i \left\|\mathbf{x}_i - \bar{\mathbf{x}}\right\|^2_{\mathbf{M}}}{\mathrm{tr}\mathbf{SM}} = \frac{w_i \sum_{k=1}^{p} (c_i^k)^2}{\sum_{k=1}^{p} \lambda_k}$$

- ainsi qu'à la variance d'une variable principale

$$\gamma_i^k = \frac{w_i (c_i^k)^2}{\lambda_k}$$

- Individu supplémentaire s de coordonnées : $\mathbf{V}_q'\mathbf{M}(\mathbf{s} - \bar{\mathbf{x}})$

$$\begin{aligned} \left\langle \mathbf{v}^k, \mathbf{V}_q\mathbf{V}_q'\mathbf{M}(\mathbf{s} - \bar{\mathbf{x}}) \right\rangle_{\mathbf{M}} &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{V}_q\mathbf{V}_q'\mathbf{M}(\mathbf{s} - \bar{\mathbf{x}}) \\ &= \mathbf{e}^{k\prime}\mathbf{V}_q'\mathbf{M}(\mathbf{s} - \bar{\mathbf{x}}) \end{aligned}$$

Projection des variables

- La variable X^j est représentée par la projection $\mathbf{D}$-orthogonale $\widehat{\mathbf{Q}_q \mathbf{x}^j}$ sur le sous-espace F_q engendré par les q premiers axes factoriels. La coordonnée de $\mathbf{x}^j$ sur $\mathbf{u}^k$ est :

$$\begin{aligned} \left\langle \mathbf{x}^j, \mathbf{u}^k \right\rangle_{\mathbf{D}} &= \mathbf{x}^{j\prime} \mathbf{D} \mathbf{u}^k = \lambda_k^{-1/2} \mathbf{x}^{j\prime} \mathbf{D} \overline{\mathbf{X}} \mathbf{M} \mathbf{v}^k \\ &= \lambda_k^{-1/2} \mathbf{e}^{j\prime} \overline{\mathbf{X}}' \mathbf{D} \overline{\mathbf{X}} \mathbf{M} \mathbf{v}^k = \sqrt{\lambda_k} v_j^k \end{aligned}$$

- Les coordonnées de la projection $\mathbf{D}$-orthogonale de $\mathbf{x}^j$ sur F_q sont les q premiers éléments de la j-ème ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$

Projection des variables

- La variable X^j est représentée par la projection $\mathbf{D}$-orthogonale $\widehat{\mathbf{Q}_q \mathbf{x}^j}$ sur le sous-espace F_q engendré par les q premiers axes factoriels. La coordonnée de $\mathbf{x}^j$ sur $\mathbf{u}^k$ est :

$$\begin{aligned} \left\langle \mathbf{x}^j, \mathbf{u}^k \right\rangle_{\mathbf{D}} &= \mathbf{x}^{j\prime} \mathbf{D} \mathbf{u}^k = \lambda_k^{-1/2} \mathbf{x}^{j\prime} \mathbf{D} \overline{\mathbf{X}} \mathbf{M} \mathbf{v}^k \\ &= \lambda_k^{-1/2} \mathbf{e}^{j\prime} \overline{\mathbf{X}}' \mathbf{D} \overline{\mathbf{X}} \mathbf{M} \mathbf{v}^k = \sqrt{\lambda_k} v_j^k \end{aligned}$$

- Les coordonnées de la projection $\mathbf{D}$-orthogonale de $\mathbf{x}^j$ sur F_q sont les q premiers éléments de la j-ème ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$

Mesure de "qualité"

- Qualité de la représentation de $\mathbf{x}^j$:

$$\left[\cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_q}\mathbf{x}^j)\right]^2 = \frac{\left\|\widehat{\mathbf{Q}_q}\mathbf{x}^j\right\|_{\mathbf{D}}^2}{\|\mathbf{x}^j\|_{\mathbf{D}}^2} = \frac{\sum_{k=1}^{q}\lambda_k (v_k^j)^2}{\sum_{k=1}^{p}\lambda_k (v_k^j)^2}.$$

- Corrélations variables$\times$facteurs : $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

$$\mathrm{cor}(X^j, C^k) = \cos\theta(\mathbf{x}^j, \mathbf{c}^k) = \cos\theta(\mathbf{x}^j, \mathbf{u}^k) = \frac{\langle \mathbf{x}^j, \mathbf{u}^k\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}} = \frac{\sqrt{\lambda_k}}{\sigma_j} v_j^k \ ;$$

- Cercle des corrélations : Les variables réduites $\widetilde{\mathbf{x}}^j = \sigma_j^{-1}\mathbf{x}^j$, $\left\|\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = 1$, sont sur une sphère unité et se projettent à l'intérieur d'un cercle :

$$\left\|\widehat{\mathbf{Q}_2}\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = \cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_2}x_j) \leq 1.$$

Mesure de "qualité"

- Qualité de la représentation de $\mathbf{x}^j$:

$$\left[\cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_q}\mathbf{x}^j)\right]^2 = \frac{\left\|\widehat{\mathbf{Q}_q}\mathbf{x}^j\right\|_{\mathbf{D}}^2}{\|\mathbf{x}^j\|_{\mathbf{D}}^2} = \frac{\sum_{k=1}^{q}\lambda_k (v_k^j)^2}{\sum_{k=1}^{p}\lambda_k (v_k^j)^2}.$$

- Corrélations variables$\times$facteurs : $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

$$\text{cor}(X^j, C^k) = \cos\theta(\mathbf{x}^j, \mathbf{c}^k) = \cos\theta(\mathbf{x}^j, \mathbf{u}^k) = \frac{\left\langle \mathbf{x}^j, \mathbf{u}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}} = \frac{\sqrt{\lambda_k}}{\sigma_j} v_j^k \ ;$$

- Cercle des corrélations : Les variables réduites $\widetilde{\mathbf{x}}^j = \sigma_j^{-1}\mathbf{x}^j$, $\left\|\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = 1$, sont sur une sphère unité et se projettent à l'intérieur d'un cercle :

$$\left\|\widehat{\mathbf{Q}_2}\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = \cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_2}x_j) \leq 1.$$

Mesure de "qualité"

- Qualité de la représentation de $\mathbf{x}^j$:

$$\left[\cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_q}\mathbf{x}^j)\right]^2 = \frac{\left\|\widehat{\mathbf{Q}_q}\mathbf{x}^j\right\|_{\mathbf{D}}^2}{\|\mathbf{x}^j\|_{\mathbf{D}}^2} = \frac{\sum_{k=1}^{q}\lambda_k (v_k^j)^2}{\sum_{k=1}^{p}\lambda_k (v_k^j)^2}.$$

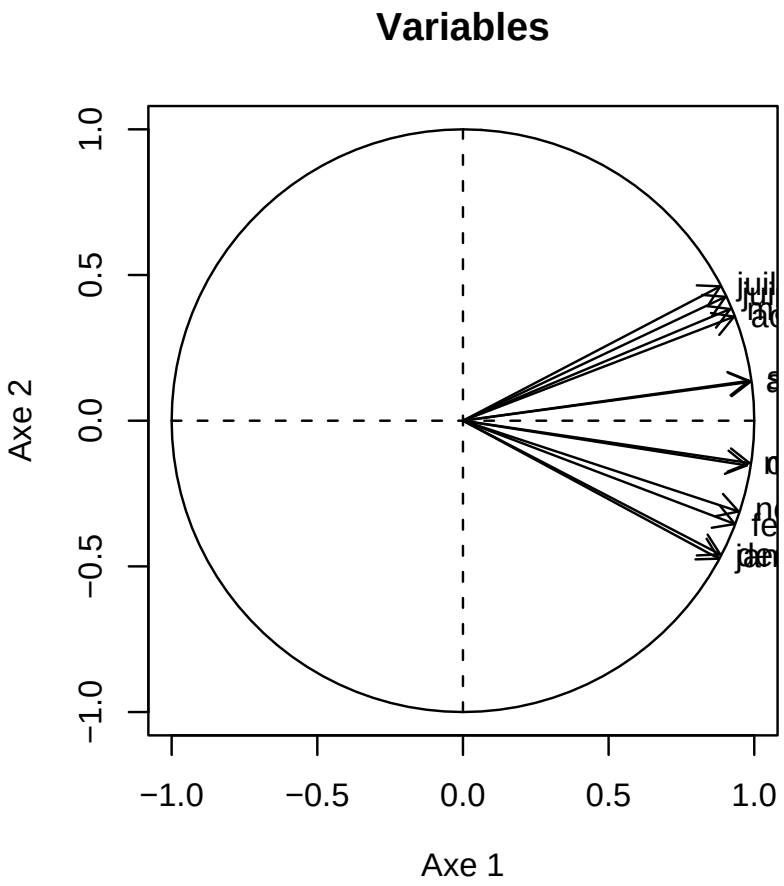
- Corrélations variables$\times$facteurs : $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

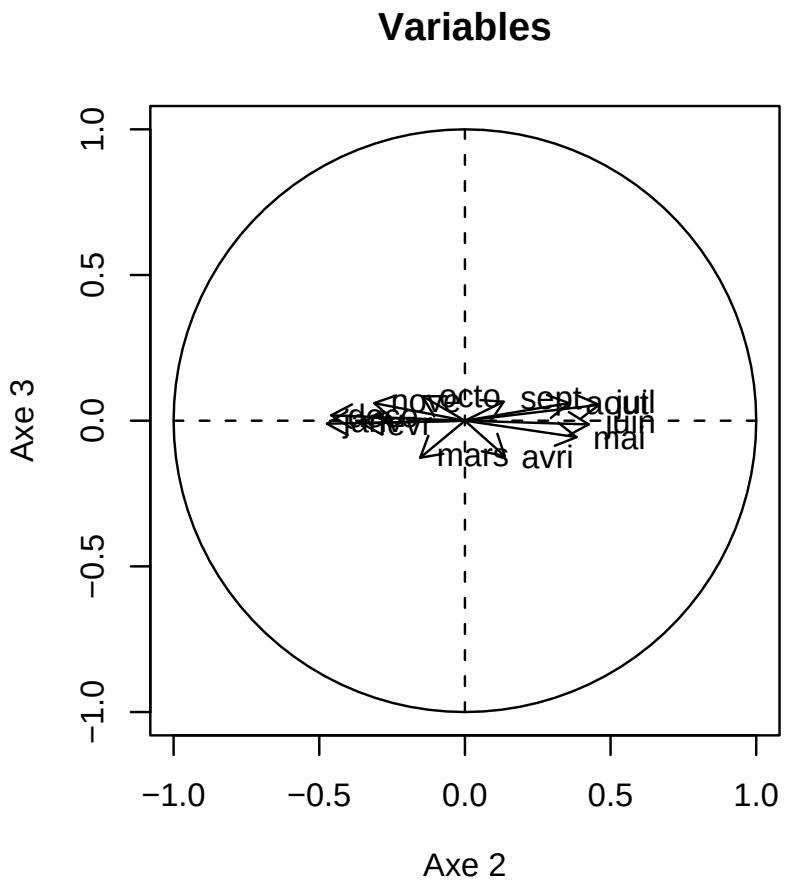
$$\text{cor}(X^j, C^k) = \cos\theta(\mathbf{x}^j, \mathbf{c}^k) = \cos\theta(\mathbf{x}^j, \mathbf{u}^k) = \frac{\left\langle \mathbf{x}^j, \mathbf{u}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}} = \frac{\sqrt{\lambda_k}}{\sigma_j} v_j^k \ ;$$

- Cercle des corrélations : Les variables réduites $\widetilde{\mathbf{x}}^j = \sigma_j^{-1}\mathbf{x}^j$, $\left\|\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = 1$, sont sur une sphère unité et se projettent à l'intérieur d'un cercle :

$$\left\|\widehat{\mathbf{Q}_2}\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = \cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_2}x_j) \leq 1.$$

Températures : graphes des variables





Interprétation du bi-plot

- la DVS de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$: $x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$
- est le produit scalaire des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j.$$

- Pour $q = 2$, $\widehat{z}_i^j$ est une approximation de la valeur prise par la variable

Coordonnées du bi-plot

Isométrique ligne matrices $\mathbf{C}$ et $\mathbf{V}$

Isométrique colonne : matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$

Souvent, en pratique : $\mathbf{C}$ et $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

Interprétation du bi-plot

- la DVS de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$: $x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$
- est le produit scalaire des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j.$$

- Pour $q = 2$, $\widehat{z_i^j}$ est une approximation de la valeur prise par la variable

Coordonnées du bi-plot

Isométrique ligne matrices $\mathbf{C}$ et $\mathbf{V}$

Isométrique colonne : matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$

Souvent, en pratique : $\mathbf{C}$ et $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

Interprétation du bi-plot

- la DVS de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$: $x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$
- est le produit scalaire des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j .$$

- Pour $q = 2$, $\widehat{z_i^j}$ est une approximation de la valeur prise par la variable

Coordonnées du bi-plot

Isométrique ligne matrices $\mathbf{C}$ et $\mathbf{V}$

Isométrique colonne : matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$

Souvent, en pratique : $\mathbf{C}$ et $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

Interprétation du bi-plot

- la DVS de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$: $x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$
- est le produit scalaire des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j.$$

- Pour $q = 2$, $\widehat{z_i^j}$ est une approximation de la valeur prise par la variable

Coordonnées du bi-plot

Isométrique ligne matrices $\mathbf{C}$ et $\mathbf{V}$

Isométrique colonne : matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$

Souvent, en pratique : $\mathbf{C}$ et $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

Interprétation du bi-plot

- la DVS de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$: $x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$
- est le produit scalaire des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j.$$

- Pour $q = 2$, $\widehat{z_i^j}$ est une approximation de la valeur prise par la variable

Coordonnées du bi-plot

Isométrique ligne matrices $\mathbf{C}$ et $\mathbf{V}$

Isométrique colonne : matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$

Souvent, en pratique : $\mathbf{C}$ et $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

Interprétation du bi-plot

- la DVS de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$: $x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$
- est le produit scalaire des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j.$$

- Pour $q = 2$, $\widehat{z_i^j}$ est une approximation de la valeur prise par la variable

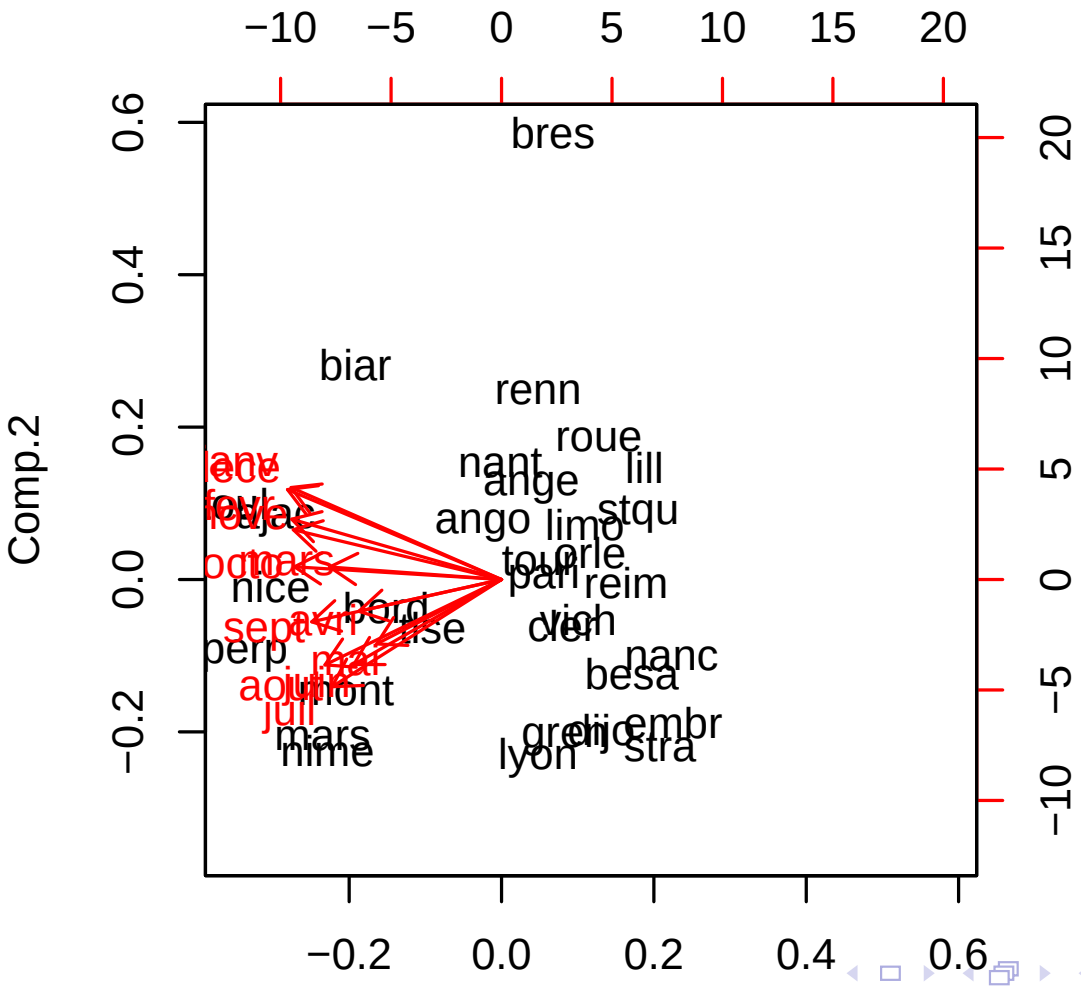
Coordonnées du bi-plot

Isométrique ligne matrices $\mathbf{C}$ et $\mathbf{V}$

Isométrique colonne : matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$

Souvent, en pratique : $\mathbf{C}$ et $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$

Températures : bi-plot



"Critères" élémentaires de choix de q

- Qualité globale

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

soit supérieure à un seuil fixé a priori

- q correspondant aux valeurs propres supérieures à leur moyenne (règle de Kaiser)
- Éboulis des valeurs propres

"Critères" élémentaires de choix de q

- Qualité globale

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

  soit supérieure à un seuil fixé a priori
- q correspondant aux valeurs propres supérieures à leur moyenne (règle de Kaiser)
- Éboulis des valeurs propres

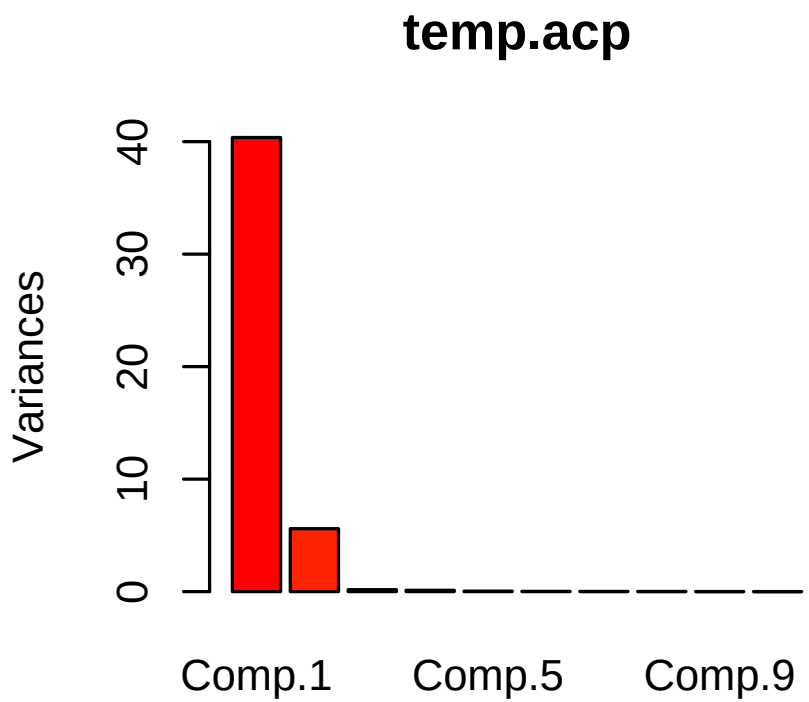
"Critères" élémentaires de choix de q

- Qualité globale

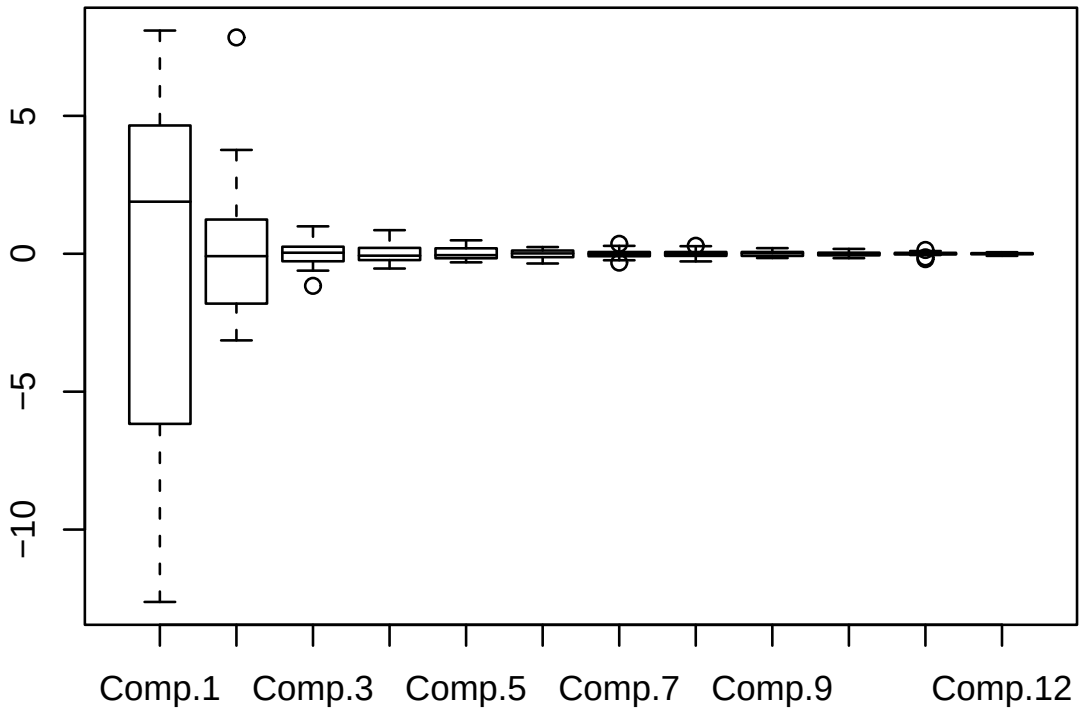
$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

  soit supérieure à un seuil fixé a priori
- q correspondant aux valeurs propres supérieures à leur moyenne (règle de Kaiser)
- Éboulis des valeurs propres

Températures : éboulis



Températures : Diagrammes-boîtes des composantes



Interprétation

- Contributions des individus
- Nombre de composantes
- Interprétation des axes par rapport aux variables initiales bien représentées
- Représentation des individus

Attention

L'ACP est une technique linéaire optimisant un critère quadratique

Interprétation

- Contributions des individus
- Nombre de composantes
- Interprétation des axes par rapport aux variables initiales bien représentées
- Représentation des individus

Attention

L'ACP est une technique linéaire optimisant un critère quadratique

Interprétation

- Contributions des individus
- Nombre de composantes
- Interprétation des axes par rapport aux variables initiales bien représentées
- Représentation des individus

Attention

L'ACP est une technique linéaire optimisant un critère quadratique

Interprétation

- Contributions des individus
- Nombre de composantes
- Interprétation des axes par rapport aux variables initiales bien représentées
- Représentation des individus

Attention

L'ACP est une technique linéaire optimisant un critère quadratique

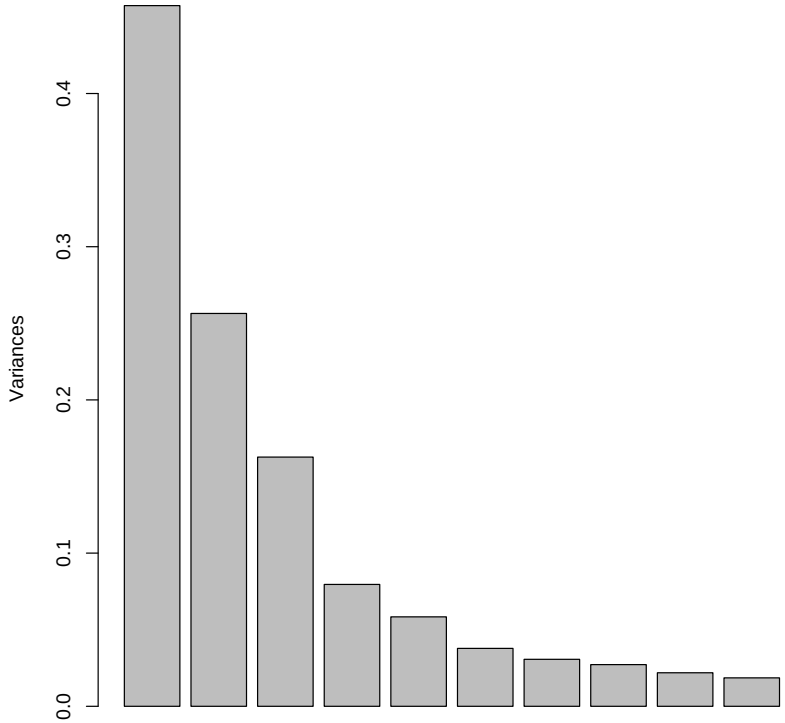
Interprétation

- Contributions des individus
- Nombre de composantes
- Interprétation des axes par rapport aux variables initiales bien représentées
- Représentation des individus

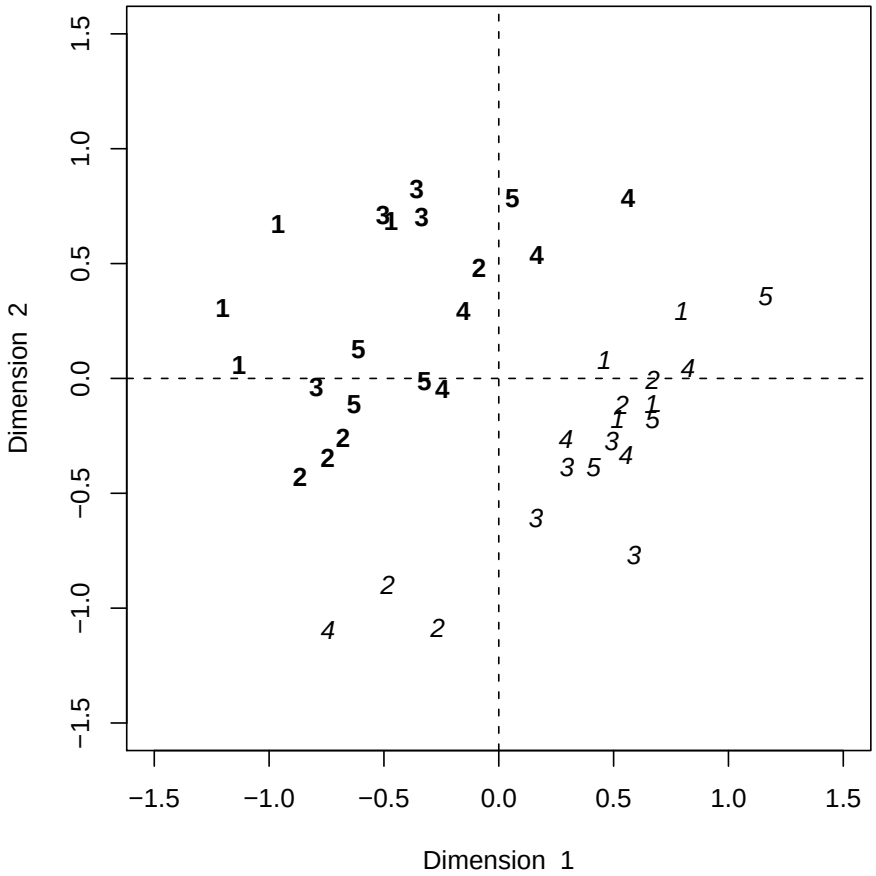
Attention

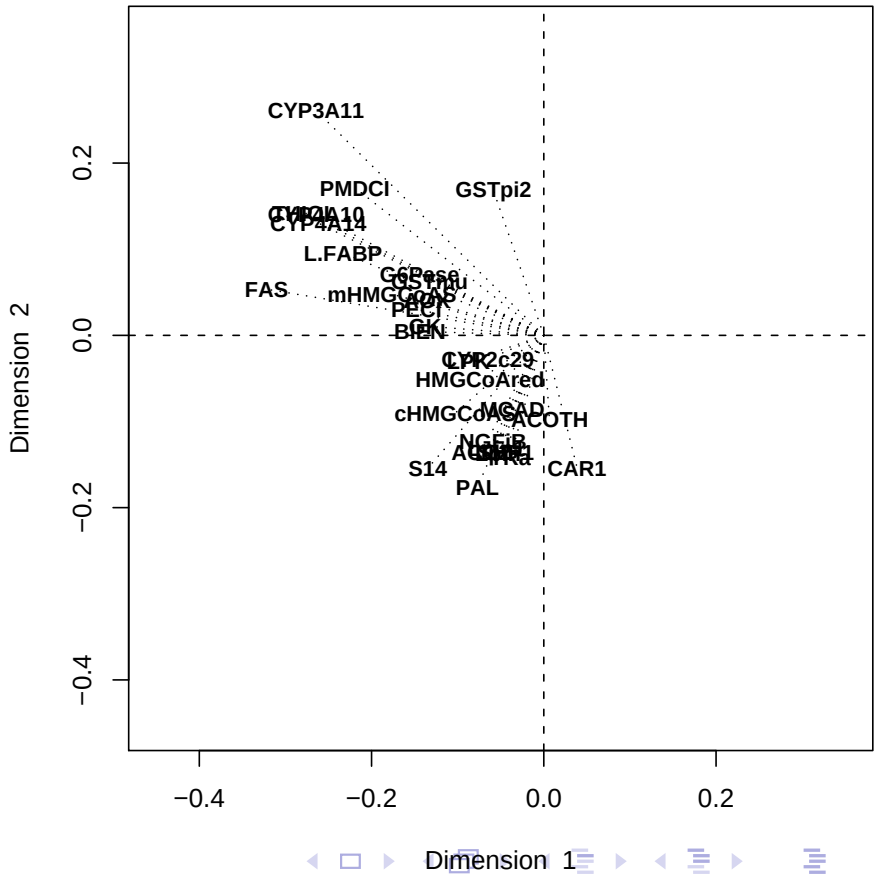
L'ACP est une technique linéaire optimisant un critère quadratique

Souris : éboulis des valeurs propres

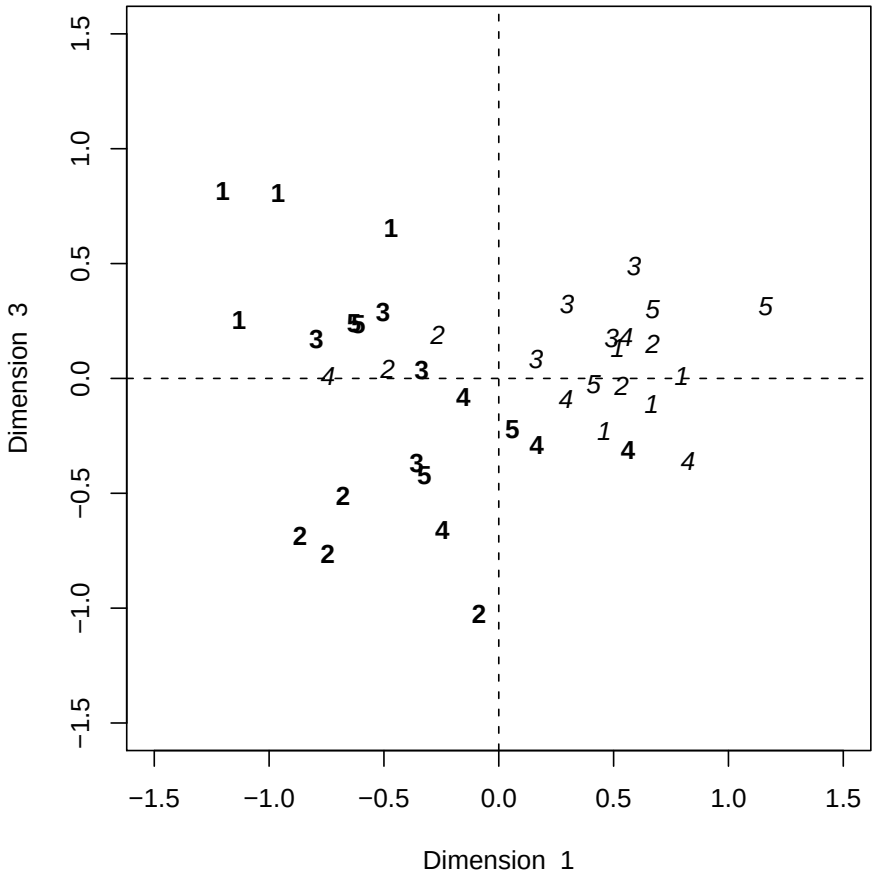


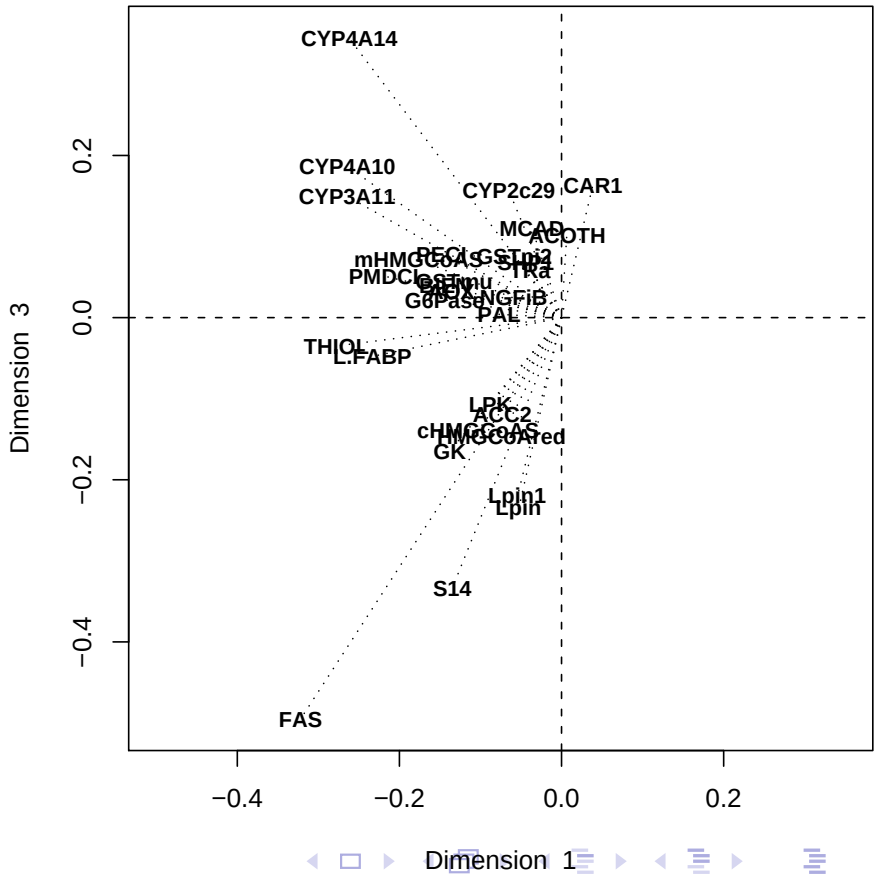
Souris : biplot





Souris : biplot





Statistique des données d'expression
Chapitre 4 : Analyse factorielle disriminante

Alain Baccini & Philippe Besse

Laboratoire de Statistique et Probabilités
Université de Toulouse

Institut de Mathématiques
math.univ-toulouse.fr/biostat

Contexte

- p variables quantitatives $X^1, \ldots, X^p$
- une variable qualitative T, à m modalités
- Ensemble Ω des n individus de poids $w_i > 0$:
  $\mathbf{D} = \text{diag}(w_i \, ; i = 1, \ldots, n); \sum_{i=1}^{n} w_i = 1$
- T engendre une partition $\{\Omega_\ell \, ; \ell = 1, \ldots, m\}$ de Ω
- $\mathbf{T}$ $(n \times m)$: matrice des indicatrices des modalités

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases} .$$

- $\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i$: poids des classes,
- $\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{diag}(\overline{w_1}, \ldots, \overline{w_m})$

Contexte

- p variables quantitatives $X^1, \ldots, X^p$
- une variable qualitative T, à m modalités
- Ensemble Ω des n individus de poids $w_i > 0$:
  $\mathbf{D} = \text{diag}(w_i\ ; i = 1, \ldots, n); \sum_{i=1}^{n} w_i = 1$
- T engendre une partition $\{\Omega_\ell\ ; \ell = 1, \ldots, m\}$ de Ω
- $\mathbf{T}$ $(n \times m)$: matrice des indicatrices des modalités

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases} .$$

- $\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i$: poids des classes,
- $\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{diag}(\overline{w_1}, \ldots, \overline{w_m})$

Contexte

- p variables quantitatives $X^1, \ldots, X^p$
- une variable qualitative T, à m modalités
- Ensemble Ω des n individus de poids $w_i > 0$:
  $\mathbf{D} = \text{diag}(w_i\ ; i = 1, \ldots, n); \sum_{i=1}^{n} w_i = 1$
- T engendre une partition $\{\Omega_\ell\ ; \ell = 1, \ldots, m\}$ de Ω
- $\mathbf{T}$ $(n \times m)$: matrice des indicatrices des modalités

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases} .$$

- $\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i$: poids des classes,
- $\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{diag}(\overline{w_1}, \ldots, \overline{w_m})$

Contexte

- p variables quantitatives $X^1, \ldots, X^p$
- une variable qualitative T, à m modalités
- Ensemble Ω des n individus de poids $w_i > 0$:
  $\mathbf{D} = \text{diag}(w_i\ ; i = 1, \ldots, n); \sum_{i=1}^{n} w_i = 1$
- T engendre une partition $\{\Omega_\ell\ ; \ell = 1, \ldots, m\}$ de Ω
- $\mathbf{T}$ $(n \times m)$: matrice des indicatrices des modalités

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases} .$$

- $\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i$: poids des classes,
- $\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{diag}(\overline{w_1}, \ldots, \overline{w_m})$

Contexte

- p variables quantitatives $X^1, \ldots, X^p$
- une variable qualitative T, à m modalités
- Ensemble Ω des n individus de poids $w_i > 0$:
  $\mathbf{D} = \text{diag}(w_i\ ; i = 1, \ldots, n); \sum_{i=1}^{n} w_i = 1$
- T engendre une partition $\{\Omega_\ell\ ; \ell = 1, \ldots, m\}$ de Ω
- $\mathbf{T}$ $(n \times m)$: matrice des indicatrices des modalités

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases} .$$

- $\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i$: poids des classes,
- $\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{diag}(\overline{w_1}, \ldots, \overline{w_m})$

Contexte

- p variables quantitatives $X^1, \ldots, X^p$
- une variable qualitative T, à m modalités
- Ensemble Ω des n individus de poids $w_i > 0$:
  $\mathbf{D} = \text{diag}(w_i \, ; i = 1, \ldots, n); \sum_{i=1}^{n} w_i = 1$
- T engendre une partition $\{\Omega_\ell \, ; \ell = 1, \ldots, m\}$ de Ω
- $\mathbf{T}$ $(n \times m)$: matrice des indicatrices des modalités

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases} .$$

- $\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i$: poids des classes,
- $\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{diag}(\overline{w_1}, \ldots, \overline{w_m})$

Contexte

- p variables quantitatives $X^1, \ldots, X^p$
- une variable qualitative T, à m modalités
- Ensemble Ω des n individus de poids $w_i > 0$:
  $\mathbf{D} = \text{diag}(w_i \, ; i = 1, \ldots, n); \sum_{i=1}^{n} w_i = 1$
- T engendre une partition $\{\Omega_\ell \, ; \ell = 1, \ldots, m\}$ de Ω
- $\mathbf{T}$ $(n \times m)$: matrice des indicatrices des modalités

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases} .$$

- $\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i$: poids des classes,
- $\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{diag}(\overline{w_1}, \ldots, \overline{w_m})$

Objectifs

- ACP représentant au mieux les m classes de T
- affectation d'un nouvel individu

Notations 1

- $\mathbf{X}$ matrice $(n \times p)$ des données
- $\mathbf{G}$ la matrice $(m \times p)$ des barycentres des classes :

$$\mathbf{G} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{X} = \begin{bmatrix} \mathbf{g}_1{}' \\ \vdots \\ \mathbf{g}_m{}' \end{bmatrix} \text{ où } \mathbf{g}_\ell = \frac{1}{w_\ell}\sum_{i\in\Omega_\ell} w_i \mathbf{x}_i$$

- $\mathbf{X}_e$ matrice $(n \times p)$ des barycentres "répétés" :
  $\mathbf{X}_e = \mathbf{T}\mathbf{G} = \mathbf{P}\mathbf{G}$ où $\mathbf{P} = \mathbf{T}\overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}$
- $\overline{\mathbf{X}} = \overline{\mathbf{X}}_r + \overline{\mathbf{X}}_e$ avec $\overline{\mathbf{X}}_r = \mathbf{X} - \mathbf{X}_e$ et $\overline{\mathbf{X}}_e = \mathbf{X}_e - \mathbf{1}_n\bar{\mathbf{x}}$

Objectifs

- ACP représentant au mieux les m classes de T
- affectation d'un nouvel individu

Notations 1

- $\mathbf{X}$ matrice $(n \times p)$ des données
- $\mathbf{G}$ la matrice $(m \times p)$ des barycentres des classes :

$$\mathbf{G} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{X} = \begin{bmatrix} \mathbf{g}_1{}' \\ \vdots \\ \mathbf{g}_m{}' \end{bmatrix} \text{ où } \mathbf{g}_\ell = \frac{1}{w_\ell}\sum_{i\in\Omega_\ell} w_i \mathbf{x}_i$$

- $\mathbf{X}_e$ matrice $(n \times p)$ des barycentres "répétés" :
  $\mathbf{X}_e = \mathbf{T}\mathbf{G} = \mathbf{P}\mathbf{G}$ où $\mathbf{P} = \mathbf{T}\overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}$
- $\overline{\mathbf{X}} = \overline{\mathbf{X}}_r + \overline{\mathbf{X}}_e$ avec $\overline{\mathbf{X}}_r = \mathbf{X} - \mathbf{X}_e$ et $\overline{\mathbf{X}}_e = \mathbf{X}_e - \mathbf{1}_n\bar{\mathbf{x}}$

Notations 2

- Matrice centrée des barycentres : $\overline{\mathbf{G}} = \mathbf{G} - \mathbf{1}_m \bar{\mathbf{x}}'$
- Variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}' \mathbf{D} \overline{\mathbf{X}}_r = \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'$$

- Variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}' \mathbf{D} \overline{\mathbf{G}} = \overline{X}_e' \mathbf{D} \overline{X}_e = \sum_{\ell=1}^{m} \overline{w_\ell} (\mathbf{g}_\ell - \bar{\mathbf{x}})(\mathbf{g}_\ell - \bar{\mathbf{x}})'$$

- $\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r$

Notations 2

- Matrice centrée des barycentres : $\overline{\mathbf{G}} = \mathbf{G} - \mathbf{1}_m \bar{\mathbf{x}}'$
- Variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}' \mathbf{D} \overline{\mathbf{X}}_r = \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'$$

- Variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}' \mathbf{D} \overline{\mathbf{G}} = \overline{X}_e' \mathbf{D} \overline{X}_e = \sum_{\ell=1}^{m} \overline{w_\ell} (\mathbf{g}_\ell - \bar{\mathbf{x}})(\mathbf{g}_\ell - \bar{\mathbf{x}})'$$

- $\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r$

Notations 2

- Matrice centrée des barycentres : $\overline{\mathbf{G}} = \mathbf{G} - \mathbf{1}_m \bar{\mathbf{x}}'$
- Variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}' \mathbf{D} \overline{\mathbf{X}}_r = \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'$$

- Variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}' \mathbf{D} \overline{\mathbf{G}} = \overline{X}_e' \mathbf{D} \overline{X}_e = \sum_{\ell=1}^{m} \overline{w_\ell} (\mathbf{g}_\ell - \bar{\mathbf{x}})(\mathbf{g}_\ell - \bar{\mathbf{x}})'$$

- $\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r$

Notations 2

- Matrice centrée des barycentres : $\overline{\mathbf{G}} = \mathbf{G} - \mathbf{1}_m \bar{\mathbf{x}}'$
- Variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}' \mathbf{D} \overline{\mathbf{X}}_r = \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'$$

- Variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}' \mathbf{D} \overline{\mathbf{G}} = \overline{X}_e' \mathbf{D} \overline{X}_e = \sum_{\ell=1}^{m} \overline{w_\ell} (\mathbf{g}_\ell - \bar{\mathbf{x}})(\mathbf{g}_\ell - \bar{\mathbf{x}})'$$

- $\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r$

Modèle

$$\mathbf{x}_i = \mathbf{z}_\ell + \boldsymbol{\varepsilon}_i \quad \forall \ell, \forall i \in \Omega_\ell$$

$$\text{avec} \begin{cases} \{\mathbf{x}_i \, ; i = 1, \ldots, n\}, \ n \text{ vect. indép. de } E \\ E(\boldsymbol{\varepsilon}_i) = 0, \ \text{var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Gamma} \\ \boldsymbol{\Gamma} \text{ régulière et inconnue,} \end{cases}$$

$\exists A_q$, sous-espace affine de dimension q de E tel que

$$\forall \ell, \mathbf{z}_\ell \in A_q, \ (q < \min(p, m-1))$$

Si $\bar{\mathbf{z}} = \sum_{\ell=1}^{m} \overline{w_\ell} \mathbf{z}_\ell$ alors $\bar{\mathbf{z}} \in A_q$
Soit E_q tel que $A_q = \bar{\mathbf{z}} + E_q$

Théorème

L'estimation des paramètres E_q et $\mathbf{z}_\ell$ du modèle est obtenue par l'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$. C'est l'Analyse Factorielle Discriminante (AFD) de $(\mathbf{X}|\mathbf{T}, \mathbf{D})$.

Preuve

- $\min_{E_q, \mathbf{z}_\ell} \left\{ \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i \left\| \mathbf{x}_i - \mathbf{z}_\ell \right\|_{\mathbf{M}}^2 ; \ \dim(E_q) = q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q \right\}$
- $\sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i \left\| \mathbf{x}_i - \mathbf{z}_\ell \right\|_{\mathbf{M}}^2 =$
  $\sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i \left\| \mathbf{x}_i - \mathbf{g}_\ell \right\|_{\mathbf{M}}^2 + + \sum_{\ell=1}^{m} \overline{w_\ell} \left\| \mathbf{g}_\ell - \mathbf{z}_\ell \right\|_{\mathbf{M}}^2$
- le problème se ramène à
  $\min_{E_q, z_\ell} \left\{ \sum_{\ell=1}^{m} \overline{w_\ell} \left\| \mathbf{g}_\ell - \mathbf{z}_\ell \right\|_{\mathbf{M}}^2 \ ; \ \dim(E_q) = q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q \right\}$
- $\sigma^2 \Gamma$ inconnue est estimée par : $\mathbf{M} = \widehat{\Gamma}^{-1} = \mathbf{S}_r^{-1}$

Théorème

L'estimation des paramètres E_q et $\mathbf{z}_\ell$ du modèle est obtenue par l'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$. C'est l'Analyse Factorielle Discriminante (AFD) de $(\mathbf{X}|\mathbf{T}, \mathbf{D})$.

Preuve

- $\min_{E_q, \mathbf{z}_\ell} \left\{ \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i \left\| \mathbf{x}_i - \mathbf{z}_\ell \right\|_{\mathbf{M}}^2 ; \ \dim(E_q) = q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q \right\}$
- $\sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i \left\| \mathbf{x}_i - \mathbf{z}_\ell \right\|_{\mathbf{M}}^2 =$
  $\sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i \left\| \mathbf{x}_i - \mathbf{g}_\ell \right\|_{\mathbf{M}}^2 + + \sum_{\ell=1}^{m} \overline{w_\ell} \left\| \mathbf{g}_\ell - \mathbf{z}_\ell \right\|_{\mathbf{M}}^2$
- le problème se ramène à
  $\min_{E_q, z_\ell} \left\{ \sum_{\ell=1}^{m} \overline{w_\ell} \left\| \mathbf{g}_\ell - \mathbf{z}_\ell \right\|_{\mathbf{M}}^2 \ ; \ \dim(E_q) = q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q \right\}$
- $\sigma^2 \mathbf{\Gamma}$ inconnue est estimée par : $\mathbf{M} = \widehat{\mathbf{\Gamma}}^{-1} = \mathbf{S}_r^{-1}$

Résultats de l'AFD

- L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à diagonaliser : $\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}$ avec $\text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p)$
- $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h)$ matrice des valeurs propres
- $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h]$, vecteurs propres $\mathbf{S}_r^{-1}$-orthonormés
- Les vecteurs $\mathbf{v}^k$ engendrent les axes discriminants
- Représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ dans les mêmes axes discriminants
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V}$ et $\overline{\mathbf{C}} = \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C}$ de l'ACP des barycentres

Résultats de l'AFD

- L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à diagonaliser : $\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}$ avec $\text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p)$
- $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h)$ matrice des valeurs propres
- $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h]$, vecteurs propres $\mathbf{S}_r^{-1}$-orthonormés
- Les vecteurs $\mathbf{v}^k$ engendrent les axes discriminants
- Représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ dans les mêmes axes discriminants
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V}$ et $\overline{\mathbf{C}} = \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C}$ de l'ACP des barycentres

Résultats de l'AFD

- L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à diagonaliser : $\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}$ avec $\text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p)$
- $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h)$ matrice des valeurs propres
- $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h]$, vecteurs propres $\mathbf{S}_r^{-1}$-orthonormés
- Les vecteurs $\mathbf{v}^k$ engendrent les axes discriminants
- Représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ dans les mêmes axes discriminants
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V}$ et $\overline{\mathbf{C}} = \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C}$ de l'ACP des barycentres

Résultats de l'AFD

- L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à diagonaliser : $\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}$ avec $\text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p)$
- $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h)$ matrice des valeurs propres
- $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h]$, vecteurs propres $\mathbf{S}_r^{-1}$-orthonormés
- Les vecteurs $\mathbf{v}^k$ engendrent les axes discriminants
- Représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ dans les mêmes axes discriminants
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V}$ et $\overline{\mathbf{C}} = \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C}$ de l'ACP des barycentres

Résultats de l'AFD

- L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à diagonaliser : $\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}$ avec $\text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p)$
- $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h)$ matrice des valeurs propres
- $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h]$, vecteurs propres $\mathbf{S}_r^{-1}$-orthonormés
- Les vecteurs $\mathbf{v}^k$ engendrent les axes discriminants
- Représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ dans les mêmes axes discriminants
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V}$ et $\overline{\mathbf{C}} = \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C}$ de l'ACP des barycentres

Résultats de l'AFD

- L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à diagonaliser : $\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}$ avec $\text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p)$
- $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h)$ matrice des valeurs propres
- $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h]$, vecteurs propres $\mathbf{S}_r^{-1}$-orthonormés
- Les vecteurs $\mathbf{v}^k$ engendrent les axes discriminants
- Représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ dans les mêmes axes discriminants
- $\mathbf{C} = \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V}$ et $\overline{\mathbf{C}} = \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C}$ de l'ACP des barycentres

Interprétation

- Qualités de représentation : issues de l'ACP
- Représentation des variables dans ($\mathbb{R}^m$, b. c., $\overline{\mathbf{D}}$) : $\mathbf{V}\mathbf{\Lambda}^{1/2}$
- Interprétations en termes d'écarts-types et de corrélations expliquées par la partition
- Matrice des corrélations expliquées variables$\times$facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$ pour interpréter les axes
- Qualité de la discrimination des classes et donc de l'explication de T par les X^j.

Interprétation

- Qualités de représentation : issues de l'ACP
- Représentation des variables dans ($\mathbb{R}^m$, b. c., $\overline{\mathbf{D}}$) : $\mathbf{V}\mathbf{\Lambda}^{1/2}$
- Interprétations en termes d'écarts-types et de corrélations expliquées par la partition
- Matrice des corrélations expliquées variables$\times$facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$ pour interpréter les axes
- Qualité de la discrimination des classes et donc de l'explication de T par les X^j.

Interprétation

- Qualités de représentation : issues de l'ACP
- Représentation des variables dans ($\mathbb{R}^m$, b. c., $\overline{\mathbf{D}}$) : $\mathbf{V}\mathbf{\Lambda}^{1/2}$
- Interprétations en termes d'écarts-types et de corrélations expliquées par la partition
- Matrice des corrélations expliquées variables×facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$ pour interpréter les axes
- Qualité de la discrimination des classes et donc de l'explication de T par les X^j.

Interprétation

- Qualités de représentation : issues de l'ACP
- Représentation des variables dans ($\mathbb{R}^m$, b. c., $\overline{\mathbf{D}}$) : $\mathbf{V}\mathbf{\Lambda}^{1/2}$
- Interprétations en termes d'écarts-types et de corrélations expliquées par la partition
- Matrice des corrélations expliquées variables$\times$facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$ pour interpréter les axes
- Qualité de la discrimination des classes et donc de l'explication de T par les X^j.

Interprétation

- Qualités de représentation : issues de l'ACP
- Représentation des variables dans $(\mathbb{R}^m$, b. c., $\overline{\mathbf{D}})$: $\mathbf{V}\mathbf{\Lambda}^{1/2}$
- Interprétations en termes d'écarts-types et de corrélations expliquées par la partition
- Matrice des corrélations expliquées variables$\times$facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$ pour interpréter les axes
- Qualité de la discrimination des classes et donc de l'explication de T par les X^j.

Expression de la variance

Les individus sont supposés de même poids

$$\mathbf{D} = \frac{1}{n}\mathbf{I}_n \text{ et } \overline{\mathbf{D}} = \frac{1}{n}\text{diag}(n_1, \ldots, n_m) \text{ où } n_\ell = \text{card}(\Omega_\ell)$$

$$\begin{aligned}
(\mathbf{S})_j^k &= \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x}^j)(x_i^k - \overline{x}^k), \\
(\mathbf{S}_e)_j^k &= \frac{1}{n}\sum_{\ell=1}^{m} n_\ell (g_\ell^j - \overline{x}^j)(g_\ell^k - \overline{x}^k), \\
(\mathbf{S}_r)_j^k &= \frac{1}{n}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(x_i^j - g_\ell^j)(x_i^k - g_\ell^k).
\end{aligned}$$

Estimation sans "biais" de la variance

Estimations sans biais en divisant par les nombres de degrés de liberté : $(n-1), (m-1), (n-m)$
Il faut remplacer

$$\mathbf{S} \quad \text{par} \quad \mathbf{S}^* = \frac{n}{n-1}\mathbf{S},$$

$$\mathbf{S}_e \quad \text{par} \quad \mathbf{S}_e^* = \mathbf{B} = \frac{n}{m-1}\mathbf{S}_e,$$

$$\mathbf{S}_r \quad \text{par} \quad \mathbf{S}_r^* = \mathbf{W} = \frac{n}{n-m}\mathbf{S}_r.$$

Variantes de l'AFD (candisc de SAS)

$$
\begin{aligned}
\mathbf{S}_e^* \mathbf{S}_r^{*-1} &= \tfrac{n-m}{m-1} \mathbf{S}_e \mathbf{S}_r^{-1}, \\
\mathbf{\Lambda}^* &= \tfrac{n-m}{m-1} \mathbf{\Lambda}, \\
\mathbf{V}^* &= \sqrt{\tfrac{n}{n-m}} \mathbf{V}, \\
\overline{\mathbf{C}}^* &= \sqrt{\tfrac{n-m}{n}} \overline{\mathbf{C}}, \\
\mathbf{V}^* \mathbf{\Lambda}^{*1/2} &= \sqrt{\tfrac{n}{m-1}} \mathbf{V} \mathbf{\Lambda}^{1/2}, \\
\mathbf{\Sigma}_e^{*-1} \mathbf{V}^* \mathbf{\Lambda}^{*1/2} &= \mathbf{\Sigma}_e^{-1} \mathbf{V} \mathbf{\Lambda}^{1/2}.
\end{aligned}
$$

Théorème

Les ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ et de $(\mathbf{G}, \mathbf{S}^{-1}, \overline{\mathbf{D}})$ partagent les mêmes vecteurs propres ; $\mathbf{S}^{-1}$ est une expression en France de la métrique de Mahalanobis.

Résultats avec métrique $\mathbf{S}^{-1}$ (logiciel SPAD)

- matrice à diagonaliser : $\mathbf{S}_e\mathbf{S}^{-1}$,
- valeurs propres : $\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}$,
- vecteurs propres : $\mathbf{V}(\mathbf{I}+\mathbf{\Lambda})^{1/2}$,
- représentation des barycentres : $\overline{\mathbf{C}}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}$,
- représentation des variables : $\mathbf{V}\mathbf{\Lambda}^{1/2}$,
- corrélations variables-facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

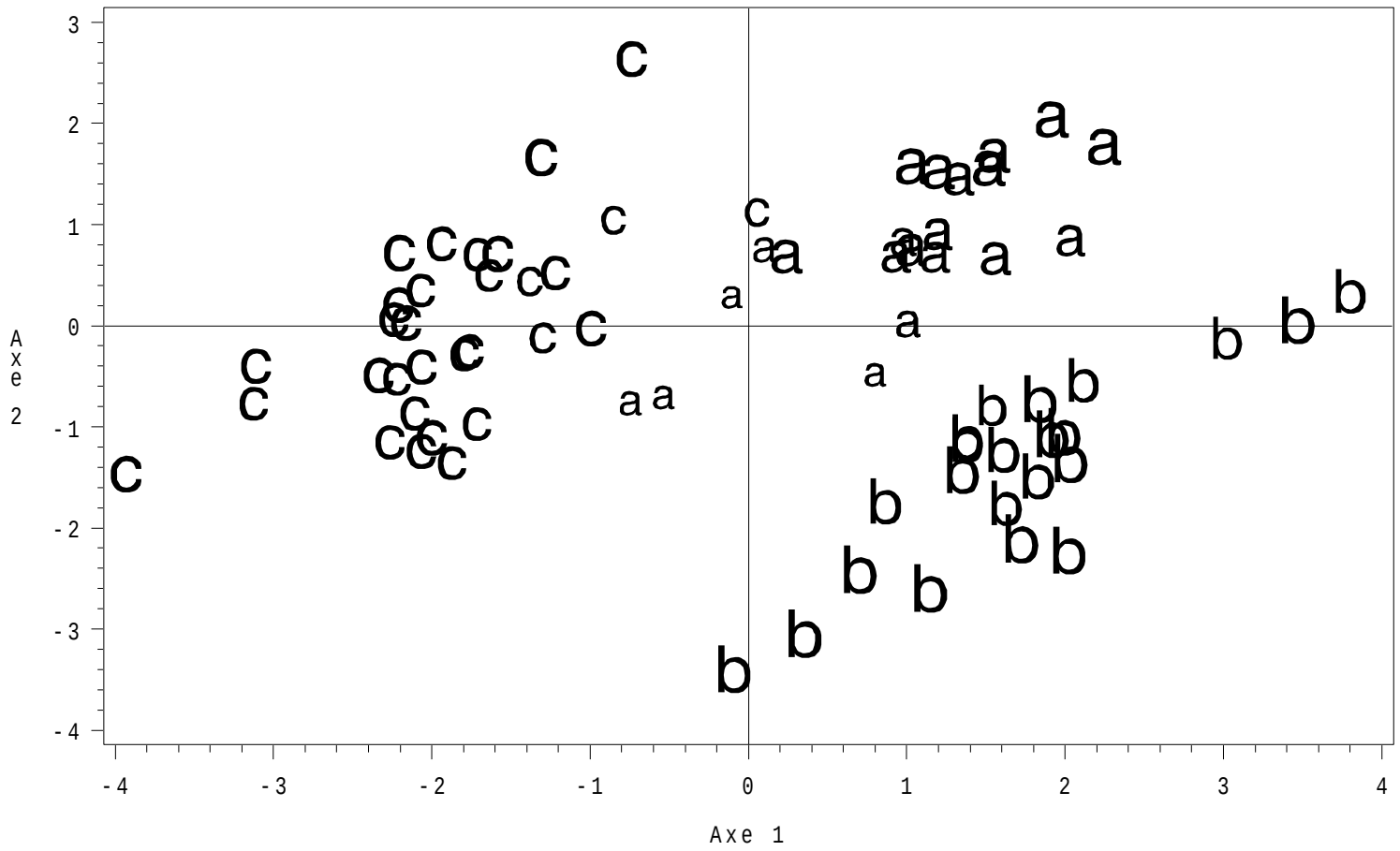
Théorème

Les ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ et de $(\mathbf{G}, \mathbf{S}^{-1}, \overline{\mathbf{D}})$ partagent les mêmes vecteurs propres ; $\mathbf{S}^{-1}$ est une expression en France de la métrique de Mahalanobis.

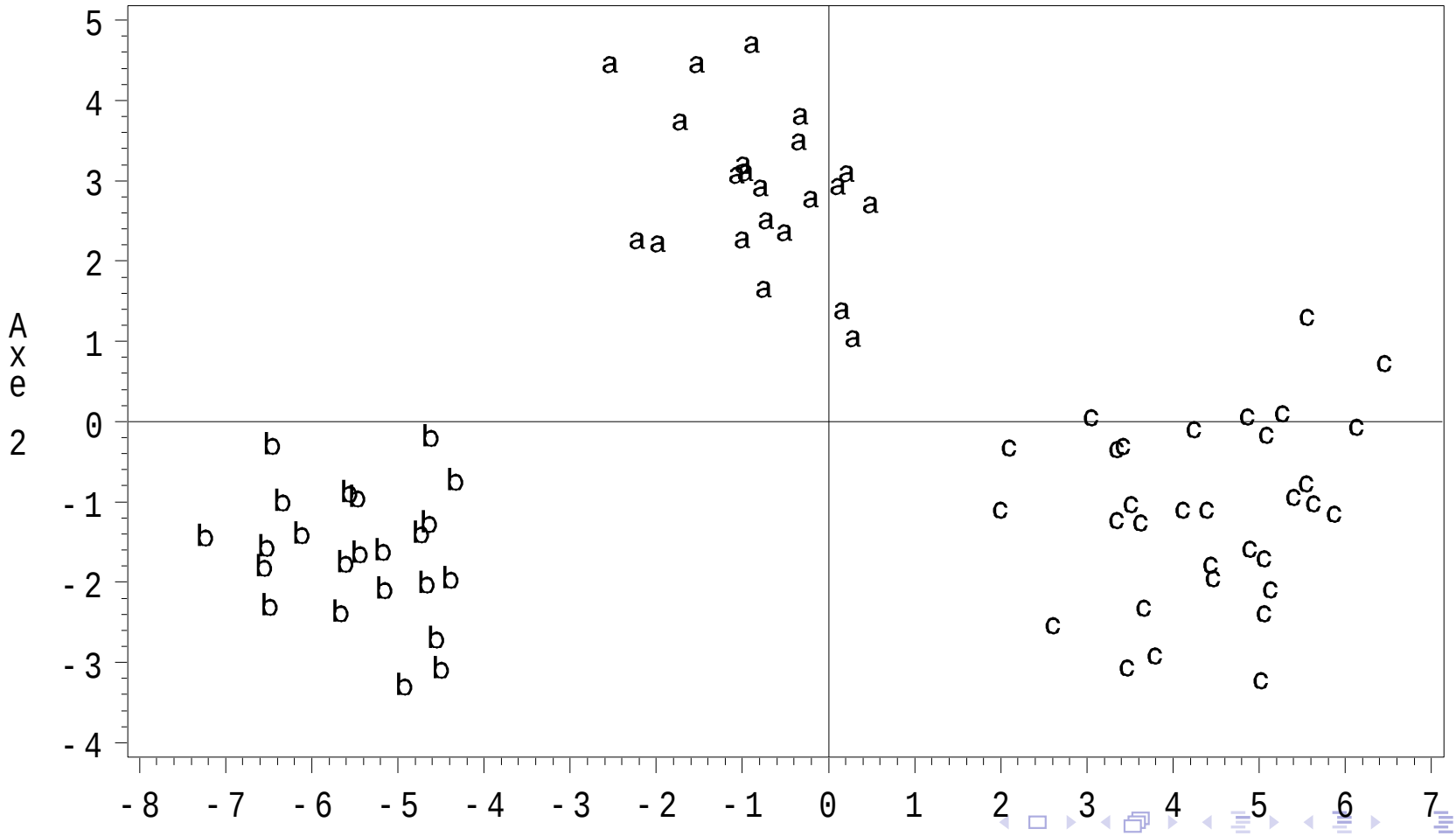
Résultats avec métrique $\mathbf{S}^{-1}$ (logiciel SPAD)

– matrice à diagonaliser : $\mathbf{S}_e\mathbf{S}^{-1}$,
– valeurs propres : $\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}$,
– vecteurs propres : $\mathbf{V}(\mathbf{I}+\mathbf{\Lambda})^{1/2}$,
– représentation des barycentres : $\overline{\mathbf{C}}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}$,
– représentation des variables : $\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

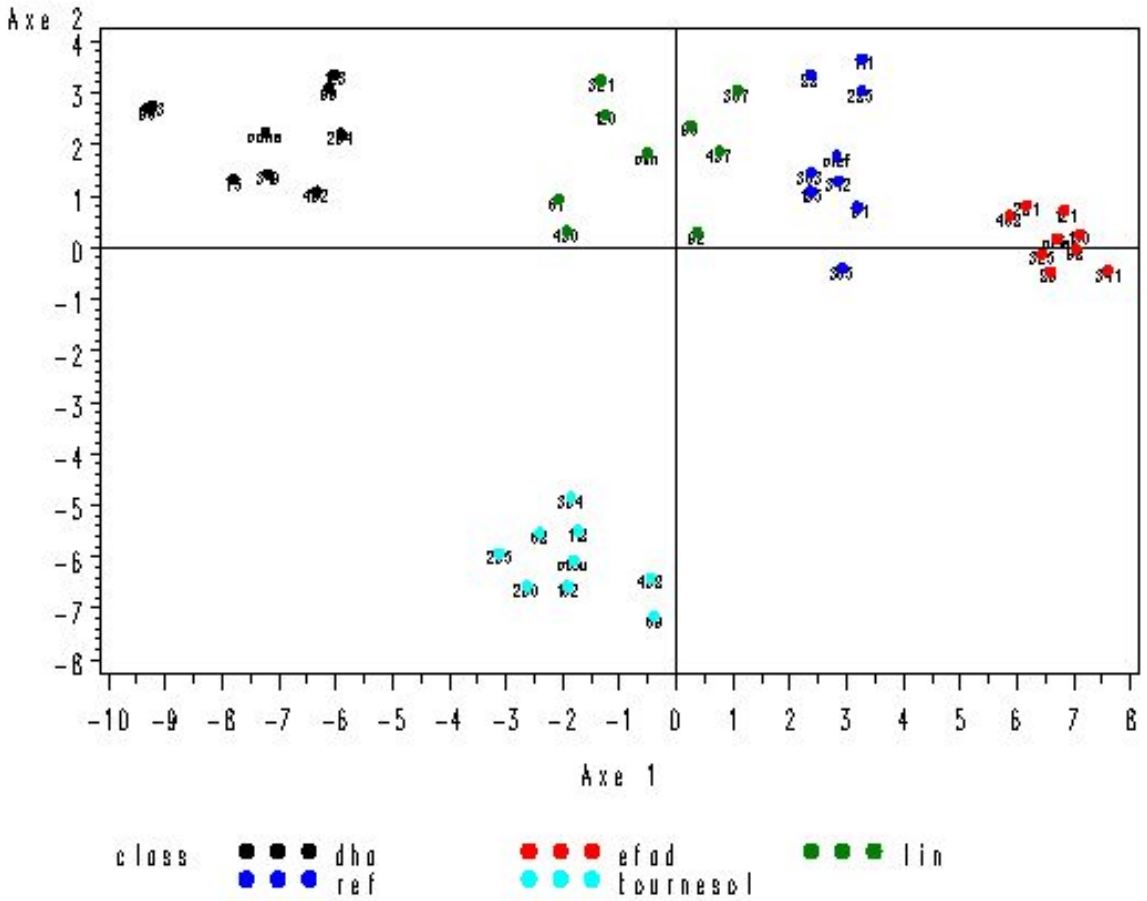
Insectes : premier plan de l'ACP



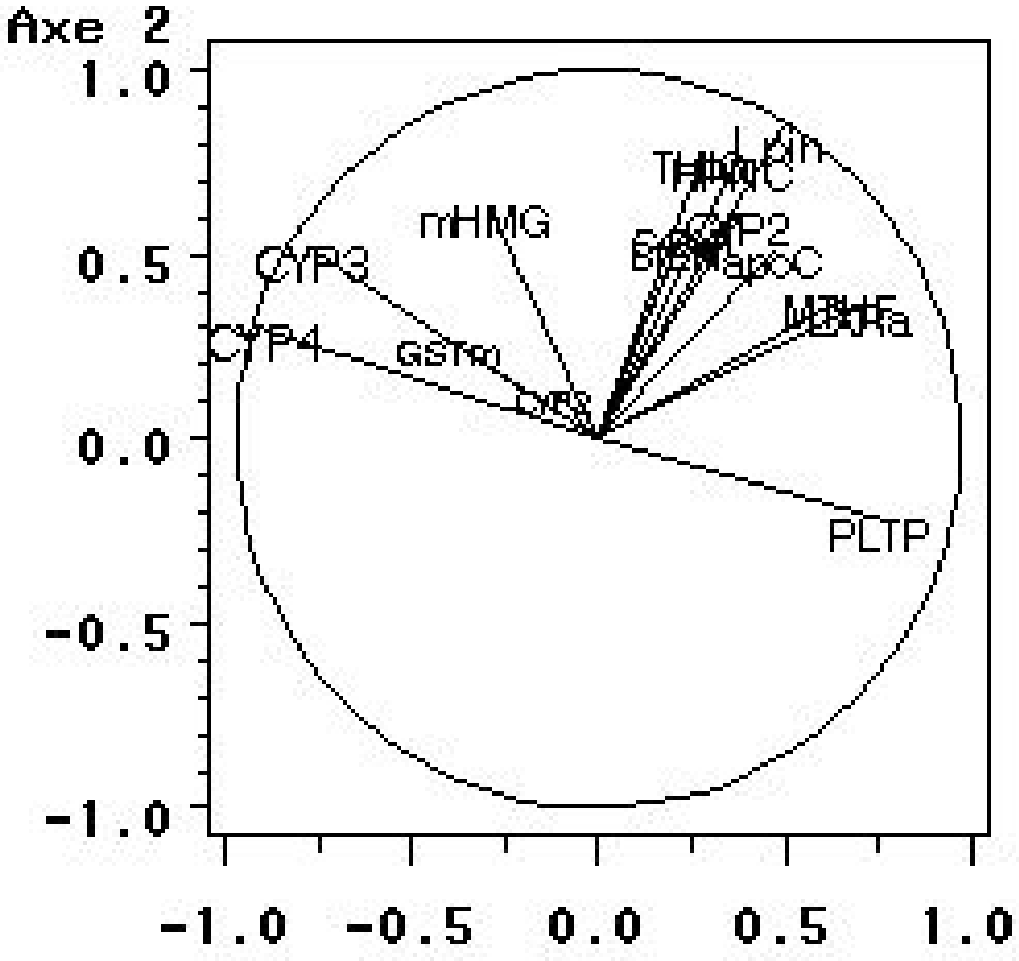
Insectes : premier plan de l'AFD



Souris sauvages : distinction des régimes



Souris sauvages : gènes les plus discriminants



Statistique des données d'expression
Chapitre 5 : Positionnement multidimensionnel

Alain Baccini & Philippe Besse

Laboratoire de Statistique et Probabilités
Université de Toulouse

Institut de Mathématiques
math.univ-toulouse.fr/biostat

Objectif

- $\mathcal{D}$ matrice $(n \times n)$ de distance ou dissimilarité entre n objets
- Objectif : représentation euclidienne approchant les indices observés dans un espace de dimension réduite
- ACP sur tableau de distances ou multidimensional scaling (MDS)
- Application aux transcrits : Quelle distance entre deux gènes X^j et X^k ?
  1. euclidienne
  2. $1 - \text{cor}(X^j, X^k)$
  3. $\sqrt{1 - \text{cor}(X^j, X^k)^2}$

Objectif

- $\mathcal{D}$ matrice $(n \times n)$ de distance ou dissimilarité entre n objets
- Objectif : représentation euclidienne approchant les indices observés dans un espace de dimension réduite
- ACP sur tableau de distances ou multidimensional scaling (MDS)
- Application aux transcrits : Quelle distance entre deux gènes X^j et X^k ?
  1. euclidienne
  2. $1 - \text{cor}(X^j, X^k)$
  3. $\sqrt{1 - \text{cor}(X^j, X^k)^2}$

Objectif

- $\mathcal{D}$ matrice $(n \times n)$ de distance ou dissimilarité entre n objets
- Objectif : représentation euclidienne approchant les indices observés dans un espace de dimension réduite
- ACP sur tableau de distances ou multidimensional scaling (MDS)
- Application aux transcrits : Quelle distance entre deux gènes X^j et X^k ?
  1. euclidienne
  2. $1 - \text{cor}(X^j, X^k)$
  3. $\sqrt{1 - \text{cor}(X^j, X^k)^2}$

Objectif

- $\mathcal{D}$ matrice $(n \times n)$ de distance ou dissimilarité entre n objets
- Objectif : représentation euclidienne approchant les indices observés dans un espace de dimension réduite
- ACP sur tableau de distances ou multidimensional scaling (MDS)
- Application aux transcrits : Quelle distance entre deux gènes X^j et X^k ?
  1. euclidienne
  2. $1 - \text{cor}(X^j, X^k)$
  3. $\sqrt{1 - \text{cor}(X^j, X^k)^2}$

Objectif

- $\mathcal{D}$ matrice $(n \times n)$ de distance ou dissimilarité entre n objets
- Objectif : représentation euclidienne approchant les indices observés dans un espace de dimension réduite
- ACP sur tableau de distances ou multidimensional scaling (MDS)
- Application aux transcrits : Quelle distance entre deux gènes X^j et X^k ?
  1. euclidienne
  2. $1 - \text{cor}(X^j, X^k)$
  3. $\sqrt{1 - \text{cor}(X^j, X^k)^2}$

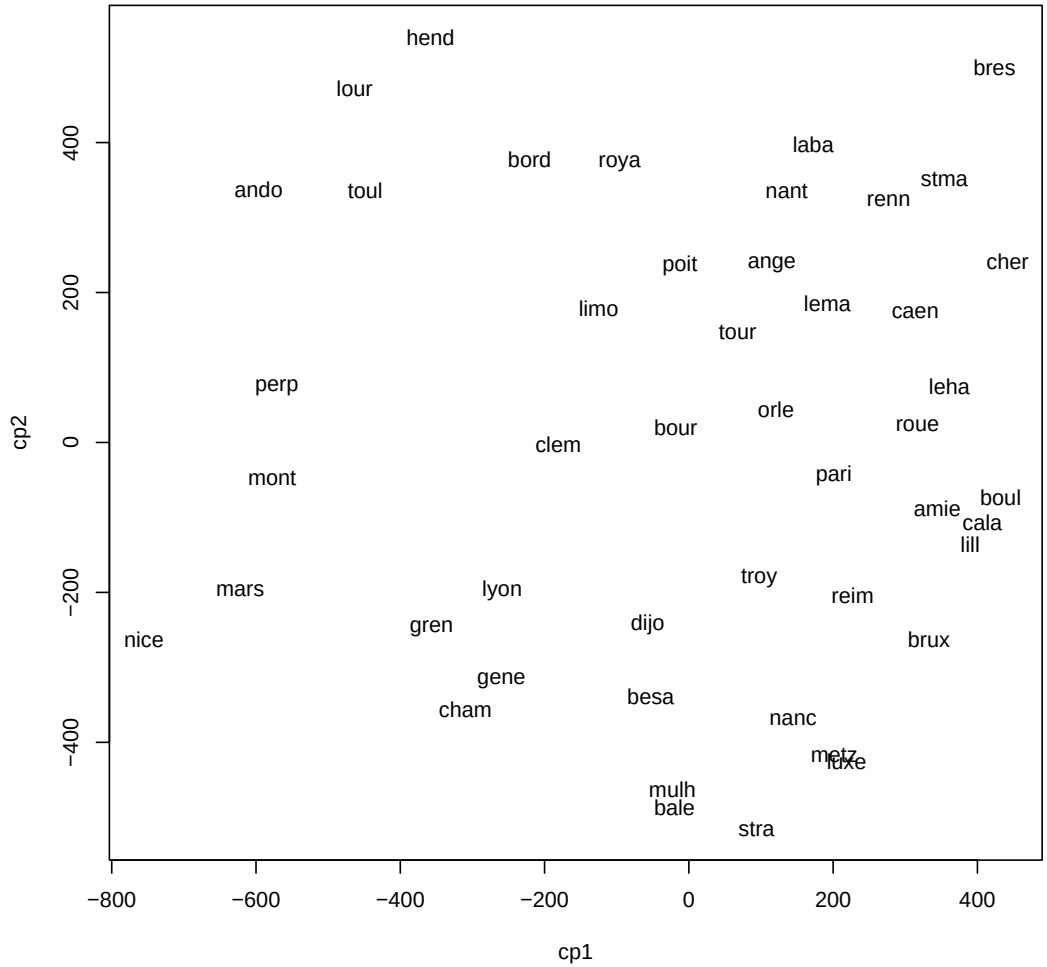
Objectif

- $\mathcal{D}$ matrice $(n \times n)$ de distance ou dissimilarité entre n objets
- Objectif : représentation euclidienne approchant les indices observés dans un espace de dimension réduite
- ACP sur tableau de distances ou multidimensional scaling (MDS)
- Application aux transcrits : Quelle distance entre deux gènes X^j et X^k ?
  1. euclidienne
  2. $1 - \text{cor}(X^j, X^k)$
  3. $\sqrt{1 - \text{cor}(X^j, X^k)^2}$

Objectif

- $\mathcal{D}$ matrice $(n \times n)$ de distance ou dissimilarité entre n objets
- Objectif : représentation euclidienne approchant les indices observés dans un espace de dimension réduite
- ACP sur tableau de distances ou multidimensional scaling (MDS)
- Application aux transcrits : Quelle distance entre deux gènes X^j et X^k ?
  1. euclidienne
  2. $1 - \text{cor}(X^j, X^k)$
  3. $\sqrt{1 - \text{cor}(X^j, X^k)^2}$

Exemple élémentaire des distances kilométriques



Définitions

- $\mathcal{D}$ $(n \times n)$: matrice de distance si elle est symétrique et si : $d_j^j = 0$ et $\forall (j,k), j \neq k, d_j^k \geq 0$
- $\mathcal{C}$ $(n \times n)$: matrice de similarité si elle est symétrique et si $\forall (j,k), c_j^k \leq c_j^j$
- Transformer des similarités en distances : $d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}$
- $\mathcal{D}$, matrice de distance est dite euclidienne si il existe une configuration de vecteurs $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ dans E de sorte que $d_j^{k^2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle$
- $\mathbf{A}$ matrice issue de $\mathcal{D}$ de terme général $d_j^k = -1/2 d_j^{k^2}$
- $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D}$: matrice de centrage

Définitions

- $\mathcal{D}$ $(n \times n)$: matrice de distance si elle est symétrique et si : $d_j^j = 0$ et $\forall (j,k), j \neq k, d_j^k \geq 0$
- $\mathcal{C}$ $(n \times n)$: matrice de similarité si elle est symétrique et si $\forall (j,k), c_j^k \leq c_j^j$
- Transformer des similarités en distances : $d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}$
- $\mathcal{D}$, matrice de distance est dite euclidienne si il existe une configuration de vecteurs $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ dans E de sorte que $d_j^{k^2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle$
- $\mathbf{A}$ matrice issue de $\mathcal{D}$ de terme général $d_j^k = -1/2 d_j^{k^2}$
- $\mathbf{H} = \mathbf{I} - \mathbf{11'D}$: matrice de centrage

Définitions

- $\mathcal{D}$ $(n \times n)$: matrice de distance si elle est symétrique et si : $d_j^j = 0$ et $\forall (j,k), j \neq k, d_j^k \geq 0$
- $\mathcal{C}$ $(n \times n)$: matrice de similarité si elle est symétrique et si $\forall (j,k), c_j^k \leq c_j^j$
- Transformer des similarités en distances : $d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}$
- $\mathcal{D}$, matrice de distance est dite euclidienne si il existe une configuration de vecteurs $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ dans E de sorte que $d_j^{k^2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle$
- $\mathbf{A}$ matrice issue de $\mathcal{D}$ de terme général $d_j^k = -1/2 d_j^{k^2}$
- $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D}$: matrice de centrage

Définitions

- $\mathcal{D}$ $(n \times n)$: matrice de distance si elle est symétrique et si : $d_j^j = 0$ et $\forall (j,k), j \neq k, d_j^k \geq 0$
- $\mathcal{C}$ $(n \times n)$: matrice de similarité si elle est symétrique et si $\forall (j,k), c_j^k \leq c_j^j$
- Transformer des similarités en distances : $d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}$
- $\mathcal{D}$, matrice de distance est dite euclidienne si il existe une configuration de vecteurs $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ dans E de sorte que $d_j^{k^2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle$
- $\mathbf{A}$ matrice issue de $\mathcal{D}$ de terme général $d_j^k = -1/2 d_j^{k^2}$
- $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D}$: matrice de centrage

Définitions

- $\mathcal{D}$ $(n \times n)$: matrice de distance si elle est symétrique et si : $d_j^j = 0$ et $\forall (j,k), j \neq k, d_j^k \geq 0$
- $\mathcal{C}$ $(n \times n)$: matrice de similarité si elle est symétrique et si $\forall (j,k), c_j^k \leq c_j^j$
- Transformer des similarités en distances : $d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}$
- $\mathcal{D}$, matrice de distance est dite euclidienne si il existe une configuration de vecteurs $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ dans E de sorte que $d_j^{k^2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle$
- $\mathbf{A}$ matrice issue de $\mathcal{D}$ de terme général $d_j^k = -1/2 d_j^{k^2}$
- $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D}$: matrice de centrage

Définitions

- $\mathcal{D}$ $(n \times n)$: matrice de distance si elle est symétrique et si : $d_j^j = 0$ et $\forall (j,k), j \neq k, d_j^k \geq 0$
- $\mathcal{C}$ $(n \times n)$: matrice de similarité si elle est symétrique et si $\forall (j,k), c_j^k \leq c_j^j$
- Transformer des similarités en distances : $d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}$
- $\mathcal{D}$, matrice de distance est dite euclidienne si il existe une configuration de vecteurs $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ dans E de sorte que $d_j^{k^2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle$
- $\mathbf{A}$ matrice issue de $\mathcal{D}$ de terme général $d_j^k = -1/2 d_j^{k^2}$
- $\mathbf{H} = \mathbf{I} - \mathbf{11'D}$: matrice de centrage

Proposition

Soit $\mathcal{D}$ une matrice de distance et $\mathbf{B}$ la matrice obtenue par double centrage de la matrice $\mathbf{A}$ issue de $\mathcal{D}$: $\mathbf{B} = \mathbf{HAH}$.

- $\mathcal{D}$ est une matrice euclidienne si et seulement si $\mathbf{B}$ est positive (toutes ses valeurs propres sont positives ou nulles)
- Si la matrice de similarité $\mathcal{C}$ est positive alors la matrice de distance $\mathcal{D}$ déduite est euclidienne

Distance entre variables quantitatives

- Objectif : Visualiser les relations d'un ensemble de variables
- X et Y observées sur n individus
- $\mathbf{x}$ et $\mathbf{y}$ vecteurs centrés de $\mathbb{R}^p, \mathbf{D}$:

$$\begin{aligned} \text{cov}(X,Y) &= \mathbf{x}'\mathbf{D}\mathbf{y} \\ \sigma_X &= \|\mathbf{x}\|_{\mathbf{D}} \\ \text{cor}(X,Y) &= \frac{\mathbf{x}'\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}} \|\mathbf{y}\|_{\mathbf{D}}}. \end{aligned}$$

- L'indice de similarité : $\text{cor}^2(X,Y)$ induit une distance euclidienne $d^2(X,Y) = 2(1 - \text{cor}^2(X,Y))$

Distance entre variables quantitatives

- Objectif : Visualiser les relations d'un ensemble de variables
- X et Y observées sur n individus
- $\mathbf{x}$ et $\mathbf{y}$ vecteurs centrés de $\mathbb{R}^p, \mathbf{D}$:

$$\begin{aligned} \text{cov}(X,Y) &= \mathbf{x}'\mathbf{D}\mathbf{y} \\ \sigma_X &= \|\mathbf{x}\|_{\mathbf{D}} \\ \text{cor}(X,Y) &= \frac{\mathbf{x}'\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}} \|\mathbf{y}\|_{\mathbf{D}}}. \end{aligned}$$

- L'indice de similarité : $\text{cor}^2(X,Y)$ induit une distance euclidienne $d^2(X,Y) = 2(1 - \text{cor}^2(X,Y))$

Distance entre variables quantitatives

- Objectif : Visualiser les relations d'un ensemble de variables
- X et Y observées sur n individus
- $\mathbf{x}$ et $\mathbf{y}$ vecteurs centrés de $\mathbb{R}^p$, $\mathbf{D}$:

$$\begin{aligned} \text{cov}(X,Y) &= \mathbf{x}'\mathbf{D}\mathbf{y} \\ \sigma_X &= \|\mathbf{x}\|_{\mathbf{D}} \\ \text{cor}(X,Y) &= \frac{\mathbf{x}'\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}}\,\|\mathbf{y}\|_{\mathbf{D}}}. \end{aligned}$$

- L'indice de similarité : $\text{cor}^2(X,Y)$ induit une distance euclidienne $d^2(X,Y) = 2(1-\text{cor}^2(X,Y))$

Distance entre variables quantitatives

- Objectif : Visualiser les relations d'un ensemble de variables
- X et Y observées sur n individus
- $\mathbf{x}$ et $\mathbf{y}$ vecteurs centrés de $\mathbb{R}^p$, $\mathbf{D}$:

$$\begin{aligned} \text{cov}(X,Y) &= \mathbf{x}'\mathbf{D}\mathbf{y} \\ \sigma_X &= \|\mathbf{x}\|_{\mathbf{D}} \\ \text{cor}(X,Y) &= \frac{\mathbf{x}'\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}}\,\|\mathbf{y}\|_{\mathbf{D}}}. \end{aligned}$$

- L'indice de similarité : $\text{cor}^2(X,Y)$ induit une distance euclidienne $d^2(X,Y) = 2(1-\text{cor}^2(X,Y))$

Proposition

La distance entre variables quantitatives $d^2(X, Y)$ est encore le carré de la distance $\|\mathbf{P_x} - \mathbf{P_y}\|_{\mathbf{D}}$ entre les projecteurs $\mathbf{D}$-orthogonaux sur les directions engendrées par les vecteurs $\mathbf{x}$ et $\mathbf{y}$

Distance entre variables qualitatives

- X et Y qualitatives à r et c modalités,
- Indices de similarité : "p-value" du χ^2, le V de Cramer, le Φ^2 de Pearson, le T de Tschuprow...
- T induit une métrique euclidienne

Proposition

$\mathbf{X}$ et $\mathbf{Y}$ matrices des indicatrices, $\mathbf{P_X}$ et $\mathbf{P_Y}$ projecteurs $\mathbf{D}$-orthogonaux :

$$\|\mathbf{P_X} - \mathbf{P_Y}\|^2_{\mathbf{D}} = 2(1 - T^2(X, Y))$$

Distance entre variables qualitatives

- X et Y qualitatives à r et c modalités,
- Indices de similarité : "p-value" du χ^2, le V de Cramer, le Φ^2 de Pearson, le T de Tschuprow...
- T induit une métrique euclidienne

Proposition

$\mathbf{X}$ et $\mathbf{Y}$ matrices des indicatrices, $\mathbf{P_X}$ et $\mathbf{P_Y}$ projecteurs $\mathbf{D}$-orthogonaux :

$$\|\mathbf{P_X} - \mathbf{P_Y}\|^2_{\mathbf{D}} = 2(1 - T^2(X, Y))$$

Distance entre une quantitative et une qualitative

- X variable quantitative associée à $\mathbf{P_x}$
- Y variable qualitative associée à $\mathbf{P_Y}$
- En pratique : utilisation très délicate

proposition

Dans le cas d'une variable quantitative X et d'une variable qualitative Y,

$$\|\mathbf{P_x} - \mathbf{P_Y}\|^2_{\mathbf{D}} = 2(1 - R^2_c(X, Y))$$

R_c désigne le rapport de corrélation : indice de similarité entre variables de types différents.

Distance entre une quantitative et une qualitative

- X variable quantitative associée à $\mathbf{P_x}$
- Y variable qualitative associée à $\mathbf{P_Y}$
- En pratique : utilisation très délicate

proposition

Dans le cas d'une variable quantitative X et d'une variable qualitative Y,

$$\|\mathbf{P_x} - \mathbf{P_Y}\|^2_{\mathbf{D}} = 2(1 - R^2_c(X, Y))$$

R_c désigne le rapport de corrélation : indice de similarité entre variables de types différents.

Objectifs du MDS

- Si $\mathcal{D}$ euclidienne, configuration de points admettant $\mathcal{D}$ comme matrice de distances
- Sinon, configuration de points fournissant la meilleure approximation à un rang q fixé au sens d'une norme sur les matrices
- Infinité de solutions : distance invariante par transformation affine
- Solution définie à une rotation et une translation près

Objectifs du MDS

- Si $\mathcal{D}$ euclidienne, configuration de points admettant $\mathcal{D}$ comme matrice de distances
- Sinon, configuration de points fournissant la meilleure approximation à un rang q fixé au sens d'une norme sur les matrices
- Infinité de solutions : distance invariante par transformation affine
- Solution définie à une rotation et une translation près

Objectifs du MDS

- Si $\mathcal{D}$ euclidienne, configuration de points admettant $\mathcal{D}$ comme matrice de distances
- Sinon, configuration de points fournissant la meilleure approximation à un rang q fixé au sens d'une norme sur les matrices
- Infinité de solutions : distance invariante par transformation affine
- Solution définie à une rotation et une translation près

Objectifs du MDS

- Si $\mathcal{D}$ euclidienne, configuration de points admettant $\mathcal{D}$ comme matrice de distances
- Sinon, configuration de points fournissant la meilleure approximation à un rang q fixé au sens d'une norme sur les matrices
- Infinité de solutions : distance invariante par transformation affine
- Solution définie à une rotation et une translation près

Theorème

$\mathcal{D}$ matrice de distance et
$\mathbf{B} = \mathbf{HAH}$ la matrice double centrée.

- Si $\mathcal{D}$ matrice de distance euclidienne de $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ alors $\mathbf{B}$ de terme général $b_j^k = (\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_i - \overline{\mathbf{x}})$ qui se met sous la forme

$$\mathbf{B} = (\mathbf{HX})(\mathbf{HX})'$$

  est positive et appelée matrice des produits scalaires

- Réciproquement, si $\mathbf{B}$ est positive de rang p, une configuration de vecteurs admettant $\mathbf{B}$ pour matrice des produits scalaires est obtenue en considérant sa décomposition spectrale $\mathbf{B} = \mathbf{U}\Delta\mathbf{U}'$. Ce sont les lignes de la matrice centrée $\mathbf{X} = \mathbf{U}\Delta^{1/2}$.

Theorème

$\mathcal{D}$ matrice de distance et
$\mathbf{B} = \mathbf{HAH}$ la matrice double centrée.

- Si $\mathcal{D}$ matrice de distance euclidienne de $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ alors $\mathbf{B}$ de terme général $b_j^k = (\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_i - \overline{\mathbf{x}})$ qui se met sous la forme

$$\mathbf{B} = (\mathbf{HX})(\mathbf{HX})'$$

  est positive et appelée matrice des produits scalaires
- Réciproquement, si $\mathbf{B}$ est positive de rang p, une configuration de vecteurs admettant $\mathbf{B}$ pour matrice des produits scalaires est obtenue en considérant sa décomposition spectrale $\mathbf{B} = \mathbf{U\Delta U}'$. Ce sont les lignes de la matrice centrée $\mathbf{X} = \mathbf{U\Delta}^{1/2}$.

Solution du MDS : cas euclidien

Si $\mathcal{D}$ euclidienne de rang q :

1. construction de $\mathbf{A}$ de terme général $-1/2d_j^{k^2}$
2. calcul de la matrice des produits scalaires par double centrage

$$\mathbf{B} = \mathbf{HAH}$$

3. diagonalisation de $\mathbf{B} = \mathbf{U}\Delta\mathbf{U}'$
4. les coordonnées d'une configuration, appelées coordonnées principales sont les lignes de la matrice

$$\mathbf{X} = \mathbf{U}\Delta^{1/2}$$

Solution du MDS : cas euclidien

Si $\mathcal{D}$ euclidienne de rang q :

1. construction de $\mathbf{A}$ de terme général $-1/2d_j^{k^2}$
2. calcul de la matrice des produits scalaires par double centrage

$$\mathbf{B} = \mathbf{HAH}$$

3. diagonalisation de $\mathbf{B} = \mathbf{U}\Delta\mathbf{U}'$
4. les coordonnées d'une configuration, appelées coordonnées principales sont les lignes de la matrice

$$\mathbf{X} = \mathbf{U}\Delta^{1/2}$$

Solution du MDS : cas euclidien

Si $\mathcal{D}$ euclidienne de rang q :

1. construction de $\mathbf{A}$ de terme général $-1/2d_j^{k^2}$
2. calcul de la matrice des produits scalaires par double centrage
$$\mathbf{B} = \mathbf{HAH}$$
3. diagonalisation de $\mathbf{B} = \mathbf{U\Delta U}'$
4. les coordonnées d'une configuration, appelées coordonnées principales sont les lignes de la matrice
$$\mathbf{X} = \mathbf{U\Delta}^{1/2}$$

Solution du MDS : cas euclidien

Si $\mathcal{D}$ euclidienne de rang q :

1. construction de $\mathbf{A}$ de terme général $-1/2d_j^{k^2}$
2. calcul de la matrice des produits scalaires par double centrage
$$\mathbf{B} = \mathbf{HAH}$$
3. diagonalisation de $\mathbf{B} = \mathbf{U\Delta U}'$
4. les coordonnées d'une configuration, appelées coordonnées principales sont les lignes de la matrice
$$\mathbf{X} = \mathbf{U\Delta}^{1/2}$$

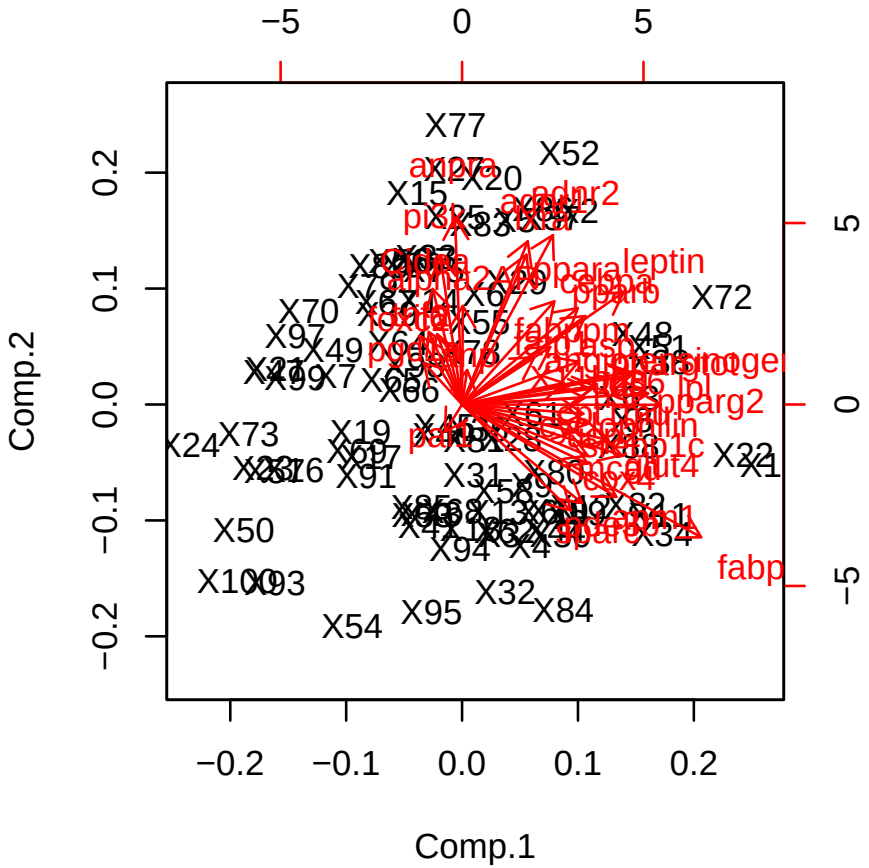
Proposition : cas euclidien

Soit $\mathbf{X}$ la matrice des données habituelles en ACP. L'ACP de $(\mathbf{X}, \mathbf{M}, 1/n\mathbf{I})$ fournit les mêmes représentations graphiques que le positionnement calculé à partir de la matrice de distances de terme général $\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}}$. Si $\mathbf{C}$ désigne la matrice des composantes principales, alors les coordonnées principales sont $\sqrt{n}\mathbf{C}$

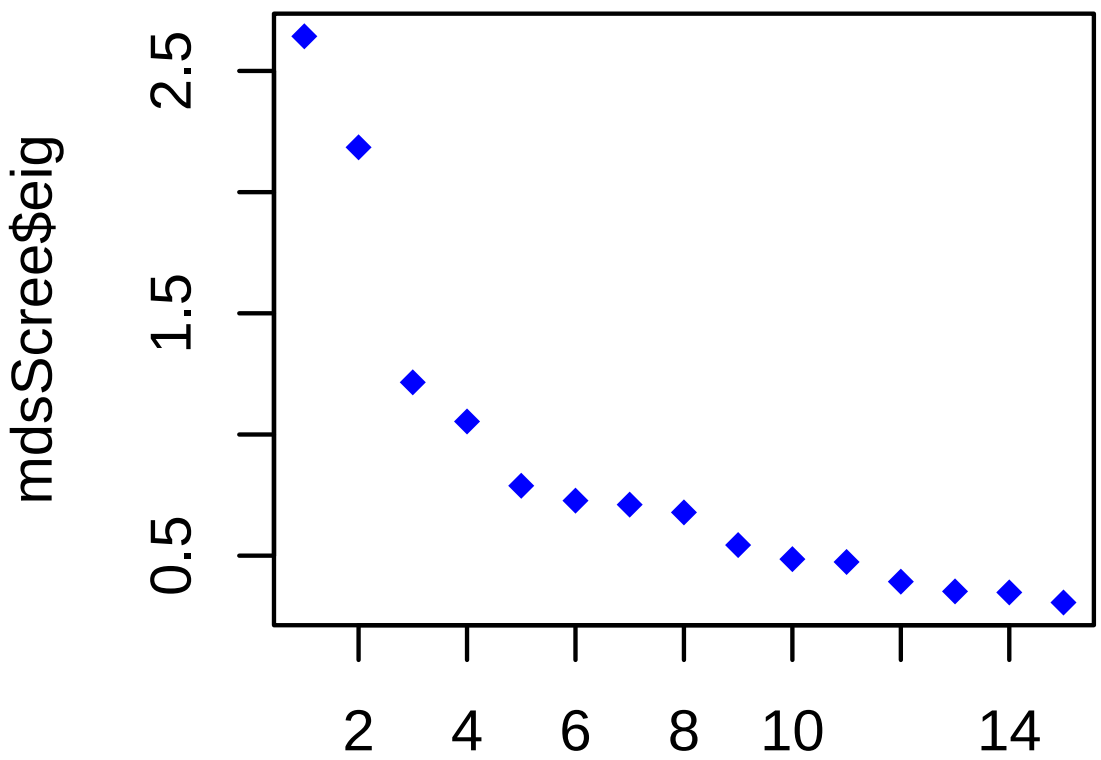
Théorème dans le cas non euclidien

Si $\mathcal{D}$ est une matrice de distance, $\mathbf{B}$ la matrice de produit scalaire associée, alors, pour une dimension q fixée, la configuration issue du MDS a une matrice de distance $\widehat{\mathcal{D}}$ qui rend $\sum_{j,k=1}^{n}(\{d_j^k\}^2 - \widehat{d_j^k}^2)$ minimum et, c'est équivalent, une matrice de produit scalaire $\widehat{\mathbf{B}}$ qui minimise $\left\|\mathbf{B} - \widehat{\mathbf{B}}\right\|^2$.

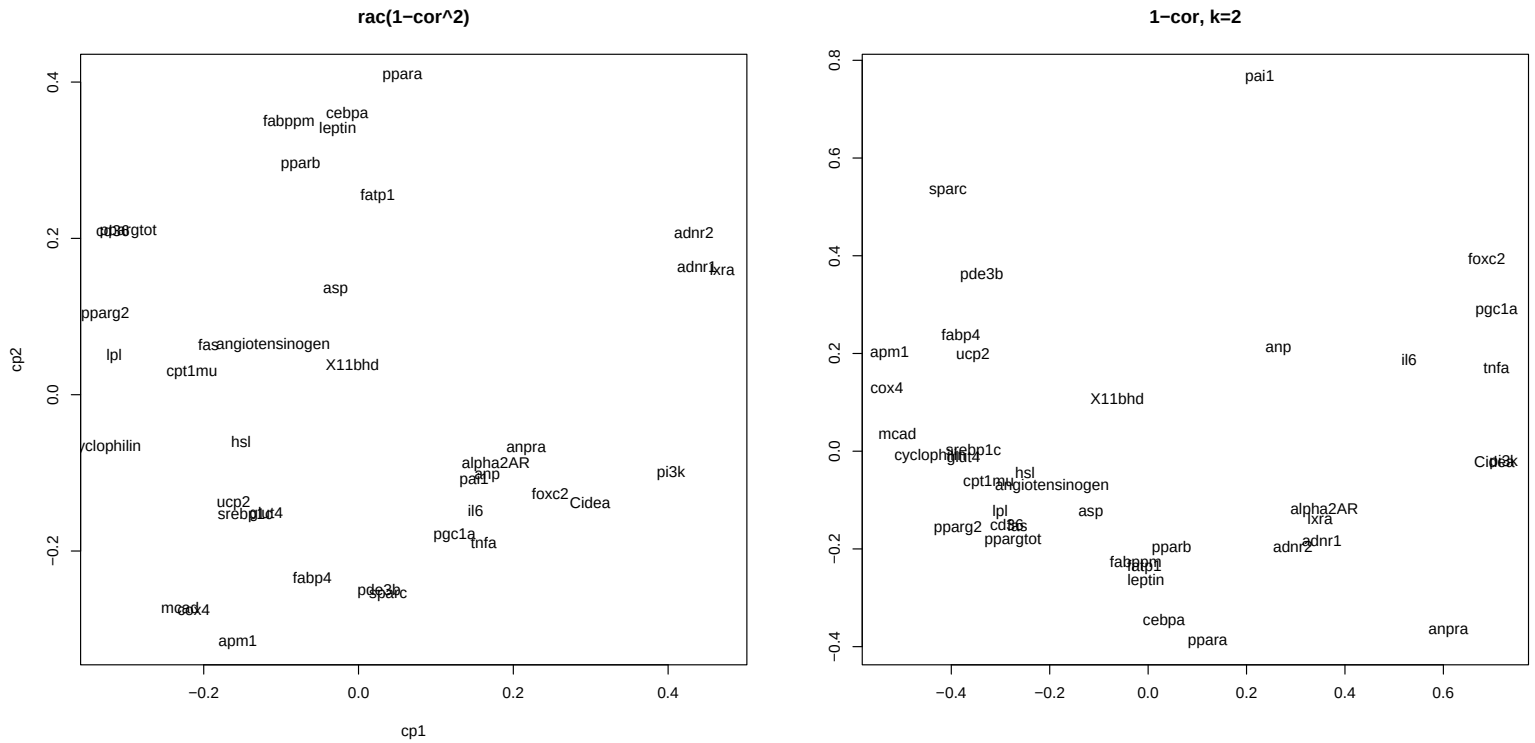
Obésité : ACP avec "effet taille"



Obésité : Décroissance des valeurs propres du MDS

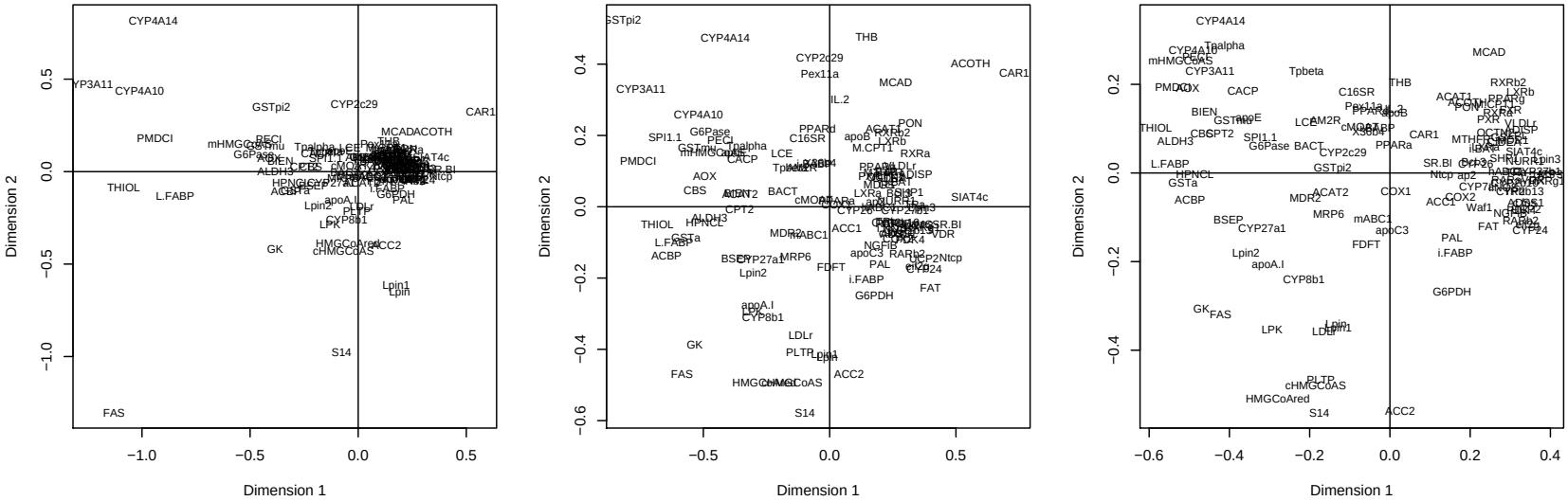


Obésité : MDS des gènes



Au sens de la corrélation au carré ou de leur corrélation

Souris : MDS des gènes



Au sens de la distance euclidienne, de leurs corrélations, de leurs corrélations carrées

Statistique des données d'expression
Chapitre 6 : Classification non supervisée

Alain Baccini & Philippe Besse

Laboratoire de Statistique et Probabilités
Université de Toulouse

Institut de Mathématiques
math.univ-toulouse.fr/biostat

Objectifs

- Matrice $\mathbf{X}(n, p)$ des observations de p variables quantitatives et/ou qualitatives sur n individus
- Tableau de distances (ou dissemblance) des individus
- Recherche d'une typologie, segmentation ou partition des individus en classes par optimisation d'un critère
- Discrimination vs. classif. *classification vs. clustering*
- La complexité impose l'exécution d'un algorithme itératif

Choix de l'utilisateur

- Mesure d'éloignement ou distance entre individus
- Critère : trace de la matrice de variance intra
- Méthode : classif. hiérarchique ou par réallocation dynamique
- Nombre de classes

Indice de ressemblance ou similarité

- $\Omega = \{i = 1, \ldots, n\}$ ensemble des individus
- s définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} s(i,j) &= s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} \\ s(i,i) &= S > 0, \forall i \in \Omega \ : \text{ ressemblance de } i \text{ avec lui-même} \\ s(i,j) &\leq S, \forall (i,j) \in \Omega \times \Omega : \text{ ressemblance majorée par } S \end{aligned}$$

- Indice de ressemblance normé s^* est défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$

Indice de ressemblance ou similarité

- $\Omega = \{i = 1, \ldots, n\}$ ensemble des individus
- s définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} s(i,j) &= s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} \\ s(i,i) &= S > 0, \forall i \in \Omega \ : \ \text{ressemblance de } i \text{ avec lui-même} \\ s(i,j) &\leq S, \forall (i,j) \in \Omega \times \Omega : \ \text{ressemblance majorée par } S \end{aligned}$$

- Indice de ressemblance normé s^* est défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$

Indice de ressemblance ou similarité

- $\Omega = \{i = 1, \ldots, n\}$ ensemble des individus
- s définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} s(i,j) &= s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} \\ s(i,i) &= S > 0, \forall i \in \Omega \ : \ \text{ressemblance de } i \text{ avec lui-même} \\ s(i,j) &\leq S, \forall (i,j) \in \Omega \times \Omega : \ \text{ressemblance majorée par } S \end{aligned}$$

- Indice de ressemblance normé s^* est défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$

Indice de dissemblance ou dissimilarité

- d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} d(i,j) &= d(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} ; \\ d(i,i) &= 0, \forall i \in \Omega \ : \end{aligned}$$

- Si s est une similarité, d est une dissimilarité

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

- Un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$

Indice de dissemblance ou dissimilarité

- d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} d(i,j) &= d(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} ; \\ d(i,i) &= 0, \forall i \in \Omega \ : \end{aligned}$$

- Si s est une similarité, d est une dissimilarité

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

- Un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$

Indice de dissemblance ou dissimilarité

- d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ avec :

$$\begin{aligned} d(i,j) &= d(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} ; \\ d(i,i) &= 0, \forall i \in \Omega \ : \end{aligned}$$

- Si s est une similarité, d est une dissimilarité

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

- Un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$

Indice de distance

- $d(i,j) = 0 \Rightarrow i = j$
- Pour éviter des incohérences entre dissemblances

Distance

- $d(i,j) = d(j,i), \forall (i,j) \in \Omega \times \Omega$;
- $d(i,i) = 0 \Leftrightarrow i = j$;
- $d(i,j) \leq d(i,k) + d(j,k); \forall (i,j,k) \in \Omega^3$.

Si Ω est fini, la distance peut être normée

Distance euclidienne

- Si Ω est muni d'un produit scalaire :
  $d(i,j) = [\langle i-j, i-j \rangle]^{1/2} = \|i-j\|$

Indice de distance

- $d(i,j) = 0 \Rightarrow i = j$
- Pour éviter des incohérences entre dissemblances

Distance

- $d(i,j) = d(j,i), \forall (i,j) \in \Omega \times \Omega$;
- $d(i,i) = 0 \Leftrightarrow i = j$;
- $d(i,j) \leq d(i,k) + d(j,k); \forall (i,j,k) \in \Omega^3$.

Si Ω est fini, la distance peut être normée

Distance euclidienne

- Si Ω est muni d'un produit scalaire :
  $d(i,j) = [\langle i-j, i-j \rangle]^{1/2} = \|i-j\|$

Indice de distance

- $d(i,j) = 0 \Rightarrow i = j$
- Pour éviter des incohérences entre dissemblances

Distance

- $d(i,j) = d(j,i), \forall (i,j) \in \Omega \times \Omega$;
- $d(i,i) = 0 \Leftrightarrow i = j$;
- $d(i,j) \leq d(i,k) + d(j,k); \forall (i,j,k) \in \Omega^3$.

Si Ω est fini, la distance peut être normée

Distance euclidienne

- Si Ω est muni d'un produit scalaire :
  $d(i,j) = [\langle i-j, i-j \rangle]^{1/2} = \|i-j\|$

Deux cas possibles

1. Soit un tableau de mesures $n \times p$
   - p variables toutes quantitatives,
   - matrice de produit scalaire sur l'espace $\mathbb{R}^P$; $\mathbf{M} = \mathbf{I}_p$
   - réduire les variables de variances hétérogènes :
2. Soit tableau $n \times n$ de dissemblances ou distances entre individus ; Attention si n grand

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

CAH : Objectif

- Agglomération itérative de 2 éléments de la partition
- Construction d'un dendrogramme ou arbre binaire
- Problème : définir $d(A, B)$ A et B deux groupes ou éléments d'une partition à partir de
  - w_A et w_B leurs pondérations
  - $d_{i,j}$ la dissemblance ou distance entre deux individus

Cas d'une dissemblance

- $d(A,B) = \min_{i \in A, j \in B}(d_{ij})$ saut minimum, single linkage
- $d(A,B) = \sup_{i \in A, j \in B}(d_{ij})$ saut maximum ou diamètre, complete linkage
- $d(A,B) = \frac{1}{\text{card}(A)\text{card}(B)} \sum_{i \in A, j \in B} d_{ij}$ saut moyen, group average linkage

Cas d'une distance euclidienne

g_A et g_B : barycentres des classes

$$d(A,B) = d(g_A,g_B) \quad \text{(distance des barycentres, centroïd)}$$

$$d(A,B) = \frac{w_A w_B}{w_A + w_B} d(g_A,g_B) \quad \text{(saut de Ward)}$$

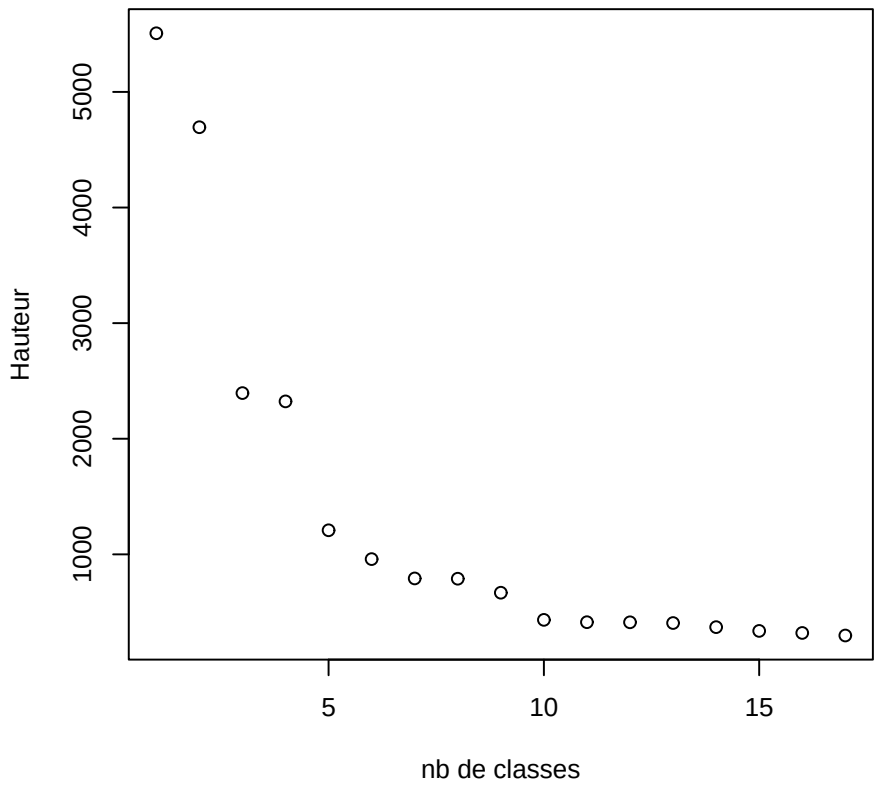
Saut de Ward et maximisation de la variance inter

Algorithme de classification ascendante hiérarchique

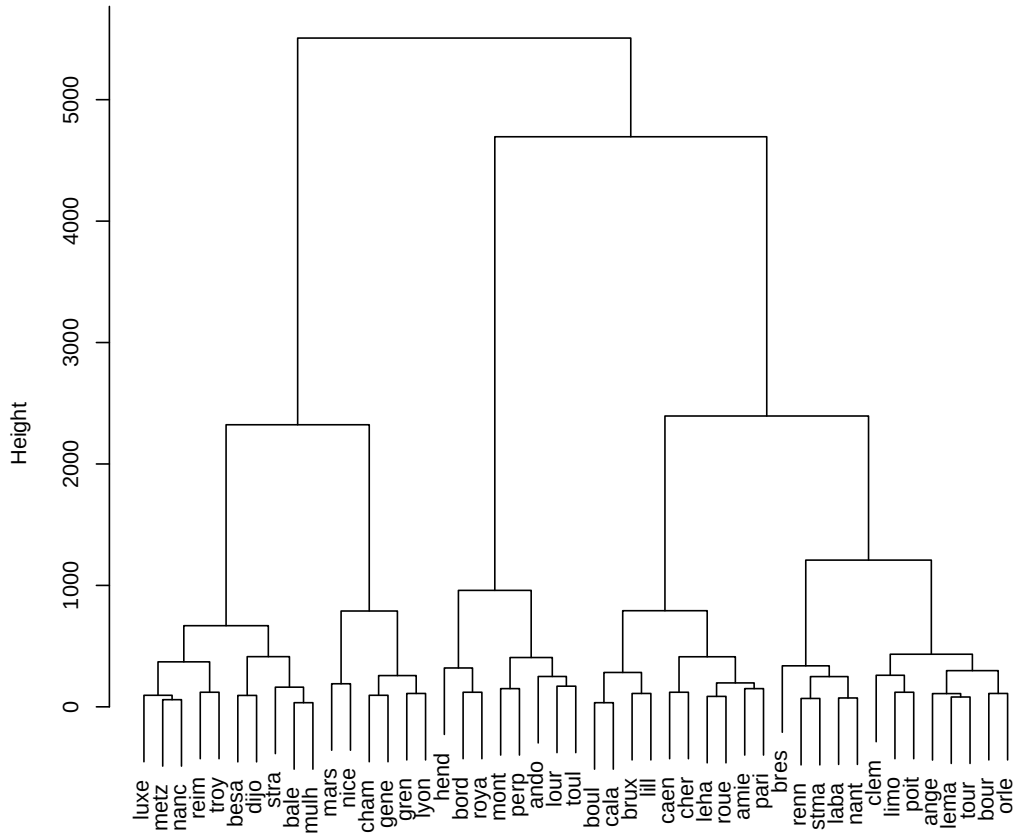
- Initialisation : singletons, calcul des distances
- Itérer jusqu'à agrégation en une seule classe :
  1. regrouper les deux classes les plus proches au sens de la "distance" entre groupes choisie
  2. mise à jour du tableau de distance en remplaçant les deux classes par la nouvelle et en calculant sa "distance" avec les autres classes

Nombre de classes : Rupture dans la décroissance du R^2 partiel (Ward)

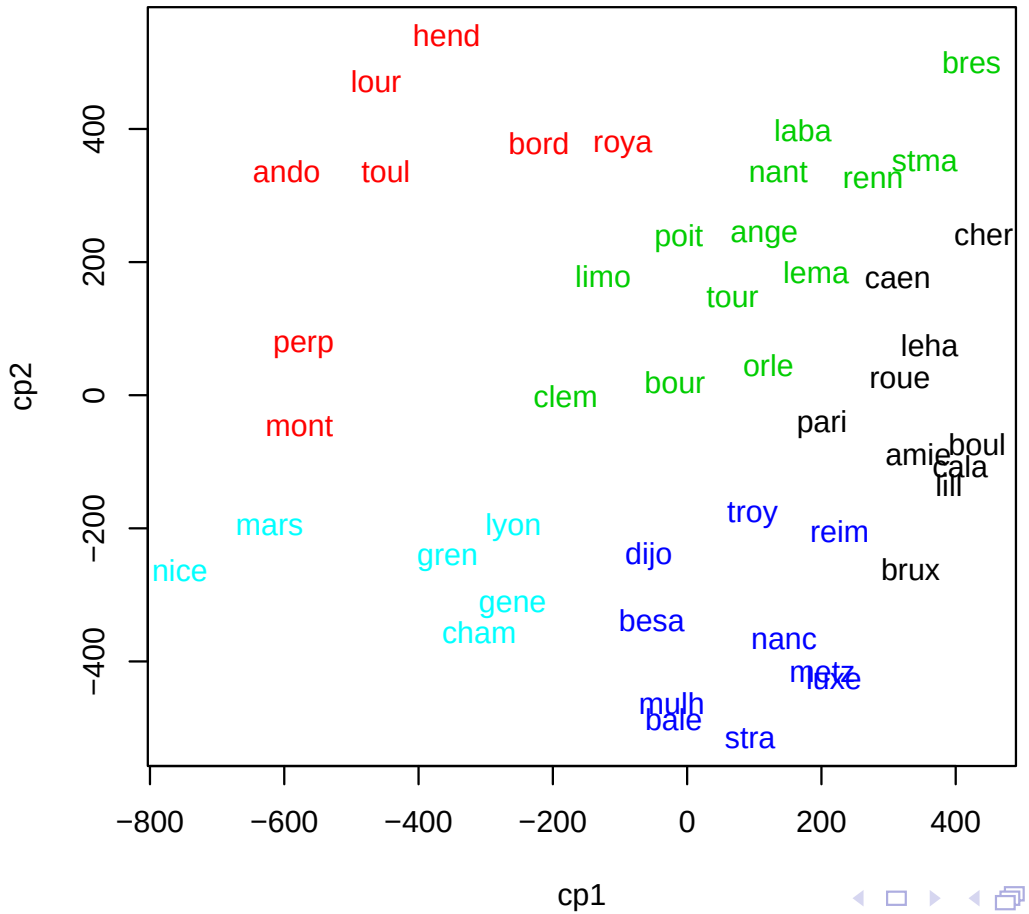
Villes : Décroissance de la variance inter classes



Villes : Exemple d'un dendogramme



Villes : Représentation des classes avec MDS



Principe des centres mobiles

- réallocation dynamique des individus à des classes
- Le nombre de classes k est fixé *a priori*

Algorithme de Forgy

- Initialisation Tirer au hasard ou sélectionner, k points dans l'espace des individus, en général k individus de l'ensemble, appelés centres ou noyaux
- Itérer jusqu'à stagnation du critère de variance inter-classe
  1. Allouer chaque individu à l'un des noyaux, c'est-à-dire à une classe
  2. Calculer le centre de gravité de chaque classe, il devient le nouveau noyau

Attention : optimum local

Principe des centres mobiles

- réallocation dynamique des individus à des classes
- Le nombre de classes k est fixé *a priori*

Algorithme de Forgy

- Initialisation Tirer au hasard ou sélectionner, k points dans l'espace des individus, en général k individus de l'ensemble, appelés centres ou noyaux
- Itérer jusqu'à stagnation du critère de variance inter-classe
  1. Allouer chaque individu à l'un des noyaux, c'est-à-dire à une classe
  2. Calculer le centre de gravité de chaque classe, il devient le nouveau noyau

Attention : optimum local

Variantes

- Algorithme $k$means : les noyaux des classes, ici les barycentres, sont recalculés à chaque allocation d'un point à une classe ; algorithme plus efficace
- Nuées dynamiques : Un centre de classes est un noyau d'éléments représentatifs d'une classe
- Partitionning Around Medoïds

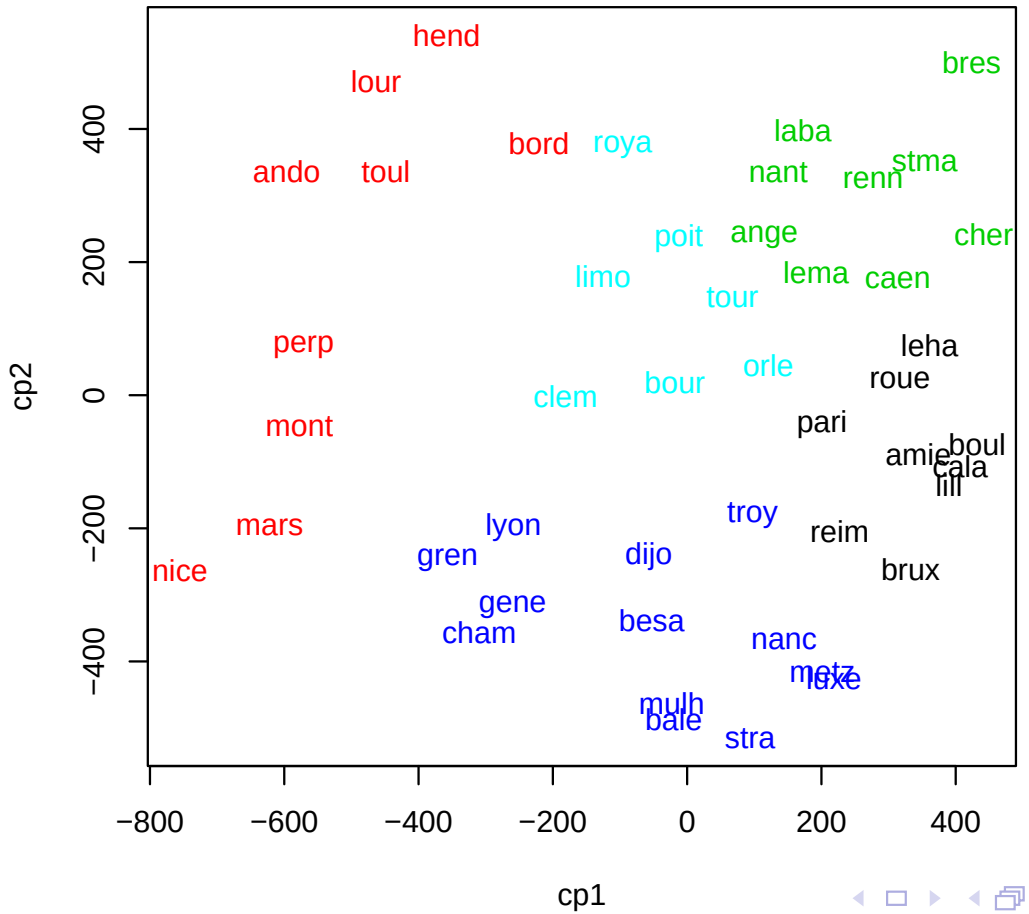
Variantes

- Algorithme $k$means : les noyaux des classes, ici les barycentres, sont recalculés à chaque allocation d'un point à une classe ; algorithme plus efficace
- Nuées dynamiques : Un centre de classes est un noyau d'éléments représentatifs d'une classe
- Partitionning Around Medoïds

Variantes

- Algorithme $k$means : les noyaux des classes, ici les barycentres, sont recalculés à chaque allocation d'un point à une classe ; algorithme plus efficace
- Nuées dynamiques : Un centre de classes est un noyau d'éléments représentatifs d'une classe
- Partitionning Around Medoïds

Villes : Classes d'un PAM avec MDS



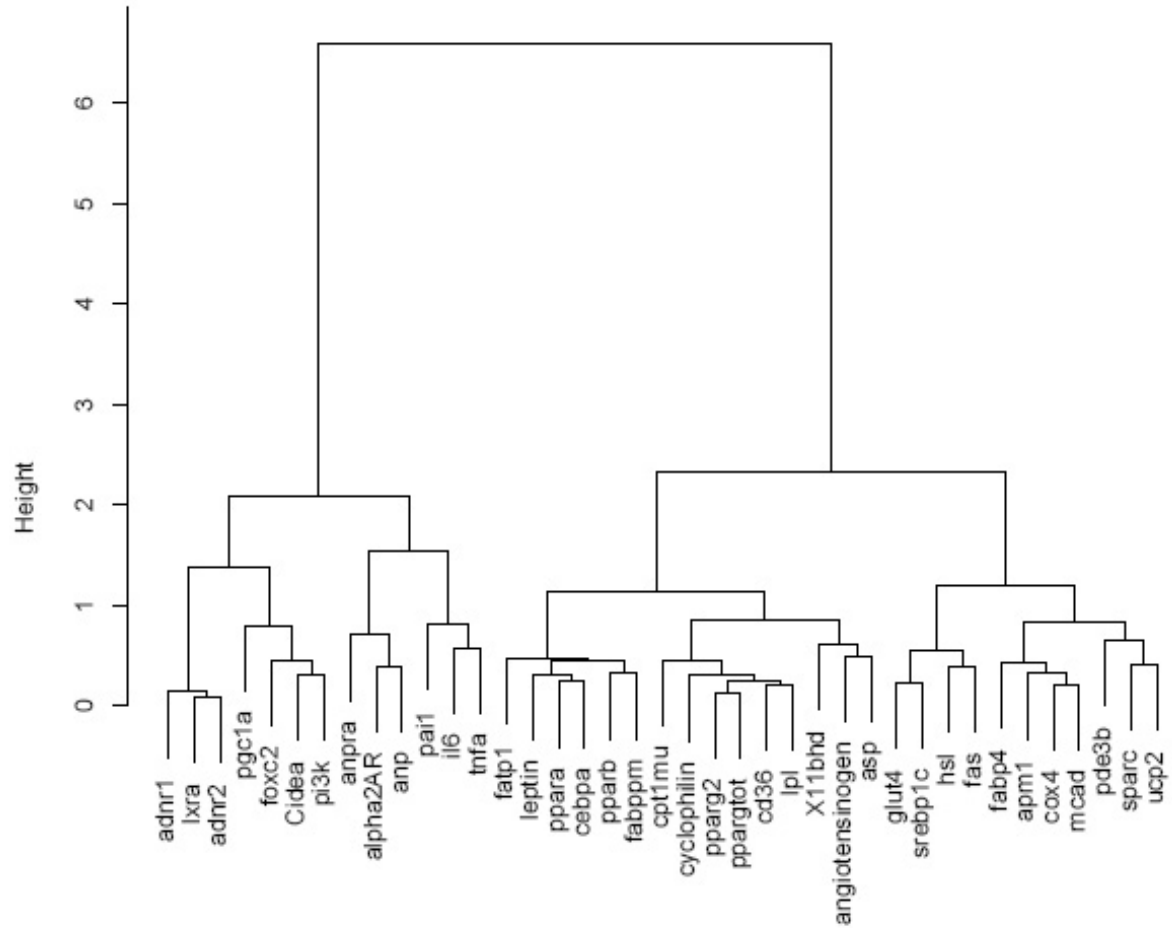
Classification de grands tableaux

1. Réallocation dynamique avec un grand nombre de classes $(n/10)$
2. Classification hiérarchique des barycentres
3. détermination d'un nombre "optimal" k de classes
4. Réallocation dynamique de l'ensemble avec k classes en choisissant pour noyaux les barycentres des classes de l'étape précédente

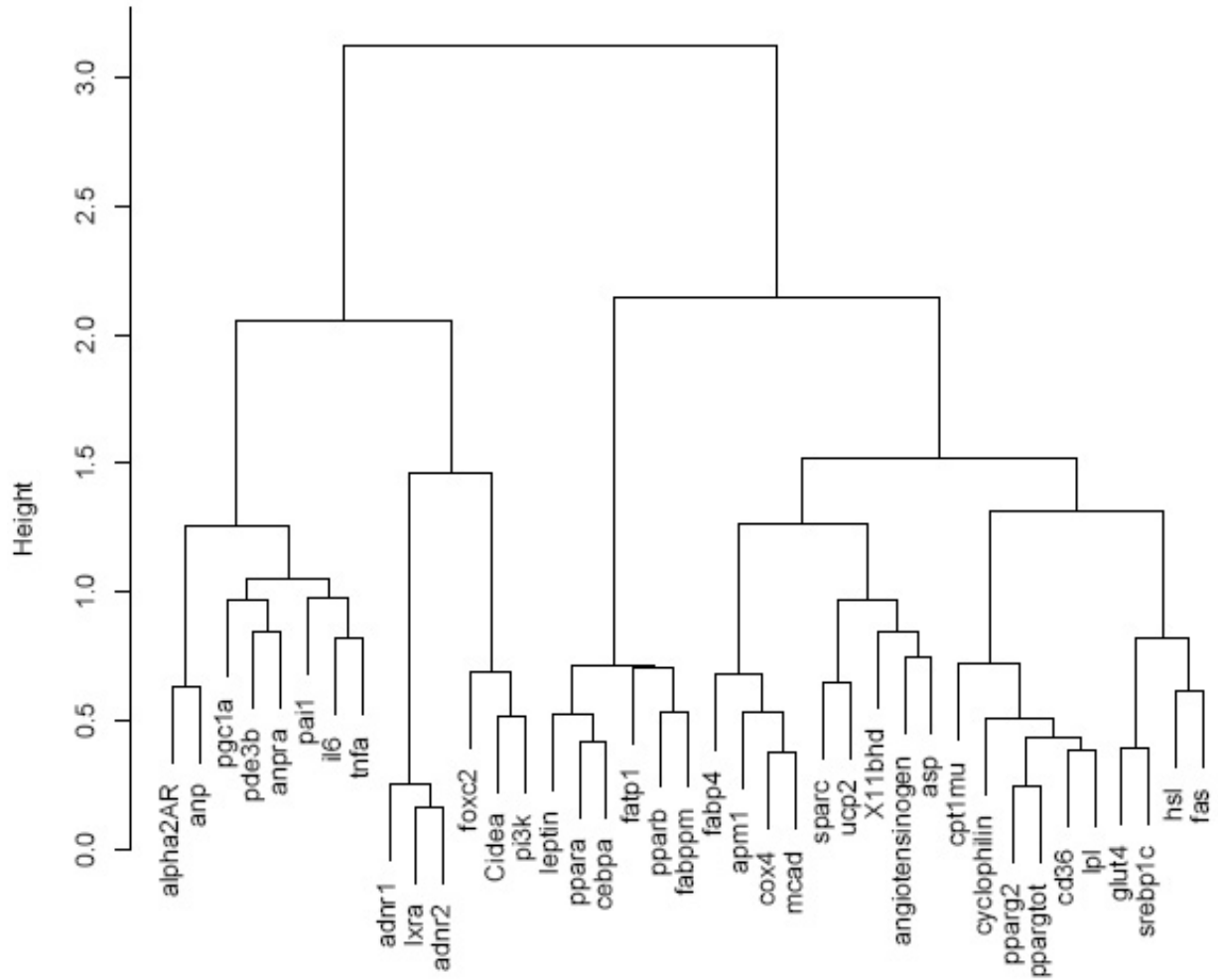
Conclusion

- Résultat : une variable qualitative T dont les modalités précisent la classe retenue pour chaque individu
- Problèmes : interprétation des classes

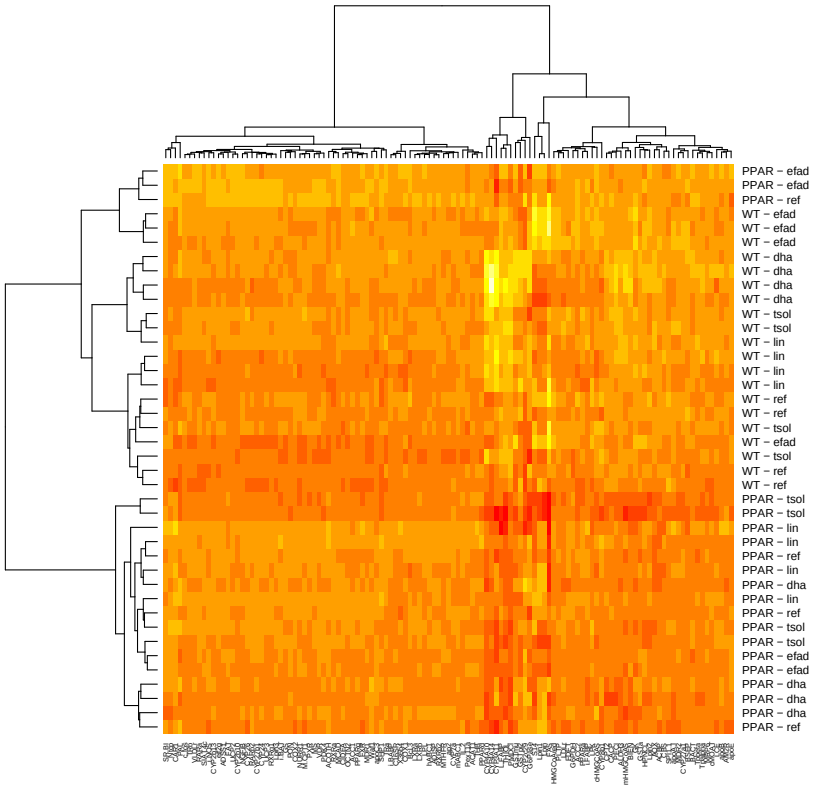
Obésité : CAH des gènes (Ward) et corrélation



Obésité : CAH des gènes (Ward) et corrélations carré



Souris : double CAH (Ward) et distance euclidienne



Souris : Classes de la CAH dans le MDS