Reconnaissance des formes

Analyse en Composantes Principales

- Chapitre 2 -

Analyse en composantes principales

Objectifs

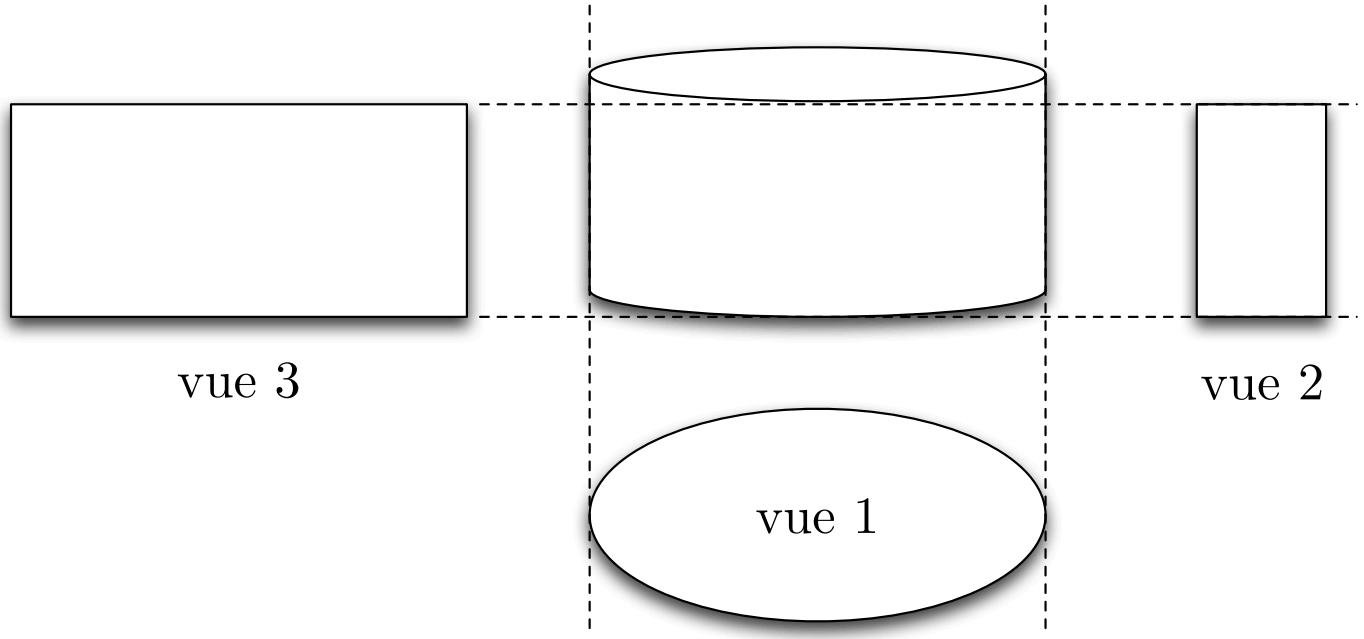
Contexte : Chaque individu $\boldsymbol{x}_i$ du tableau $\boldsymbol{X}$ est considéré comme un point d'un espace vectoriel $\mathcal{E}$ de dimension p.
L'ensemble des individus constitue un nuage de points dans $\mathcal{E}$.
On note $\boldsymbol{m}$ son point moyen, appelé aussi *centre de gravité*.

Objectif : On cherche à réduire le nombre p de variables en préservant au maximum la structure du nuage, afin de le représenter le plus fidèlement possible.

Principe : L'ACP vise à projeter les données dans un sous-espace approprié, de dimension plus faible, préservant la topologie du nuage.

Analyse en composantes principales

Exemple



Analyse en composantes principales

Exemple plus réaliste

cidre	odeur	sucre	acide	amer	astringence	suffocante	piquante	alcool	parfum	fruité
1	2.14	1.86	3.29	2.29	2	0.14	2.29	1.86	1.29	1.29
2	2.43	0.79	2.71	2.57	2	0.43	2.57	2.86	0.43	0.14
3	2.71	3.14	2.57	2.57	1.43	0.14	2.14	0.86	2.29	1.71
4	3	3.71	2.14	2.07	1.57	0	1.29	1	3.14	3.14
5	3.43	1.29	2.86	3.14	2.17	1	1.86	2.86	1.14	0.29
6	3.14	0.86	2.86	3.79	2.57	0.14	1.71	3.29	0.14	0
7	3.14	1.14	2.86	2.86	2	0.43	1.71	1.86	0.14	0
8	2.43	3.71	3.21	1.57	1.71	0	1	0.57	2.57	2.86
9	5.1	2.86	2.86	3.07	1.79	1.71	0.43	1.43	0.57	2.71
10	3.07	3.14	2.57	3	2	0	0.43	1.29	2.57	3.07

Notes obtenues pour 10 cidres selon 10 critères lors d'un concours agricole.

Espace des individus

Métrique

Métrique : L'ACP repose sur les distances entre individus dans $\mathcal{E}$. Le choix de la métrique a une influence fondamentale sur le résultat de l'analyse.

Définition : Soit $\boldsymbol{M}$ une matrice définie positive de dimension p. La fonction suivante $d_M : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ définit une métrique

$$d_M^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_M^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)$$

Cette distance est appelée distance de Mahalanobis lorsque $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$, où $\boldsymbol{\Sigma}$ est la matrice de variance-covariance des données.

Produit scalaire : La métrique définie ci-dessus dérive du produit scalaire

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_M = \boldsymbol{x}_i^\top \boldsymbol{M}\, \boldsymbol{x}_j$$

On dit que $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ sont M-orthogonaux si $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_M = 0$.

Analyse des données

Métrique

Utiliser la métrique $\boldsymbol{M} = \boldsymbol{T}^\top \boldsymbol{T}$ sur le tableau de données $\boldsymbol{X}$ est équivalent à travailler avec la métrique euclidienne sur le tableau transformé $\boldsymbol{X}\boldsymbol{T}^\top$.

Tableau transformé : Lorsqu'on travaille sur le tableau transformé comme ci-dessus, il est alors possible d'utiliser la norme euclidienne. En effet,

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_M = \boldsymbol{x}_i^\top (\boldsymbol{T}^\top \boldsymbol{T})\, \boldsymbol{x}_j = (\boldsymbol{T}\boldsymbol{x}_i)^\top (\boldsymbol{T}\boldsymbol{x}_j) = \langle \boldsymbol{T}\boldsymbol{x}_i, \boldsymbol{T}\boldsymbol{x}_j \rangle_I$$

Réciproque : Pour toute matrice définie positive $\boldsymbol{M}$, il existe une matrice définie positive $\boldsymbol{T}$ telle que $\boldsymbol{M} = \boldsymbol{T}^\top \boldsymbol{T}$. On notera improprement $\boldsymbol{T} = \boldsymbol{M}^{\frac{1}{2}}$.

Appliquer préalablement la transformation $\boldsymbol{X}\boldsymbol{T}^\top \to \boldsymbol{X}$ permet de simplifier les traitements.

Analyse des données

Métriques particulières

Métrique euclidienne : Elle est obtenue pour $\boldsymbol{M} = \boldsymbol{I}$.

L'une des difficultés rencontrées avec la métrique euclidienne est qu'elle privilégie les variables les plus dispersées et dépend donc de leur unité de mesure.

Métrique réduite : Elle consiste à prendre $\boldsymbol{M} = \boldsymbol{D}_{1/\sigma^2}$, où $\boldsymbol{D}_{1/\sigma^2}$ est la matrice diagonale de termes diagonaux les inverses $\frac{1}{\sigma_i^2}$ des variances des variables.

$$\boldsymbol{D}_{1/\sigma^2} = \begin{pmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_p^2} \end{pmatrix}$$

Cette métrique permet de s'affranchir de l'unité de mesure des variables, et de donner la même importance à chaque variable dans le calcul de la distance.

Analyse des données

Tableau de données centrées réduites

Utiliser la métrique $\boldsymbol{M} = \boldsymbol{D}_{1/\sigma^2} = \boldsymbol{D}_{1/\sigma}^\top \boldsymbol{D}_{1/\sigma}$ sur le tableau de données $\boldsymbol{X}$ revient à travailler avec la métrique euclidienne sur le tableau transformé $\boldsymbol{X}\boldsymbol{D}_{1/\sigma}^\top$.

En effet :

$$\begin{aligned} d_M^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_M^2 &= (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{D}_{1/\sigma}^\top \boldsymbol{D}_{1/\sigma} (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &= (\boldsymbol{D}_{1/\sigma}\, \boldsymbol{x}_i - \boldsymbol{D}_{1/\sigma}\, \boldsymbol{x}_j)^\top (\boldsymbol{D}_{1/\sigma}\, \boldsymbol{x}_i - \boldsymbol{D}_{1/\sigma}\, \boldsymbol{x}_j) \end{aligned}$$

Il est équivalent de travailler avec la métrique $\boldsymbol{D}_{1/\sigma^2}$ sur le tableau $\boldsymbol{X}$, ou avec la métrique euclidienne $\boldsymbol{I}$ sur le tableau centré réduit $\boldsymbol{Z}$ composé des données :

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

Le tableau de données centré réduit $\boldsymbol{Z}$ se calcule matriciellement ainsi :

$$\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{D}_{1/\sigma} = (\boldsymbol{X} - \boldsymbol{1}\,\boldsymbol{m}^\top)\,\boldsymbol{D}_{1/\sigma}.$$

Inerties

Inertie par rapport à un point

Définition : L'inertie du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ en un point quelconque $\boldsymbol{a}$ est donnée par

$$I_a = \sum_{i=1}^{n} p_i \|\boldsymbol{x}_i - \boldsymbol{a}\|_M^2$$

Propriété : L'inertie du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ en son point moyen $\boldsymbol{m}$, ou *centre de gravité*, est

$$I_m = \sum_{i=1}^{n} p_i \|\boldsymbol{x}_i - \boldsymbol{m}\|_M^2 = \frac{1}{2} \sum_{i,j=1}^{n} p_i\, p_j \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_M^2 = \text{Trace}(\boldsymbol{\Sigma M})$$

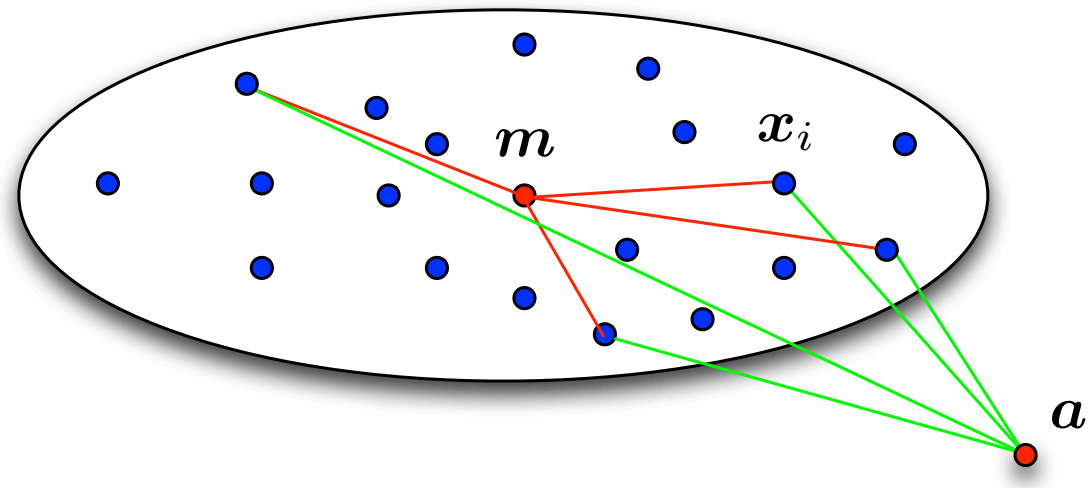
Cette propriété a été démontrée au chapitre précédent.

Propriété : Pour un tableau de données centrées réduites, on a

$$I_m = \text{Trace}(\boldsymbol{R}) = p$$

Inerties

Inertie par rapport à un point



Inerties

Théorème de Huygens

Propriété : Soit $\boldsymbol{m}$ le centre de gravité du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, et $\boldsymbol{a}$ un point quelconque de $\mathbb{R}^p$. L'inertie I_a du nuage au point $\boldsymbol{a}$ est donnée par

$$\boxed{I_a = I_m + d_M^2(\boldsymbol{a}, \boldsymbol{m})}$$

En conséquence, I_a est minimum pour $\boldsymbol{a} = \boldsymbol{m}$.

Démonstration :

$$\begin{aligned} I_a &= \sum_{i=1}^{n} p_i \, \|(\boldsymbol{x}_i - \boldsymbol{m}) - (\boldsymbol{a} - \boldsymbol{m})\|_M^2 \\ &= \sum_{i=1}^{n} p_i \, \|\boldsymbol{x}_i - \boldsymbol{m}\|_M^2 + \sum_{i=1}^{n} p_i \, \|(\boldsymbol{a} - \boldsymbol{m})\|_M^2 - 2 \sum_{i=1}^{n} p_i \, \langle \boldsymbol{x}_i - \boldsymbol{m}, \boldsymbol{a} - \boldsymbol{m} \rangle_M \\ &= I_m + d_M^2(\boldsymbol{a}, \boldsymbol{m}) \end{aligned}$$

Inerties

Inertie par rapport à un axe

Définition : L'inertie du nuage de points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ par rapport à un axe Δ est définie par

$$I_\Delta = \sum_{i=1}^{n} p_i \, d_M^2(\boldsymbol{x}_i, \Delta)$$

Cette inertie quantifie la dispersion du nuage des individus autour de Δ.

Rappels : Soit $\Delta(\boldsymbol{a}, \boldsymbol{u})$ un axe de vecteur directeur normé $\boldsymbol{u}$ passant par $\boldsymbol{a}$. On rappelle que

$$d_M^2(\boldsymbol{x}_i, \Delta) = \min_{v_i} d^2(\boldsymbol{z}_i, \boldsymbol{a} + v_i \, \boldsymbol{u})$$

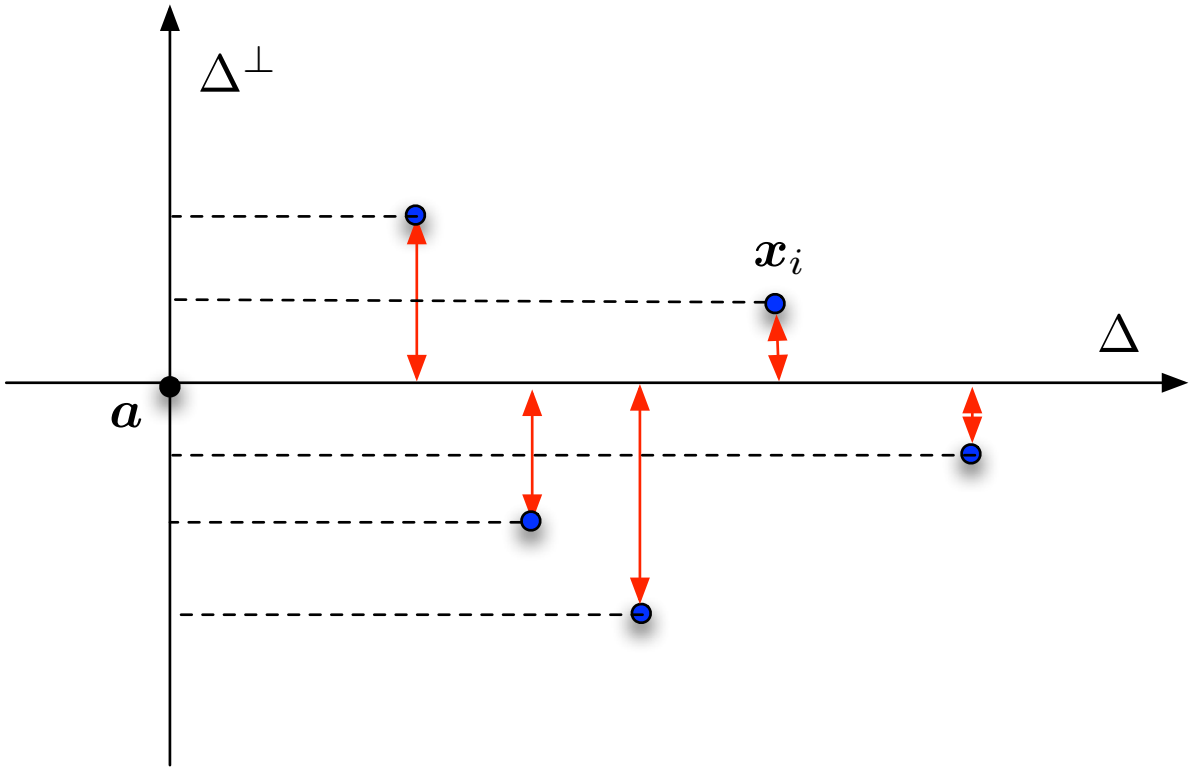
En conséquence, l'inertie du nuage $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ par rapport à $\Delta(\boldsymbol{a}, \boldsymbol{u})$ est définie par

$$I_{\Delta(\boldsymbol{a},\boldsymbol{u})} = \min_{\boldsymbol{v}} \sum_{i=1}^{n} p_i \, d_M^2(\boldsymbol{x}_i, \boldsymbol{a} + v_i \, \boldsymbol{u})$$

avec $\boldsymbol{v} = [v_1, \ldots, v_n]^\top$.

Inerties

Inertie par rapport à un axe



Analyse en Composantes Principales

Détermination du premier axe

Définition du problème : L'axe Δ recherché est celui représentant le mieux les données au sens de l'inertie I_Δ minimum. Il est donc obtenu par

$$\min_{\boldsymbol{u},\boldsymbol{a}} I_{\Delta(\boldsymbol{a},\boldsymbol{u})} \quad \text{avec} \quad I_{\Delta(\boldsymbol{a},\boldsymbol{u})} = \min_{\boldsymbol{v}} \sum_{i=1}^{n} p_i \, d_M^2(\boldsymbol{x}_i, \boldsymbol{a} + v_i \, \boldsymbol{u})$$

où Δ est défini par un point $\boldsymbol{a}$ et un vecteur directeur normé $\boldsymbol{u}$.

Détermination de $\boldsymbol{a}$: En appliquant le Théorème de Huygens, on trouve

$$I_{\Delta(\boldsymbol{a},\boldsymbol{u})} = I_{\Delta(\boldsymbol{m},\boldsymbol{u})} + d_M^2(\boldsymbol{a}, \boldsymbol{m})$$

En conséquence, $I_{\Delta(\boldsymbol{a},\boldsymbol{u})}$ est minimum pour $\boldsymbol{a} = \boldsymbol{m}$.
L'axe recherché passe par $\boldsymbol{m}$.

Analyse en Composantes Principales

Centrage et normalisation des données

Nous considérerons les données centrées et normalisées

$$\boldsymbol{Z} = (\boldsymbol{X} - \mathbf{1}\,\boldsymbol{m}^\top)\,\boldsymbol{M}^{\frac{1}{2}}$$

pour :

- ▷ faire correspondre $\boldsymbol{m}$ avec l'origine $\mathbf{0}$ du repère
- ▷ travailler avec la métrique euclidienne sur le nuage de points $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$

Analyse en Composantes Principales

Détermination du premier axe

Redéfinition du problème : L'axe Δ recherché passe par l'origine $\mathbf{0}$ et est solution du problème

$$\min_{\boldsymbol{u}} I_{\Delta(\boldsymbol{u})} \quad \text{avec} \quad I_{\Delta(\boldsymbol{u})} = \min_{\boldsymbol{v}} \sum_{i=1}^{n} p_i \, d^2(\boldsymbol{z}_i, v_i \, \boldsymbol{u})$$

Calcul du gradient : En développant la fonction coût, notée $J_\Delta(\boldsymbol{u}, \boldsymbol{v})$, on obtient

$$J_\Delta(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{n} p_i \, [\|\boldsymbol{z}_i\|^2 - 2 \, v_i \langle \boldsymbol{z}_i, \boldsymbol{u} \rangle + v_i^2 \|\boldsymbol{u}\|^2]$$

En conséquence

$$\nabla_u J_\Delta(\boldsymbol{u}, \boldsymbol{v}) = -2\boldsymbol{Z}^\top \boldsymbol{D} \, \boldsymbol{v} + 2\|\boldsymbol{v}\|_D^2 \, \boldsymbol{u}$$

$$\nabla_v J_\Delta(\boldsymbol{u}, \boldsymbol{v}) = -2\boldsymbol{D}\boldsymbol{Z}\boldsymbol{u} + 2\|\boldsymbol{u}\|^2 \boldsymbol{D} \, \boldsymbol{v}$$

avec $\boldsymbol{D}$ la matrice diagonal de termes diagonaux p_i.

Analyse en Composantes Principales

Détermination du premier axe

Conditions d'optimalité : Celles-ci s'écrivent

$$\begin{cases} \boldsymbol{Z}^\top \boldsymbol{D}\, \boldsymbol{v} = \|\boldsymbol{v}\|_D^2\, \boldsymbol{u} \\ \boldsymbol{D}\boldsymbol{Z}\boldsymbol{u} = \|\boldsymbol{u}\|^2 \boldsymbol{D}\, \boldsymbol{v} \end{cases}$$

En combinant les 2 équations, on obtient

$$\boldsymbol{Z}^\top \boldsymbol{D}\boldsymbol{Z}\boldsymbol{u} = \underbrace{\|\boldsymbol{u}\|^2 \|\boldsymbol{v}\|_D^2}_{\lambda}\, \boldsymbol{u}$$

La solution est un vecteur propre $\boldsymbol{u}$ de la matrice de covariance $\boldsymbol{\Sigma} = \boldsymbol{Z}^\top \boldsymbol{D}\boldsymbol{Z}$

Lequel ?

Analyse en Composantes Principales

Détermination du premier axe

Lequel ? En injectant le résultat dans la fonction coût, on obtient

$$\begin{aligned} I_{\Delta(\boldsymbol{u})} &= \sum_{i=1}^{n} p_i \left[\|\boldsymbol{z}_i\|^2 - 2\, v_i \langle \boldsymbol{z}_i, \boldsymbol{u} \rangle + v_i^2 \|\boldsymbol{u}\|^2 \right] \\ &= \text{Trace}(\boldsymbol{\Sigma}) - 2\, \boldsymbol{u}^\top \boldsymbol{Z}^\top \boldsymbol{D}\, \boldsymbol{v} + \|\boldsymbol{v}\|_D^2 \, \|\boldsymbol{u}\|^2 \\ &= \text{Trace}(\boldsymbol{\Sigma}) - 2\, \|\boldsymbol{v}\|_D^2 \, \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|_D^2 \, \|\boldsymbol{u}\|^2 \\ &= \text{Trace}(\boldsymbol{\Sigma}) - \lambda(\boldsymbol{u}) \end{aligned}$$

où $\lambda(\boldsymbol{u})$ est la valeur propre associée à $\boldsymbol{u}$

Le vecteur propre recherché est donc celui associé à la plus grande valeur propre

Analyse en Composantes Principales

Détermination du premier axe

Soient $\lambda_1 \geq \ldots \geq \lambda_p$ les valeurs propres de la matrice de covariance $\mathbf{\Sigma}$, de vecteurs propres unitaires associés $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$

Choix du premier axe : L'inertie $I_{\Delta(\boldsymbol{u})} = \text{Trace}(\mathbf{\Sigma}) - \lambda(\boldsymbol{u})$ est minimum pour le couple $(\boldsymbol{u}_1, \lambda_1)$

Inertie correspondante : La valeur λ_1 est, par définition, l'inertie expliquée par l'axe $\Delta(\boldsymbol{u}_1)$.

Inerties

Théorème d'Inclusion

▷ Soient $\mathcal{E}_1$ et $\mathcal{E}_1^\perp$ les sous-espaces affines de directions respectives Δ et $\Delta^\perp$ passant par l'origine.

▷ L'inertie du nuage par rapport à $\mathcal{E}_1$ et $\mathcal{E}_1^\perp$ est définie par

$$I_{\mathcal{E}_1} = \sum_{i=1}^{n} p_i \, d^2(\boldsymbol{z}_i, \mathcal{E}_1)$$

$$I_{\mathcal{E}_1^\perp} = \sum_{i=1}^{n} p_i \, d^2(\boldsymbol{z}_i, \mathcal{E}_1^\perp)$$

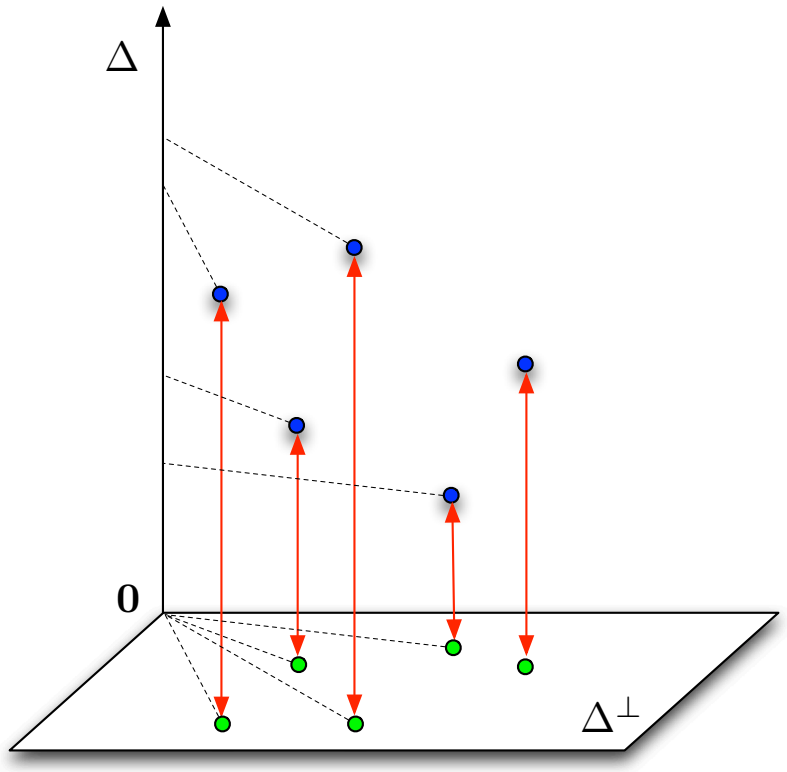
Voir le schéma ci-après. Par le Théorème de Pythagore, on a

$$\boxed{I_0 = I_{\mathcal{E}_1} + I_{\mathcal{E}_1^\perp}}$$

où I_0 est l'inertie du nuage par rapport à l'origine.
On en déduit le Théorème d'Inclusion ci-après.

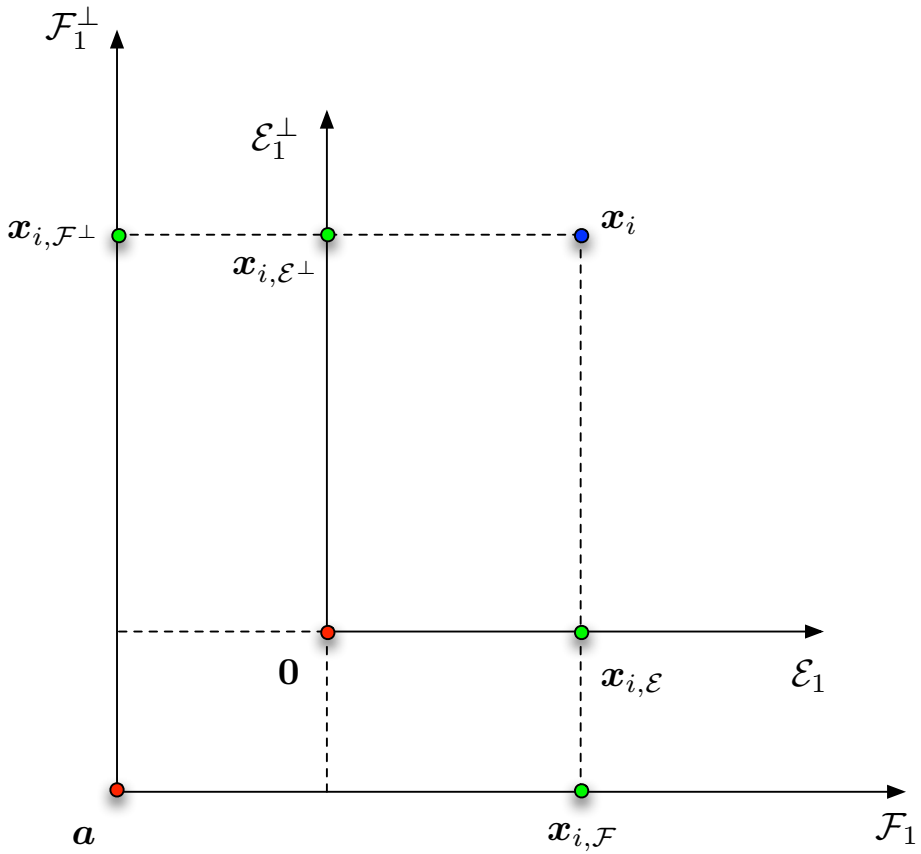
Inerties

Inertie par rapport à un sous-espace affine



Inerties

Théorème d'Inclusion



Analyse en Composantes Principales

Théorème d'Inclusion

Théorème d'Inclusion : Soit $\mathcal{E}_k$ un sous-espace, de dimension k, d'inertie expliquée maximum.

Le sous-espace $\mathcal{E}_{k+1}$ de dimension $k+1$, d'inertie expliquée maximum, est la somme directe de $\mathcal{E}_k$ et du sous-espace de dimension 1 de $\mathcal{E}_k^{\perp}$ dont l'inertie expliquée y est maximum.

Analyse en Composantes Principales

Théorème d'Inclusion

Recherches des axes : Etant donné le Théorème d'Inclusion, pour déterminer le sous-espace optimum $\mathcal{E}_k$, il est possible de procéder séquentiellement :

- recherche du premier axe d'inertie portée maximum
- recherche du deuxième axe, orthogonal au premier, d'inertie portée maximum
- etc.

Analyse en Composantes Principales

Détermination du deuxième axe

▷ On construit le tableau de données résiduel $\boldsymbol{Z}_1 = \boldsymbol{Z} - \boldsymbol{v}_1 \boldsymbol{u}_1^\top$. Il est aisé de montrer que les données correspondantes appartiennent à $\Delta^\perp$

$$\begin{aligned} \boldsymbol{Z}_1 \, \boldsymbol{u}_1 &= \boldsymbol{Z} \, \boldsymbol{u}_1 - \|\boldsymbol{u}_1\|^2 \, \boldsymbol{v}_1 \\ &= 0 \end{aligned}$$

d'après la deuxième condition d'optimalité, car $\boldsymbol{D}$ est inversible.

▷ La matrice $\boldsymbol{Z}_1^\top \boldsymbol{D} \boldsymbol{Z}_1$ admet les mêmes couples valeurs/vecteurs propres $(\lambda_k, \boldsymbol{u}_k)$ que $\boldsymbol{Z}^\top \boldsymbol{D} \boldsymbol{Z}$, pour $k \neq 1$,

$$\begin{aligned} (\boldsymbol{Z}_1^\top \boldsymbol{D} \boldsymbol{Z}_1) \, \boldsymbol{u}_k &= (\boldsymbol{Z}^\top \boldsymbol{D} \boldsymbol{Z}) \, \boldsymbol{u}_k \quad k \neq 1 \\ &= \lambda_k \, \boldsymbol{u}_k \end{aligned}$$

car $(\boldsymbol{u}_1 \perp \boldsymbol{u}_k)$, excepté λ_1 qui devient nulle

$$\begin{aligned} (\boldsymbol{Z}_1^\top \boldsymbol{D} \boldsymbol{Z}_1) \, \boldsymbol{u}_1 &= [\boldsymbol{Z} - \boldsymbol{v}_1 \boldsymbol{u}_1^\top]^\top \boldsymbol{D} \, [\boldsymbol{Z} - \boldsymbol{v}_1 \boldsymbol{u}_1^\top] \\ &= (1 - 2 + 1) \, \|\boldsymbol{u}_1\|^2 \|\boldsymbol{v}_1\|_D^2 \, \boldsymbol{u}_1 = 0 \end{aligned}$$

Analyse en Composantes Principales

Détermination du deuxième axe

Pour déterminer le deuxième axe, on pose le même problème que pour le premier, avec le tableau $\boldsymbol{Z}_1$. On aboutit alors à $\boldsymbol{u}_2$. En effet

Choix du deuxième axe : L'inertie $I_{\Delta(\boldsymbol{u})} = \text{Trace}(\boldsymbol{\Sigma}_1) - \lambda(\boldsymbol{u})$ est minimum pour le couple $(\boldsymbol{u}_2, \lambda_2)$

Remarque : $\text{Trace}(\boldsymbol{\Sigma}_1) = \text{Trace}(\boldsymbol{\Sigma}) - \lambda_1$

Inerties correspondantes : L'inertie expliquée par l'axe $\Delta(\boldsymbol{u}_2)$ vaut λ_2. Pour l'espace affine de direction $\Delta_{\boldsymbol{u}_1} + \Delta_{\boldsymbol{u}_2}$, par le Théorème de Pythagore, elle vaut $\lambda_1 + \lambda_2$.

et ainsi de suite...

Analyse en Composantes Principales

Axes principaux et inertie expliquée

- Les axes $\Delta(\boldsymbol{u}_k)$ sont appelés **axes factoriels** ou **axes principaux**
- L'inertie expliquée par l'axe $\Delta(\boldsymbol{u}_k)$ est la valeur propre λ_k de $\boldsymbol{\Sigma}$
- L'inertie expliquée par le sous-espace factoriel $\mathcal{E}_{k_0}$ engendré par les k_0 premiers axes principaux est

$$\lambda_1 + \ldots + \lambda_{k_0}$$

- Le **taux d'inertie expliquée** par $\mathcal{E}_{k_0}$ est donné par

$$\frac{\lambda_1 + \ldots + \lambda_{k_0}}{\sum_{k=1}^{p} \lambda_k}$$

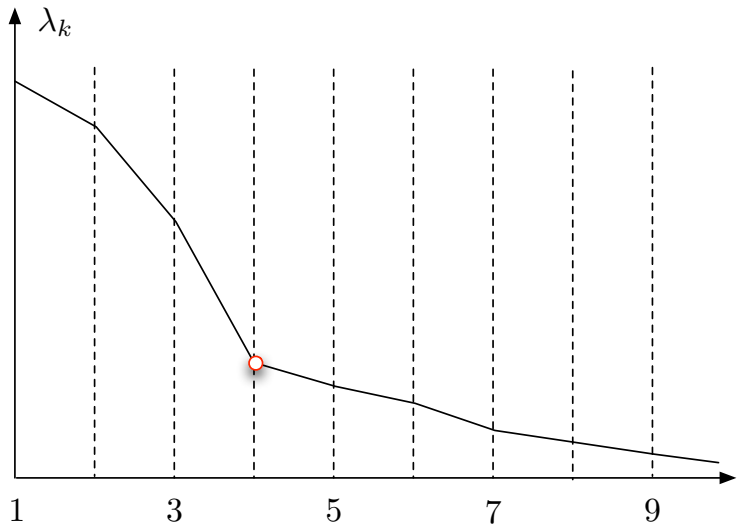
car $I_0 = \text{Trace}(\boldsymbol{\Sigma}) = \sum_{k=1}^{p} \lambda_k$

Analyse en Composantes Principales

Nombre d'axes à retenir

L'ACP vise à réduire la dimension de l'espace des individus. On souhaite conserver aussi peu d'axes principaux que possible. Des critères empiriques :

- ▷ Retenir les axes offrant une corrélation suffisante entre composantes principales et variables initiales. Voir plus loin.
- ▷ Fixer un taux minimum d'inertie expliquée, par exemple 0.80
- ▷ Rechercher un coude dans l'éboulis des valeurs propres λ_k



Analyse en Composantes Principales

Composantes principales

Coordonnées des individus : La coordonnée $c_{\ell k}$ de l'individu $\boldsymbol{z}_\ell$ sur l'axe principal $\Delta(\boldsymbol{u}_k)$ est donnée par sa projection sur $\boldsymbol{u}_k$

$$c_{\ell k} = \boldsymbol{z}_\ell^\top \boldsymbol{u}_k$$

Composantes principales : Il s'agit des vecteurs $\boldsymbol{c}_k$ des coordonnées des individus sur l'axe principal $\Delta(\boldsymbol{u}_k)$, c'est à dire

$$\boldsymbol{c}_k = \boldsymbol{Z}\, \boldsymbol{u}_k$$

Analyse en Composantes Principales

Propriétés des composantes principales

Moyenne arithmétique :

$$\bar{\boldsymbol{c}}_k = \mathbf{1}^\top \boldsymbol{D}\, \boldsymbol{c}_k = \mathbf{1}^\top \boldsymbol{D}\, \boldsymbol{Z}\boldsymbol{u}_k = \mathbf{0}$$

car $\mathbf{1}^\top \boldsymbol{D}\, \boldsymbol{Z} = \mathbf{0}$, le tableau étant centré

Variance :

$$\mathrm{var}(\boldsymbol{c}_k) = \boldsymbol{c}_k^\top \boldsymbol{D}\, \boldsymbol{c}_k = \boldsymbol{u}_k^\top \boldsymbol{Z}^\top \boldsymbol{D}\boldsymbol{Z}\boldsymbol{u}_k = \lambda_k$$

car $\boldsymbol{u}_k$ est vecteur propre de $\boldsymbol{\Sigma} = \boldsymbol{Z}^\top \boldsymbol{D}\boldsymbol{Z}$

Covariance : Il s'agit des vecteurs $\boldsymbol{c}_k$ des coordonnées des individus sur l'axe principal $\Delta(\boldsymbol{u}_k)$, c'est à dire

$$\mathrm{cov}(\boldsymbol{c}_k, \boldsymbol{c}_\ell) = \boldsymbol{c}_k^\top \boldsymbol{D}\, \boldsymbol{c}_\ell = \boldsymbol{u}_k^\top \boldsymbol{Z}^\top \boldsymbol{D}\boldsymbol{Z}\boldsymbol{u}_\ell = 0$$

car $(\boldsymbol{u}_k \perp \boldsymbol{u}_\ell)$. Les composantes principales sont donc décorrélées.

Analyse en Composantes Principales

Qualité de la représentation d'un individu

Définition : La qualité de la représentation d'un individu ℓ sur l'axe principal k est donnée par le cosinus de leur angle, soit

$$\cos(\boldsymbol{z}_\ell, \boldsymbol{u}_k) = \frac{\boldsymbol{z}_\ell^\top \boldsymbol{u}_k}{\|\boldsymbol{z}_\ell\|}$$

Comme $\boldsymbol{z}_\ell^\top \boldsymbol{u}_k = c_{\ell k}$, et $\{\boldsymbol{u}_k\}_k$ est une base orthonormée de $\mathbb{R}^p$, on a

$$\cos^2(\boldsymbol{z}_\ell, \boldsymbol{u}_k) = \frac{c_{\ell k}^2}{\sum_{k=1}^p c_{\ell k}^2}$$

Analyse en Composantes Principales

Contribution d'un individu à une composante principale

On a vu que $\text{var}(\boldsymbol{c}_k) = \lambda_k = \sum_{i=1}^{n} p_i \, c_{ik}^2$.

Définition : On définit la contribution de l'individu ℓ à un axe principal k par

$$\theta_{\ell k} = \frac{p_\ell c_{\ell k}^2}{\lambda_k}$$

Interprétation : La contribution d'un individu ℓ doit être relativisée par rapport à son poids p_ℓ. Ainsi, si $\theta_{\ell k} > \alpha \, p_i$ avec $\alpha > 2$, on peut juger la contribution de l'individu ℓ importante.

Sur-représentation : L'individu ℓ peut être considéré comme sur-représenté par l'axe principal k si $\theta_{\ell k} > 0.25$. Il a eu une influence trop importante dans la détermination de celui-ci, risquant de perturber la définition des autres axes.

Analyse en Composantes Principales

Espace des variables

On s'intéresse à présent aux colonnes $\boldsymbol{z}^j$ du tableau $\boldsymbol{Z}$, représentant les variables.

Métrique $\boldsymbol{D}$: On munit l'espace des variables d'une métrique naturelle, $\boldsymbol{D}$, celle des poids des individus

$$\langle \boldsymbol{z}^i, \boldsymbol{z}^j \rangle_D = (\boldsymbol{z}^i)^\top \boldsymbol{D}\, \boldsymbol{z}^j \qquad \|\boldsymbol{z}^i\|_D^2 = (\boldsymbol{z}^i)^\top \boldsymbol{D}\, \boldsymbol{z}^i$$

Interprétation : Pour des variables centrées, ce qui est le cas ici, on a

$$\mathrm{cor}(\boldsymbol{z}^i, \boldsymbol{z}^j) = \frac{\langle \boldsymbol{z}^i, \boldsymbol{z}^j \rangle_D}{\|\boldsymbol{z}^i\|_D \, \|\boldsymbol{z}^j\|_D} = \cos(\boldsymbol{z}^i, \boldsymbol{z}^j)$$

Remarque : Les composantes principales normalisée $\boldsymbol{c}_k/\sqrt{\lambda_k}$ forment une base $\boldsymbol{D}$-orthonormale :

$$\left\langle \frac{\boldsymbol{c}_i}{\sqrt{\lambda_i}}, \frac{\boldsymbol{c}_j}{\sqrt{\lambda_j}} \right\rangle_D = \mathrm{cor}\left(\frac{\boldsymbol{c}_i}{\sqrt{\lambda_i}}, \frac{\boldsymbol{c}_j}{\sqrt{\lambda_j}} \right) = \delta_{ij}$$

Analyse en Composantes Principales

Corrélation des composantes principales et variables

▷ Le coefficient de corrélation entre $\boldsymbol{c}_k$ et $\boldsymbol{z}^i$ est donné par

$$\text{cor}(\boldsymbol{c}_k, \boldsymbol{z}^i) = \frac{\langle \boldsymbol{c}_k, \boldsymbol{z}^i \rangle_D}{\sqrt{\lambda_k}}$$

▷ Les coefficients de corrélation entre $\boldsymbol{c}_k$ et l'ensemble des variables de $\boldsymbol{Z}$ peuvent être calculés ainsi

$$\text{cor}(\boldsymbol{c}_k, \boldsymbol{Z}) = \frac{\boldsymbol{Z}^\top \boldsymbol{D}\, \boldsymbol{c}_k}{\sqrt{\lambda_k}}$$

▷ Comme $\boldsymbol{Z}^\top \boldsymbol{D}\, \boldsymbol{c}_k = \boldsymbol{Z}^\top \boldsymbol{D}\, \boldsymbol{Z} \boldsymbol{u}_k = \boldsymbol{\Sigma} \boldsymbol{u}_k$, on a directement

$$\text{cor}(\boldsymbol{c}_k, \boldsymbol{Z}) = \sqrt{\lambda_k}\, \boldsymbol{u}_k$$

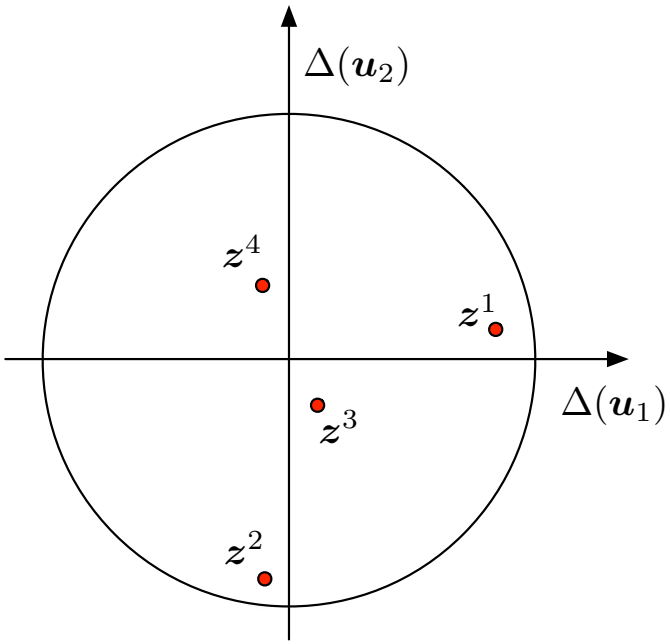
Analyse en Composantes Principales

Cercle des corrélations

On représente traditionnellement chaque variable $\boldsymbol{z}^i$ par un point de coordonnées

$$(\mathrm{cor}(\boldsymbol{z}^i, \boldsymbol{c}_1), \mathrm{cor}(\boldsymbol{z}^i, \boldsymbol{c}_2), \mathrm{cor}(\boldsymbol{z}^i, \boldsymbol{c}_3), \ldots)$$

afin d'interpréter les axes principaux obtenus.



Analyse en Composantes Principales

En résumé

- Centrer et réduire les données $\boldsymbol{X} \to \boldsymbol{Z}$
- Calculer la matrice de covariance $\boldsymbol{\Sigma} = \boldsymbol{Z}^\top \boldsymbol{D} \boldsymbol{Z}$
- Diagonaliser $\boldsymbol{\Sigma}$ pour obtenir $\{(\boldsymbol{u}_k, \lambda_k)\}_{k=1,\ldots,p}$
- Choisir le nombre k_0 d'axes à retenir
- Préciser le taux d'inertie expliqué par chacun des axes retenus
- Calculer les composantes principales $\{\boldsymbol{c}_k\}_{k=1,\ldots,k_0}$
- Représenter les individus dans le(s) repère(s) des axes principaux
- Tracer les cercles de corrélation
- Interpréter