Analyse de données et apprentissage

Analyse en composantes principales

El-Hadi Djermoune

10 février 2026

Université de Lorraine
Faculté des Sciences et Technologies
M1 Ingénierie de Systèmes Complexes

CRAN UL CNRS – département BioSiS
Bureau: 422, Bât. Henri Poincaré 4e étage
email: el-hadi.djermoune@univ-lorraine.fr

Table des matières

Objectifs

Analyse en composantes principales

- L'ACP (*principal component analysis* – PCA) est une technique statistique permettant de simplifier des données.
- Elle a été développée par Pearson en 1901 et Hotelling en 1933, voir [Jolliffe, 2002].
- Le but de la méthode est de réduire la dimensionalité de données multivariables en préservant autant que possible l'information principale que les données contiennent.
- L'ACP est une transformation linéaire qui transforme les données dans un nouveau système d'axes de telle sorte que les nouvelles variables (appelée « composantes principales ») soient :
  1. décorrélées ;
  2. la plus grande variance vienne avec la première composante, la deuxième plus grande variance avec la deuxième composante, etc.

Objectifs du chapitre

- Comprendre l'ACP.
- Savoir utiliser et interpréter les résultats d'une ACP.

Table des matières

Statistiques empiriques

- La moyenne empirique d'un *échantillon* $\{x_i\}_{i=1}^n$ d'une *population* est :

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

- La variance empirique[1] d'un échantillon $\{x_i\}_{i=1}^n$ s'écrit :

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2}$$

- L'écart-type d'un échantillon $\{x_i\}_{i=1}^n$ est la racine carrée de la variance :

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{3}$$

Exemples :

1. $\{0, 8, 12, 20\}$: $\bar{x} = 20$, $\sigma_x = 7.2$.
2. $\{8, 9, 11, 12\}$: $\bar{x} = 20$, $\sigma_x = 1.6$.

1. Un autre formule d'estimation de la variance, très courante dans la littérature statistique, est donnée par : $\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Elle est appelée *estimateur non biaisé*.

Statistiques empiriques

- La covariance entre deux échantillons de même taille $\{x_i\}_{i=1}^n$ et $\{y_i\}_{i=1}^n$ est une mesure bidimensionnelle :

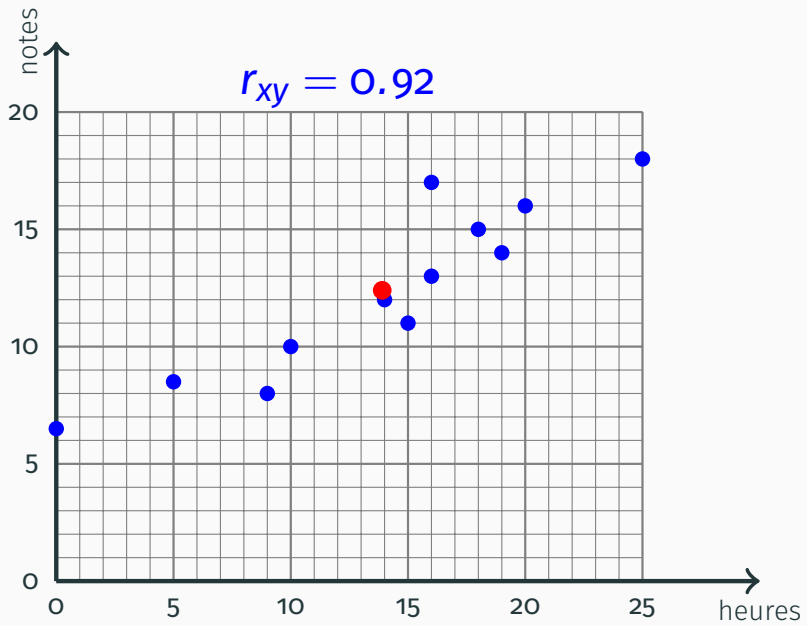
$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{4}$$

- Le coefficient de corrélation linéaire est la covariance normalisée :

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} \in [-1, 1]. \tag{5}$$

Exemple : données sur une classe de 12 étudiants.

N° étudiant	Heures	Notes
1	9	8
2	15	11
3	25	18
4	14	12
5	10	10
6	18	15
7	0	6.5
8	16	17
9	5	8.5
10	19	14
11	16	13
12	20	16
Moyenne	13.9	12.4
Ecart-type	6.61	3.56
Variance	43.7	12.7
Covariance	21.7	



- $-1 \leq r_{xy} \leq 1$
- le signe indique le sens de la dépendance (+ veut dire que quand x augmente, y augmente)
- 0 : x et y sont linéairement indépendants
- ± 1 : x et y sont complètement dépendants (relation affine)

Statistiques empiriques

- La covariance est une quantité qui se se calcule pour 2 ensembles de données (ou données bidimensionnelles).
- Pour les données composées de p échantillons, notés $\mathbf{x}_1, \ldots, \mathbf{x}_p$, on calcule la covariance entre chaque combinaison de 2 échantillons, ce qui revient à déterminer $\binom{p}{2} = \frac{p!}{2!(p-2)!}$ valeurs de covariance :

$$\mathbf{V} = \begin{bmatrix} \operatorname{cov}(\mathbf{x}_1, \mathbf{x}_1) & \operatorname{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \operatorname{cov}(\mathbf{x}_1, \mathbf{x}_p) \\ \operatorname{cov}(\mathbf{x}_2, \mathbf{x}_1) & \operatorname{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \operatorname{cov}(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(\mathbf{x}_p, \mathbf{x}_1) & \operatorname{cov}(\mathbf{x}_p, \mathbf{x}_2) & \cdots & \operatorname{cov}(\mathbf{x}_p, \mathbf{x}_p) \end{bmatrix} \triangleq \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{bmatrix}$$

- La matrice de variance-covariance possède les propriétés suivantes :
  - elle est symétrique pour des données réelles et hermitienne pour des données complexes ;
  - elle est diagonalisable[2] et ses *valeurs propres* sont réelles et positives ;
  - sa diagonale contient les variances individuelles, réelles et positives.
- Pour le jeu de données précédent, on trouve :

$$\mathbf{V} = \begin{bmatrix} 43.7 & 21.7 \\ 21.7 & 12.7 \end{bmatrix}$$

2. Une matrice diagonalisable est une matrice dont les vecteurs propres forment une base.

Vecteur propre

Un vecteur $\mathbf{v}$ est un vecteur propre pour une matrice $\mathbf{A}$ si sa transformation par $\mathbf{A}$ (*i.e.* e produit $\mathbf{A} \cdot \mathbf{v}$) est colinéaire à $\mathbf{v}$, c'est-à-dire $\mathbf{Av} = \lambda\mathbf{v}$. Les vecteurs propres sont associés (« propres ») à une matrice carrée.

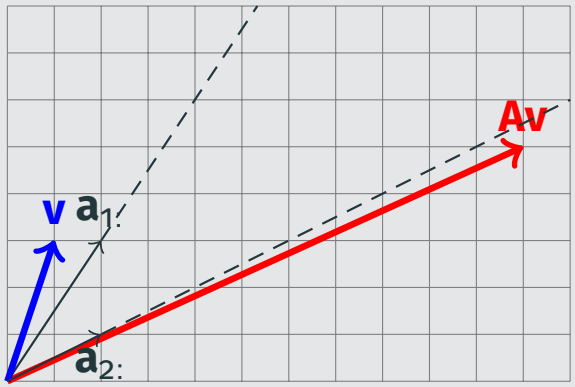
Une matrice $p \times p$ peut avoir, au plus, p vecteurs propres.

Exemple : on considère la matrice 2×2

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{1:} \\ \mathbf{a}_{2:} \end{bmatrix}$$

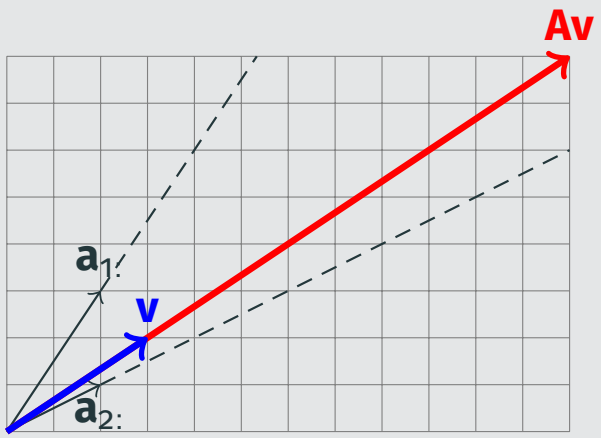
1. $\mathbf{v} = [1, 3]^\top$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 11 \\ 5 \end{bmatrix}$$



2. $\mathbf{v} = [3, 2]^\top$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$



Valeur propre

Si $\mathbf{v}$ est un vecteur propre de $\mathbf{A}$, alors $\mathbf{Av} = \lambda\mathbf{v}$ et λ est une *valeur propre*. Pour l'exemple précédent, le résultat de la multiplication de la matrice par le vecteur propre $[3, 2]^\top$ est un vecteur 4 fois plus grand : on dit que $\lambda = 4$ est une valeur propre associée au vecteur propre $[3, 2]^\top$ de $\mathbf{A}$.

Décomposition en valeurs propres

La décomposition en valeurs propres est la procédure qui permet de calculer les valeurs et vecteurs propres d'une matrice $\mathbf{A}$.

1. **Valeurs propres :** trouver les racines du polynôme caractéristique

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0 \tag{6}$$

   où $\mathbf{I}$ est la matrice identité. Ce polynôme possède p racines dans $\mathbb{C}$.

2. **Vecteurs propres :** pour chacune des valeurs propres λ_i, résoudre :

$$\mathbf{Av}_i = \lambda_i\mathbf{v}_i. \tag{7}$$

   pour $\mathbf{v}_i \neq 0$. Si $\mathbf{v}$ est un vecteur propre de $\mathbf{A}$, il en est de même de $\alpha\mathbf{v}$, $\forall\alpha \neq 0$.

Table des matières

Les données

- Matrice des p variables et n individus :

$$\mathbf{X} = \begin{array}{c} \\ \mathbf{x}_{1:} \\ \mathbf{x}_{2:} \\ \\ \mathbf{x}_{i:} \\ \\ \mathbf{x}_{n:} \end{array} \overset{\text{individus}}{\overbrace{\begin{array}{c} \text{variables / caractéristiques} \\ \begin{array}{cccccc} \mathbf{x}_{:1} & \mathbf{x}_{:2} & & \mathbf{x}_{:j} & & \mathbf{x}_{:p} \\ \downarrow & \downarrow & & \downarrow & & \downarrow \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \end{array}} \tag{8}$$

- On suppose que les données x_{ij} sont réelles[3].
- Chaque ligne i de $\mathbf{X}$ représente les valeurs prises par l'individu i sur les p variables.
- Chaque colonne j de $\mathbf{X}$ représente les valeurs de la variable j associées aux n individus.
- L'individu i sera identifié par le vecteur $\mathbf{x}_{i:} = [x_{i1}, \ldots, x_{ip}]^\top$ tandis que la variable j sera identifiée par le vecteur $\mathbf{x}_{:j} = [x_{1j}, \ldots, x_{nj}]^\top$.

3. C'est le cas en pratique, mais le passage aux complexes ne pose aucune difficulté.

Centrage des données

- **Individus moyen :** le vecteur $\bar{\mathbf{x}}$ est le point moyen du nuage des individus ; on l'appelle également l'individu moyen

$$\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]^\top, \qquad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$

  Les données centrées $\mathbf{Y}$ sont telles que :

$$\mathbf{y}_{:j} = \mathbf{x}_{:j} - \bar{x}_j \iff \mathbf{Y} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^\top.$$

- **Matrice de covariance :** contient les covariances entre les var. j et j'

$$v_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \iff \mathbf{V} = [v_{jj'}] = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y}.$$

- **ACP non normée :** c'est la méthode d'analyse qui utilise la matrice de covariance $\mathbf{V}$ pour déterminer les composantes principales.
- **ACP normée :** c'est la méthode d'analyse qui utilise la matrice de corrélation $\mathbf{R}$. On pose $\sigma_j^2 = v_{jj}$, alors (noter que $r_{jj} = 1$)

$$r_{jj'} = \frac{v_{jj'}}{\sigma_j \sigma_{j'}} \iff \mathbf{R} = [r_{jj'}] = \mathbf{D}_{\frac{1}{\sigma}}\mathbf{V}\mathbf{D}_{\frac{1}{\sigma}} \quad \text{avec} \quad \mathbf{D}_{\frac{1}{\sigma}} = \begin{bmatrix} \frac{1}{\sigma_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_p} \end{bmatrix}$$

Espace des individus

- Dans l'espace des individus $\mathbb{R}^p$, chaque individu est représenté par un point, formant ainsi un nuage de points.
- L'objectif de l'ACP est de visualiser ce nuage dans un espace de faible dimension le plus fidèlement possible.
- L'analyse repose sur la distance entre individus ; plusieurs distances (métriques) sont possibles.
- La distance euclidienne entre individus est la plus utilisée :

$$d^2(i, i') = d^2(\mathbf{x}_{i:}, \mathbf{x}_{i':}) = ||\mathbf{x}_{i:} - \mathbf{x}_{i':}||_2^2 = (\mathbf{x}_{i:} - \mathbf{x}_{i':})^\top (\mathbf{x}_{i:} - \mathbf{x}_{i':})$$

$$\Longrightarrow d(i, i') = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}.$$

- L'inertie totale (ou variance totale) du nuage des individus mesure la dispersion des points autour du **centre de gravité $\bar{\mathbf{x}}$** du nuage :

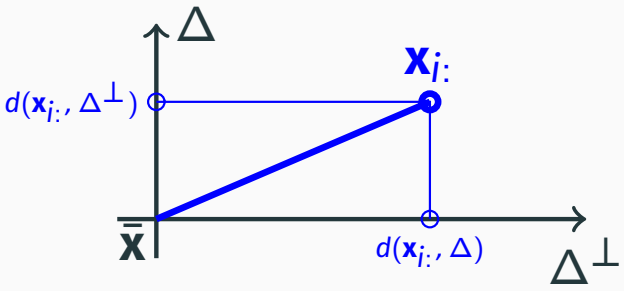
$$I = \frac{1}{n} \sum_{i=1}^{n} d^2(\mathbf{x}_{i:}, \bar{\mathbf{x}}) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{p} v_{jj} = \text{trace}(\mathbf{V}).$$

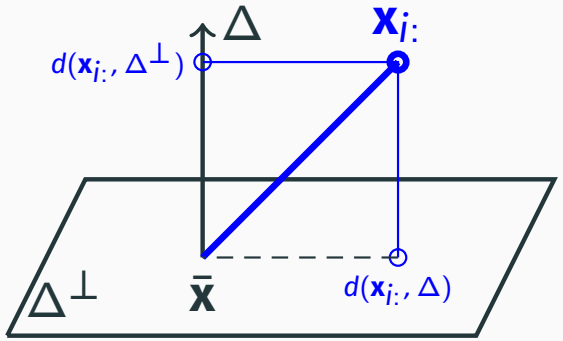
- Plus l'inertie est grande, plus le nuage est dispersé et, au contraire, plus elle est petite, plus le nuage est concentré sur son centre de gravité.

Espace des individus

- L'inertie des individus par rapport à un axe Δ passant par $\bar{\mathbf{x}}$ est :

$$I_\Delta = \frac{1}{n}\sum_{i=1}^{n} d^2(\mathbf{x}_{i:}, \Delta).$$





- En projettant le nuage sur Δ, on perd l'inertie I_Δ et on conserve[4] $I_{\Delta^\perp}$.
  - $I_{\Delta^\perp}$ est l'**inertie expliquée** par l'axe Δ.
  - I_Δ est l'**inertie perdue** en projettant les individus sur l'axe Δ.
- Théorème de Huygens :

$$I_\Delta + I_{\Delta^\perp} = I.$$

- Soit $\mathbf{u}$ le vecteur directeur de Δ, la projection des individus sur Δ est :

$$I_{\Delta^\perp} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_{i:}\cdot\mathbf{u})^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}^\top\mathbf{x}_{i:}\mathbf{x}_{i:}^\top\mathbf{u} = \mathbf{u}^\top\mathbf{V}\mathbf{u} \quad \text{(si } \mathbf{X} \text{ centrée)}.$$

4. $\Delta^\perp$ est le complément orthogonal de Δ dans $\mathbb{R}^p$.

Espace des variables

- Chaque variable est représentée par un vecteur dans un espace de dimension n.
- Le produit scalaire entre deux variables centrées $\mathbf{y}_{:j}$ et $\mathbf{y}_{:j'}$ est :

$$\langle \mathbf{y}_{:j}, \mathbf{y}_{:j'} \rangle = \sum_{i=1}^{n} y_{ij} y_{ij'} = n \cdot v_{jj'}.$$

- Le carré de la norme d'une variable est égal à un multiple de sa variance :

$$||\mathbf{y}_{:j}||^2 = n \cdot v_{jj} = n \cdot \sigma_{jj}^2.$$

- Le cosinus de l'angle $\theta_{jj'}$ entre deux variables $\mathbf{y}_{:j}$ et $\mathbf{y}_{:j'}$ est leur coefficient de corrélation linéaire :

$$\cos(\theta_{jj'}) = \frac{\langle \mathbf{y}_{:j}, \mathbf{y}_{:j'} \rangle}{||\mathbf{y}_{:j}||_2 \cdot ||\mathbf{y}_{:j'}||_2} = \frac{v_{jj'}}{\sigma_j \sigma_{j'}} = r_{jj'}.$$

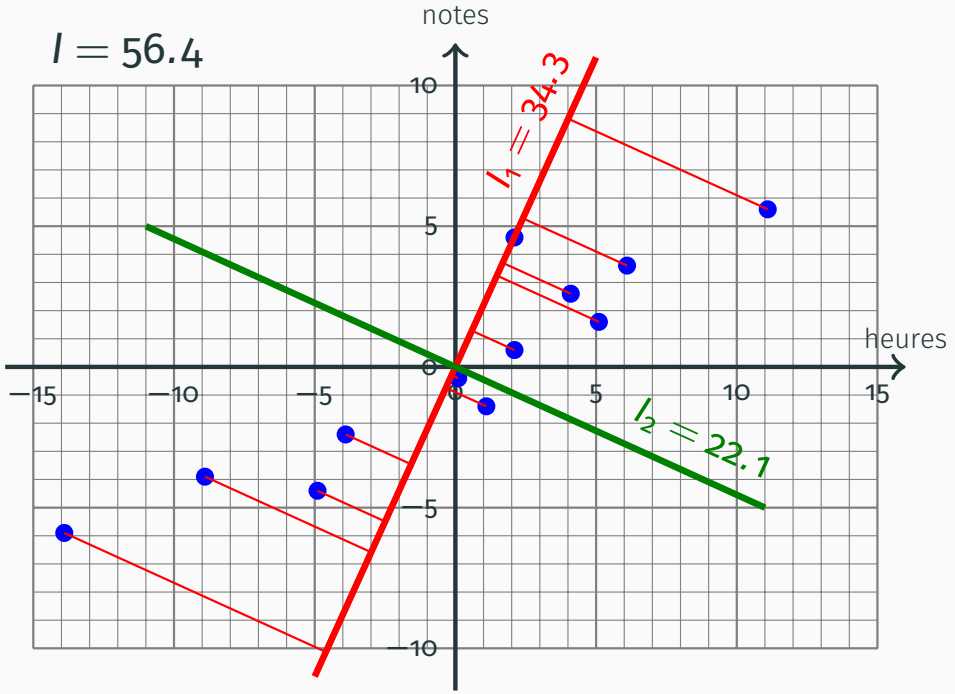
- Dans l'espace des variables, nous nous intéressons aux angles entre variables plutôt qu'aux distances, et on représente les variables comme des vecteurs et non des points.

Table des matières

Recherche des axes portant l'inertie maximale

Réduction de dimension

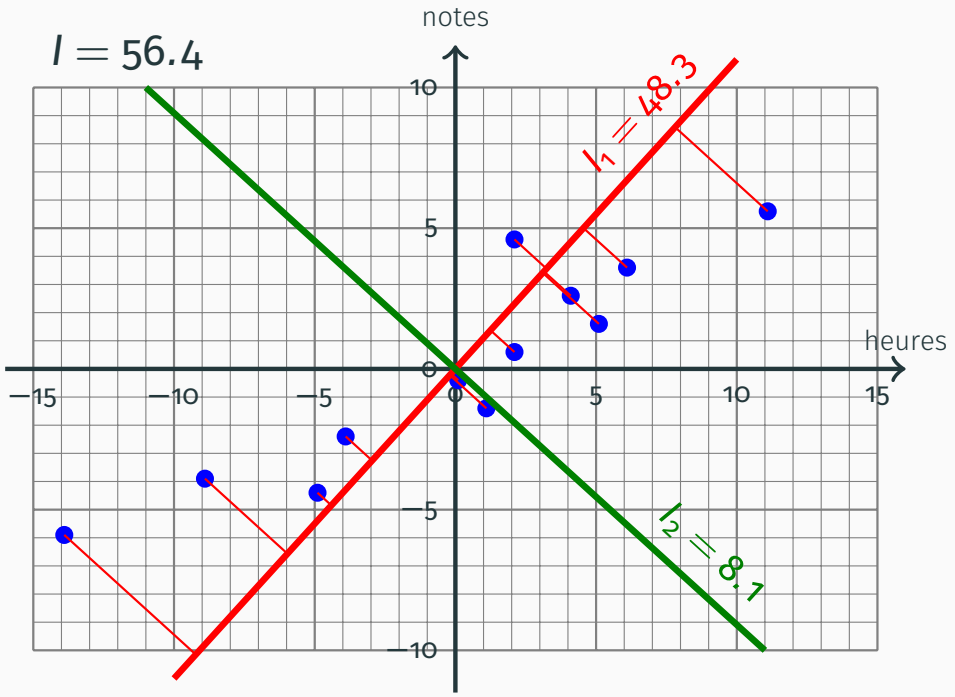
- L'ACP consiste à rechercher un sous-espace F_k de dim. $k < p$ tel que le nuage, une fois projeté dans ce sous-espace, soit au minimum déformé.
- Comme la projection diminue les distances, on cherche le sous-espace F_k qui maximise la distance entre individus (en moyenne).
- On cherche donc le sous-espace qui porte la plus grande part de l'inertie du nuage des individus.



Recherche des axes portant l'inertie maximale

Réduction de dimension

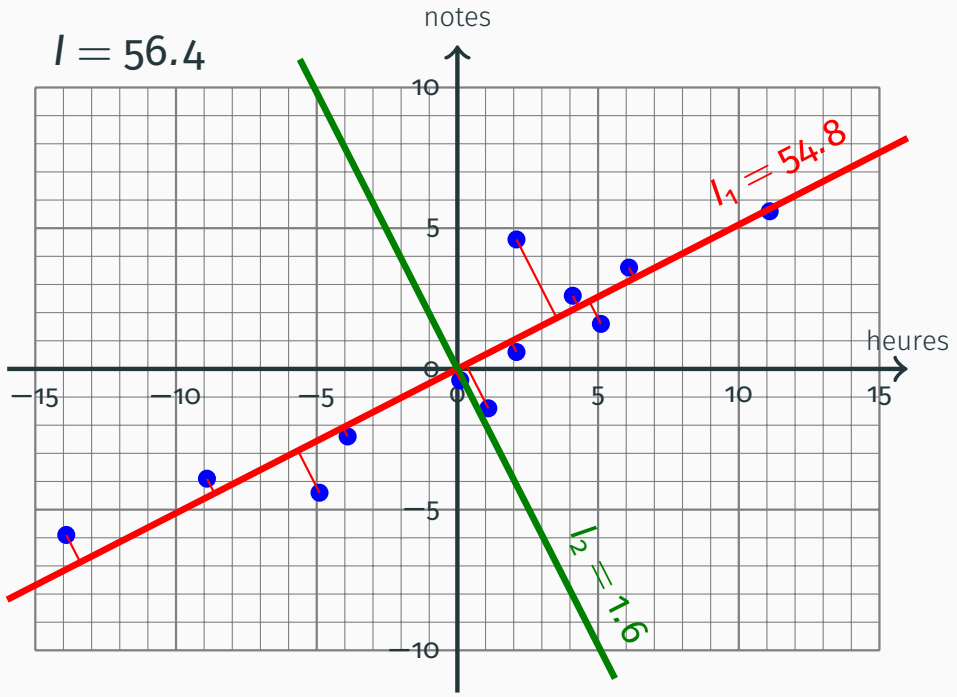
- L'ACP consiste à rechercher un sous-espace F_k de dim. $k < p$ tel que le nuage, une fois projeté dans ce sous-espace, soit au minimum déformé.
- Comme la projection diminue les distances, on cherche le sous-espace F_k qui maximise la distance entre individus (en moyenne).
- On cherche donc le sous-espace qui porte la plus grande part de l'inertie du nuage des individus.



Recherche des axes portant l'inertie maximale

Réduction de dimension

- L'ACP consiste à rechercher un sous-espace F_k de dim. $k < p$ tel que le nuage, une fois projeté dans ce sous-espace, soit au minimum déformé.
- Comme la projection diminue les distances, on cherche le sous-espace F_k qui maximise la distance entre individus (en moyenne).
- On cherche donc le sous-espace qui porte la plus grande part de l'inertie du nuage des individus.



Recherche des axes portant l'inertie maximale

Théorème

Soit F_k un sous-espace portant l'inertie maximale, alors le sous-espace F_{k+1} portant l'inertie maximale est la somme directe de F_k et du sous-espace de dimension 1 orthogonal à F_k portant l'inertie maximale.

Analyse en composante principale

- L'ACP consiste à trouver, un par un, les axes Δ_k portant l'inertie max.
- L'inertie expliquée par Δ_1 est la plus grande part de l'inertie du nuage, l'inertie expliqué par Δ_2 est la plus grande de l'inertie restante, ...
- Les axes Δ_k sont tous orthogonaux 2 à 2.
- La combinaison des K axes portant chacun l'inertie maximale forme le sous-espace de dimension $K \leq p$ recherché par l'ACP.
- **Rappel :** on cherche le 1er des axes (Δ_1), de vecteur directeur $\mathbf{u}_1$, qui vérifie :

$$\mathbf{u}_1 = \arg\max_{\mathbf{u} \in \mathbb{R}^p} \mathbf{u}^\top \mathbf{V} \mathbf{u} \quad \text{sous la contrainte} \quad \mathbf{u}^\top \mathbf{u} = ||\mathbf{u}||^2 = 1.$$

- Nous allons voir que la solution à ce problème d'optimisation revient à cherche le vecteur propre de $\mathbf{V}$ associé à la plus grande valeur propre.

Recherche du 1er axe portant l'inertie maximale

- $\mathbf{V}$ est symétrique, elle est donc diagonalisable et ses valeurs propres sont positives :

$$\mathbf{V} = \mathbf{P}\Lambda\mathbf{P}^\top = \sum_{j=1}^{p} \lambda_j \mathbf{p}_{:j}\mathbf{p}_{:j}^\top$$

  Λ : matrice diag. des valeurs propres ; $\mathbf{P}$: vecteurs propres (colonnes).

- Si on range les valeurs propres (**réelles et positives**) par ordre décroissant ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$), on peut écrire :

$$\mathbf{u}^\top \mathbf{V}\mathbf{u} = \sum_{j=1}^{p} \lambda_j \mathbf{u}^\top \mathbf{p}_{:j}\mathbf{p}_{:j}^\top \mathbf{u} = \sum_{j=1}^{p} \lambda_j \langle \mathbf{u}, \mathbf{p}_{:j} \rangle^2 = \sum_{j=1}^{p} \lambda_j \underbrace{||\mathbf{p}_{:j}||^2 ||\mathbf{u}||^2}_{=1} \cos^2(\theta_j)$$

$$\leq \lambda_1 \sum_{j=1}^{p} \cos^2(\theta_j) \leq \lambda_1$$

  $$\Longrightarrow \mathbf{u}_1 = \arg\max_{\mathbf{u} \in \mathbb{R}^p} \mathbf{u}^\top \mathbf{V}\mathbf{u} = \mathbf{p}_{:1}, \quad \text{(pour } \theta_1 = 0\text{)}.$$

- Le 1er vecteur $\mathbf{u}_1$ maximisant l'inertie expliquée est le vecteur propre $\mathbf{p}_{:1}$ associé à la plus grande valeur propre λ_1. L'inertie expliquée par $\mathbf{u}_1$ est :

$$I^\perp_{\mathbf{u}_1} = \lambda_1.$$

Recherche des autres axes portant l'inertie maximale

- Le 2^{e} axe portant l'inertie maximale est un axe orthogonale au premier (i.e. au vecteur $\mathbf{u_1} = \mathbf{p_1}$) tel que :

$$\mathbf{u}^\top \mathbf{V}\mathbf{u} = \sum_{j=2}^{p} \lambda_j \cos^2(\theta_j) \leq \lambda_2 \sum_{j=2}^{p} \cos^2(\theta_j) \leq \lambda_2$$

$$\Longrightarrow \mathbf{u_2} = \mathbf{p}_{:2} \quad \text{et} \quad \mathbf{u_2} \perp \mathbf{u_1}.$$

- L'inertie expliquée par $\mathbf{u_2}$ vaut :

$$I^{\perp}_{\mathbf{u}_2} = \lambda_2.$$

Théorème

Le sous-espace F_K de dimension K portant l'inertie maximale est engendré par les K vecteurs propres associés aux K plus grandes valeurs propres de la matrice de covariance $\mathbf{V}$ du nuage des individus.

Axes principaux et composantes principales

- Les axes $\mathbf{u}_j = \mathbf{p}_{:j}$ sont appelés **axes factoriels** ou **axes principaux**.
- L'inertie expliquée par l'axe $\mathbf{u}_j$ est la valeur propre λ_j.
- L'**inertie expliquée** par le sous-espace factoriel F_K est :

$$I_{F_K^\perp} = \lambda_1 + \cdots + \lambda_K. \tag{9}$$

- Le **pourcentage d'inertie** expliquée par le sous-espace F_K est :

$$\frac{\lambda_1 + \cdots + \lambda_K}{\lambda_1 + \cdots + \lambda_p}. \tag{10}$$

- On appelle j-ième **composante principale** les coordonnées $\mathbf{c}_{:j} \in \mathbb{R}^n$ des n individus sur l'axe factoriel $\mathbf{u}_j$, qui sont les projections des individus sur ces axes :

$$\mathbf{c}_{:j} = \mathbf{Y}\mathbf{u}_j. \tag{11}$$

- Chaque ligne i de la matrice des composantes principales $\mathbf{C} = [\mathbf{c}_{:1}, \ldots, \mathbf{c}_{:K}]$ peut être vue comme la représentation de l'individu i dans un nouvel espace de dimesnion $K \leq p$.