Analyse de données

Javier Cuadrado <javiercuadrado@cnrs.fr>| Ritvikmath

2023-01-26

On note x les scalaires, $\bar{x}$ les vecteurs et $\bar{\bar{x}}$ les matrices.

1 Visualisation

1.1 Le problème de la projection

2 ACP des individus

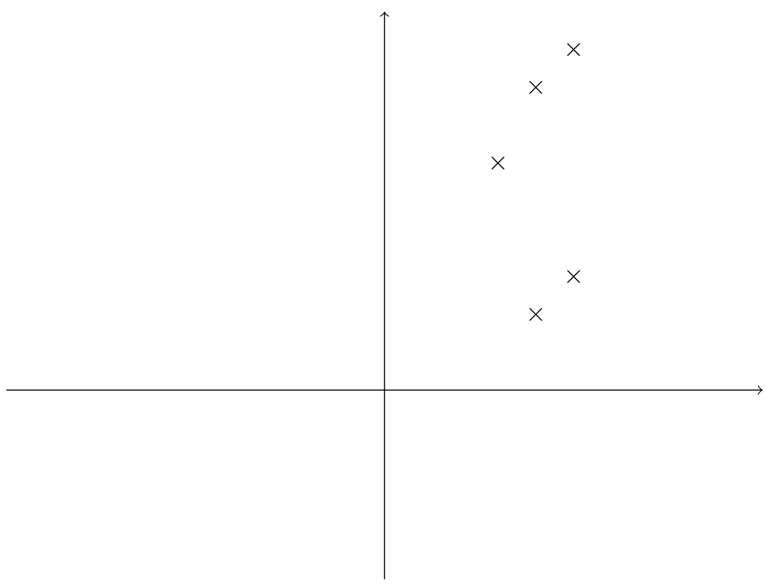

Figure 1:

On prend $\{x_n, n \in [\![1, N]\!]\} \in \mathbb{R}^{[\![1,N]\!]}, N \in \mathbb{N}$

On cherche à avoir la projection gardant le plus d'informations

$$\begin{aligned} \mathrm{Pro}_{S\bar{u}_i}(\bar{x}_i) &:= (\bar{u}^\top \bar{x})\bar{u}_i \\ \text{moyenne} &:= (\bar{u}_i^\top \langle \bar{x} \rangle)\bar{u}_i \end{aligned}$$

La variance de la projection est

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}(\bar{u}_i^\top \bar{x}_n - \bar{u}_i^\top \langle \bar{x}_i \rangle)^2 &= \frac{1}{N}\sum_{n=1}^{N}\bar{u}_i^\top (\bar{x}_n - \langle \bar{x} \rangle)^2 \bar{u}_i \\ &= \bar{u}_i^\top \left(\underbrace{\frac{1}{N}\sum_{n=1}^{N}(\bar{x}_n - \langle \bar{x} \rangle)^2}_{\bar{\bar{\Sigma}}} \right) \bar{u}_i\end{aligned}$$

Maximisation de $\bar{u}_i \bar{\bar{\Sigma}} \bar{u}_i$ On prend $\lambda \in \mathbb{R}$ et on utilise Lagrange (cf 2e année lol)

$$\begin{aligned}\bar{u}_i^* &= \max_{\bar{u}_i} \bar{u}_i^\top \bar{\bar{\Sigma}} \bar{u}_i + \lambda(1 - \bar{u}_i^\top \bar{u}_i) \\ \iff 2\bar{\bar{\Sigma}}\bar{u}_i^* - \lambda 2 \bar{u}_i^* &= 0 && \text{en dérivant selon } \bar{u}_i \\ \iff \bar{\bar{\Sigma}}\bar{u}_i^* &= \lambda \bar{u}_i^*\end{aligned}$$

On obtient une *équation aux vecteurs propres.*
On centre et on réduit les données:

$$x \leftarrow \frac{x - \langle x \rangle}{\sqrt{\sigma^2}}$$

Réduire les données peut être dangereux, car ça donne la même importance à toutes les variables

- Obligatoire si les variables ont des unités différentes
- Sinon, pas forcément pertinent

2.1 Exercice 1

Table 1: Données

	v_1	v_2	v_3
x_1	1	0	-1
x_2	2	1	-3
x_3	-1	-2	3
x_4	0	-1	1
x_5	-1	2	-1
x_6	-2	1	1
x_7	1	0	-1
x_8	0	1	1

$$
\begin{aligned}
\langle \bar{x} \rangle &= \frac{1}{N} \sum_{i=1}^{N} \bar{x}_i \\
&= \frac{1}{8} \sum_{i=1}^{N} \bar{x}_i \\
&= \begin{pmatrix} 2 & 3 & 4 \end{pmatrix}
\end{aligned}
$$

$$
\Sigma: \ \sigma_{XY} = (X - E(X))^\top (Y - E(Y))
$$

$$
\begin{aligned}
\Sigma &= \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i^\top \bar{Y}_i \\
&= \frac{1}{N} \bar{Y}^\top \bar{Y} \\
&= \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ -1 & -1 & 2 \end{pmatrix}
\end{aligned}
$$

Avec $Y = \left(\quad \text{(les données)} \quad \right)$

On trouve les éléments propres par un polynôme caractéristique

$$
\begin{aligned}
\chi_\Sigma &= 0 - \left(\frac{3}{2}\right)^3 \begin{vmatrix} 1-\lambda' & 0 & -1 \\ 0 & 1-\lambda' & -1 \\ -1 & -1 & 2-\lambda' \end{vmatrix} \qquad \text{avec } \lambda' = \frac{2}{3}\lambda \\
&= 0
\end{aligned}
$$

On en déduit $\mathrm{Sp}(\Sigma) = \frac{1}{2}\{9, 3, 0\}$

Puis $\begin{cases} \bar{u}_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} -1 & -1 & 2 \end{pmatrix} \\ \bar{u}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 & 0 \end{pmatrix} \\ \bar{u}_3 = \frac{1}{\sqrt{3}} \begin{pmatrix} -1 & -1 & -1 \end{pmatrix} \end{cases}$

On a $\bar{Y}_2 = \begin{pmatrix} 2 & 1 & -3 \end{pmatrix}$ $\bar{u}_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} -1 & -1 & 2 \end{pmatrix}$

Donc $\mathrm{CP}_i\,|_{X_2} = \bar{y}_2' \cdot \bar{u}_1 = -3.674$

$$
\mathrm{CT}_i\,|_{X_n} = \frac{-(\mathrm{CP}_i\,|_{X_n})^2}{\sum_{i=1}^{N} (CP_i|_{X_n})}
$$