Analyse en Composantes Principales (ACP)

Rachid EL MAAZOUZ

Août 2018

1 Approche géométrique

1.1

L'intérêt de centrer et réduire les données de chaque variable est d'étudier leurs variations par rapport a une valeur de référence(Moyennes et variances).

Soit A = $(a_{ij})_{1\leq i\leq n, 1\leq j\leq p} \in \mathbb{R}^{n\times p}$ la matrice des données où la i-ème ligne représente les données du i-ème pays pour chaque $1 \leq i \leq n$.

On définit la matrice identité I_n de $\mathbb{R}^{n\times n}$ et la matrice $J_n \in \mathbb{R}^{n\times n}$ comme suit:

$$J = \begin{pmatrix} 1/n & 1/n & \dots & 1/n \\ 1/n & 1/n & \dots & 1/n \\ \vdots & \vdots & \ddots & \vdots \\ 1/n & 1/n & \dots & 1/n \end{pmatrix}$$

L'idée est de centrer et réduire chaque colonne de la matrice A. Pour ce faire un ensemble de transformations seront appliquées sur la matrice de données A.

Soit I la matrice identité de taille n : $I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$, l'idée et de centrer les valeurs de chaque variable (colonne) autour de la moyenne des valeurs de cette variable, pour cela on doit calculer une matrice T qui comprend les moyennes de chaque variable sur la colonne qui lui correspond.

Soit J la matrice de taille n dont tous les coefficients sont 1/n : $J = \begin{pmatrix} 1/n & 1/n & \dots & 1/n \\ 1/n & 1/n & \dots & 1/n \\ \vdots & \vdots & \ddots & \vdots \\ 1/n & 1/n & \dots & 1/n \end{pmatrix}$

La matrice J permet de calculer les moyennes sur chaque colonne. En effet, la matrice:

$$\text{J*A} = \begin{pmatrix} 1/n & 1/n & \dots & 1/n \\ 1/n & 1/n & \dots & 1/n \\ \vdots & \vdots & \vdots & \vdots \\ 1/n & 1/n & \dots & 1/n \end{pmatrix} * \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \dots & \dots & a_{np} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n a_{k1}/n & \sum_{k=1}^n a_{k2}/n & \dots & \sum_{k=1}^n a_{kp}/n \\ \sum_{k=1}^n a_{k1}/n & \sum_{k=1}^n a_{k2}/n & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n a_{k1}/n & \sum_{k=1}^n a_{k2}/n & \dots & \sum_{k=1}^n a_{kp}/n \end{pmatrix}$$

donne une matrice de taille (n,p) dont les colonnes portes les moyennes de chaque variable (colonnes

de A).
La nouvelle matrice centrée à considérer $M = A - J * A = (I - J) * A$

Maintenant procédons à la réduction de la nouvelle matrice M. Il suffit de deviser ses valeurs par l'écart type calculé sur chaque colonne.

Soit N la matrice carrée de taille d, définie par: $N = \underset{i=1,\, j=1}{\overset{n,\, p}{(e_{ij})}}$ tel que: $e_{ii} = 1/\sqrt{\sum_{k=1}^{n} m_{ii}^2}$, 0 sinon
Calculons le produit $X = M * N$:

$$X = M*N = \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1d} \\ m_{21} & m_{22} & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & \dots & \dots & m_{nd} \end{pmatrix} * \begin{pmatrix} e_{11} & 0 & \dots & 0 \\ 0 & e_{22} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & e_{dd} \end{pmatrix} = \begin{pmatrix} m_{11} * e_{11} & m_{12} * e_{22} & \dots & m_{1d} * e_{dd} \\ m_{21} * e_{11} & m_{22} * e_{22} & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ m_{n1} * e_{11} & m_{n2} * e_{22} & \dots & m_{nd} * e_{dd} \end{pmatrix}$$

Au final, les colonnes de la matrice X sont les colonnes de la matrice M divisées par l'écart type calculé sur chaque colonne correspondante. X est la matrice centrée réduite de la matrice originale A.

1.2

Calculons $I_{D_u}(X) = \sum_{k=1}^{n} \|\vec{p}_{D_u}(x_i)\|^2$

$$\begin{aligned} I_{D_u}(X) &= \sum_{k=1}^{n} \|\vec{p}_{D_u}(x_i)\|^2 \\ &= \sum_{k=1}^{n} \langle x_i, u \rangle^2 \\ &= \sum_{k=1}^{n} (u^{\mathrm{t}}.x_i)^2 \\ &= \sum_{k=1}^{n} (u^{\mathrm{t}}.x_i)(u^{\mathrm{t}}.x_i) \\ &= \sum_{k=1}^{n} (u^{\mathrm{t}}.x_i)(x_i^{\mathrm{t}}.u) \\ &= \sum_{k=1}^{n} u^{\mathrm{t}}.(x_i.x_i^{\mathrm{t}}).u \\ &= u^{\mathrm{t}}.[\sum_{k=1}^{n} (x_i.x_i^{\mathrm{t}})].u \end{aligned} \tag{1}$$

Avec la matrice $\Sigma = \sum_{k=1}^{n} (x_i.x_i^{\mathrm{t}})$, l'équation (1) peut s'écrire comme suit:
$I_{D_u}(X) = u^{\mathrm{t}}.\Sigma.u$

L'intérêt de calcul de la projection $I_{D_u}(X)$ est de savoir les valeurs de u qui permettent de maximiser $I_{D_u}(X)$.

Considérons le Lagrangien $L(\lambda, u) = I_{D_u}(X) - \lambda.(u^t.u - 1)$, avec la condition $u^t.u = 1$.

Soit $L(\lambda, u) = u^t.\Sigma.u - \lambda.(u^t.u - 1)$:

$$\frac{\partial L(\lambda, u)}{\partial u} = 0 \Rightarrow 2.\Sigma.u - 2.\lambda.u = 0$$
$$\Rightarrow \Sigma.u = \lambda.u$$

Donc la valeur maximale sous contrainte $u^t.u = 1$ de $I_{D_u}(X)$ est atteinte sur les vecteurs propres de la matrice Σ puisque diagonalisable et ses valeurs propres sont positives ou nulles(matrice symétrique et définie positive donc). Ces vecteurs propres correspondent aux valeurs propres λ_i de la matrice Σ.

Pour visualiser les points dans un plan R^3, on prend les 3 premières projections qui correspondent aux 3 grandes valeurs propres de la matrice Σ.

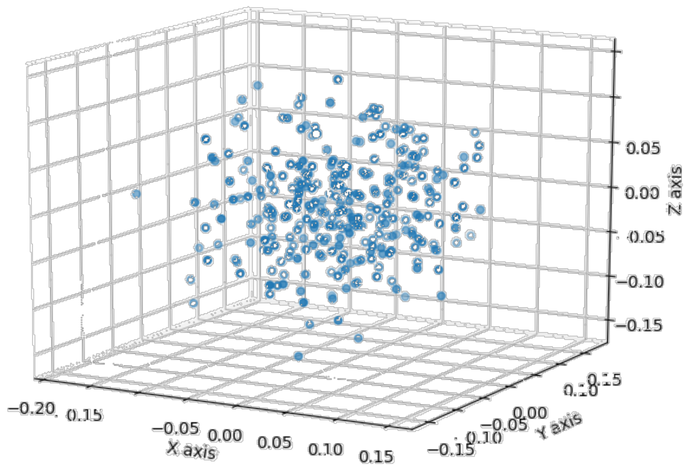

Figure 1: Représentation en base PCA

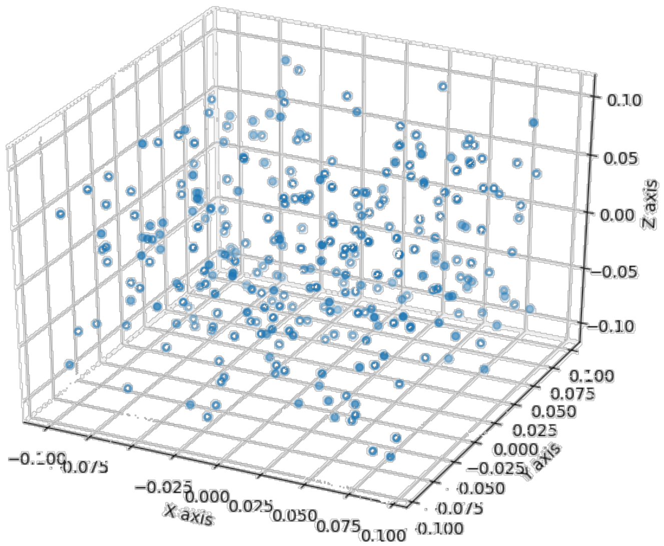

Figure 2: Représentation en base canonique de R^3

La dispersion des points sur la base canonique est plus forte que sur celle de base générée par les 3 premiers vecteurs propres de la matrice Σ, e qui confirme le principe de la méthode PCA: visualiser le maximum de points dans un espace minimal.