Compléments sur l'ACP

I] Décomposition en valeurs singulières :

__Théorème :__ A matrice $n \times p$ à valeurs réelles.

. $r = \text{rang}(A) \leq \min(n, p)$.

. $U = \left(u_1 \middle| u_2 \middle| \cdots \middle| u_r \right) \updownarrow n$ matrice de taille $n \times r$

constituée d'une B.O.N (Base orthonormée) de vecteurs propres de $X . X^T$ (matrice $n \times n$ symétrique positive de rang r) associés aux r valeurs propres > 0 rangées dans l'ordre décroissant $(\lambda_1 \geq \lambda_2 \cdots \geq \lambda_r > 0)$

. $V = \left(v_1 \middle| \cdots \middle| v_r \right) \updownarrow p$ matrice $p \times r$,

constituée d'une BON de vecteurs propres de $X^T . X$ $(p \times p)$ associées à $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_r > 0$ (mêmes valeurs propres pour XX^T et X^TX)

. $D = \begin{pmatrix} \sigma_1(A) & & 0 \\ & \ddots & \\ 0 & & \sigma_r(A) \end{pmatrix}$ où $\sigma_i(A) = \sqrt{\lambda_i}$

. Alors $A = U . D . V^T$

$$= \sum_{i=1}^{r} \sigma_i(A) \, u_i . v_i^T.$$

II] THÉORÈME d'APPROXIMATION MATRICIELLE :

Pour une matrice A $n \times p$ de colonnes $a_1, -, a_p$

On note $\|A\|_F^2 = \sum_{j=1}^{p} \|a_j\|^2 = \sum_{i,j=1}^{p} A_{i,j}^2 = Tr(A.A^T)$

$\|\cdot\|_F$ est appelée norme de Frobenius de la matrice A.

(et il s'agit bien d'une norme). Ainsi, pour

2 matrices A et B de même taille

$$\|A - B\|_F^2 = \sum_{j=1}^{p} \|a_j - b_j\|^2$$

où a_j et b_j désignent les colonnes de A et B.

<u>Remarque</u> : Si $\tilde{a}_i$ et $\tilde{b}_i$, $i = 1, -, n$, correspondent aux lignes de A et B, on a aussi

$$\|A - B\|_F^2 = \sum_{i=1}^{n} \|\tilde{a}_i - \tilde{b}_i\|^2$$

<u>THÉORÈME</u> : A matrice $n \times p$ de rang r, $d < r$.

$$\mathcal{Z}_d = \{ Z \text{ de taille } n \times p,\ rg(Z) \leq d \}.$$

Alors

$$\inf_{Z \in \mathcal{Z}_d} \|A - Z\|_F^2 = \sum_{k=d+1}^{r} \sigma_k^2(A)$$

De plus, l'inf est atteint ~~de~~ en :

$$A_d = \sum_{k=1}^{d} \sigma_k(A) . u_k . v_k^T$$

<u>PREUVE</u> : Voir Giraud Théorème C.5.

III. APPLICATION à l'ACP :

X matrice des observations de taille $n \times p$.

. On note $X^{(1)}, \dots, X^{(n)}$ ses lignes.

. S_d : sous-espace vectoriel de dimension $\leqslant d$ de $\mathbb{R}^p$.

. $P_{S_d}(X) = \begin{pmatrix} P_{S_d}(X^{(1)}) \\ \hline \vdots \\ \hline P_{S_d}(X^{(n)}) \end{pmatrix}$ matrice des projections orthogonales des lignes de X

. $P_{S_d}(X)$ est de rang $\leqslant d$ donc appartient à $\mathcal{L}_d$.

Comme de plus, on a

$$\sum_{i=1}^{n} \| X^{(i)} - P_{S_d}(X^{(i)}) \|^2 = \| X - P_{S_d}(X) \|_F^2$$

on déduit du théorème précédent le corollaire suivant :

COROLLAIRE :

$$\inf_{\substack{\text{sev } S_d \\ \text{de dim} \leq d}} \sum_{i=1}^{n} \| X^{(i)} - P_{S_d}(X^{(i)}) \|^2 \text{ est atteint}$$

en $X_d = \sum_{k=1}^{d} \sigma_k(X) \cdot u_k \cdot v_k^T$

Preuve du Corollaire :

D'après le théorème, on a clairement :

$$\inf_{\substack{\text{sev } S_d \\ \text{de dim} \leq d}} \sum_{i=1}^{n} \| X^{(i)} - P_{S_d}(X^{(i)}) \|^2 \geq \inf_{Z \in \mathcal{L}_d} \| X - Z \|_F^2 .$$

Or,

$$X_d \overset{(*)}{=} \begin{pmatrix} P_{S_d^*}(X^{(1)}) \\ \hline \vdots \\ \hline P_{S_d^*}(X^{(n)}) \end{pmatrix} = P_{S_d^*}(X)$$

où $S_{d^*} = \mathrm{Vect}\{v_1, \ldots, v_d\}$ d'où le résultat.

$(*)$: Prouvons cette égalité. Par construction,

$$X^{(i)} = \sum_{j=1}^{r} c_j(i) \cdot v_j$$

où $\boxed{c_j(i) := \sigma_j(X) . u_j(i)}$

et

$$X_d^{(i)} = \sum_{j=1}^{d} c_j(i) . v_j$$

Ligne i de la matrice X_d

Ainsi $X_d^{(i)} \in \mathrm{Vect}\{v_1, \ldots, v_d\}$

et $X^{(i)} - X_d^{(i)} \in \mathrm{Vect}\{v_{d+1}, \ldots, v_r\}$

Comme la famille $\{v_1, \ldots, v_r\}$ est orthogonale, il vient

$$X_d^{(i)} \perp X^{(i)} - X_d^{(i)} \implies X_d^{(i)} = P_{S_{d^*}}(X^{(i)}).$$

- Les vecteurs propres $v_1, \dots, v_d$ sont appelés les AXES PRINCIPAUX.
- Les vecteurs (propres) $u_1, \dots, u_d$ sont appelés les FACTEURS PRINCIPAUX.
- $\sigma_1, \dots, \sigma_r$ sont appelés les valeurs singulières.
- $c_1, \dots, c_r$ sont les composantes principales.

$$c_j = \sigma_j(X)\, u_j = X v_j$$

(car $X v_j = X \lambda_j^{-1/2} X^T u_j = \sqrt{\lambda_j}\, u_j$)

- Part d'INERTIE : $$r_d = \frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$
- Règle d'ARRET : EBOULIS des valeurs propres.
- VOIR PAR EXEMPLE slides GADAT

http://perso.math.univ-toulouse.fr/gadat/files/2012/12/04Acp.pdf

04_Acp.pdf