ANALYSE en COMPOSANTES PRINCIPALES

1. Cas général de l'ACP
2. ACP bipondérée
3. ACP simple (covariances)
4. ACP standard (corrélations)

1 ACP : Cas général

Protocole multivarié de K variables sur (J, n_J) : $(x^J k)_{k \in K}$

Remplacer les K variables par L' nouvelles variables

1.1 Du protocole multivarié au nuage de points

Nuage de points (M^J, n_J) dans un espace affin $\mathcal{U}$ de dimension K muni du repère cartésien $(\mathrm{O}, (\overrightarrow{\delta_k})_{k \in K})$.

Le profil $x^{jK} = (x^{jk})_{k \in K}$ de j est représenté, par le point pondéré (M^j, n_j) : $\overrightarrow{\mathrm{OM}}^j = \sum\limits_{k \in K} x^{jk} \overrightarrow{\delta_k}$.

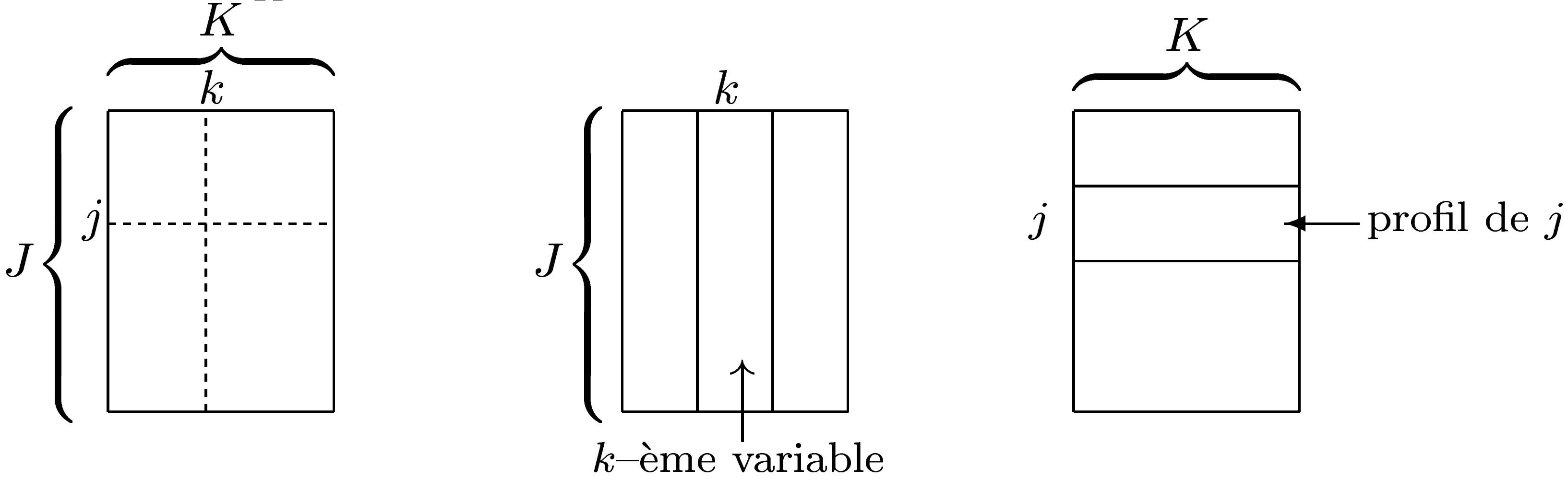
Nuage euclidien : $(\mathrm{M}^j \mathrm{M}^{j'})^2 = \sum\limits_{k \in K} \sum\limits_{k' \in K} q_{kk'} (x^{jk} - x^{j'k})(x^{jk'} - x^{j'k'})$

avec $q_{kk'} = \langle \delta_k^K | \delta_{k'}^K \rangle$

Effectuer l'*analyse en composantes principales* du protocole multivarié x^{JK} c'est déterminer les variables principales du nuage euclidien (M^J, n_J), selon les méthodes présentées au chapitre NUAGE.

2 ACP bipondérée

Protocole de notes x^{JK}, muni de la mesure–effectifs n_J et de la pondération ϖ_K



Distance entre profils : $d(j,j') = \Big(\sum_{k\in K} \varpi_k (x^{jk} - x^{j'k})^2\Big)^{1/2}$

Espace affin $\mathcal{U}$ de dimension K muni du repère cartésien $(\mathrm{O}, (\overrightarrow{\delta_k})_{k\in K})$

$\overrightarrow{\mathrm{OM}^j} = \sum_{j\in J} x^{jk}\overrightarrow{\delta_k}$

Point moyen $\mathrm{G} = \sum_{j\in J} f_j \mathrm{M}^j$, d'où $\overrightarrow{\mathrm{OG}} = \sum_{i\in I} f_j \overrightarrow{\mathrm{OM}}^j = \sum_{k\in K} \overline{x}^k \overrightarrow{\delta_k}$

$(\mathrm{M}^j\mathrm{M}^{j'})^2 = \sum_k \varpi_k (x^{jk} - x^{j'k})^2$

Pour $k \neq k'$, on a $\overrightarrow{\delta_k} \perp \overrightarrow{\delta_{k'}}$ et $\|\overrightarrow{\delta_k}\| = \sqrt{\varpi_k}$.

La matrice $\mathbf{Q}$ des produits scalaires est diagonale, notée $\mathbf{\Omega}_K$, avec $q_{kk} = \varpi_k$ et $q_{kk'} = 0$ pour $k \neq k'$.

Variance du nuage : $\sum_{k \in K} \varpi_k \operatorname{Var} x^{Jk}$.

Contribution de la variable x^{Jk} à la variance du nuage :
$\mathrm{Cta}_k = \varpi_k \operatorname{Var} x^{Jk}$.

Nous appellerons l'ACP d'un tel protocole : ACP *pondérée*.

2.1 Directions et variables principales

Formules de passage

$$Vac : \overrightarrow{\delta_k} \mapsto \left(\langle \overrightarrow{\delta}_k | \overrightarrow{\mathrm{GM}}^j \rangle\right)_{j \in J} = \varpi_k x_0^{Jk}$$

$$Vac^\star : \delta_J^j \mapsto f_j \overrightarrow{\mathrm{GM}^j} = \sum_{k_i n K} ((f_j\, x_0^{jk}) \delta_J^j$$

$$Som = Vac^\star \circ Vac : \overrightarrow{\delta}_k \mapsto \sum_{k' \in K} \varpi_k v^{kk'} \overrightarrow{\delta}^{k'}$$

$$\begin{tikzcd}
 & \mathbb{R}_J \arrow[dl, "Eff"'] \\
\mathcal{V} \arrow[dr, shift right, "Vac"'] & \\
 & \mathbb{R}^J \arrow[ul, shift right, "Vac^*"'] \arrow[uu, "f_J"']
\end{tikzcd}$$

Variable $a^K = \sum_{k \in K} a^k\, \delta_k^K \longrightarrow$ vecteur géométrique $\overrightarrow{\alpha} = \sum_{k \in K} a^k \overrightarrow{\delta_k}$.

Equation aux directions et valeurs propres :

$$Som(\overrightarrow{\alpha}) = \lambda \overrightarrow{\alpha} \text{ soit } \sum_{k' \in K} v^{kk'} a_{k'} = \lambda a^k \quad \text{avec : } a_k = a^k \varpi_k$$

Variable principale calibrée (formule de passage) :

$y_\ell^J = \sum_{k \in K} a_{k\ell}\, x_0^{Jk}$, avec $\sum_{k \in K} \varpi_k (a_\ell^k)^2 = \sum_{k \in K} \frac{a_{k\ell}^2}{\varpi_k} = 1$

y_ℓ^j est la coordonnée du point M^j sur l'axe principal ℓ

$$\mathrm{Moy}\, y_\ell^J = 0 \text{ et } \mathrm{Var}\, y_\ell^J = \lambda_\ell$$

2.2 Formules de reconstitution

Reconstitution d'ordre L' du tableau des variables centrées :

$$\widetilde{x}^{jk}_{L'} = \sum_{\ell=1}^{L'} a^k_\ell y^j_\ell \qquad \text{avec } \sum_{k\in K} \varpi_k (a^k_\ell)^2 = 1$$

Reconstitution des distances :

$$(\mathrm{GM}^j)^2 = \sum_{k\in K} \varpi_k (x^{jk} - \overline{x}^k)^2 = \sum_{\ell=1}^{L} (y^j_\ell)^2$$

Variance du nuage : $\mathrm{Var}\,\mathrm{M}^J = \sum_{k\in K} \varpi_k \,\mathrm{Var}\, x^{Jk} = \sum_{\ell=1}^{L} \lambda_\ell$

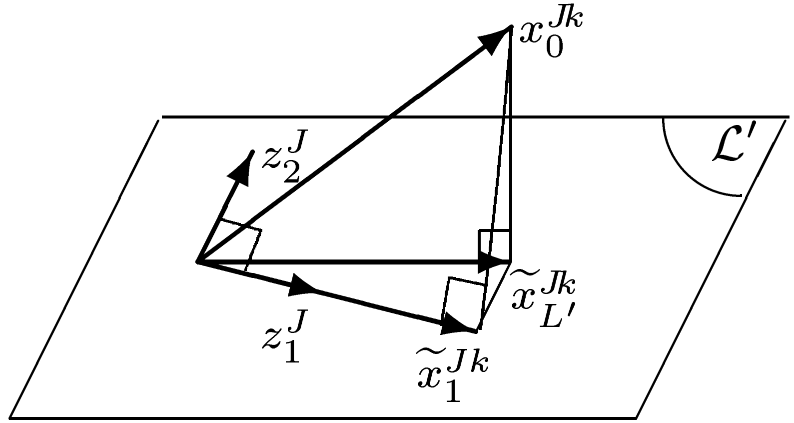
2.3 Espace des variables : caractérisation statistique

En termes statistiques, le problème de l'ACP pondérée se formule :
Problème : parmi les variables $\sum\limits_{k\in K} a_k\, x^{Jk}$, combinaisons linéaires des variables initiales $(x^{Jk})_{k\in K}$, chercher celle(s) qui vérifie(nt) :
$\mathrm{Var}(\sum\limits_{k\in K} a_k\, x^{Jk})/(\sum\limits_{k\in K} a_k^2/\varpi_k)$ maximum.

Chercher $\mathrm{Var}(\sum\limits_{k\in K} a_k\, x^{Jk})/(\sum\limits_{k\in K} a_k^2/\varpi_k)$ maximum revient à déterminer la première variable principale, puisque cette statistique est la variance du nuage dans la direction $\overrightarrow{\alpha}$.

2.4 Espace des variables

Régression des variables initiales sur les variables principales.



Formule de reconstitution : $x_0^{Jk} = \sum_{\ell=1}^{L} b_\ell^k z_\ell^J$ avec $b_\ell^k = a_\ell^k \xi_\ell$

Le coefficient de régression partielle de x_0^{Jk} est le coefficient de régression simple $b_\ell^k = \mathrm{Cov}(x^{Jk}|z_\ell^J)$, avec $\mathrm{Var}\, x^{Jk} = \sum_{\ell=1}^{L} (b_\ell^k)^2$

Corrélation.

$r_\ell^k = \text{Corr}\,(x^{Jk}|z_\ell^J) = b_\ell^k /\, \text{Ety}\, x^{Jk}$

$\sum_{\ell=1}^{L} (r_\ell^k)^2 = 1$

r_ℓ^k = au cosinus de l'angle entre les variables (vecteurs) x_0^{Jk} et z_ℓ^J

Corrélation multiple R entre x^{Jk} et les L' premières variables principales : $R^2 = \sum_{\ell=1}^{L'} (r_\ell^k)^2 = \sum_{\ell=1}^{L'} (b_\ell^k)^2 / \text{Var}\, x^{Jk}$;

2.4.1 Variable moyenne

Posons $\varpi = \sum\limits_{k \in K} \varpi_k$ et notons $\widehat{x}^j$ la ϖ_K–moyenne des notes de l'individu j, d'où la variable des moyennes, ou *variable moyenne* $\widehat{x}^J = (\widehat{x}^j)_{j\in J}$ avec : $\widehat{x}^j = \sum\limits_{k\in K} \varpi_k\, x^{jk}/\varpi$.

On a : $\mathrm{Moy}\,\widehat{x}^J = \sum\limits_{k\in K} \varpi_k \widehat{x}^k/\varpi$, et $\mathrm{Var}\,\widehat{x}^J = \sum\limits_{k\in K}\sum\limits_{k'\in K} v^{kk'}\varpi_k\varpi_{k'}/\varpi^2$.

La variable moyenne (centrée) est *principale* si et seulement si $\mathrm{Cov}(x^{Jk}|\widehat{x}^J)$ ne dépend pas de k.

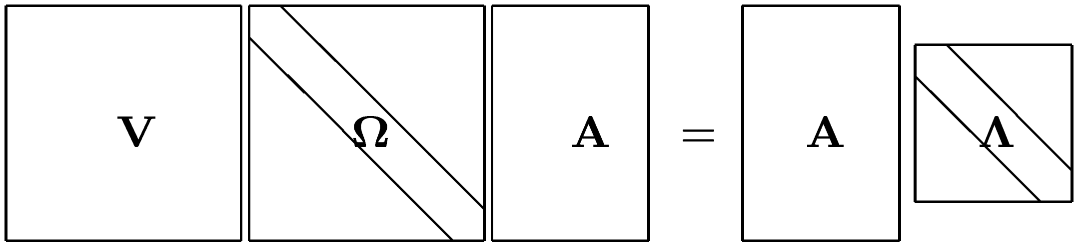
La variable principale calibrée est alors $y_\ell^J = \sqrt{\varpi}\,\widehat{x}_0^J$, associée à la valeur propre $\lambda_\ell = \varpi\,\mathrm{Var}\,\widehat{x}^J$.

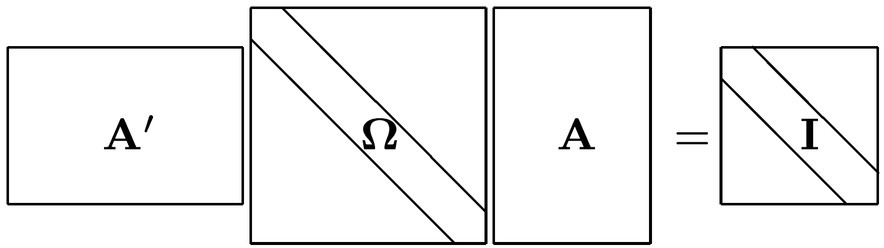
Si $\widehat{x}_0^J$ est variable principale associée à la valeur propre λ_ℓ, toute variable principale associée à $\lambda_{\ell'} \neq \lambda_\ell$ est l' *effet d'un contraste* sur K appliqué aux K variables initiales.

2.5 Formulaire : *diagrammes résumés des formules principales*

Equation aux directions et valeurs propres :

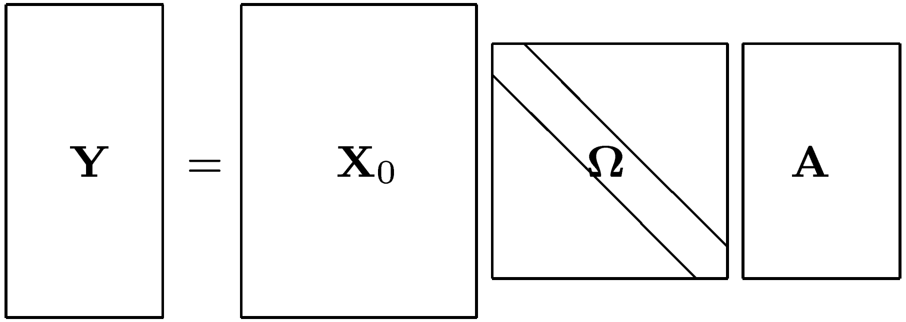
$$\sum_{k' \in K} v^{kk'} \left(\varpi_{k'} a_\ell^{k'}\right) = \lambda_\ell a_\ell^k \qquad \text{avec} \qquad \sum_{k \in K} \varpi_k (a_\ell^k)^2 = 1$$





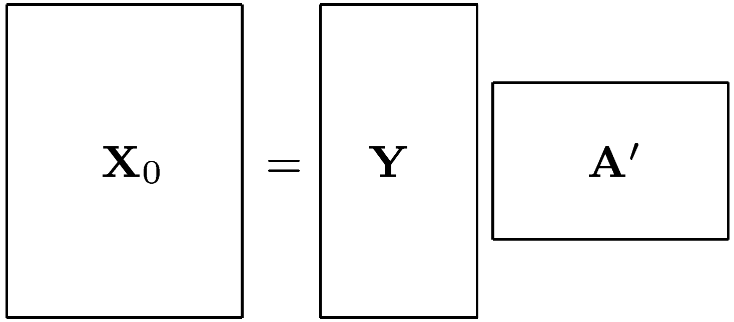
Coordonnées principales

$$y_\ell^J = \sum_{k \in K} x_0^{Jk} \left(\varpi_k a_\ell^k\right)$$

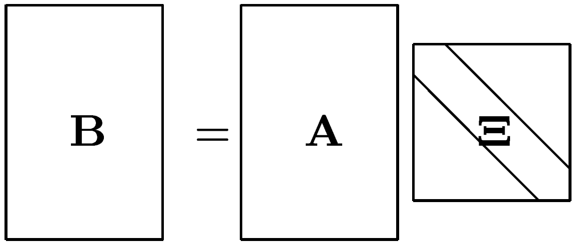


Formules de reconstitution

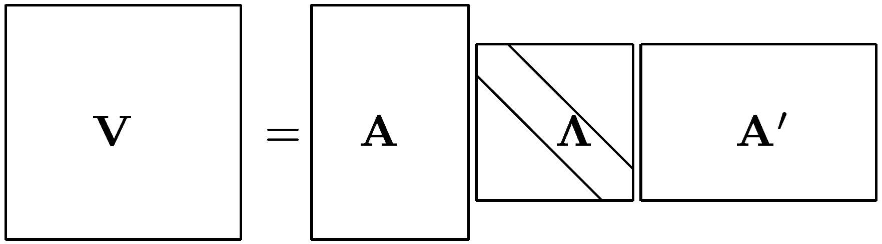
$$x_0^{Jk} = \sum_{\ell=1}^{L} a_\ell^k y_\ell^J$$



$$b_\ell^k = \xi_\ell a_\ell^k$$

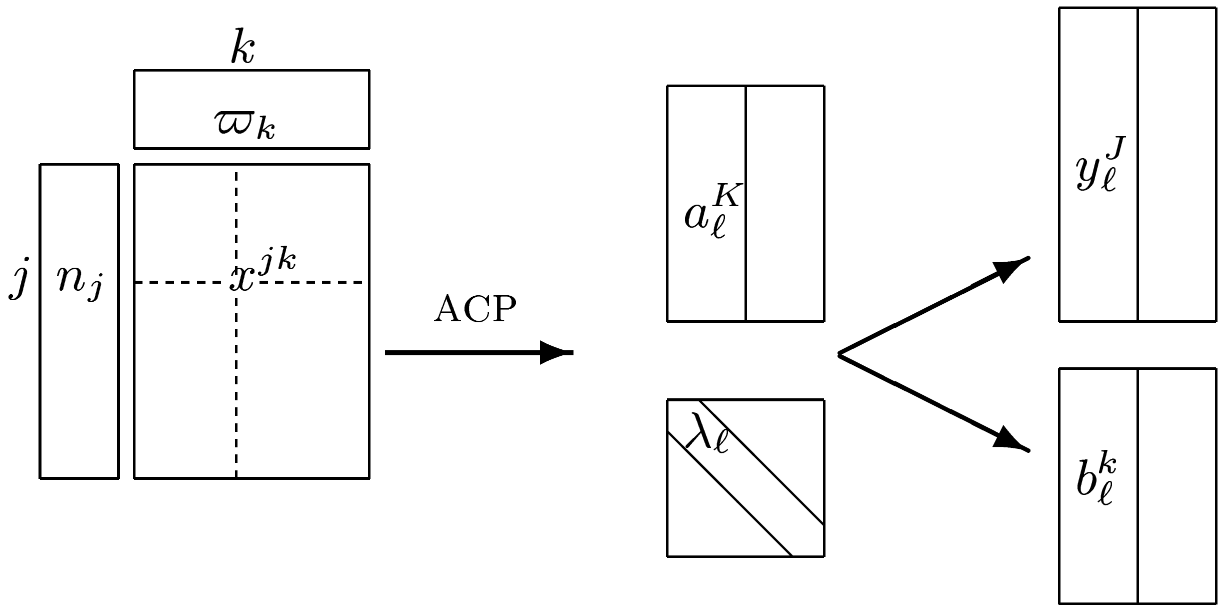


$$\upsilon^{kk'} = \sum_{\ell=1}^{L} \lambda_\ell a_\ell^k a_\ell^{k'}$$



2.6 Principaux résultats

- vecteurs et valeurs propres λ_ℓ ;
- coordonnées (y^j_ℓ) des $/J/$ points du nuage sur les L' $(L' \leq L)$ axes principaux retenus pour l'interprétation ;
- coefficients de régression (b^k_ℓ) des $/K/$ variables initiales sur les L' premières variables principales réduites.

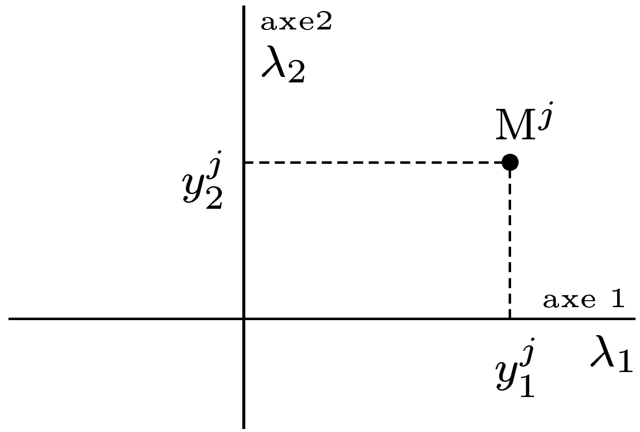


$$y_\ell^J = \sum_{k \in K} (\varpi_k a_\ell^k) x_0^{Jk}$$

$$b_\ell^K = \sqrt{\lambda_\ell}\, a_\ell^K$$

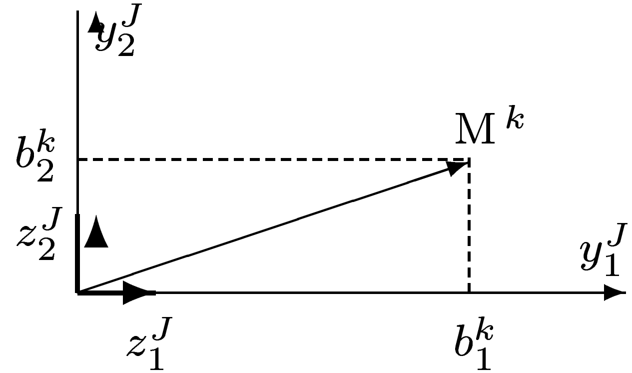
2.7 Représentations graphiques

Espace d'observables, projection du nuage des individus sur les premiers axes 1, 2, etc., ou dans les plans principaux 1-2, 1-3, 2-3, etc.
M^j est représenté par le point d'abscisse y_1^j et d'ordonnée y_2^j, etc.



Espace des variables

Représentation des régressions des K variables initiales centrées variables principales 1, 2, 3, etc. Coordonnées : (b_1^k, b_2^k) par rapport à (z_1^J, z_2^J).



Procédure de calcul.

1. On calcule la matrice $\mathbf{S}$ de terme général $s_{kk'} = \sqrt{\varpi_k\,\varpi_{k'}}\,v^{kk'}$.
2. On diagonalise $\mathbf{S}$, d'où : $(\lambda_\ell, (c_{k\ell})_{k\in K})_{\ell=1,\dots L}$, avec $\sum\limits_{k\in K} c_{k\ell}^2 = 1$.
3. Pour $\ell = 1, \dots L'$ $(\leq L)$ et $j \in J$, on calcule :
   $y_\ell^j = \sum\limits_{k\in K} \sqrt{\varpi_k}\,c_{k\ell}\,x_0^{jk}$.
4. Pour $\ell = 1, \dots L'$ et $k \in K$, on calcule : $b_\ell^k = \sqrt{\lambda_\ell}\,c_{k\ell}/\sqrt{\varpi_k}$.[a]

[a] Un programme d'ACP pondérée, réalisé avec P. Bonnet, est disponible auprès des auteurs.

Formulations matricielles. $\mathbf{c}_\ell = K$–colonne $(c_{k\ell})_{k\in K}$,

$\mathbf{C}$: matrice $K \times L'$ des K–colonnes $(\mathbf{c}_\ell)_{\ell=1,\dots L'}$;

$\mathbf{y}_\ell$ la J–colonne des coordonnées principales $(y^j_\ell)_{j\in J}$;

$\mathbf{b}_\ell$ la K–colonne des coefficients de régression $(b^k_\ell)_{k\in K}$;

$\boldsymbol{\Omega}_K^{1/2}$ la matrice diagonale $(\sqrt{\omega_k})_k \in K$; on a :

1. $\mathbf{S} = \boldsymbol{\Omega}_K^{1/2}\mathbf{V}\boldsymbol{\Omega}_K^{1/2}$.
2. $\mathbf{SC} = \mathbf{C}\boldsymbol{\Lambda}$ avec $\mathbf{C}'\mathbf{C} = \mathbf{I}$.
3. Pour $\ell = 1, \dots L'$: $\mathbf{y}_\ell = \mathbf{X}_0\boldsymbol{\Omega}_K^{1/2}\mathbf{c}_\ell$.
4. Pour $\ell = 1, \dots L'$: $\mathbf{b}_\ell = \sqrt{\lambda_\ell}\,\boldsymbol{\Omega}_K^{-1/2}\mathbf{c}_\ell$.

3 ACP simple ou des covariances

Toutes les variables sont affectées de poids ϖ_k égaux à 1 (J

$(\mathrm{M}^j\mathrm{M}^{j'})^2 = \sum_{k\in K}(x^{jk} - x^{j'k})^2$ (métrique euclidienne

élémentaire). La base $(\overrightarrow{\delta_k})_{k\in K}$ est orthonormée ($\mathbf{Q} = \mathbf{I}_K$).

Variance du nuage = somme des variances des variables initiales ($\mathrm{Var}\,\mathrm{M}^J = \sum_{k\in K} \mathrm{Var}\, x^{Jk}$)

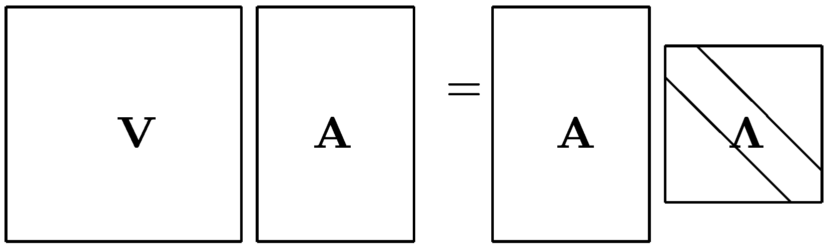
Equation aux directions et valeurs propres : $\sum_{k'\in K} v^{kk'} a_{k'\ell} = \lambda_\ell a_\ell^k$

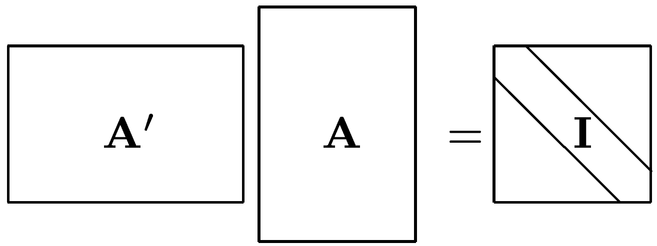
Dans l'ACP simple, directions principales et valeurs propres s'obtiennent par diagonalisation de la matrice **V** *des variances et covariances entre les variables initiales : l'ACP simple est l'*analyse des covariances.

3.1 Formulaire de l'ACP simple

Equation aux directions et valeurs propres :

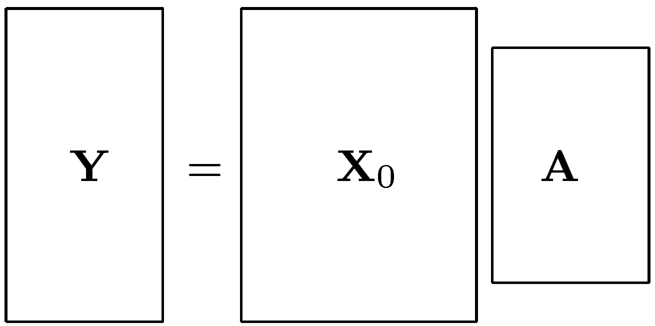
$$\sum_{k' \in K} v^{kk'} a_{k'\ell} = \lambda_\ell a_\ell^k \qquad \text{avec} \qquad a_\ell^k = a_{k\ell} \ \text{ et } \ \sum_{k \in K} (a_\ell^k)^2 = 1$$





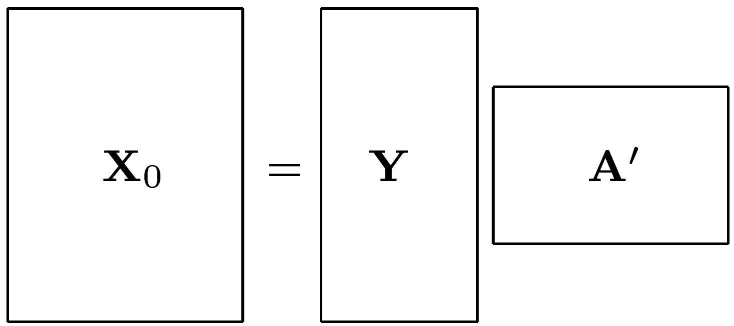
Coordonnées principales

$$y_\ell^J = \sum_{k \in K} a_{k\ell} x_0^{Jk}$$

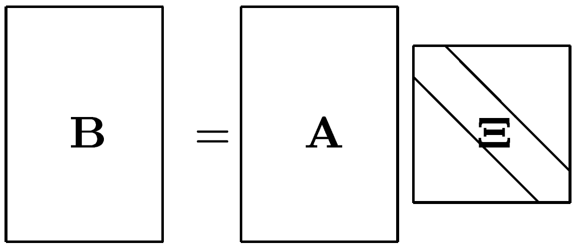


Formules de reconstitution

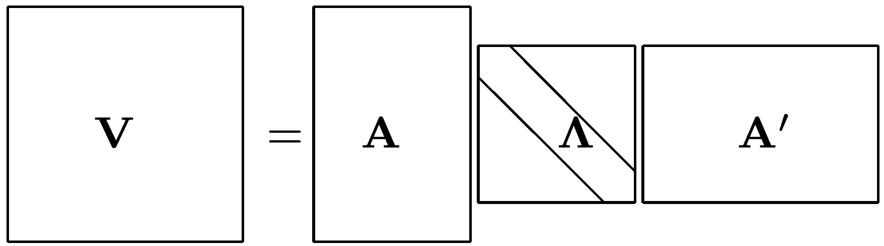
$$x_0^{Jk} = \sum_{\ell=1}^{L} a_\ell^k y_\ell^J$$



$$b_\ell^k = \xi_\ell a_\ell^k$$



$$v^{kk'} = \sum_{\ell=1}^{L} \lambda_\ell a_\ell^k a_\ell^{k'}$$



3.2 Procédure de calcul

1. On calcule la matrice $\mathbf{V}$ des covariances.
2. On diagonalise $\mathbf{V}$, d'où $(\lambda_\ell, (a_{k\ell})_{k\in K})_{\ell=1,\dots L}$ avec $\sum_{k\in K}(a_{k\ell})^2 = 1$.
3. Pour $\ell = 1, 2, \dots L'$ $(\leq L)$, on calcule $\forall j \in J$: $y_\ell^j = \sum_{k\in K} a_{k\ell} x_0^{jk}$.
4. Pour $\ell = 1, 2, \dots L'$ $(\leq L)$, on calcule $\forall k \in K$: $b_\ell^k = \sqrt{\lambda_\ell}\, a_{k\ell}$.

4 ACP standard ou des corrélations

Variables hétérogènes : ramener à une même échelle par une procédure de solidarisation

L'ACP standard revient donc d'une part à *solidariser par réduction* les échelles des variables, d'autre part à donner à ces variables des *poids égaux à 1* ; les variables ont donc la même contribution à la variance du nuage.

Dans l'ACP standard, directions principales et valeurs propres s'obtiennent par diagonalisation de la matrice des corrélations entre les variables initiales : l'ACP standard est l'*analyse des corrélations*.

4.1 Propriétés de l'ACP standard

• La variance du nuage = somme des valeurs propres = nombre de variables :

• Les coefficients de régression $b_\ell^k = \xi_\ell a_\ell^k$ sont les coefficients de corrélation entre variables initiales et variables principales : $b_\ell^k = r_\ell^k$.

Les coefficients a_ℓ^k vérifient donc la double propriété :

$$\forall \ell \ : \ \sum_{k \in K} (a_\ell^k)^2 = 1 \text{ et } \forall k \ : \ \sum_{\ell=1}^{L} \lambda_\ell (a_\ell^k)^2 = 1 (= \sum_{\ell=1}^{L} (b_\ell^k)^2)$$

On a $(a_\ell^k)^2 = \mathrm{Ctr}_k^\ell$ (contribution relative de la variable k à l'axe ℓ) ; et $(b_\ell^k)^2 = \lambda_\ell (a_\ell^k)^2 = \mathrm{Ctr}_\ell^k$ (contribution relative de l'axe ℓ à la variable k).

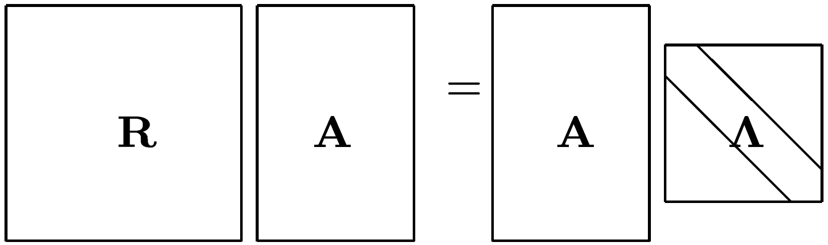
• Le *coefficient de corrélation multiple* $R_{L'}$ entre la variable initiale x^{Jk} et les L' premières variables principales est donné par

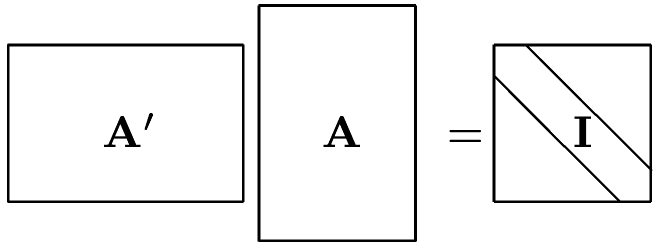
$$R_{L'}^2 = \sum_{\ell=1}^{L'} (r_\ell^k)^2 = \sum_{\ell=1}^{L'} (b_\ell^k)^2.$$

4.2 Formulaire de l'ACP standard

Equation aux directions et valeurs propres :

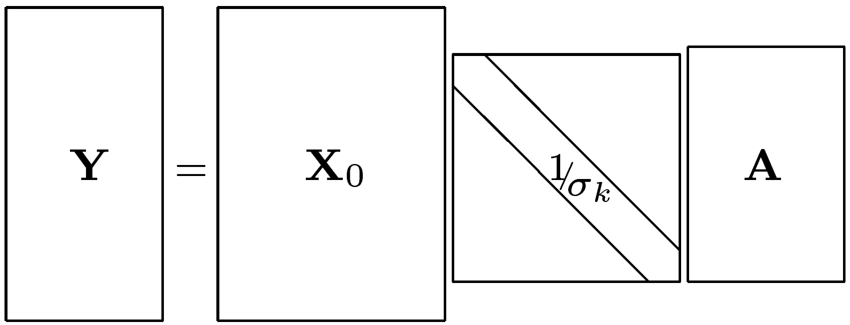
$$\sum_{k' \in K} r^{kk'} a_{k'\ell} = \lambda_\ell a_\ell^k \qquad \text{avec} \qquad \sum_{k \in K} (a_\ell^k)^2 = \sum_{k \in K} (a_{k\ell})^2 = 1$$





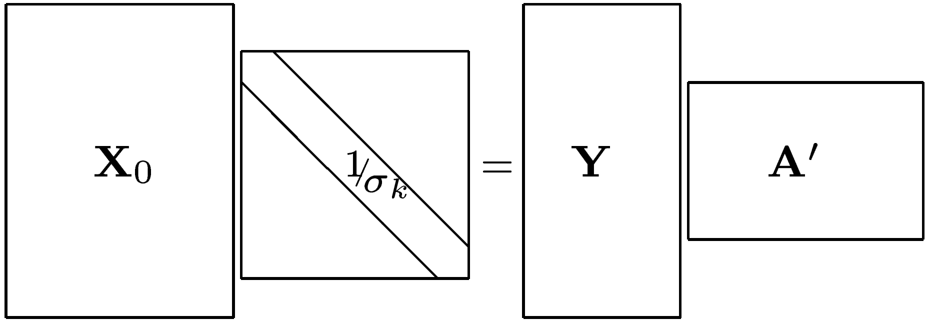
Coordonnées principales

$$y_\ell^J = \sum_{k \in K} a_{k\ell} \frac{x_0^{Jk}}{\sigma_k}$$

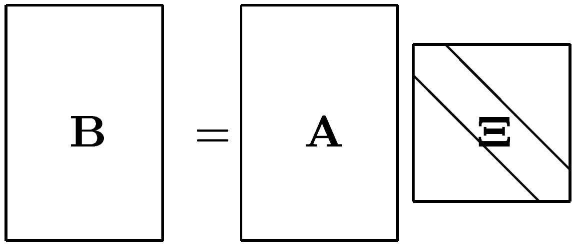


Formules de reconstitution

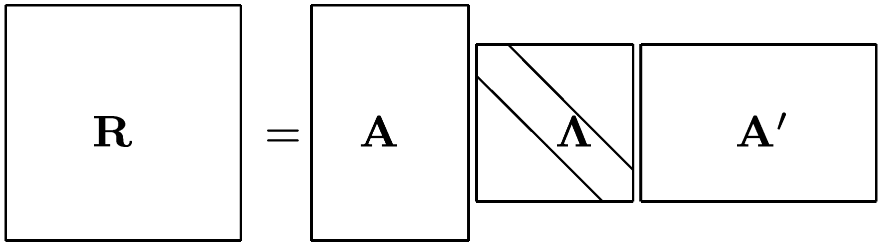
$$\frac{x_0^{Jk}}{\sigma_k} = \sum_{\ell=1}^{L} a_\ell^k y_\ell^J$$



$$b_\ell^k = r_\ell^k = \xi_\ell a_\ell^k$$



$$r^{kk'} = \sum_{\ell=1}^{L} \lambda_\ell a_\ell^k a_\ell^{k'} = \sum_{\ell=1}^{L} b_\ell^k b_\ell^{k'}$$



4.3 Procédure de calcul

1. On calcule la matrice $\mathbf{R}$ des corrélations.

2. On diagonalise $\mathbf{R}$, d'où $(\lambda_\ell, (a_{k\ell})_{k\in K})_{\ell=1,\dots L}$ avec $\sum_{k\in K}(a_{k\ell})^2 = 1$.

3. Pour $\ell = 1, \dots L'$ $(\leq L)$ et $\forall j \in J$, on calcule :
   $y_\ell^j = \sum_{k\in K} a_{k\ell} x_0^{jk}/\sigma_k$.

4. Pour $\ell = 1, \dots L'$ $(\leq L)$ et $\forall k \in K$, on calcule : $b_\ell^k = \sqrt{\lambda_\ell}\, a_{k\ell}$.

5 Méthodologie et interprétation

5.1 Analyse globale du nuage

- Contributions des individus, contributions des variables.
- Contributions des axes

On retient au moins tous les axes dont la contribution est supérieure à une contribution moyenne (e.g. à $\mathrm{Var}\,\mathrm{M}^J/\mathtt{K}$, ou $\mathrm{Var}\,\mathrm{M}^J/L$) ; on s'assure ensuite que les L' axes retenus prennent en compte une part importante de la variance du nuage.

- *Facteur de taille* : sur le cercle des corrélations, les variables sont toutes situées d'un même côté.

Corrélation de la variable moyenne avec la variable principale importante = facteur de taille