Partie IV. L'analyse en composantes principales

Inertie

Définition l'inertie en un point $\mathbf{v}$ du nuage de points est

$$I_{\mathbf{v}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{v}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{v})' \mathbf{M} (\mathbf{e}_i - \mathbf{v}).$$

Inertie totale La plus petite inertie possible est $I_{\mathbf{g}}$, donnée par

$$I_{\mathbf{g}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{g}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{g})' \mathbf{M} (\mathbf{e}_i - \mathbf{g})$$

qui est la seule intéressante puisque $I_{\mathbf{v}} = I_{\mathbf{g}} + \|\mathbf{v} - \mathbf{g}\|^2_M$.

Autres relations $I_{\mathbf{g}}$ mesure la moyenne des carrés des distances entre les individus

$$2I_{\mathbf{g}} = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \|\mathbf{e}_i - \mathbf{e}_j\|^2_{\mathbf{M}}.$$

Interprétation L'inertie totale mesure l'étalement du nuage de points

Calcul de l'inertie

Forme matricielle L'inertie totale est aussi donnée par la trace de la matrice $\mathbf{VM}$ (ou $\mathbf{MV}$)

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{VM}) = \mathrm{Tr}(\mathbf{MV})$$

Métrique usuelle $\mathbf{M} = \mathbf{I}_p$ correspond au produit scalaire usuel et

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{V}) = \sum_{j=1}^{p} \sigma_i^2$$

Métrique réduite obtenue quand $\mathbf{M} = \mathbf{D}_{1/\boldsymbol{\sigma}^2} = \mathbf{D}^2_{1/\boldsymbol{\sigma}}$

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{D}_{1/\boldsymbol{\sigma}^2}\mathbf{V}) = \mathrm{Tr}(\mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{V}\mathbf{D}_{1/\boldsymbol{\sigma}}) = \mathrm{Tr}(\mathbf{R}) = p.$$

Variables centrées réduites On se retrouve encore dans le cas où

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{R}) = p.$$

L'analyse de composantes principales (version 2)

Principe on cherche à projeter orthogonalement le nuage de points sur un espace F_k de dimension $k < p$, sous la forme

$$\mathbf{e}_i^* - \mathbf{g} = c_{i1}\mathbf{a}_1 + c_{i2}\mathbf{a}_2 + \cdots + c_{ik}\mathbf{a}_k$$

Les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ définissent l'espace F_k et les $c_{i\ell}$ sont les coordonnées de $\mathbf{e}_i^*$.

Critère on veut que la moyenne des carrés des distances entre les points $\mathbf{e}_i$ et leur projetés $\mathbf{e}_i^*$ soit minimale. Comme on a toujours (théorème de Pythagore)

$$\|\mathbf{e}_i - \mathbf{g}\|^2 = \|\mathbf{e}_i - \mathbf{e}_i^*\|^2 + \|\mathbf{e}_i^* - \mathbf{g}\|^2,$$

cela revient à maximiser l'inertie du nuage projeté.

On cherche donc F_k, sous espace de dimension k de F_p, qui maximise l'inertie du nuage projeté sur F_k.

Valeurs propres et vecteurs propres : un exemple

Données une matrice et trois vecteurs

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & -1 \\ 2 & 4 & -2 \\ 1 & -1 & 3 \end{bmatrix}$$

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \ \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Vecteurs propres on peut vérifier que

$$\mathbf{A}\mathbf{v}_1 = 2\mathbf{v}_1, \ \mathbf{A}\mathbf{v}_2 = 4\mathbf{v}_2 \text{ et } \mathbf{A}\mathbf{v}_3 = 6\mathbf{v}_3.$$

On dit que $\mathbf{v}_1$, $\mathbf{v}_2$ et $\mathbf{v}_3$ sont vecteurs propres de $\mathbf{A}$ associés aux valeurs propres $\lambda_1 = 2$, $\lambda_2 = 4$ et $\lambda_3 = 6$.

Propriétés (valables en général)

- $-\mathbf{v}_1$ ou $3\mathbf{v}_1$ sont aussi vecteurs propres de $\mathbf{A}$ associés à λ_1 ;
- On a $\mathrm{Tr}(\mathbf{A}) = 5 + 4 + 3 = 12 = \lambda_1 + \lambda_2 + \lambda_3$.

Résultat principal (admis)

Propriété Il existe p réels $\lambda_1, \ldots, \lambda_p$ et p vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_p$, tels que

$$\mathbf{VM}\mathbf{a}_k = \lambda_k \mathbf{a}_k.$$

- Les $\lambda_k \geq 0$ sont les *valeurs propres* de $\mathbf{VM}$ et sont classées par ordre décroissant :

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0.$$

- Les $\mathbf{a}_k$ sont les vecteurs propres de $\mathbf{VM}$ et sont « $\mathbf{M}$-orthonormaux » :

$$\langle \mathbf{a}_k, \mathbf{a}_k \rangle_{\mathbf{M}} = 1, \qquad \langle \mathbf{a}_k, \mathbf{a}_\ell \rangle_{\mathbf{M}} = 0 \text{ si } k \neq \ell.$$

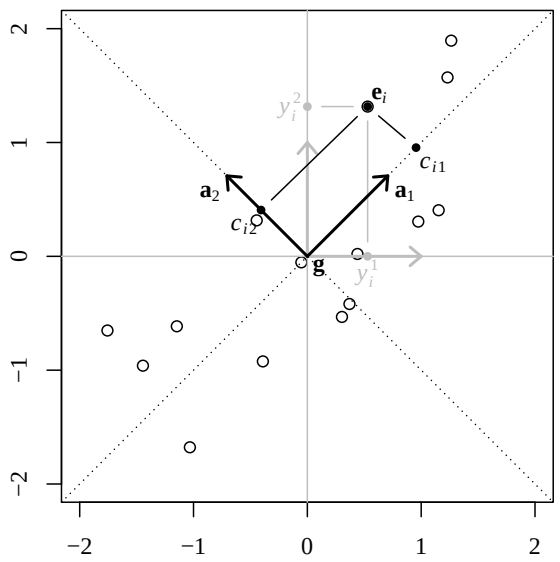
Théorème principal *La projection sur k variables est obtenue en considérant les k premières valeurs propres $\lambda_1, \ldots, \lambda_k$ et les $\mathbf{a}_1, \ldots, \mathbf{a}_k$ correspondants, appelés* axes principaux.

Le calcul ne dépend pas du nombre de variables retenues.

Idée du lien avec l'inertie on sait que $I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{VM}) = \lambda_1 + \cdots + \lambda_p$. Si on ne garde que les données relatives à $\mathbf{a}_1, \ldots, \mathbf{a}_k$, on gardera l'inertie $\lambda_1 + \cdots + \lambda_k$, et c'est le mieux qu'on puisse faire.

Partie V. Les éléments de l'ACP

Changement de coordonnées



$$\mathbf{e}_i - \mathbf{g} = (y_i^1, y_i^2)' = y_i^1(1,0)' + y_i^2(0,1)' = c_{i1}\mathbf{a}_1 + c_{i2}\mathbf{a_2}$$

Les composantes principales

Coordonnées des individus supposons que $\mathbf{e}_i - \mathbf{g} = \sum_{\ell=1}^{p} c_{i\ell}\mathbf{a}_\ell$, alors

$$\langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k\rangle_{\mathbf{M}} = \sum_{\ell=1}^{p} c_{i\ell}\langle \mathbf{a}_\ell, \mathbf{a}_k\rangle_{\mathbf{M}} = c_{ik}$$

La coordonnée de l'individu centré $\mathbf{e}_i - \mathbf{g}$ sur l'axe principal $\mathbf{a}_k$ est donc donné par la projection $\mathbf{M}$-orthogonale

$$c_{ik} = \langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k\rangle_{\mathbf{M}} = (\mathbf{e}_i - \mathbf{g})'\mathbf{M}\mathbf{a}_k.$$

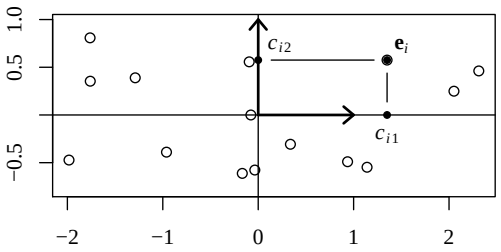
Composantes principales ce sont les variables $\mathbf{c}_k = (c_{1k}, \ldots, c_{nk})$ de taille n définies par

$$\mathbf{c}_k = \mathbf{Y}\mathbf{M}\mathbf{a}_k.$$

Chaque $\mathbf{c}_k$ contient les coordonnées des projections $\mathbf{M}$-orthogonales des individus centrés sur l'axe défini par les $\mathbf{a}_k$.

Représentation des individus dans un plan principal

Qu'est-ce que c'est ? pour deux composantes principales $\mathbf{c}_1$ et $\mathbf{c}_2$, on représente chaque individu i par un point d'abscisse c_{i1} et d'ordonnée c_{i2}.



Quand ? Elle est utile pour des individus discernables.

Propriétés des composantes principales

Moyenne arithmétique les composantes principales sont centrées :

$$\bar{c}_k = \mathbf{c}_k'\mathbf{D_p}\mathbf{1}_n = \mathbf{a}_k'\mathbf{M}\mathbf{Y}'\mathbf{D_p}\mathbf{1}_n = 0$$

car $\mathbf{Y}'\mathbf{D_p}\mathbf{1}_n = \mathbf{0}$ (les colonnes de $\mathbf{Y}$ sont centrées).

Variance la variance de $\mathbf{c}_k$ est λ_k car

$$\begin{aligned}\mathrm{var}(\mathbf{c}_k) &= \mathbf{c}_k'\mathbf{D_p}\mathbf{c}_k = \mathbf{a}_k'\mathbf{M}\mathbf{Y}'\mathbf{D_p}\mathbf{Y}\mathbf{M}\mathbf{a}_k \\ &= \mathbf{a}_k'\mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a}_k = \lambda_k\mathbf{a}_k'\mathbf{M}\mathbf{a}_k = \lambda_k.\end{aligned}$$

Par conséquent on a toujours $\lambda_k \geq 0$

Covariance de même, pour $k \neq \ell$,

$$\mathrm{cov}(\mathbf{c}_k, \mathbf{c}_\ell) = \mathbf{c}_k'\mathbf{D_p}\mathbf{c}_\ell = \cdots = \lambda_\ell\mathbf{a}_k'\mathbf{M}\mathbf{a}_\ell = 0.$$

Les composantes principales ne sont pas corrélées entre elles.

Facteurs principaux

Définition on associe à $\mathbf{a}_k$ le facteur principal $\mathbf{u}_k = \mathbf{M}\mathbf{a}_k$ de taille p. C'est un vecteur propre de $\mathbf{MV}$ car

$$\mathbf{M}\mathbf{V}\mathbf{u}_k = \mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a}_k = \lambda_k\mathbf{M}\mathbf{a}_k = \lambda_k\mathbf{u}_k$$

Calcul en pratique, on calcule les $\mathbf{u}_k$ par diagonalisation de $\mathbf{MV}$, puis on obtient les $\mathbf{c}_k = \mathbf{Y}\mathbf{u}_k$. Les $\mathbf{a}_k$ ne sont pas intéressants.

Interprétation Si on pose $\mathbf{u}_k' = (u_{1k}, \ldots, u_{pk})$, on voit que la matrice des u_{jk} sert de matrice de passage entre la nouvelle base et l'ancienne

$$c_{ik} = \sum_{j=1}^{p} y_i^j u_{jk}, \qquad \mathbf{c}_k = \sum_{j=1}^{p} \mathbf{y}^j u_{jk} \qquad \mathbf{c}_k = \mathbf{Y}\mathbf{u}_k$$

Formules de reconstitution

Reconstitution Par définition des $\mathbf{c}_k$, on a $\mathbf{e}_i - \mathbf{g} = \sum_{k=1}^{p} c_{ik}\mathbf{a}_k$, et donc

$$y_i^j = \sum_{k=1}^{p} c_{ik}a_{kj}, \qquad \mathbf{y}^j = \sum_{k=1}^{p} \mathbf{c}_k a_{kj}, \qquad \mathbf{Y} = \sum_{k=1}^{p} \mathbf{c}_k\mathbf{a}_k'$$

Les a_{kj} forment de matrice de passage entre l'ancienne base et la nouvelle.

Approximation Les k premiers termes fournissent la meilleure approximation de $\mathbf{Y}$ par une matrice de rang k au sens des moindres carrés (théorème de Eckart-Young).