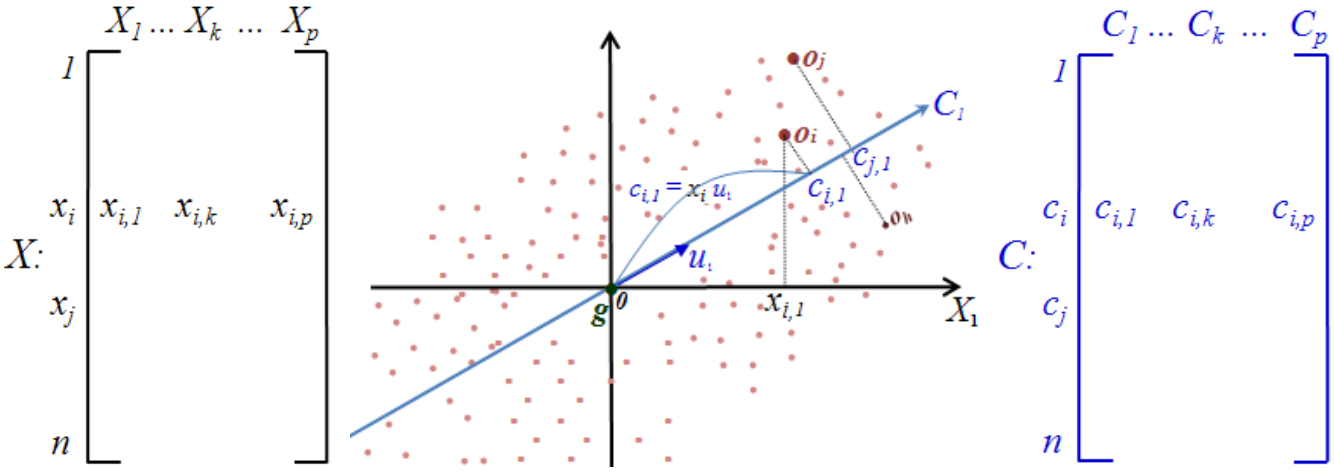
MASTER : Traitement de l'Information et Exploitation des Données
Fouad Badran, Cécile Mallet, Carlos, Mejia,
Charles Sorror, Sylvie Thiria

Rappels de cours sur l'analyse en composante principale (ACP)

Position du problème :



Soit X un ensemble de données $x_{i,k}$ de n individus représentés par p variables**.** Lorsque la dimension p de l'espace devient > 3, la représentation graphique n'est plus possible. Dès lors, on cherche une représentation de X avec un nombre plus réduit de dimension. Il s'agit donc de réaliser une opération de projection dont on sait qu'elle raccourcit les distances. Il faut trouver un ou plusieurs axes de projection tels que la moyenne des carrés des distances entre les points en projection soit maximale. Le problème revient à déterminer des axes unitaires $\boldsymbol{u_k}$ tel que le produit $X\boldsymbol{u_k}$, qui réalise la projection orthogonale, réponde à cette contrainte. Cela consiste à effectuer un changement de base où les vecteurs $\boldsymbol{u_k}$ doivent être les vecteurs unitaires de la nouvelle base ($\|\boldsymbol{u_k}\| = 1\ \forall k$).
En ACP, les vecteurs $\boldsymbol{u_k}$ sont appelée **axes principaux** ; et les nouvelles coordonnées $C_k = X\boldsymbol{u_k}$ sont des combinaisons linéaire des caractères initiaux. Les C_k sont appelées **composantes principales**, elles remplacent les anciennes variables.

Résolution :

Pour faire une ACP, on utilise des données **centrées** ; c'est ce que nous supposerons par la suite pour la matrice des données X. La moyenne est alors placée au centre de gravité du nuage, elle vaut donc 0. Pour un axe k, les distances des points projetés par rapport à l'origine correspondent à leurs coordonnées ($c_{i,k}$). C'est la moyenne de ces distances élevées au carré $c^2_{i,k}$ que l'on veux maximiser, c'est-à-dire, Max$[\frac{1}{n}\sum_{i=1}^{n} c^2_{i,k}]$. Les données étant centrées ce terme correspond à la variance, l'ACP revient à maximiser cette variance. On a vu que les termes C_k sont des

combinaisons linéaires de X et de u_k. Maximiser le terme $\frac{1}{n}\sum_{i=1}^{n} c_{i,k}^2$, revient à maximiser $\boldsymbol{u}^t_k X^t X \boldsymbol{u}_k$

(le terme $\frac{1}{n} X^t X$ correspond à la matrice de variance-covariance S[1])

Ce problème est équivalent (non démontré ici), à rechercher $Max[{}^t\boldsymbol{u}_k X^t X \boldsymbol{u}_k - \lambda_k {}^t\boldsymbol{u}_k \boldsymbol{u}_k]$ où λ_k est un multiplicateur de Lagrange. Par dérivation, on trouve : $X^t X \boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k$. Les solutions $\boldsymbol{u}_k$ recherchée sont les vecteurs propres de la matrice de variance-covariance S. (Les vecteurs sont les directions dans lesquelles la matrice agit, les valeurs propres sont les facteurs multiplicatifs associés à ces directions).

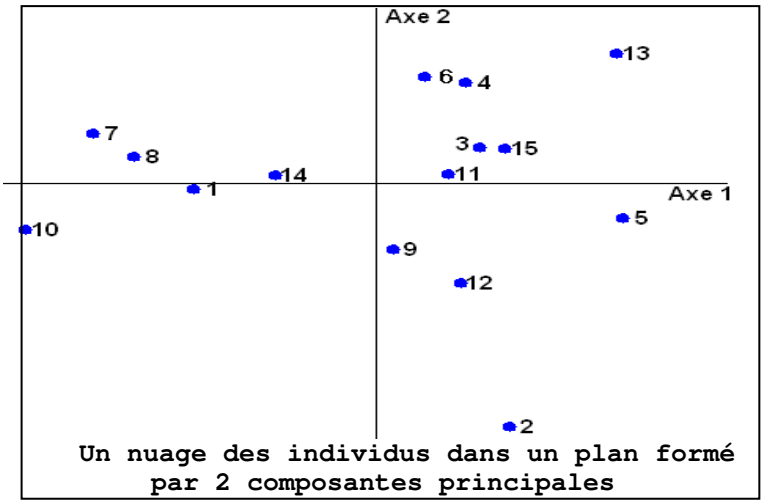
On a donc Max = $\boldsymbol{u}_k {}^t X^t X \boldsymbol{u}_k = \lambda_k \boldsymbol{u}^t_k \boldsymbol{u}_k = \lambda_k$ car le vecteur $\boldsymbol{u}_k$ est de norme 1. λ_k correspond donc à la variance maximum sur l'axe de projection k.

La matrice de variance-covariance S est symétrique et définie positive, on sait qu'elle possède p vecteurs propres orthogonaux deux à deux et ses p valeurs propres sont toutes positives ou nulles. On choisit alors les axes $\boldsymbol{u}_k$ dans l'ordre d'importance des p valeurs λ_k qui leur sont associées en les indiçant de *1* à p : ($\lambda_1 > \lambda_2, \ldots > \lambda_p$).

Nuage des individus :

Ayant trouvé les facteurs $\boldsymbol{u}_k$, rangés dans une matrice $\boldsymbol{U}$, on peut dès lors calculer les coordonnées des individus sur les composantes principales de la nouvelle base : $\boldsymbol{C = XU}$.

Les composantes de $\boldsymbol{C}$ ***(C1, C2, ..., Cp)*** peuvent être représentées 2 à 2 (sur un plan orthogonal) par un nuage des individus comme celui de la figure ci-contre. Il permet de voir comment se sont organisés les individus en projection dans la nouvelle base. Des formes particulières ou des regroupements apparaissent-ils ? On peut également tenir compte de la proximité des individus par rapport à l'origine (centre de gravité) qui correspond à la valeur moyenne. Les individus peuvent contribuer plus ou moins fortement à la direction des axes principaux : ces points seront repérés grâce aux calculs d'indicateurs comme la qualité de représentation ou la contribution de chaque individu dans l'analyse. C'est en examinant ces différents éléments de la disposition des individus sur les plans factoriels que pourra être menée une interprétation.



Un nuage des individus dans un plan formé par 2 composantes principales

[1] Plus précisément, on a la matrice de variance-covariance $S = X^t Q X$ avec Q une matrice (d'ordre n) diagonale de poids. L'égalité se vérifie lorsque les termes diagonaux valent *1/n*. Dans la pratique, quand on utilise un échantillon on estime la matrice de variance-covariance en prenant *1/(n-1)* pour avoir une estimation non biaisée. Ici, par soucis de simplification de la présentation nous laissons tomber la matrice de poids Q (Q=Identité).

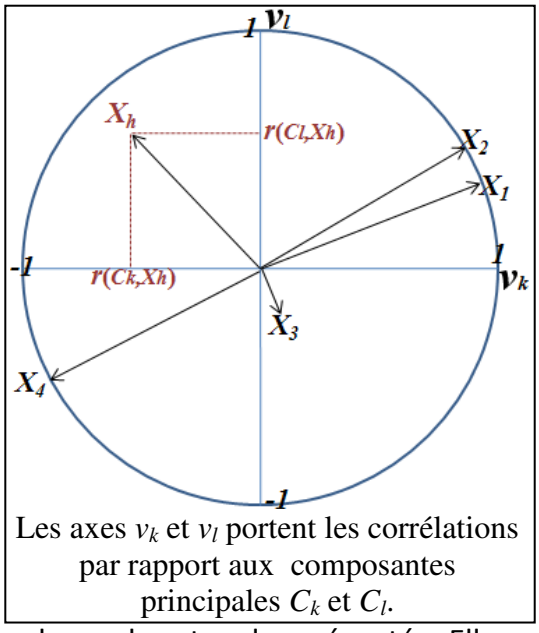
Inertie et variance expliquée :

L'**inertie** d'un nuage de points (notée Ig), est la moyenne des carrés des distances de ses points $\boldsymbol{o}_i$ à son centre de gravité $\boldsymbol{g}$ ($g = (\bar{x}1, \bar{x}2, ..., \bar{x}p)$) éventuellement pondérés par des poids q_i : $Ig = \Sigma_i q_i d^2(\boldsymbol{o}_i, \boldsymbol{g})$ est une mesure de la dispersion autour du point $\boldsymbol{g}$. Avec des pondérations unitaires et des données centrées ($\boldsymbol{g}=0$) on obtient $Ig = (1/n)\Sigma_i \boldsymbol{x}_i^2$: c'est la variance. La somme des valeurs propres permet de calculer la variance des données initiales : $Ig = \Sigma_i \lambda$. Le rapport λ_k / Ig est appelé part d'inertie ou **variance expliquée** par l'axe k. On calcule aussi une part de variance cumulée : $\Sigma_{k=1 \text{ à } m} \lambda_k / Ig$ ($avec\ m<=p$).

Le choix du nombre de composantes à utiliser pour l'interprétation dépend de la question que l'on cherche à résoudre. On essaye en général de former des plans qui totalisent une variance cumulée suffisamment « significative » ou qui semblent présenter un intérêt particulier. Le critère du coude peut aussi être utilisé.

Cercle des corrélations :

Les corrélations entre les nouvelles et les anciennes variables sont déterminantes pour l'interprétation de l'ACP. Le cercle des corrélations (figure ci contre) permet une représentation des corrélations entre les variables initiales et les composantes principales prises 2 à 2. On note $r(C_k,X_h)$ le coefficient de corrélation linéaire[2] entre la variable X_h et la composante C_k. On représente donc les variables initiales par un point dont les coordonnées sont ses corrélations avec 2 composantes C_k et C_l,. Ce point s'inscrit dans un cercle de rayon 1[3].



Les axes v_k et v_l portent les corrélations par rapport aux composantes principales C_k et C_l.

- Une variable proche du bord du cercle indique qu'elle est bien représentée (c'est le cas pour X_1 et X_2 sur la figure). A l'inverse, une variable, comme X_3 qui est proche du centre du cercle est mal représentée. Elle ne peut pas être utilisée valablement pour interpréter les axes du plan choisi.

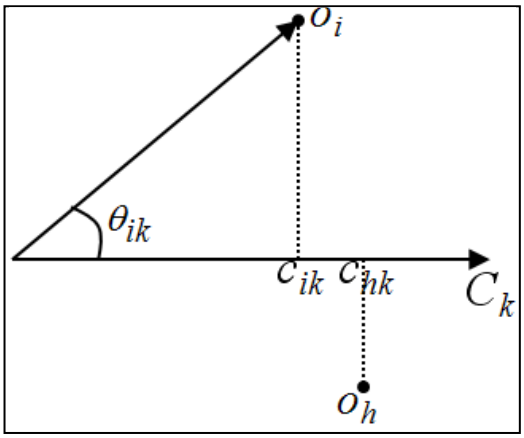
- Deux variables proches du bord du cercle et proche l'une de l'autre indique une forte corrélation (linéaire) entre les deux variables. Deux variables proches d'un bord opposé sont anti-corrélées. Sur la figure, X_4 est anti-corrélé à X_1 et X_2 ; si elles sont orthogonales elles sont décorrélées.

---

[2] $r(C_k,X_h)=cov(X_h,C_k) / sqrt(Var(X_h).Var(C_k))$. On peut également exprimer ce coefficient en termes d'angle **:** Soit θ l'angle entre C_k,et X_h, on a $cos(\theta) = < C_k ; X_h > / \|C_k,X_h\| \|C_k,X_h\| = r(C_k,X_h)$.

[3] Les composantes principales étant orthogonales 2 à 2 et on peut montrer que $r^2(Ck,Xh) + r^2(Ck,Xh) \leq 1$

Qualité de représentation des individus (qlt) :



Elle permet de savoir si un individu ($\boldsymbol{o}_i$), est ou n'est pas bien représenté par un axe k. Pour cela, on considère l'angle θik entre le vecteur de l'individu $\boldsymbol{o}_i$ et sa projection c_{ik} sur l'axe k. On calcule donc :

$$qlt_k(\boldsymbol{o}_i) = cos^2(\theta_{ik}) = c_{ik}^2 \;/\; \|\boldsymbol{o}_i\|^2$$

$$avec\ \|\boldsymbol{o}_i\|^2 = \Sigma_{k=1,p}\, x_{ik}^2 = \Sigma_{k=1,p}\, c_{ik}^2$$

Lorsque l'angle θ est proche de 0, le cosinus, et donc la qualité de représentation (qlt) est proche de 1, l'individu est alors bien représenté pour l'axe k considéré (il est proche de l'axe).
Inversement, un angle θ important induit une faible valeur pour qlt; dans ce cas, l'individu i est mal représenté par l'axe k. Incidemment, on remarque sur la figure, que 2 individus proches sur la projection ne le sont pas nécessairement dans l'espace initial.

Contribution :

Question : quels individus $\boldsymbol{o}_i$ caractérisent le plus fortement un axe k ? : C'est la notion de contribution des individus à la détermination de la direction de l'axe k. On la définie comme suit : $ctr_k(\boldsymbol{o}_i) = q_i\, c_{ik}^2/\lambda_k$ où q_i est le poids de l'individu i. C'est la part de la variance de $\boldsymbol{c}_k$ due à l'individu i. Plus un point est éloigné de l'origine sur un axe du nuage, plus forte est la contribution de ce point pour cet axe. La valeur de la contribution permet de détecter des points trop importants qu'il faut alors retirer pour recommencer l'ACP. Ces points pourraient éventuellement être mis en individus supplémentaires[4].

ACP normée

Lorsque les données sont hétérogènes en moyenne ou en dispersion, on peut les réduire en les divisant par leur écart-type. On parle alors d'ACP normée. Les avantages de la réduction sont que les variables deviennent sans dimension, elles sont invariantes par changement d'unité de mesure, de cette manière toutes les variables ont la même importance dans l'analyse (variance égale à 1).
Dans le cas de données $X(n,p)$ qui sont centrées-réduites, on a :

- $\frac{1}{n} X^t X = cov(X) = R$ (la matrice de corrélation)
- $r(C_k, X_h) = \boldsymbol{u}_{h,k}\ \sqrt{\lambda_k}$
- $Ig = Trace(R) = p$ (constante qui ne dépend pas des données, mais que du nombre de variables)

[4] Un individu supplémentaire ($\boldsymbol{x}_s$) est un individu qui ne participe pas aux calculs de l'ACP mais que l'on peut projeter sur les plans principaux : $\boldsymbol{c}_s = \boldsymbol{x}_s.\boldsymbol{u}$. De même, on peut avoir des variables supplémentaires, qui ne sont pas utilisées par l'ACP, mais que l'on souhaite représenter sur les axes principaux. Avant la projection, il importe de faire subir aux individus supplémentaires les mêmes prétraitements (centrage réduction) que ceux appliqués aux données initiales

Compléments

Nous avons essayé de présenter l'ACP le plus simplement possible. On trouve différentes présentations dans laquelle on fait intervenir une matrice M appelé Métrique. Elle sert à choisir une mesure de distance entre les données. M = I par exemple correspond à la distance euclidienne. M diagonale de terme $1/Var(X)$ est équivalent à prendre des données centrées-réduites, ce qui nous ramène au cas d'une ACP normée qui est la pratique la plus usuelle.

Liens

Voici quelques liens, trouvés sur internet qui nous ont parus intéressant, il en existe évidemment beaucoup d'autres qui traitent de l'ACP.

C. Duby, S. Robin, Institut National Agronomique Paris – Grignon :
http://www.agroparistech.fr/IMG/pdf/ACP2006.pdf
Laurence Reboul, Maitre de conférences en Statistique à l'Université de Poitiers :
http://iml.univ-mrs.fr/~reboul/ADD2-MAB.pdf
http://iml.univ-mrs.fr/~reboul/ADD3-MAB.pdf
Jean-Marc Lasgouttes - INRIA :
https://who.rocq.inria.fr/Jean-Marc.Lasgouttes/ana-donnees/cours-acp.pdf