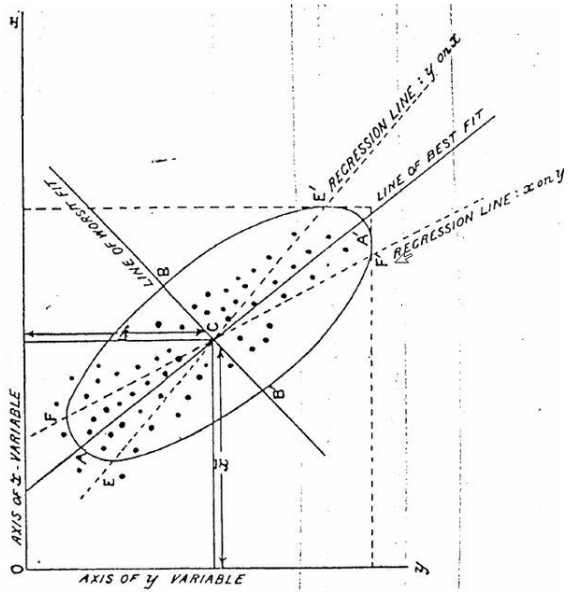
Analyse en composantes principales

Pour les articles homonymes, voir ACP, PCA et KLT (homonymie).

L'**analyse en composantes principales** (**ACP** ou **PCA** en anglais), ou selon le domaine d'application la **transformation de Karhunen–Loève** (**KLT**)[1], est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

Il s'agit d'une approche à la fois géométrique[2] (les variables étant représentées dans un nouvel espace, selon des directions d'inertie maximale) et statistique (la recherche portant sur des axes indépendants expliquant au mieux la variabilité — la variance — des données). Lorsqu'on veut *compresser* un ensemble de N variables aléatoires, les n *premiers axes* de l'analyse en composantes principales sont un *meilleur choix*, du point de vue de l'inertie ou de la variance.

1 Histoire



Extrait de l'article de Pearson de 1901 : la recherche de la « droite du meilleur ajustement ».

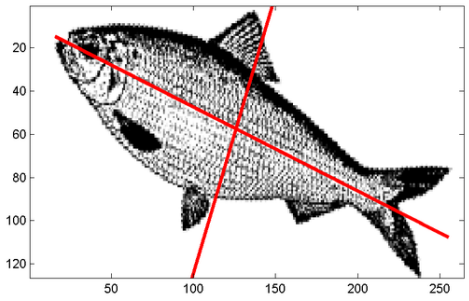
L'ACP prend sa source dans un article de Karl Pearson publié en 1901[3]. Le père du test du χ^2 y prolonge ses travaux dans le domaine de la régression et des corrélations entre plusieurs variables. Pearson utilise ces corrélations non plus pour expliquer une variable à partir des autres (comme en régression), mais pour décrire et résumer l'information contenue dans ces variables.

Encore connue sous le nom de transformée de Karhunen-Loève ou de transformée de Hotelling, l'ACP a été de nouveau développée et formalisée dans les années 1930 par Harold Hotelling[4]. La puissance mathématique de l'économiste et statisticien américain le conduira aussi à développer l'analyse canonique, généralisation des analyses factorielles dont fait partie l'ACP.

Les champs d'application sont aujourd'hui multiples, allant de la biologie à la recherche économique et sociale, et plus récemment le traitement d'images. L'ACP est majoritairement utilisée pour :

- décrire et visualiser des données ;
- les décorréler ; la nouvelle base est constituée d'axes qui ne sont pas corrélés entre eux ;
- les débruiter, en considérant que les axes que l'on décide d'*oublier* sont des axes *bruités*.

2 Exemples introductifs



Les deux axes d'une ACP sur la photo d'un poisson.

Premier exemple

Dans le cas d'une image, comme dans la figure ci-contre, les pixels sont représentés dans un plan et considérés comme une variable aléatoire à deux dimensions. L'ACP va déterminer les deux axes qui expliquent le mieux la dispersion de l'objet, interprété comme un nuage de points.

Elle va aussi les ordonner par inertie expliquée, le second axe étant perpendiculaire au premier.

Second exemple

Dans une école imaginaire, on n'enseigne que deux matières sur lesquelles les élèves sont notés : le français et les mathématiques. En appliquant l'ACP au tableau de notes, on dégagera probablement en premier axe des valeurs par élève très proches de leur moyenne générale dans les deux matières. C'est cet axe qui résumera au mieux la variabilité des résultats selon les élèves. Mais un professeur voulant pousser l'analyse des résultats, s'intéressa aussi au second axe, qui ordonne les élèves selon l'ampleur de leurs écarts entre les deux notes, et indépendamment du premier axe.

On comprend l'intérêt de la méthode d'ACP quand on étend l'analyse à 10 matières enseignées : la méthode va calculer pour chaque élève 10 nouvelles valeurs, selon 10 axes, chacun étant indépendant des autres. Les derniers axes apporteront très peu d'information sur le plan statistique : ils mettront probablement en évidence quelques élèves au profil singulier. Selon son point de vue d'analyse, le professeur, dans sa pratique quotidienne, veillera donc plus particulièrement à ces élèves qui auront été mis en évidence par les derniers axes de la méthode ACP, et/ou corrigera peut-être une erreur qui se serait glissée dans son tableau de notes, mais à l'inverse, il ne prendra pas en compte ces derniers axes s'il mène une réflexion globale s'intéressant aux caractéristiques pédagogiques majeures, ou autrement dit, principales. Si on prend pour exemple une classe de 1re S, on a de fortes chances pour avoir comme axe principal un regroupement des matières scientifiques, et comme second axe les matières littéraires. Ces deux variables expliquent les notes obtenues par les élèves de la classe.

La puissance de l'ACP est qu'elle sait aussi prendre en compte des données de nature hétérogène : par exemple un tableau des différents pays du monde avec le PNB par habitant, le taux d'alphabétisation, le taux d'équipement en téléphones portables, le prix moyen du hamburger, etc. Elle permet d'avoir une intuition rapide des effets conjoints entre ces variables.

3 Échantillon

On applique usuellement une ACP sur un ensemble de *N* variables aléatoires X_1, …, *XN* connues à partir d'un échantillon de K réalisations conjointes de ces variables.

Cet échantillon de ces *N* variables aléatoires peut être structuré dans une matrice *M*, à *K* lignes et *N* colonnes.

$$M = \begin{bmatrix} X_{1,1} & \cdots & X_{1,N} \\ \vdots & \ddots & \vdots \\ X_{K,1} & \cdots & X_{K,N} \end{bmatrix}$$

Chaque variable aléatoire *Xn*, dont *X1*, ₙ, …, *XK*, *n* sont des réalisations indépendantes, a une moyenne $\bar{X}_n$ et un écart type σ*Xn*.

3.1 Poids

Si les réalisations (les éléments de la matrice *M*) sont à probabilités égales alors chaque réalisation (un élément $X_{i,j}$ de la matrice) a la même importance $1/K$ dans le calcul des caractéristiques de l'échantillon. On peut aussi appliquer un poids p_i différent à chaque réalisation conjointe des variables (cas des échantillons redressés, des données regroupées, ...). Ces poids, qui sont des nombres positifs de somme 1 sont représentés par une matrice diagonale *D* de taille *K* :

$$D = \begin{bmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_K \end{bmatrix}$$

Dans le cas le plus courant de poids égaux, $D = \frac{1}{K} I$ où I est la matrice identité.

3.2 Transformations de l'échantillon

Le vecteur $(\bar{X}_1, \cdots, \bar{X}_N)$ est le centre de gravité du nuage de points ; on le note souvent **g**. On a $g = M^T D \tilde{1}$ où $\tilde{1}$ désigne le vecteur de $\mathbb{R}^K$ dont toutes les composantes sont égales à 1.

La matrice *M* est généralement centrée sur le centre de gravité :

$$\bar{M} = \begin{bmatrix} X_{1,1} - \bar{X}_1 & \cdots & X_{1,N} - \bar{X}_N \\ \vdots & \ddots & \vdots \\ X_{K,1} - \bar{X}_1 & \cdots & X_{K,N} - \bar{X}_N \end{bmatrix} = M - \tilde{1} g^T$$

Elle peut être aussi **réduite** :

$$\tilde{M} = \begin{bmatrix} \frac{X_{1,1} - \bar{X}_1}{\sigma(X_1)} & \cdots & \frac{X_{1,N} - \bar{X}_N}{\sigma(X_N)} \\ \vdots & \ddots & \vdots \\ \frac{X_{K,1} - \bar{X}_1}{\sigma(X_1)} & \cdots & \frac{X_{K,N} - \bar{X}_N}{\sigma(X_N)} \end{bmatrix}$$

Le choix de réduire ou non le nuage de points (i.e. les *K* réalisations de la variable aléatoire (X_1, …, *XN*)) est un choix de modèle :

- si on ne réduit pas le nuage : une variable à forte variance va « tirer » tout l'effet de l'ACP à elle ;
- si on réduit le nuage : une variable qui n'est qu'un bruit va se retrouver avec une variance apparente égale à une variable informative.

3.3 Calcul de covariances et de corrélations

Une fois la matrice M transformée en $\bar{M}$ ou $\tilde{M}$, il suffit de la multiplier par sa transposée pour obtenir :

- la matrice de variance-covariance des X_1, ..., XN si M n'est pas réduite : Covariances $= 1/K \cdot \bar{M}^T \cdot \bar{M}$;
- la matrice de corrélation des X_1, ..., XN si M est réduite : Corrélations $= 1/K \cdot \tilde{M}^T \cdot \tilde{M}$.

Ces deux matrices sont carrées (de taille N), symétriques, et réelles. Elles sont donc diagonalisables dans une base orthonormée en vertu du théorème spectral.

De façon plus générale, la matrice de variance-covariance s'écrit $V = M^T DM - gg^T = \bar{M}^T \cdot D \cdot \bar{M}$.

De plus, si l'on note $D_{1/s}$ la matrice diagonale des inverses des écarts-types :

$$D_{1/s} = \begin{bmatrix} 1/s_1 & & 0 \\ & \ddots & \\ 0 & & 1/s_N \end{bmatrix} = \begin{bmatrix} 1/\sigma(X_1) & & 0 \\ & \ddots & \\ 0 & & 1/\sigma(X_N) \end{bmatrix}$$

alors on a :

$$\tilde{M} = \bar{M} \cdot D_{1/s}$$

La matrice des coefficients de corrélation linéaire entre les N variables prises deux à deux, notée ***R***, s'écrit :

$$R = \tilde{M}^T \cdot D \cdot \tilde{M} = D_{1/s} V D_{1/s}$$

4 Critère d'inertie

Dans la suite de cet article, nous considèrerons que le nuage est transformé (centré et réduit si besoin est). Chaque Xn est donc remplacé par $X_n - \bar{X}_n$ ou $(X_n - \bar{X}_n)/\sigma(X_n)$. Nous utiliserons donc la matrice M pour noter $\bar{M}$ ou $\tilde{M}$ suivant le cas.

Le principe de l'ACP est de trouver un axe u, issu d'une combinaison linéaire des Xn, tel que la variance du nuage autour de cet axe soit maximale.

Pour bien comprendre, imaginons que la variance de u soit égale à la variance du nuage ; on aurait alors trouvé une combinaison des Xn qui contient toute la *diversité* du nuage original (en tout cas toute la part de sa diversité captée par la variance).

Un critère couramment utilisé est la variance de l'échantillon (on veut maximiser la variance expliquée par le vecteur u). Pour les physiciens, cela a plutôt le sens de maximiser l'inertie expliquée par u (c'est-à-dire minimiser l'inertie du nuage autour de u).

4.1 Projection

Finalement, nous cherchons le vecteur u tel que la projection du nuage sur u ait une variance maximale. La projection de l'échantillon des X sur u s'écrit :

$$\pi_u(M) = M \cdot u$$

la variance empirique de $\pi u(M)$ vaut donc :

$$\pi_u(M)^T \cdot 1/K \cdot \pi_u(M) = u^T \cdot \underbrace{M^T \cdot 1/K \cdot M}_{C} \cdot u$$

où C est la matrice de covariance.

Comme nous avons vu plus haut que C est diagonalisable dans une base orthonormée, notons P le changement de base associé et $\Delta = \mathrm{Diag}(\lambda_1, \ldots, \lambda_N)$ la matrice diagonale formée de son spectre :

$$\pi_u(M)^T \cdot 1/K \cdot \pi_u(M) = u^T P^T \Delta P u = (Pu)^T \Delta \underbrace{(Pu)}_{v}$$

Les valeurs $(\lambda_1, \ldots, \lambda_N)$ de la diagonale de Δ sont rangées en ordre décroissant. Le vecteur unitaire u qui maximise $v^T \Delta v$ est un vecteur propre de C associé à la valeur propre λ_1 ; on a alors :

$$v^T \cdot \Delta \cdot v = \lambda_1$$

La valeur propre λ_1 est la variance empirique sur le premier axe de l'ACP.

Il est aussi possible de démontrer ce résultat en maximisant la variance empirique des données projetées sur u sous la contrainte que u soit de norme 1 (par un multiplicateur de Lagrange α) :

$$L(u, \alpha) = u^T \cdot C \cdot u - \alpha(u^T u - 1)$$

On continue la recherche du deuxième axe de projection w sur le même principe en imposant qu'il soit orthogonal à u.

4.2 Diagonalisation

La diagonalisation de la matrice de corrélation (ou de covariance si on se place dans un modèle non réduit), nous a permis d'écrire que le vecteur qui explique le plus d'inertie du nuage est le premier vecteur propre. De même le deuxième vecteur qui explique la plus grande part de l'inertie restante est le deuxième vecteur propre, etc.

Nous avons vu en outre que la variance expliquée par le k-ième vecteur propre vaut λk.

Finalement, la question de l'ACP se ramène à un problème de diagonalisation de la matrice de corrélation.

4.3 Optimisation numérique

Numériquement, la matrice M étant rectangulaire, il peut être plus économique de la décomposer en valeurs singulières, puis de recombiner la décomposition obtenue, plutôt que de diagonaliser $M'M$.

5 ACP et variables qualitatives

En ACP, il est fréquent que l'on veuille introduire des variables qualitatives en supplémentaire. Par exemple, on a mesuré de nombreuses variables quantitatives sur des plantes. Pour ces plantes, on dispose aussi de variables qualitatives, par exemple l'espèce à laquelle appartient la plante. On soumet ces données à une ACP des variables quantitatives. Lors de l'analyse des résultats, il est naturel de chercher à relier les composantes principales à la variable qualitative *espèce*. Pour cela on produit les résultats suivant.

- Identification, sur les plans factoriels, des différentes espèces en les représentant par exemple par des couleurs différentes.
- Représentation, sur les plans factoriels, des centres de gravité des plantes appartenant à une même espèce.
- Indication, pour chaque centre de gravité et pour chaque axe, d'une probabilité critique pour juger de la significativité de l'écart entre un centre de gravité et l'origine.

Tous ces résultats constituent ce que l'on appelle *introduire une variable qualitative en supplémentaire*. Cette procédure est détaillée dans et Escofier&Pagès 2008, Husson, Lê & Pagès 2009 et Pagès 2013. Peu de logiciels offrent cette possibilité de façon « automatique ». C'est le cas de SPAD qui historiquement, à la suite des travaux de Ludovic Lebart, a été le premier logiciel à le proposer, et du package R FactoMineR.

6 Résultats théoriques

Si les sections précédentes ont travaillé sur un échantillon issu de la loi conjointe suivie par X_1, …, XN, que dire de la validité de nos conclusions sur n'importe quel autre échantillon issu de la même loi ?

Plusieurs résultats théoriques permettent de répondre au moins partiellement à cette question, essentiellement en se positionnant par rapport à une distribution gaussienne comme référence.

7 Méthodes voisines et extensions : la famille factorielle

L'analyse en composantes principales est la plus connue des méthodes factorielles ; d'autres méthodes factorielles existent pour analyser d'autres types de tableau. À chaque fois, le principe général est le même.

- On considère deux nuages de points, l'un associé aux lignes du tableau analysé et l'autre aux colonnes de ce tableau.
- Ces deux nuages sont liés par des relations de dualité : ils ont la même inertie totale ;
- Chacun de ces nuages est projeté sur ses directions d'inertie maximum.
- D'un nuage à l'autre, les directions d'inertie de même rang sont liées par des relations de dualité (ou de transition) : elles ont la même inertie et les coordonnées des projections sur l'une se déduisent des coordonnées des projections sur l'autre.

7.1 Analyse factorielle des correspondances (AFC)

Elle s'applique à des tableaux de contingence c'est-à-dire des tableaux croisant deux variables qualitatives. Ce type de tableau est très différent de celui analysé par ACP : en particulier, les lignes et les colonnes jouent des rôles symétriques alors que la distinction entre lignes et colonnes (c'est-à-dire entre individus et variables) est majeure en ACP.

7.2 Analyse des correspondances multiples (ACM)

Elle s'applique à des tableaux dans lesquels un ensemble d'individus est décrit par un ensemble de variables qualitatives. Ce type de tableau est donc voisin de celui analysé en ACP, les variables quantitatives étant remplacées par des variables qualitatives. L'ACM est souvent vue comme un cas particulier de l'AFC mais ce point de vue est très réducteur. L'ACM possède suffisamment de propriétés spécifiques pour être considérée comme une méthode à part entière.

On peut aussi présenter l'ACM à partir de l'ACP comme cela est fait dans Pagès 2013. L'intérêt est de relier entre eux les ressorts de l'ACP et ceux de l'ACM ce qui ouvre la voie au traitement simultané des deux types de variables (cf. AFDM et AFM ci-après)

7.3 Analyse factorielle de données mixtes (AFDM)

Les données sont constituées par un ensemble d'individus pour lesquels on dispose de plusieurs variables, comme en ACP ou en ACM. Mais, ici, les variables sont aussi bien quantitatives que qualitatives. L'analyse factorielle de données mixtes traite simultanément les deux types de variables en leur faisant jouer un rôle actif. L'AFDM est décrite dans Pagès 2013 et Escofier&Pagès 2008.

7.4 Analyse factorielle multiple (AFM)

Les données sont, ici encore, constituées par un ensemble d'individus pour lesquels on dispose de plusieurs variables. Mais cette fois, outre qu'elles peuvent être quantitatives et/ou qualitatives, les variables sont structurées en groupes. Ce peut être, par exemple, les différents thèmes d'un questionnaire. L'AFM prend en compte cette structure en groupes dans l'analyse de ces données. L'AFM est décrite en détail dans Pagès 2013 et Escofier&Pagès 2008.

8 Applications

8.1 Compression

L'Analyse en Composantes Principales est usuellement utilisée comme outil de compression linéaire. Le principe est alors de ne retenir que les *n* premiers vecteurs propres issus de la diagonalisation de la matrice de corrélation (ou covariance), lorsque l'inertie du nuage projeté sur ces *n* vecteurs représente *qn* pourcents de l'inertie du nuage original, on dit qu'on a un taux de compression de 1 - *qn* pourcents, ou que l'on a compressé à *qn* pourcents. Un taux de compression usuel est de 20 %.

Les autres méthodes de compressions statistiques habituelles sont :

- l'analyse en composantes indépendantes ;
- les cartes auto-adaptatives (SOM, *self organizing maps* en anglais) ; appelées aussi cartes de Kohonen ;
- l'analyse en composantes curvilignes ;
- la compression par ondelettes.

Il est possible d'utiliser le résultat d'une ACP pour construire une classification statistique des variables aléatoires X_1, …, *XN*, en utilisant la distance suivante (*Cn, n'* est la corrélation entre *Xn* et *Xn'*) :

$$d(X_n, X_{n'}) = \sqrt{2\,(1 - C_{n,n'})}$$

8.2 Analyse de séries dynamiques d'images

L'ACP, désignée en général dans le milieu du traitement du signal et de l'analyse d'images plutôt sous son nom de Transformée de Karhunen-Loève (TKL) est utilisée pour analyser les séries dynamiques d'images[5], c'est-à-dire une succession d'images représentant la cartographie d'une grandeur physique, comme les scintigraphies dynamiques en médecine nucléaire, qui permettent d'observer par gamma-caméra le fonctionnement d'organes comme le cœur ou les reins.

Dans une série de P images, chaque pixel est considéré comme un point d'un espace affine de dimension P dont les coordonnées sont la valeur du pixel pour chacune des P images au cours du temps. Le nuage ainsi formé par tous les points de l'image peut être analysé par l'ACP, (il forme un hyper-ellipsoïde à P dimensions) ce qui permet de déterminer ses axes principaux.

En exprimant tous les points dans le repère orthogonal à P dimensions des axes de l'ACP, on passe ainsi de la série temporelle d'origine (les pixels représentent la valeur en fonction du temps) à une nouvelle série (également de P images) dans l'espace de Karhunen-Loève : c'est la Transformée de Karhunen-Loève, qui est une opération réversible : on parle de « TKL » et de « TKL inverse » ou « TLK^{-1} ».

La compression est possible car l'information est contenue presque entièrement sur les premiers axes de l'ACP. Mais la notion de « compression » sous-entend que les autres images correspondant aux autres axes sont volontairement ignorées. La TKL étant réversible, la suppression arbitraire des axes les moins énergétiques constitue alors un filtrage permettant de réduire le bruit temporel de la série d'images.

Concrètement, l'application de TKL + suppression des axes les moins significatifs + TKL^{-1} permet de supprimer le fourmillement apparent (bruit temporel) d'une série animée d'images.

En imagerie médicale fonctionnelle, on améliore ainsi la qualité visuelle de la visualisation scintigraphique du cycle cardiaque moyen.

Par ailleurs, l'analyse de l'importance respective des valeurs propres de l'ACP permet d'approcher le nombre de fonctionnements physiologiques différents. On a ainsi pu montrer que le cœur sain peut être entièrement représenté avec 2 images (2 axes de l'ACP contiennent toute l'information utile), alors que pour certaines pathologies l'information utile s'étale sur 3 images[6].

8.3 Analyse d'images multi-spectrales

Comme pour l'application précédente, la longueur d'onde remplaçant juste le temps, la TKL a été proposée à plusieurs reprises pour extraire l'information utile d'une

série d'images monochromes représentant les intensités pour des longueurs d'ondes différentes. De telles images peuvent être issues de microscopie optique classique, confocale ou SNOM (Microscope optique en champ proche)[7].

8.4 Évolution de la topographie

De la même manière, la TKL permet de mettre en évidence des cinétiques différentes lors de l'analyse topographique dynamique, c'est-à-dire l'analyse de l'évolution du relief au cours du temps. Elle permet alors de déceler des phénomènes invisibles par simple observation visuelle, mais se distinguant par une cinétique légèrement différente (par exemple pollution d'une surface rugueuse par un dépôt)[8].

9 Logiciels

Il y a de très nombreux logiciels incluant l'ACP. Le package R FactoMineR, est probablement le logiciel libre le plus complet dans le domaine de l'analyse des données (incluant, en particulier, outre l'ACP, toutes les extensions décrites plus haut : AFC, ACM, AFDM et AFM). Ce logiciel est relié au livre Husson, Lê & Pagès 2009.

10 Notes

[1] http://tcts.fpms.ac.be/cours/1005-07-08/codage/codage/xcodim2.pdf

[2] Une vidéo d'introduction à l'ACP fondée sur la géométrie est accessible ici.

[3] (en) Pearson, K., « On Lines and Planes of Closest Fit to Systems of Points in Space », *Philosophical Magazine*, vol. 2, n° 6, 1901, p. 559–572 (lire en ligne [PDF])

[4] (en) « Analysis of a Complex of Statistical Variables with Principal Components », 1933, *Journal of Educational Psychology*.

[5] Évaluation de la perfusion et de la fonction contractile du myocarde à l'aide de l'analyse de Karhunen-Loève en tomographie d'émission monophotonique myocardique synchronisée à l'ECG par P. Berthout, R. Sabbah, L. Comas, J. Verdenet, O. Blagosklonov, J.-C. Cardot et M. Baud dans *Médecine nucléaire* Volume 31, Volume 12, décembre 2007, Pages 638-646.

[6] Baud, Cardot, Verdenet et al, Service de médecine nucléaire, Hôpital Jean-Minjoz, boulevard Fleming, 25030 Besançon cedex, France (nombreuses publications sur plus de 30 ans)

[7] Analysis of optical near-field images by Karhunen—Loève transformation Daniel Charraut, Daniel Courjon, Claudine Bainier, and Laurent Moulinier, *Applied Optics*, Vol. 35, Issue 20, p. 3853-3861 (1996)

[8] (en) Jean-Yves Catherin, Measure in 2D, visualise in 3D and understand in 4D dans *Micronora Informations* juin 2008, page 3

11 Voir aussi

- Valeurs propres
- Compression statistique
- Équilibre biais / variance
- Analyse de la variance
- Partitionnement de données
- Exploration de données
- Iconographie des corrélations
- Michel Loève
- Kari Karhunen
- Théorème de Karhunen-Loève (en)
- Analyse discriminante linéaire

12 Références

- Jean-Paul Benzécri ; *Analyse des données. T2* (leçons sur l'analyse factorielle et la reconnaissance des formes et travaux du Laboratoire de statistique de l'Université de Paris 6. T. 2 : l'analyse des correspondances), Dunod Paris Bruxelles Montréal, 1973
- Jean-Paul Benzécri et Al. *Pratique de l'analyse des données. T1* (analyse des correspondances. Exposé élémentaire), Dunod Paris, 1984,
- Jean-Paul Benzécri et Al. *Pratique de l'analyse des données. T2* (abrégé théorique. Études de cas modèle), Dunod Paris, 1984
- Brigitte Escofier et Jérôme Pagès, Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation, Dunod, Paris, 2008, 4e éd. (1re éd. 1988), 318 p. (ISBN 978-2-10-051932-3)
- François Husson, Sébastien Lê et Jérôme Pagès, Analyse des données avec R, Presses Universitaires de Rennes, 2009, 224 p. (ISBN 978-2-7535-0938-2)
- Ludovic Lebart, Morineau Alain, Piron Marie ; *Statistique exploratoire multidimensionnelle*, Dunod Paris, 1995
- Jérôme Pagès, Analyse factorielle multiple avec R, EDP sciences, Paris, 2013, 253 p. (ISBN 978-2-7598-0963-9)

- Mathieu Rouaud ; Probabilités, statistiques et analyses multicritères Un livre de 290 pages qui traite de l'ACP (les principes et de nombreux exemples concernant, entre autres, les isolants thermiques et les eaux minérales). Version numérique libre et gratuite.
- Michel Volle, *Analyse des données*, Economica, 4e édition, 1997, (ISBN 2-7178-3212-2)

12.1 Liens externes

- FactoMineR, une bibliothèque de fonctions R destinée à l'analyse des données
-  Portail des probabilités et de la statistique

13 Sources, contributeurs et licences du texte et de l'image

13.1 Texte

- **Analyse en composantes principales** *Source :* https://fr.wikipedia.org/wiki/Analyse_en_composantes_principales?oldid=133703462 *Contributeurs :* Cdang, HB, Phe-bot, MaCRoEco, Vincnet, Nbrouard, Erasmus, Gdgourou, Orel'jan, Ripounet, Michel Volle, Tdoune, GrdScarabe, Krom17, Azerty694, Gzen92, EdC, RobotQuistnix, YurikBot, Frelaur, Jean-Luc W, Pautard, Sylenius, Maxxtwayne, Lehalle, Agua, Gbdivers, PieRRoBoT, Moineau44, Yopai, Thijs !bot, Jarfe, Kyle the bot, Arkanosis, Fmbot, Sebleouf, Jamcib, TouristeCatégorisant, Salebot, TXiKiBoT, Fluti, Godix, Ptbotgourou, Xic667, SieBot, Louperibot, Ambigraphe, DumZiBoT, DeepBot, SniperMaské, ToePeu.bot, DragonBot, Desiderius Severus, WikiCleanerBot, SilvonenBot, ZetudBot, Univmaths, Bruce rennes~frwiki, Bub's wikibot, Lesty, Arnaud.trebaol, JackPotte, Guadalou, Luckas-bot, Micbot, Visualnumerics, Anne Bauval, Aadri, SebGR, Berlascalp, Xqbot, Obersachsebot, JackBot, Romainbrasselet, RB117, BenzolBot, Sebculture, Lomita, Jackverr, Chefsoleil, Saison, KamikazeBot, Monto1843, ZéroBot, Francoishusson, Xerti, Bugmenot1992, Bli, Marion.cuny, Tony17455, Racinaire, Jimmy-jambe, OrlodrimBot, Elopash, FDo64, Poupoul2, Ramzan, Roll-Morton, Addbot, Baha2490, Hermine00G, JuL789, Statistix35, Do not follow, HeyCat, Gzen92Bot et Anonyme : 68

13.2 Images

- **Fichier:Disambig_colour.svg** *Source :* https://upload.wikimedia.org/wikipedia/commons/3/3e/Disambig_colour.svg *Licence :* Public domain *Contributeurs :* Travail personnel *Artiste d'origine :* Bub's
- **Fichier:Karl_Pearson_line_of_best_fit_diagramm_from_philosophical_magazine_1901_2_559-572.jpg** *Source :* https://upload.wikimedia.org/wikipedia/commons/7/73/Karl_Pearson_line_of_best_fit_diagramm_from_philosophical_magazine_1901_2_559-572.jpg *Licence :* Public domain *Contributeurs :* Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2 :559-572 *Artiste d'origine :* Selfmade extract from the above article page 566
- **Fichier:Logo_proba_4.svg** *Source :* https://upload.wikimedia.org/wikipedia/commons/f/f7/Logo_proba_4.svg *Licence :* CC BY-SA 3.0 *Contributeurs :* Travail personnel *Artiste d'origine :* Ipipipourax
- **Fichier:PCA_fish.png** *Source :* https://upload.wikimedia.org/wikipedia/commons/9/90/PCA_fish.png *Licence :* CC BY-SA 2.5 *Contributeurs :* own work by Lehalle, moved from French Wikipedia *Artiste d'origine :* Lehalle

13.3 Licence du contenu

- Creative Commons Attribution-Share Alike 3.0