

Méthodologie générale d'une étude ACP : Généralités, concepts et exemples

H. Zaki^{1*}, .M. Benlyas¹, Y. Filali Zegzouti¹, M. Bouachrine²

¹ Equipe de Recherche Biologie, Environnement & Santé, Faculty of Science and Technics, University Moulay Ismail, Meknes, Morocco

² Equipe de Recherche Matériaux, Environnement & Modélisation. ESTM, University Moulay Ismail, Meknes, Morocco

*Corresponding Author: H. Zaki (mmezakihanane@gmail.com)

Résumé: *Le développement des modèles statistiques des relations quantitatives structure-activité/propriété joue un rôle important dans la conception de nouveaux produits chimiques spécialement pharmaceutiques. La prédiction des propriétés biologiques des produits non testés passe d'abord par une analyse ACP. Il s'agit de l'analyse en composante principale : c'est une méthode basée sur des statistiques descriptives multidimensionnelles permettant de traiter simultanément un nombre quelconque de variables quantitatives. L'objectif est de visualiser et résumer l'information contenue dans les différentes données afin d'avoir une représentation permettant plus facilement l'interprétation. Dans ce papier, nous décrivons des généralités et les principaux concepts et techniques les plus pertinents pour la réalisation d'une étude ACP.*

Mots clés: ACP, Propriétés biologiques, descripteurs moléculaires, statistiques

1- Introduction

L'analyse en composante principale ACP (Jolliffe, 1986) est une méthode basée sur des statistiques descriptives multidimensionnelles permettant de traiter simultanément un nombre quelconque de variables quantitatives. Le cas de plusieurs individus (n individus) mesurés par rapport à un grand nombre de variables numériques. Ces variables sont la plupart du temps corrélées entre elles. Elle consiste à rechercher des facteurs en nombre restreint en résumant le mieux possible les données considérées. Elle aboutit à des représentations graphiques des données (des individus comme des variables) par rapport à ces facteurs représentés comme des axes. Ces représentations graphiques sont du type nuage de points. Proposée par Hotelling en 1933 mais elle n'est devenue une technique opérationnelle qu'à partir des années 60 avec le développement des outils informatiques. Cette méthode a été réinterprétée sous un formalisme probabiliste par Tipping et Bishop en 1999, elle a de nombreuses applications comprennent la compression de données, le

traitement de l'image, la visualisation, l'analyse exploratoire des données, la reconnaissance des formes et la prévision des séries chronologiques. (Tipping et al 1999).

Objectif de la méthode

L'objectif est de visualiser et résumer l'information contenue dans les différentes données afin d'avoir une représentation permettant plus facilement l'interprétation.

Principe de la méthode

Le principe de l'ACP est de réduire la dimension des données initiales (qui est (p) si l'on considère p variables quantitatives), en remplaçant les p variables initiales par (q) facteurs appropriés ($q < p$). Les q facteurs cherchés sont des moyennes pondérées des variables initiales. Leur choix se fait en maximisant la dispersion des individus selon ces facteurs (variance maximum). Des techniques mathématiques appropriées permettent de réaliser tout cela de façon automatique et optimale. (Anderson 1963).

2- Méthodologie de l'ACP :

Les données à analyser

L'ACP est appliquée sur p variables quantitatives notées $X_1, \dots, X_j, \dots, X_p$ observées sur n individus notés $1, \dots, i, \dots, n$. L'observation de la variable X_j observées sur l'individu i est $x_{j,i}$

Donc l'ensemble des informations se représente de la manière suivante

	X_1	X_j	X_p
1	$x_{1,1}$	$x_{j,1}$	$x_{p,1}$
.....
i	$x_{1,i}$	$x_{j,i}$	$x_{p,i}$
.....
n	$x_{1,n}$	$x_{j,n}$	$x_{p,n}$

Critère d'inertie

Les q ($q=2$ ou $q=3$) facteurs que l'on va définir, pour résumer l'information contenue dans le tableau initial, doivent maximiser la dispersion du nuage des observations. Généralement, lorsqu'on dispose d'un nuage d'observations en plusieurs dimensions, on parle d'inertie (somme des variances des variables considérées).

En passant de la dimension initiale p à la dimension réduite q , on perd, obligatoirement, de la dispersion, de l'inertie. L'idée est de choisir les facteurs convenables pour perdre le moins possible la dispersion.

Transformation des données

On cherche des combinaisons linéaires des variables initiales, appelées facteurs, ou encore composantes principales, s'écrivant sous la forme suivante :

$$C_1 = a_{1,1} X_1 + a_{2,1} X_2 + \dots + a_{p,1} X_p$$

$$C_2 = a_{1,2} X_1 + a_{2,2} X_2 + \dots + a_{p,2} X_p$$

.....

3- Résultats et exemple

Résultats concernant les variables

L'ACP permet de calculer les corrélations variables-facteurs, autrement dit les coefficients de corrélations linéaires entre chaque variable initiale et chaque facteur retenu. Les corrélations variables-facteurs permettent de réaliser les graphiques des variables dont l'étude détaillée conduit à préciser la signification des axes.

Résultats concernant les individus

L'ACP permet aussi de calculer les coordonnées des individus sur les axes, leurs contributions à la dispersion selon chacun de ces axes et les cosinus carres. Les coordonnées permettent de réaliser les graphiques des individus (1 ou 3 graphiques, selon que l'on a choisi $q = 2$ ou $q = 3$).

Exemple :

Le cas que nous allons étudier croise 22 sujets et 14 variables, les sujets sont 22 molécules dérivés de indole-2-carboxylate et les variables sont les différents descripteurs électroniques et topologiques. L'Analyse en composante principale va se faire sur le logiciel XLStat . L'ACP sur XLStat se fait d'une manière simple : il suffit de suivre les étapes dans le tutoriel sur le lien suivant : <https://help.xlstat.com/>.

*Les résultats **d'analyses en composante principale** sont représentés sur le tableau 1 sous forme d'une matrice de corrélation (Tabelau 1) : Ce premier résultat intéressant à analyser est la matrice des corrélations, on remarque par exemple à travers la lecture de cette matrice de corrélation, d'une part, que l'énergie de répulsion et l'énergie électronique sont parfaitement corrélées ($r = -1$) ces deux paramètres sont donc redondantes. D'autre part, on remarque que le moment dipolaire est peu corrélé avec l'indice topologique de la molécule qui signifie que le moment dipolaire ne dépend pas de ce descripteur. En générale plus le coefficient de corrélation tend vers 1 plus la corrélation est forte entre les descripteurs.*

Tableau 1 : Matrice de corrélation

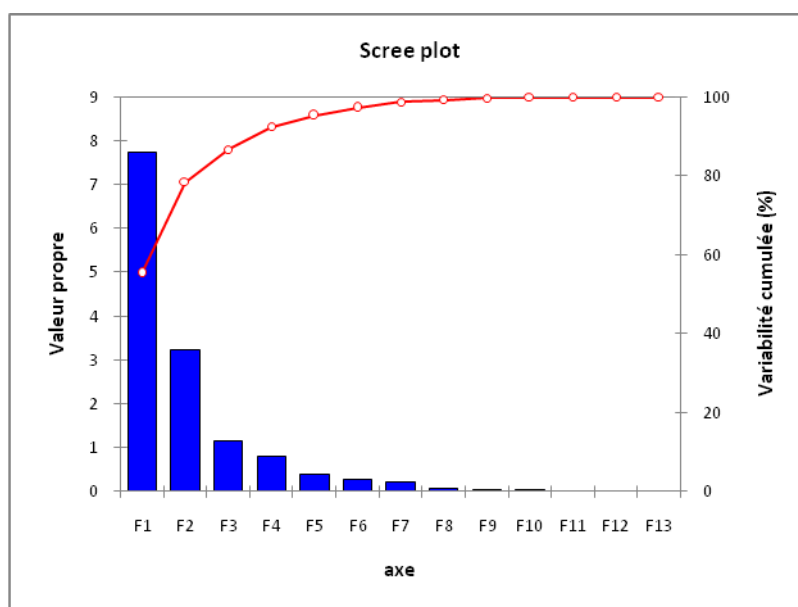
Matrice de corrélation (Pearson (n)) :														
Variables	lar topologica	energy (Kca	onique energ	log P	wiener index	weight (Atom	oefficient (octa	ovality	ulsion Energy	HOMO (eV)	LUMO (eV)	Gap (eV)	nt dipolaire (Ital	Energy (u.a.)
molecular to	1	-0,276	-0,973	0,757	0,614	0,983	0,787	0,864	0,972	-0,238	0,302	-0,292	0,081	-0,900
stretch energ	-0,276	1	0,364	-0,209	-0,200	-0,366	-0,266	-0,381	-0,361	0,031	-0,056	0,048	-0,198	0,426
electronique	-0,973	0,364	1	-0,702	-0,633	-0,996	-0,762	-0,886	-1,000	0,198	-0,281	0,262	-0,166	0,948
log P	0,757	-0,209	-0,702	1	0,311	0,717	0,966	0,545	0,706	-0,299	0,454	-0,414	-0,344	-0,625
wiener inde	0,614	-0,200	-0,633	0,311	1	0,624	0,376	0,549	0,631	0,296	-0,344	0,343	0,262	-0,585
Molecular W	0,983	-0,366	-0,996	0,717	0,624	1	0,770	0,901	0,995	-0,222	0,284	-0,274	0,147	-0,950
Partition Coe	0,787	-0,266	-0,762	0,966	0,376	0,770	1	0,653	0,767	-0,363	0,422	-0,421	-0,268	-0,675
ovality	0,864	-0,381	-0,886	0,545	0,549	0,901	0,653	1	0,882	-0,319	0,203	-0,264	0,166	-0,867
Repulsion Er	0,972	-0,361	-1,000	0,706	0,631	0,995	0,767	0,882	1	-0,195	0,283	-0,262	0,165	-0,945
HOMO (eV)	-0,238	0,031	0,198	-0,299	0,296	-0,222	-0,363	-0,319	-0,195	1	-0,783	0,920	0,317	0,222
LUMO (eV)	0,302	-0,056	-0,281	0,454	-0,344	0,284	0,422	0,203	0,283	-0,783	1	-0,964	-0,432	-0,271
Gap (eV)	-0,292	0,048	0,262	-0,414	0,343	-0,274	-0,421	-0,264	-0,262	0,920	-0,964	1	0,408	0,265
Moment dip	0,081	-0,198	-0,166	-0,344	0,262	0,147	-0,268	0,166	0,165	0,317	-0,432	0,408	1	-0,256
Total Energy	-0,900	0,426	0,948	-0,625	-0,585	-0,950	-0,675	-0,867	-0,945	0,222	-0,271	0,265	-0,256	1
Les valeurs en gras sont différentes de 0 à un niveau de signification alpha=0,05														

Le tableau 2 et le graphique associé sont liés à un objet mathématique, les valeurs propres, qui sont heureusement liées à un concept très simple : la qualité de la projection lorsque l'on passe de N dimensions (N étant le nombre de variables, ici 14) à un nombre plus faible de dimensions. Dans notre cas, on voit que la première valeur propre vaut 7.747 et représente 55.334% de la variabilité. Cela signifie que si l'on représente les données sur un seul axe, alors on aura toujours 55% de la variabilité totale qui sera préservée.

A chaque valeur propre correspond un facteur. Chaque facteur est en fait une combinaison linéaire des variables de départ. Les facteurs ont la particularité de ne pas être corrélés entre eux. Les valeurs propres et les facteurs sont triés par ordre décroissant de variabilité représentée.

Tableau 2 : Valeurs propres et variabilités

Valeurs propres :						
	F1	F2	F3	F4	F5	F6
Valeur propre	7,747	3,241	1,158	0,806	0,391	0,274
Variabilité (%)	55,334	23,153	8,272	5,757	2,796	1,955
% cumulé	55,334	78,487	86,759	92,516	95,312	97,267



Idéalement, les deux premières valeurs propres correspondent à un % élevé de la variabilité, si bien que la représentation sur les deux premiers axes factoriels est de bonne qualité.

Le premier graphique particulier à la méthode est le cercle des corrélations (voir ci-dessous le cercle sur les axes F1 et F2). Il correspond à une projection des variables initiales sur un

plan à deux dimensions constitué par les deux premiers facteurs. Lorsque deux variables sont loin du centre du graphique, alors si elles sont : proches les unes par rapport aux autres, alors elles sont significativement positivement corrélées (r proche de 1), orthogonales les unes par rapport aux autres, alors elles sont significativement non-corrélées (r proche de 0), symétriquement opposées par rapport au centre, alors elles sont significativement négativement corrélées (r proche de -1).

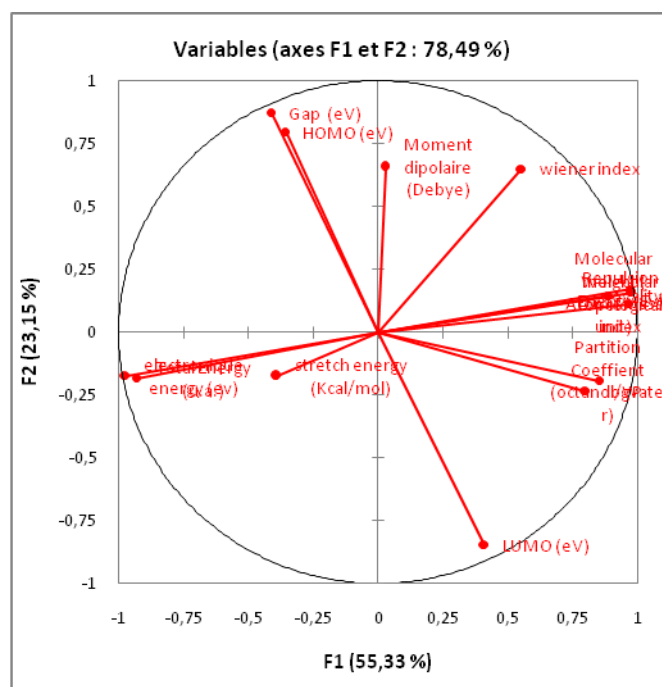


Figure 1 : Cercle de corrélation

D'une manière plus simple, l'axe horizontale est la première dimension de l'ACP (F1) et l'axe verticale est la deuxième dimension (F2), chaque ligne rouge représente une variable, chaque deux variables séparées par un angle aigu par rapport à un axe sont corrélées positivement par exemple l'énergie Gap et l'énergie HOMO sont positivement liées, les angles droites reflète l'indépendance, on peut dire par exemple que l'indice de Winner n'est pas lié à l'énergie HOMO et l'énergie Gap, les grandes angles reflètent la corrélation négative par exemple l'énergie LUMO est corrélée négativement à l'énergie Gap.

Dans la figure 2, nous avons présenté le graphique des observations qui correspond à l'un des objectifs de l'ACP. Il permet de représenter les individus sur une carte à deux dimensions, et ainsi d'identifier des tendances

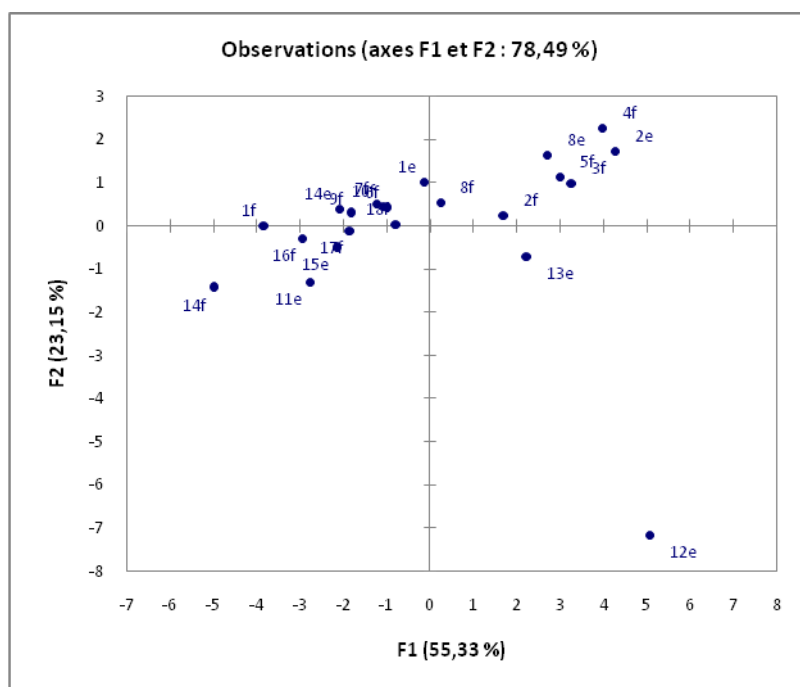


Figure 2 : graphique des observations

Dans notre exemple, on remarque que les molécules 12e, 14f et 13e ont des caractéristiques particulières du fait qu'elles sortent du groupe, alors que par exemple 5f, 3f, et 8e ont plus de caractères en commun puisque elles se regroupent.

Si on fait la superposition du cercle de corrélation (Figure 1) et le graphique des observations (Figure 2), on peut extraire plus d'informations à savoir les caractéristiques de chaque groupe d'individu, par exemple toutes les molécules situées en haut de l'axe horizontal F1 et à gauche de l'axe verticale F2 ont l'énergie Gap et l'énergie HOMO corrélés positivement. Généralement les informations rassemblées des deux figures sont résumées dans le graphique des Biplot (Figure 3), ce dernier présente des difficultés d'interprétation du à la superposition des individus et les descripteurs.

Conclusion

L'ACP nous a permis d'extraire rapidement une quantité d'informations très intéressantes à partir d'un jeu de données multidimensionnelles, grâce à des graphiques simples. L'ACP fonctionne sur des tableaux de variables quantitatives et d'autres méthodes équivalentes sont disponibles pour d'autres types de variables notamment l'analyse des correspondances multiples pour les variables qualitatives. Cette méthode est très utilisée par notre groupe de recherche pour étudier les corrélations structure-activités de molécules possédant des activités biologiques (insecticides, bactériologiques, toxicologiques, pharmacologiques.....)

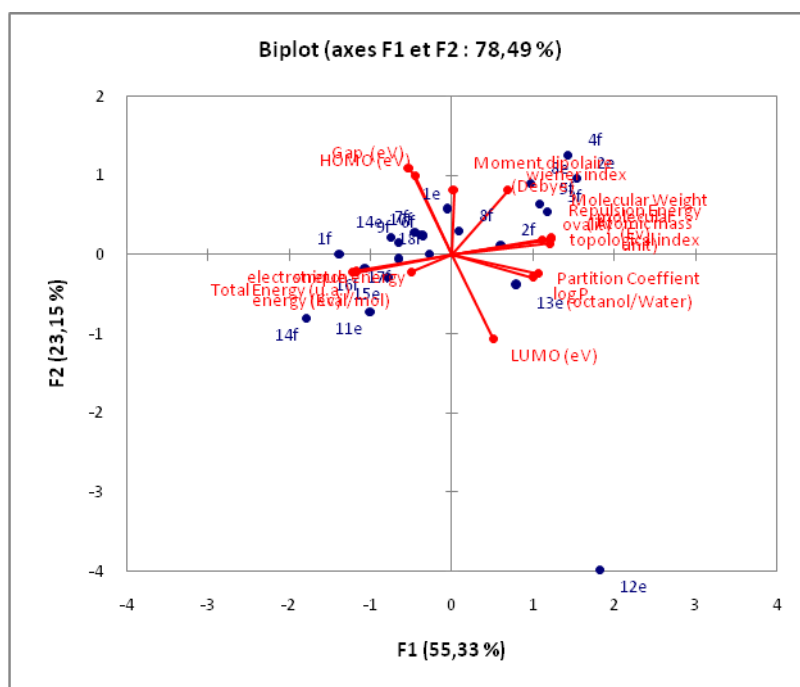


Figure 3 : Graphique des Biplot

Références

- Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* 34, 122–148
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24, 417–441.
- Jolliffe, I. T. (1986). *Principal component analysis*. Springer-verlag.
- Michael E. Tipping, christopher m. Bishop (1999). *Probabilistic principal component analysis* microsoft research.