Exemple illustratif pour l'A.C.P.

Lorsqu'on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4 !), comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.
C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire, par un procédé mathématique approprié, les facteurs que l'on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques désirés dans cet espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus telle qu'elle est définie par l'ensemble des variables initiales (ainsi remplacées par les facteurs).
C'est l'interprétation de ces graphiques qui permettra de comprendre la structure des données analysées. Cette interprétation sera guidée par un certain nombre d'indicateurs numériques, appelés aides à l'interprétation, qui sont là pour aider l'utilisateur à faire l'interprétation la plus juste et la plus objective possible.

Sur le plan théorique, l'Analyse en Composantes Principales est une méthode relativement complexe, dans la mesure où elle fait appel à des notions mathématiques non élémentaires : celles de matrices, d'éléments propres…. Pour faciliter la tâche du lecteur, nous avons choisi de présenter l'A.C.P. à travers son déroulement sur un exemple fictif, très simple, et qui parlera à tout le monde : les notes obtenues par des élèves dans diverses disciplines.

1. Présentation

Considérons les notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines (mathématiques, physique, français, anglais) :

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

On sait comment analyser séparément chacune de ces 4 variables, soit en faisant un graphique, soit en calculant des résumés numériques. Nous savons également qu'on peut regarder les liaisons entre 2 variables (par exemple mathématiques et français), soit en faisant un graphique du type nuage de points, soit en calculant leur coefficient de corrélation linéaire, voire en réalisant la régression de l'une sur l'autre Mais, comment faire une étude simultanée des 4 variables, ne serait-ce qu'en réalisant un graphique? La difficulté vient de ce que les individus (les élèves) ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension 4 (chaque élève étant caractérisé par les 4 notes qu'il a obtenues). L'objectif de l'Analyse en Composantes Principales est de revenir à un espace de dimension

réduite (par exemple, ici, 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent des données initiales.
Par analogie, on peut penser au photographe qui cherche le meilleur angle de vue pour transcrire en dimension 2 (le plan de sa photo) une scène située en dimension 3 (notre espace ambiant). La méthode mathématique va se charger de trouver l'angle de vue optimal, se substituant ainsi au coup d'œil du photographe...
Nous présentons ci-dessous quelques résultats de l'A.C.P. réalisée, sur ces données. Cela va permettre de se rendre compte des possibilités de la méthode. On notera que l'on s'est limité à 2 décimales dans les résultats.

2. Résultats

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

La matrice de corrélation

Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

Remarquons que toutes les corrélations linéaires sont positives (ce qui signifie que toutes les variables varient, en moyenne, dans le même sens), certaines étant très fortes (0.98 et 0.95), d'autres moyennes (0.65 et 0.51), d'autres enfin plutôt faibles (0.40 et 0.23).

3. Résultats généraux

Continuons l'examen des sorties de cette analyse par l'étude de la matrice des variances covariances, matrice de même nature que celle des corrélations, bien que moins parlante (nous verrons néanmoins plus loin comment elle est utilisée concrètement). La diagonale de cette matrice fournit les variances des 4 variables considérées (on notera qu'au niveau des calculs, il est plus commode de manipuler la variance que l'écart-type ; pour cette raison, dans de nombreuses méthodes statistiques, comme l'A.C.P., on utilise la variance pour prendre en compte la dispersion d'une variable quantitative).

Matrice variance - covariance

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Les valeurs propres données ci-dessous sont celles de la matrice des variances-covariances.

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00
	40.30	1.00	

Interprétation

Chaque ligne du tableau ci-dessus correspond à une variable virtuelle (voilà les facteurs !) dont la colonne val. pr. (valeur propre) fournit la variance (en fait, chaque valeur propre représente la variance du facteur correspondant). La colonne pct. var, ou pourcentage de variance, correspond au pourcentage de variance de chaque ligne par rapport au total. La colonne pct. cum., ou pourcentage cumulé représente le cumul de ces pourcentages.
Additionnons maintenant les variances des 4 variables initiales (diagonale de la matrice des variances-covariances) : 11,39 + 8,94 + 12,06 + 7,91 = 40,30: La dispersion totale des individus considérés, en dimension 4, est ainsi égale à 40,30.
Additionnons par ailleurs les 4 valeurs propres obtenues : 28,23+12,03+0,03+0,01 = 40,30: Le nuage de points en dimension 4 est toujours le même et sa dispersion globale n'a pas changé. C'est la répartition de cette dispersion, selon les nouvelles variables que sont les facteurs, ou composantes principales, qui se trouvent modifié : les 2 premiers facteurs restituent à eux seuls la quasi-totalité de la dispersion du nuage, ce qui permet de négliger les 2 autres.
Par conséquent, les graphiques en dimension 2 présentés ci-dessous résument presque parfaitement la configuration réelle des données qui se trouvent en dimension 4 : l'objectif (résumé pertinent des données en petite dimension) est donc atteint.

4. Résultats sur les variables

Le résultat fondamental concernant les coordonnées des variables
Les deux premières colonnes de ce tableau permettent, tout d'abord, de réaliser le graphique des variables donné par la Fig. 1.1.
Mais, ces deux colonnes permettent également de donner une signification aux facteurs (donc aux axes des graphiques).

FACTEURS -->	F1	F2	F3	F4
MATH	0.81	-0.58	0.01	-0.02
PHYS	0.90	-0.43	-0.03	0.02
FRAN	0.75	0.66	-0.02	-0.01
ANGL	0.91	0.40	0.05	0.01

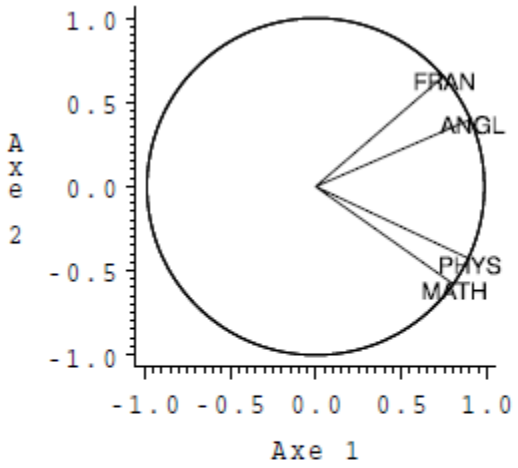

FIG. 1.1 – *Représentation des variables*

On notera que les deux dernières colonnes ne seront pas utilisées puisqu'on ne retient que deux dimensions pour interpréter l'analyse.

Interprétation

Ainsi, on voit que le premier facteur est corrélé positivement, et assez fortement, avec chacune des 4 variables initiales : plus un élève obtient de bonnes notes dans chacune des 4 disciplines, plus il a un score élevé sur l'axe 1 ; réciproquement, plus ses notes sont mauvaises, plus son score est négatif ; l'axe 1 représente donc, en quelques sortes, le résultat global (dans l'ensemble des 4 disciplines considérées) des élèves. En ce qui concerne l'axe 2, il oppose, d'une part, le français et l'anglais (corrélations positives), d'autre part, les mathématiques et la physique (corrélations négatives). Il s'agit donc d'un axe d'opposition entre disciplines littéraires et disciplines scientifiques, surtout marqué par l'opposition entre le français et les mathématiques. Cette interprétation, qui est déjà assez claire, peut être précisée avec graphiques et tableaux relatifs aux individus. Nous les présentons maintenant.

5. Résultats sur les individus

Le tableau donné ci-dessous contient tous les résultats importants de l'A.C.P. sur les individus.

Coordonnées des individus ; contributions ; cosinus carrés

	POIDS	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
jean	0.11	-8.61	-1.41	20.99	29.19	1.83	0.97	0.03
alan	0.11	-3.88	-0.50	4.22	5.92	0.23	0.98	0.02
anni	0.11	-3.21	3.47	6.17	4.06	11.11	0.46	0.54
moni	0.11	9.85	0.60	26.86	38.19	0.33	1.00	0.00
didi	0.11	6.41	-2.05	12.48	16.15	3.87	0.91	0.09
andr	0.11	-3.03	-4.92	9.22	3.62	22.37	0.28	0.72
pier	0.11	-1.03	6.38	11.51	0.41	37.56	0.03	0.97
brig	0.11	1.95	-4.20	5.93	1.50	16.29	0.18	0.82
evel	0.11	1.55	2.63	2.63	0.95	6.41	0.25	0.73

On notera que chaque individu représente 1 élément sur 9, d'où un poids (une pondération) de 1/9 = 0,11, ce qui est fourni par la première colonne du tableau.
Les 2 colonnes suivantes fournissent les coordonnées des individus (les élèves) sur les deux premiers axes (les facteurs) et ont donc permis de réaliser le graphique des individus. Ce dernier (Fig. 1.2) permet de préciser la signification des axes, donc des facteurs.

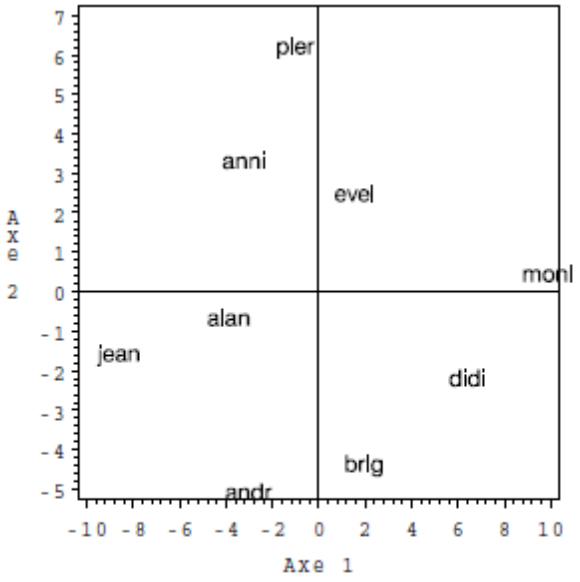

FIG. 1.2 – *Représentation des individus*

La signification et l'utilisation des dernières colonnes du tableau seront explicitées un peu plus loin.

Interprétation

On confirme ainsi que l'axe 1 représente le résultat d'ensemble des élèves : si on prend leur score { ou coordonnée { sur l'axe 1, on obtient le même classement que si on prend leur moyenne générale. Par ailleurs, l'élève le plus haut sur le graphique, celui qui a la coordonnée la plus élevée sur l'axe 2, est Pierre dont les résultats sont les plus contrastés en faveur des disciplines littéraires (14 et 11.5 contre 7

et 5.5). C'est exactement le contraire pour André qui obtient la moyenne dans les disciplines scientifiques (11 et 10) mais des résultats très faibles dans les disciplines littéraires (7 et 5.5). On notera que Monique et Alain ont un score voisin de 0 sur l'axe 2 car ils ont des résultats très homogènes dans les 4 disciplines (mais à des niveaux très distincts, ce qu'a déjà révélé l'axe 1). L'axe 2 oppose bien les littéraires (en haut) aux scientifiques (en bas).
Les 3 colonnes suivantes du tableau fournissent des contributions des individus à diverses dispersions : cont1 et cont2 donnent les contributions (en pourcentages) des individus à la variance selon les axes 1 et 2 (rappelons que l'on utilise ici la variance pour mesurer la dispersion) ;
contg donne les contributions générales, c'est - à -dire à la dispersion en dimension 4 (il s'agit de ce que l'on appelle l'inertie du nuage des élèves ; la notion d'inertie généralise celle de variance en dimension quelconque, la variance étant toujours relative à une seule variable). Ces contributions sont fournies en pourcentages (chaque colonne somme à 100) et permettent de repérer les individus les plus importants au niveau de chaque axe (ou du nuage en dimension 4). Elles servent en général à affiner l'interprétation des résultats de l'analyse.

Jean lui seul, cet individu représente près de 30 % (29,19) de la variance : il est prépondérant (au même titre que Monique) dans la définition de l'axe 1 ; cela provient du fait qu'il a le résultat le plus faible, Monique ayant, à l'opposé, le résultat le meilleur.
Enfin, les 2 dernières colonnes du tableau sont des cosinus carrés qui fournissent la qualité de la représentation de chaque individu sur chaque axe. Ces quantités s'additionnent axe par axe, de sorte que, en dimension 2, Evelyne est représentée à 98 % (0.25 + 0.73), tandis que les 8 autres individus le sont à 100 %.
Précisons un peu cette notion. Lorsqu'on considère les données initiales, chaque individu (chaque élève) est représenté par un vecteur dans un espace de dimension 4 (les éléments - ou coordonnées - de ce vecteur sont les notes obtenues dans les 4 disciplines). Lorsqu'on résume les données en dimension 2, et donc qu'on les représente dans un plan, chaque individu est alors représenté par la projection du vecteur initial sur le plan en question. Le cosinus carré relativement aux deux premières dimensions (par exemple, pour Evelyne, 0.98 ou 98 %) est celui de l'angle formé par le vecteur initial et sa projection dans le plan. Plus le vecteur initial est proche du plan, plus l'angle en question est petit et plus le cosinus, et son carré, sont proches de 1 (ou de 100 %) : la représentation est alors très bonne. Au contraire, plus le vecteur initial est loin du plan, plus l'angle en question est grand (proche de 90 degrés) et plus le cosinus, et son carré, sont proches de 0 (ou de 0 %) : la représentation est alors très mauvaise. On utilise les carrés des cosinus parce qu'ils s'additionnent suivant les différentes dimensions, contrairement à leurs racines.