Introduction à l'Analyse en Composantes Principales (ACP)

Résumé

Introduction élémentaire aux techniques factorielles de réduction de dimension lors de l'étude de p variables quantitatives. Meilleures représentations planes des individus et des variables. Valeurs propres et vecteurs propres d'une matrice de variances ou corrélation et composantes principales.

Retour au plan.

1 Introduction

1.1 objectif

La description des liaisons entre deux variables par des techniques statistiques bidimensionnelles conduisent à se poser la question de la représentation simultanées de données en dimension plus grande que 2. Quelle graphique permettrait de "généraliser" le nuage de points tracé dans le cas de deux variables permettant d'aborder la structure de corrélation présente entre plus de 2 variables. L'outil utilisé est alors l'analyse en composantes principales.

Mathématiquement, l'analyse en composantes principales est un simple *changement de base* : passer d'une représentation dans la base canonique des variables initiales à une représentation dans la base des *facteurs* définis par les vecteurs propres de la matrice des corrélations.

1.2 Exemple jouet

Une présentation très élémentaire de cette démarche est proposée sur un exemple jouet de données. 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

Nous savons 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 (chacun 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.

2 Descriptions uni et bivariée

Tout logiciel fournit la moyenne, l'écart-type, le minimum et le maximum de chaque variable. Il s'agit donc, pour l'instant, d'études univariées.

Statistiques élémentaires

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

Notons au passage la grande homogénéité des 4 variables considérées : même ordre de grandeur pour les moyennes, les écarts-types, les minima et les maxima.

Le tableau suivant est la *matrice des corrélations*. Elle donne les coefficients de corrélation linéaire des variables prises deux à deux. C'est une succession d'analyses bivariées, constituant un premier pas vers l'*analyse multivariée*.

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 Décomposition spectrale de la matrice des covariances

3.1 Résultats numériques

Continuons l'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 en A.C.P., on utilise la variance pour prendre en compte la dispersion d'une variable quantitative).

Matrice des variances-covariances

	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	

3.2 Interprétation statistique

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). Un facteur est une combinaison linéaire des variables initiales dans laquelle les coefficients sont données par les coordonnées des vecteurs propres (changement de base).

L'ACP peut être définie comme la recherche des *combinaisons linéaires de plus grande variance, des variables initiales* (les valeurs propres).

La colonne PCT. VAR, ou pourcentage de variance, correspond au pourcentage de variance de chaque ligne par rapport au total. La colonne PCT. CUM. représente le cumul de ces pourcentages en dimension 1, 2... 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. Il s'agit d'un simple changement de base dans un espace vectoriel.

C'est la répartition de cette dispersion, selon les nouvelles variables de plus grande dispersion, que sont les facteurs ou *composantes principales*, qui se trouve modifiée : 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 donnée en petite dimension) est donc atteint.

3.3 Interprétation géométrique

Une autre interprétation est d'ordre géométrique (cf. figure 1). Chaque individu x_i (resp. variable x^j) est considéré comme un vecteur à p (resp. n) composantes dans un espace vectoriel. L'ACP est la recherche du meilleur plan (ou sous-espace) de projection : le plus proche au sens des moindres carrés, pour obtenir la représentation la plus fidèle, ou la moins déformée, des individus (resp. des variables) dans un sous-espace de dimension réduite.

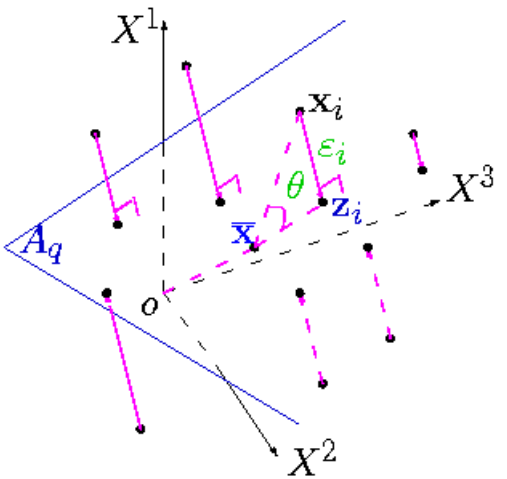

FIGURE 1 – *Interprétation géométrique de l'ACP comme la recherche du meilleur sous-espace de représentation.*

4 Étude des variables

4.1 Résultats numériques

Le résultat fondamental concernant les variables est le tableau des **corrélations variables-facteurs**. Il s'agit des coefficients de corrélation linéaire entre les variables initiales et les facteurs. Ce sont ces corrélations qui vont permettre de donner un sens aux facteurs (de les interpréter).

Corrélations variables-facteurs

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

Les deux premières colonnes de ce tableau permettent, tout d'abord, de réaliser le *graphique des variables* (version SAS) de la figure 2.

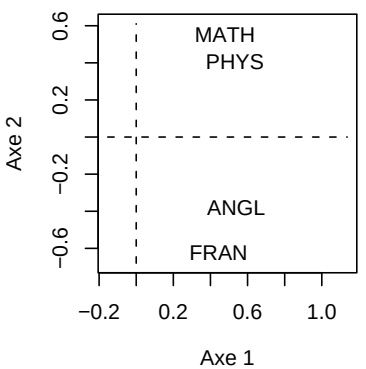

FIGURE 2 – *Données fictives : Représentation des variables*

Mais, ces deux colonnes permettent également de donner une signification aux facteurs (donc aux axes des graphiques).

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

4.2 Interprétation

Par construction, le cosinus de l'angle de deux vecteurs variables approche le coefficient de corrélation entre ces variables. Ainsi, on lit (cf. figure 2) 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. Le premier facteur représente approximativement la note moyenne (centrée sur la moyenne de la classe) de chaque élève. 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. L'axe 2 approche donc la moyenne des matières scientifique moins la moyenne des matières littéraires.

Cette interprétation peut être précisée avec les graphiques et tableaux relatifs aux individus que nous présentons maintenant.

5 Étude des individus

5.1 Résultats numériques

Le tableau ci-dessous contient tous les résultats importants sur les individus.

	Coordonnées des individus et cosinus carrés				
	POIDS	FACT1	FACT2	COSCA1	COSCA2
jean	0.11	-8.61	-1.41	0.97	0.03
alan	0.11	-3.88	-0.50	0.98	0.02
anni	0.11	-3.21	3.47	0.46	0.54
moni	0.11	9.85	0.60	1.00	0.00
didi	0.11	6.41	-2.05	0.91	0.09
andr	0.11	-3.03	-4.92	0.28	0.72
pier	0.11	-1.03	6.38	0.03	0.97
brig	0.11	1.95	-4.20	0.18	0.82

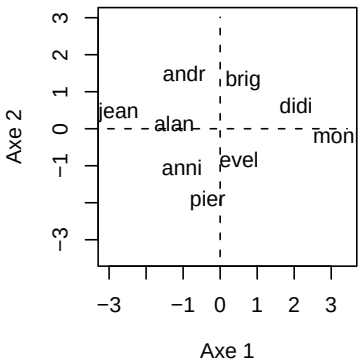

FIGURE 3 – *Données fictives : Représentation des individus*

evel	0.11	1.55	2.63	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 ci-dessus.

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 permet de préciser la signification des axes, donc des facteurs.

5.2 Interprétation

On peut ainsi voir 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).

5.3 Compléments à l'interprétation

Des logiciels comme SPAD fournissent d'autres résultats d'aide à l'interprétation.

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, Évelyne est représentée à 98 % (0.25 + 0.73), tandis que les 8 autres individus le sont à 100 %.

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 Évelyne, 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.

6 Représentation simultanée

Un troisième type de représentation graphique associant individus et variables (le *biplot*) est détaillé dans le document décrivant plus précisément l'analyse en composantes principales. Ce graphe associant des vecteurs individus et variables appartenant à des espaces vectoriels différents nécessite un développement plus détailler pour en justifier la construction et l'interprétation.

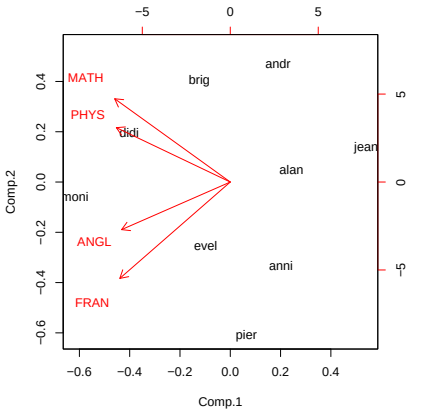

FIGURE 4 – *Données fictives : Représentation simultanée*