

# Analyse en composantes principales

L'analyse en composantes principales (ACP), développée en France dans les années 1960 par J-P. Benzécri, est une méthode de statistique exploratoire permettant de décrire un grand tableau de données de type individus / variables. Lorsque les individus sont décrits par un nombre important de variables, aucune représentation graphique simple ne permet de visualiser le nuage de points formé par les données. L'ACP propose une représentation dans un espace de dimension réduite, permettant ainsi de mettre en évidence d'éventuelles structures au sein des données. Pour cela, nous recherchons les sous-espaces dans lesquels la projection du nuage déforme le moins possible le nuage initial.

Dans la suite, nous identifions un vecteur  $\mathbf{x} \in \mathbb{R}^p$  avec la matrice colonne de ses coordonnées dans la base canonique, que nous notons  $\mathbf{x} = (x_1, \dots, x_p)^t$ , où pour une matrice  $A$ ,  $A^t$  désigne sa transposée.

## 1 Les données

Les données sont sous la forme d'un tableau à  $n$  lignes et  $p$  colonnes, que l'on stocke sous la forme d'une matrice  $X$  de taille  $n \times p$  :

- chaque ligne  $(x_i^1, \dots, x_i^p)$  de  $X$  représente les valeurs prises par l'**individu**  $i$  sur les  $p$  variables,
- de même chaque colonne  $(x_1^j, \dots, x_n^j)^t$  de  $X$  représente les valeurs de la **variable**  $j$  pour les  $n$  individus.

Par simplicité de langage, l'individu  $i$  sera indentifié au vecteur  $\mathbf{x}_i = (x_i^1, \dots, x_i^p)^t$  de  $\mathbb{R}^p$  tandis que la variable  $j$  sera identifiée au vecteur  $\mathbf{x}^j = (x_1^j, \dots, x_n^j)^t$  de  $\mathbb{R}^n$ .

L'exemple que nous traiterons tout au long de ce document est le suivant : lors d'un concours agricole, un jury a donné des notes à 10 marques de cidres relativement à 10 critères de dégustation. Le Tableau 1 reprend ces notes. Dans notre formalisme d'ACP, les marques de cidres sont les individus et les critères gustatifs sont les variables.

### 1.1 Point moyen du nuage des individus

Le vecteur  $g$  des moyennes arithmétiques de chacune des  $p$  variables définit le **point moyen** du nuage des individus :

$$g = (\bar{x}^1, \dots, \bar{x}^p)^t \quad (1)$$

où  $\bar{x}^j = \frac{1}{n} \sum_{i=1}^n x_i^j$  est la moyenne<sup>1</sup> des valeurs prises par la  $j$ -ème variable.

En retranchant  $\bar{x}^j$  aux valeurs de la variable  $j$  prises par les individus, on construit le tableau  $Y$  des **données centrées** :  $y_i^j = x_i^j - \bar{x}^j$ . La matrice  $Y$  correspondante s'écrit en fonction de  $X$  de la façon

---

<sup>1</sup>Notez que nous supposons ici que tous les individus ont le même poids, ce qui n'est pas systématique en ACP.

cidre	odeur	sucré	acide	amer	astringence	suffocante	piquante	alcool	parfum	fruitée
1	2,14	1,86	3,29	2,29	2	0,14	2,29	1,86	1,29	1,29
2	2,43	0,79	2,71	2,57	2	0,43	2,57	2,86	0,43	0,14
3	2,71	3,14	2,57	2,57	1,43	0,14	2,14	0,86	2,29	1,71
4	3	3,71	2,14	2,07	1,57	0	1,29	1	3,14	3,14
5	3,43	1,29	2,86	3,14	2,17	1	1,86	2,86	1,14	0,29
6	3,14	0,86	2,86	3,79	2,57	0,14	1,71	3,29	0,14	0
7	3,14	1,14	2,86	2,86	2	0,43	1,71	1,86	0,14	0
8	2,43	3,71	3,21	1,57	1,71	0	1	0,57	2,57	2,86
9	5,1	2,86	2,86	3,07	1,79	1,71	0,43	1,43	0,57	2,71
10	3,07	3,14	2,57	3	2	0	0,43	1,29	2,57	3,07

TAB. 1 – Notes obtenues par 10 marques de cidres sur 10 critères lors d’un concours agricole.

suivante :  $Y = X - \mathbf{1}g^t$  où  $\mathbf{1}$  est le vecteur de  $\mathbb{R}^n$  dont toutes les composantes valent 1. Le terme *centré* signifie que les moyennes des variables  $y^j$  sont nulles.

## 1.2 Matrice de variance

35 On note  $s_j^2 = \frac{1}{n} \sum_{i=1}^n (x_i^j - \bar{x}^j)^2$  la **variance** de la variable  $j$ , sa racine carrée  $s_j$  est l’**écart-type** et  $v_{kl} = \frac{1}{n} \sum_{i=1}^n (x_i^k - \bar{x}^k)(x_i^l - \bar{x}^l)$  est la **covariance** des variables  $k$  et  $l$ .  
On appelle **matrice de variance** la matrice symétrique  $V$  contenant les variances  $s_j^2$  sur la diagonale et les covariances  $v_{kl}$  en dehors de la diagonale (ligne  $k$  colonne  $l$  pour  $v_{kl}$ ). Cette matrice s’écrit :

$$V = \frac{1}{n} X^t X - gg^t = \frac{1}{n} Y^t Y. \quad (2)$$

40 De même, on définit le **coefficient de corrélation linéaire** entre les variables  $k$  et  $l$  par  $r_{kl} = \frac{v_{kl}}{s_k s_l}$ . Ce coefficient exprime le niveau de corrélation (linéaire) entre les variables  $k$  et  $l$  : plus il est proche de 1, plus les variables sont corrélées positivement, plus il est proche de -1, plus elles sont corrélées négativement. Un coefficient de corrélation nul indique l’absence de corrélation linéaire.

En divisant chaque colonne  $j$  du tableau centré  $Y$  par l’écart-type  $s_j$  de la variable  $j$ , on construit le tableau  $Z$  des **données centrées réduites** :  $z_i^j = (x_i^j - \bar{x}^j)/s_j$ . La matrice  $Z$  s’exprime en fonction de  $X$  par  $Z = (X - \mathbf{1}g^t)D_{1/s}$  où  $D_{1/s}$  est la matrice diagonale contenant  $\frac{1}{s_1}, \dots, \frac{1}{s_p}$  sur sa diagonale. Le terme *réduit* signifie que les variances des variables  $z^j$  sont égales à 1.

La matrice  $R = D_{1/s} V D_{1/s}$  est dite de **corrélations**. Regroupant les coefficients de corrélation linéaire entre les  $p$  variables prises deux à deux, elle résume la structure des dépendances linéaires entre les  $p$  variables. Elle est symétrique et sa diagonale est composée de 1.

La matrice de corrélation pour l’exemple des cidres est la suivante (Table 2, les corrélations importantes apparaissent en gras).

	odeur	sucré	acide	amer	astringence	suffocante	piquante	alcool	parfum	fruitée
odeur	1,00	0,08	-0,16	0,49	0,04	<b>0,84</b>	-0,61	0,03	-0,29	0,18
sucré	0,08	1,00	-0,29	-0,60	<b>-0,77</b>	-0,19	-0,61	<b>-0,92</b>	<b>0,87</b>	<b>0,95</b>
acide	-0,16	-0,29	1,00	-0,08	0,34	0,14	0,14	0,15	-0,40	-0,27
amer	0,49	-0,60	-0,08	1,00	<b>0,71</b>	0,38	-0,03	<b>0,70</b>	<b>-0,63</b>	-0,50
astringence	0,04	<b>-0,77</b>	0,34	<b>0,71</b>	1,00	0,07	0,14	<b>0,86</b>	<b>-0,66</b>	<b>-0,64</b>
suffocante	<b>0,84</b>	-0,19	0,14	0,38	0,07	1,00	-0,23	0,22	-0,50	-0,10
piquante	-0,61	-0,61	0,14	-0,03	0,14	-0,23	1,00	0,48	-0,33	<b>-0,73</b>
alcool	0,03	<b>-0,92</b>	0,15	<b>0,70</b>	<b>0,86</b>	0,22	0,48	1,00	<b>-0,76</b>	<b>-0,83</b>
parfum	-0,29	<b>0,87</b>	-0,40	<b>-0,63</b>	<b>-0,66</b>	-0,50	-0,33	<b>-0,76</b>	1,00	<b>0,80</b>
fruitée	0,18	<b>0,95</b>	-0,27	-0,50	<b>-0,64</b>	-0,10	<b>-0,73</b>	<b>-0,83</b>	<b>0,80</b>	1,00

TAB. 2 – Matrice de corrélation des critères notés sur les cidres.

## 2 Les espaces des individus et des variables

### 2.1 L'espace des individus

#### 2.1.1 La métrique

Dans l'espace des individus, que nous supposons ici être  $\mathbb{R}^p$ , chaque individu est représenté par un point, formant ainsi un nuage de points. L'objectif de l'ACP est de visualiser ce nuage de point dans un espace de faible dimension le plus fidèlement possible. L'analyse repose donc sur les distances entre individus dans  $\mathbb{R}^p$ , et le choix de la métrique aura donc une influence sur le résultat de l'ACP. La distance  $d(i, j)$  entre deux individus  $i$  et  $j$  peut s'écrire de façon générale :

$$d^2(i, j) = d^2(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_M^2 = \langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j \rangle_M = (\mathbf{x}_i - \mathbf{x}_j)^t M (\mathbf{x}_i - \mathbf{x}_j). \quad (3)$$

où  $M$  est une matrice symétrique définie positive de taille  $p$  spécifiant la distance choisie,  $\langle \cdot, \cdot \rangle_M$  et  $\|\cdot\|_M$  étant le produit scalaire et la norme associés à la métrique  $M$ .

Lorsque toutes les variables sont exprimées dans la même unité (ce qui est le cas dans l'exemple des cidres) et que les variances ne sont pas trop différentes (de sorte que chaque variable ait la même importance dans le calcul des distances entre points), la distance euclidienne convient. La métrique est alors la matrice identité  $M = I$  et la distance euclidienne s'écrit classiquement :

$$d(i, j) = \left( \sum_{k=1}^p (x_i^k - x_j^k)^2 \right)^{\frac{1}{2}}. \quad (4)$$

Dans le cas contraire, la métrique la plus couramment utilisée est la métrique des inverses des variances  $M = D_{1/s^2}$ , où  $D_{1/s^2}$  est la matrice diagonale contenant  $\frac{1}{s_1^2}, \dots, \frac{1}{s_p^2}$  sur sa diagonale. Cette métrique permet à la fois de s'affranchir de l'unité de mesure et de donner à chaque variable la même importance dans le calcul de la distance. C'est cette métrique que nous utiliserons dans ce document.

**Remarque 2.1.** La matrice de la métrique des inverses des variances peut s'écrire  $M = D_{1/s^2} = D_{1/s} D_{1/s}$ . Ainsi, la distance entre deux individus  $i$  et  $j$  est

$$d^2(i, j) = (\mathbf{x}_i - \mathbf{x}_j)^t D_{1/s} D_{1/s} (\mathbf{x}_i - \mathbf{x}_j) = (D_{1/s} \mathbf{x}_i - D_{1/s} \mathbf{x}_j)^t (D_{1/s} \mathbf{x}_i - D_{1/s} \mathbf{x}_j). \quad (5)$$

Travailler avec la métrique  $M = D_{1/s^2}$  sur le tableau  $X$  revient donc à travailler avec la métrique identité sur le tableau réduit  $D_{1/s}X$ , ou encore sur le tableau centré réduit  $Z$ . L'ACP usuelle revient donc à centrer et réduire les variables puis à utiliser la métrique identité : c'est ce qu'on appelle une **ACP normée**.

### 2.1.2 L'inertie

L'inertie est une notion fondamentale en ACP, puisqu'elle est une mesure de la dispersion du nuage de points autour de son centre de gravité  $g$ . L'**inertie totale** du nuage autour de son centre de gravité  $g$  est :

$$I_g = \frac{1}{n} \sum_{i=1}^n d^2(g, \mathbf{x}_i) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - g)^t M (\mathbf{x}_i - g). \quad (6)$$

Plus l'inertie est grande, plus le nuage est dispersé, et au contraire plus elle est petite, plus le nuage est concentré sur son centre de gravité.

Lorsque  $M = I$ , l'inertie totale peut s'écrire

$$I_g = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p (x_i^j - \bar{x}^j)^2 = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n (x_i^j - \bar{x}^j)^2 = \sum_{j=1}^p s_j^2 \quad (7)$$

qui n'est autre que la trace de la matrice de variance  $V$

$$I_g = \text{tr}(V). \quad (8)$$

Ce résultat se généralise au cas d'une métrique  $M$  quelconque en  $I_g = \text{tr}(VM)$  (cf. Annexe 5.1). Dans le cas de l'ACP normée, la matrice de variance du tableau  $Z$  est la matrice de corrélation  $R$  et sa trace est donc égale au nombre  $p$  de variables  $I_g = p$ .

L'**inertie du nuage par rapport à un axe**  $\Delta$  passant par le centre de gravité  $g$  est

$$I_\Delta = \frac{1}{n} \sum_{i=1}^n d^2(\Delta, \mathbf{x}_i) = \frac{1}{n} \sum_{i=1}^n d^2(h_{\Delta i}, \mathbf{x}_i) \quad (9)$$

où  $h_{\Delta i}$  est la projection orthogonale de  $\mathbf{x}_i$  sur l'axe  $\Delta$ . Cette inertie mesure la proximité du nuage des individus à l'axe.

De même, l'**inertie du nuage par rapport à un sous-espace vectoriel**  $F$  passant par le centre de gravité  $g$  est

$$I_F = \frac{1}{n} \sum_{i=1}^n d^2(h_{F i}, \mathbf{x}_i) \quad (10)$$

où  $h_{F i}$  est la projection orthogonale de  $\mathbf{x}_i$  sur le sous-espace  $F$ .

Soit  $F^\perp$  le supplémentaire orthogonal de  $F$  dans  $\mathbb{R}^p$ , alors d'après le théorème de Pythagore (cf. Figure 1 en dimension 3) on a

$$d^2(h_{F i}, \mathbf{x}_i) + d^2(h_{F^\perp i}, \mathbf{x}_i) = d^2(g, \mathbf{x}_i) = d^2(g, h_{F i}) + d^2(g, h_{F^\perp i}). \quad (11)$$

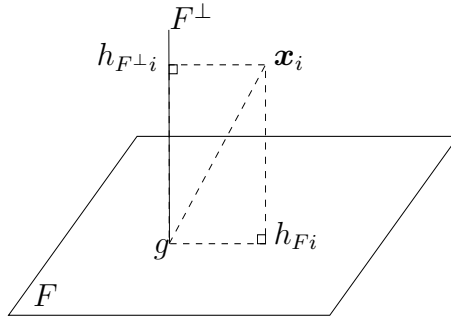


FIG. 1 –

On en déduit le théorème de Huygens :

$$I_F + I_{F^\perp} = I_g. \quad (12)$$

Lorsque  $F$  est un axe, l'inertie  $I_{F^\perp}$  par rapport à son supplémentaire orthogonal mesure alors l'allongement du nuage de points selon cet axe  $F$ , et est appelée l'**inertie portée par l'axe** ou inertie **expliquée** par l'axe. Ainsi, en projetant le nuage des individus sur un sous-espace  $F$ , on perd l'inertie mesurée par  $I_F$  et on ne conserve que  $I_{F^\perp}$ .

Attention à ne pas confondre  $I_F$ , l'inertie par rapport à l'axe  $F$ , et  $I_{F^\perp}$  l'inertie expliquée ou portée par ce même axe, qui n'est autre que l'inertie par rapport à son orthogonal.

**Expression de l'inertie  $I_{\Delta_u^\perp}$  portée par un axe  $\Delta_u$**  L'inertie portée par l'axe  $\Delta_u$  de vecteur directeur  $\mathbf{u}$  est

$$I_{\Delta_u^\perp} = \frac{1}{n} \sum_{i=1}^n d^2(h_{\Delta_u^\perp, i}, \mathbf{x}_i) \quad (13)$$

Comme on peut s'en convaincre sur la Figure 2,  $d(h_{\Delta_u^\perp, i}, \mathbf{x}_i) = \langle \mathbf{x}_i, \mathbf{u} \rangle_M$ .

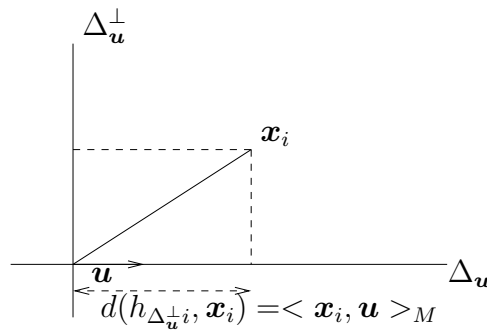


FIG. 2 –

Ainsi ,

$$I_{\Delta_u^\perp} = \frac{1}{n} \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{u} \rangle_M^2 = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^t M \mathbf{u})^2 = \frac{1}{n} \sum_{i=1}^n \mathbf{u}^t M \mathbf{x}_i \mathbf{x}_i^t M \mathbf{u} \quad (14)$$

car  $(\mathbf{x}_i^t M \mathbf{u})^t = \mathbf{x}_i^t M \mathbf{u}$  puisque  $c$ 'est un scalaire et que  $M$  est symétrique,

$$I_{\Delta_u^\perp} = \mathbf{u}^t M \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^t \right) M \mathbf{u} = \frac{1}{n} \mathbf{u}^t M X^t X M \mathbf{u}. \quad (15)$$

Ainsi, si  $X$  est centrée de matrice de variance  $V$ , on a

$$I_{\Delta_u^\perp} = \mathbf{u}^t M V M \mathbf{u} \quad (16)$$

125

## 2.2 L'espace des variables

Chaque variable est considérée comme un vecteur d'un espace de dimension  $n$ , l'espace des variables. La métrique utilisée pour le calcul des distances entre variables est la métrique identité, car on suppose que tous les individus ont le même poids. Si ce n'est pas le cas une métrique spécifique doit alors être utilisée.

130

Soit  $\mathbf{y}^1, \dots, \mathbf{y}^p$  les variables  $\mathbf{x}^1, \dots, \mathbf{x}^p$  centrées. On a les propriétés suivantes :

- le produit scalaire entre deux variables  $\mathbf{y}^k$  et  $\mathbf{y}^l$  (sous la métrique identité) est

$$\langle \mathbf{y}^k, \mathbf{y}^l \rangle = \sum_{i=1}^n y_i^k y_i^l = v_{kl}, \quad (17)$$

- le carré de la norme d'une variable est égale à sa variance

$$\|\mathbf{y}^k\|^2 = s_k^2 \quad (18)$$

et l'écart-type de la variance représente donc sa longueur,

- le cosinus de l'angle  $\theta_{kl}$  entre deux variables  $\mathbf{y}^k$  et  $\mathbf{y}^l$  est leur coefficient de corrélation linéaire :

$$\cos(\theta_{kl}) = \frac{\langle \mathbf{y}^k, \mathbf{y}^l \rangle}{\|\mathbf{y}^k\| \|\mathbf{y}^l\|} = \frac{v_{kl}}{s_k s_l} = r_{kl}. \quad (19)$$

Nous nous intéressons donc, dans l'espace des variables, aux angles entre variables plutôt qu'aux distances, et on représente les variables comme des vecteurs et non des points.

140

## 3 L'analyse en composantes principales

Nous travaillons désormais en ACP normée, avec le tableau de données centrées  $Z = (X - \mathbf{1}g^t)D_{1/s}$  et la métrique identité. Les individus et les variables sont désormais les  $\mathbf{z}_i$  et  $\mathbf{z}^j$ .

L'ACP consiste à chercher un sous-espace  $F_k$  de dimension  $k$  inférieure à celle de l'espace de départ, tel que le nuage, une fois projeté dans ce sous-espace, soit au minimum déformé. Comme la projection diminue nécessairement les distances, on cherche le sous-espace  $F_k$  qui maximise la moyenne des carrés des distances entre individus :

145

$$\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d^2(i, j) = 2I_g. \quad (20)$$

Il faut donc que l'inertie du nuage projeté soit maximale.

150

**Théorème 3.1.** Soit  $F_k$  un sous-espace portant l'inertie maximale, alors le sous-espace de dimension  $k + 1$  portant l'inertie maximale est la somme directe de  $F_k$  et du sous-espace de dimension 1 orthogonal à  $F_k$  portant l'inertie maximale.

La démonstration de ce théorème figure en annexe 5.2.

155 Il est donc possible de rechercher le sous espace  $F_k$  séquentiellement axe par axe, en cherchant tout d'abord l'axe portant l'inertie maximale, puis celui orthogonal à ce dernier portant l'inertie maximale, et ainsi de suite.

### 3.1 Recherche de l'axe portant l'inertie maximale

160 On cherche l'axe  $\Delta_{\mathbf{u}}$  tel que l'inertie  $I_{\Delta_{\mathbf{u}}^\perp}$  expliquée par cet axe soit maximale. Cela revient à chercher  $\Delta_{\mathbf{u}}$  tel que  $I_{\Delta_{\mathbf{u}}}$  soit minimale d'après le théorème de Huygens (12). Puisqu'en projetant sur l'axe  $\Delta_{\mathbf{u}}$  on perd l'inertie  $I_{\Delta_{\mathbf{u}}}$ , on aura bien une inertie restante  $I_{\Delta_{\mathbf{u}}^\perp}$  maximale, ce qui revient à déformer le moins possible le nuage des individus.

165 Comme on est en ACP normée la métrique  $M$  est l'identité et la matrice de variance  $V$  est égale à la matrice de corrélation  $R$  (cf. remarque 2.1). L'inertie expliquée par l'axe  $\Delta_{\mathbf{u}}$  est alors  $I_{\Delta_{\mathbf{u}}^\perp} = \mathbf{u}^t R \mathbf{u}$  d'après (16). Il faut donc trouver le vecteur unitaire  $\mathbf{u}$  solution du problème d'optimisation sous contrainte suivant :

$$\begin{cases} \max_{\mathbf{u}} \mathbf{u}^t R \mathbf{u} \\ \mathbf{u}^t \mathbf{u} = 1 \end{cases} \quad (21)$$

170 La matrice de corrélation  $R$  étant symétrique elle est diagonalisable :  $R = P \Delta P^t$ , où  $\Delta$  est la matrice diagonale composée des valeurs propres  $\lambda_1 \geq \dots \geq \lambda_j \geq \dots \geq \lambda_p$ , et où la matrice de passage  $P$  est la matrice orthogonale dont les colonnes sont les vecteurs propres  $\mathbf{v}_j$  de  $R$ . Il vient alors que

$$\mathbf{u}^t R \mathbf{u} = \sum_{j=1}^p \lambda_j \langle \mathbf{u}, \mathbf{v}_j \rangle^2. \quad (22)$$

En nommant  $u_j$  les composantes du vecteur  $\mathbf{u}$  (normé) dans la base des vecteurs propres, on a

$$\mathbf{u}^t R \mathbf{u} = \sum_{j=1}^p \lambda_j u_j^2 \leq \lambda_1 \underbrace{(u_1^2 + \dots + u_p^2)}_{=\mathbf{u}^t \mathbf{u} = 1} \leq \lambda_1. \quad (23)$$

175 Le vecteur  $\mathbf{u}$  maximisant cette quantité n'est autre que  $\mathbf{v}_1$ , le vecteur propre associé à la plus grande valeur propre  $\lambda_1$  de  $R$ , pour lequel on a donc  $I_{\Delta_{\mathbf{v}_1}^\perp} = \lambda_1$ .

### 3.2 Recherche des axes suivants

Nous cherchons cette fois un vecteur unitaire  $\mathbf{u}$ , orthogonal au précédent ( $\mathbf{v}_1$ ), et maximisant la quantité  $\mathbf{u}^t R \mathbf{u}$ . En nommant encore  $u_j$  les composantes de ce vecteur  $\mathbf{u}$  dans la base des vecteurs propres, on a :

$$\mathbf{u}^t R \mathbf{u} = \sum_{j=2}^p \lambda_j u_j^2 \leq \lambda_2 (u_2^2 + \dots + u_p^2) \leq \lambda_2. \quad (24)$$

Le vecteur  $u$  maximisant cette quantité n'est autre que  $v_2$ , le vecteur propre associé à la seconde plus grande valeur propre  $\lambda_2$  de  $R$ , pour lequel on a donc  $I_{\Delta_{v_2}^\perp} = \lambda_2$ .  
Et ainsi de suite. On en déduit ainsi le théorème suivant.

**Théorème 3.2.** *Le sous-espace  $F_K$  de dimension  $K$  portant l'inertie maximale est engendré par les  $K$  vecteurs propres associés aux  $K$  plus grandes valeurs propres de la matrice de corrélation  $R$  du nuage des individus.*

### 3.3 Notations

- Les axes  $\Delta_{v_k}$  sont appelés **axes factoriels** ou **axes principaux**.
- L'inertie expliquée par l'axe  $\Delta_{v_k}$  est la valeur propre  $\lambda_k$
- L'**inertie expliquée** par le sous-espace factoriel  $F_K$  engendré par les axes factoriels est

$$I_{F_K^\perp} = \lambda_1 + \dots + \lambda_K, \quad (25)$$

et le pourcentage d'inertie expliqué par ce sous-espace est donc  $\frac{\lambda_1 + \dots + \lambda_K}{p}$ .

- On appelle  $k$ -ième **composante principale** les coordonnées  $c^k \in \mathbb{R}^n$  des  $n$  individus sur l'axe factoriel  $\Delta_{v_k}$ , qui sont les projections des individus sur ces axes :

$$c^k = Zv_k. \quad (26)$$

Les composantes principales peuvent être vue comme les nouvelles variables dans le sous-espace factoriel.

## 4 Interprétation des résultats d'une ACP

### 4.1 Qualité de représentation sur les axes factoriels

Un des points les plus délicats de l'ACP est d'apprécier la perte d'information engendrée par la réduction de la dimension. On a vu précédemment que la qualité de représentation du nuage dans le sous-espace factoriel est exprimée par le pourcentage d'inertie expliquée  $\frac{\lambda_1 + \dots + \lambda_K}{p}$ . Cette mesure globale doit être complétée par d'autres considérations comme la représentation des individus et des variables dans ce sous-espace factoriel. Il peut en effet arriver que les axes retenus, bien que représentant une part importante de l'inertie globale du nuage, ne suffisent pas à expliquer correctement certains individus ou variables : deux individus peuvent être proches en projection sur les axes factoriels retenus tout en étant très éloignés en réalité, si leurs représentations sur ces axes ne sont pas de bonne qualité.

La **qualité de représentation d'un individu**  $i$  sur un axe factoriel est mesuré par le **cosinus carré** de l'angle entre l'axe factoriel et le vecteur  $z_i$ . Plus le cosinus est grand, plus  $z_i$  sera proche de l'axe factoriel et donc sera bien représenté sur cet axe.

La **qualité de représentation d'une variable**  $j$  sur le  $k$ -ième axe factoriel est exprimée par le coefficient de corrélation linéaire  $r(c^k, z^j)$  entre la variable initiale  $z^j$  et la nouvelle variable, composante principale,  $c^k$ . La valeur de cette corrélation sera également très importante pour interpréter les nouveaux axes factoriels en fonction des variables initiales. Le calcul de cette corrélation, non développé dans ce document, montre que

$$r(c^k, z^j) = \sqrt{\lambda_k} v_k^j, \quad (27)$$



où  $v_k^j$  est la  $j$ -ème coordonnée du vecteur  $v_k$ .

220 Ainsi, bien qu'ayant retenu un sous-espace factoriel expliquant une part importante de l'inertie totale, il est possible que certaines variables ou individus d'intérêt soient mal représentés dans ce sous-espace. Il sera alors intéressant de compléter le sous-espace factoriel en ajoutant des axes factoriels supplémentaires de sorte que ces variables ou individus d'intérêt soient bien représentés.

## 4.2 Sélection du nombre $k$ d'axes factoriels à retenir

225 La sélection du nombre d'axes à retenir, abordée précédemment, est une étape importante d'une ACP. Les critères utilisés sont en général empiriques : la méthode du coude qui consiste à détecter un coude sur le diagramme des valeurs propres, ou le critère de Kaiser. Ce dernier consiste à ne retenir que les valeurs propres supérieures à la moyenne : puisqu'on est en ACP normée la somme des valeurs propres est égale au nombre  $p$  de ces valeurs propres, et la moyenne des valeurs propres est donc égale à 1. On ne retiendra donc par le critère de Kaiser que les valeurs propres supérieures à 1.

## 230 4.3 Interprétation

La méthode la plus naturelle pour interpréter les composantes principales (ou les axes factoriels) est de les relier aux variables initiales, en utilisant les corrélations précédemment introduites. Généralement, pour un couple de composantes principales  $c^1$  et  $c^2$ , on représente les corrélations dans une figure appelée cercle de corrélation (Figure 3), où chaque variable  $z^j$  est représentée par un point de coordonnées  $(r(c^1, z^j), r(c^2, z^j))$ .

235 Une fois les composantes principales interprétées, on représente les individus dans les plans factoriels formés en croisant deux à deux les axes factoriels retenus (Figure 4). Chaque représentation est interprétée en essayant de repérer des groupes d'individus et en donnant à ces groupes une signification en fonction de leur place dans les plans factoriels.

## 240 5 Application à l'exemple des cidres

Une ACP normée a été réalisée sur les données de cidres précédemment introduites, et nous présentons une analyse synthétique des résultats dans cette section.

La première étape consiste à sélectionner le nombre d'axes factoriels que l'on souhaite conserver. En utilisant le critère de Kaiser, nous sélectionnons les 3 premières valeurs propres (Tableau 3), qui expliquent ensemble plus de 87% de l'inertie totale du nuage de points. Néanmoins, comme le troisième

	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	$\lambda_5$	$\lambda_6$	$\lambda_7$	$\lambda_8$	$\lambda_9$	$\lambda_{10}$
valeur propre	5,154	2,502	1,097	0,834	0,194	0,14	0,049	0,024	0,006	0
inertie expliquée (%)	51,54	25,02	10,97	8,34	1,94	1,40	0,49	0,24	0,06	0
inertie expl. cumulée (%)	51,54	76,56	87,53	95,87	97,81	99,21	99,70	99,94	100	100

TAB. 3 – Tableau des valeurs propres.

245 axe n'est corrélé significativement qu'avec une seule variable, la variable acidité, nous ne le considérons pas dans l'interprétation synthétique présentée dans ce document.

Le cercle des corrélations pour le plan formé des deux premiers axes factoriels est représenté Figure 3. Exceptée la variable acidité, qui est fortement corrélée avec le troisième axe factoriel, toutes les

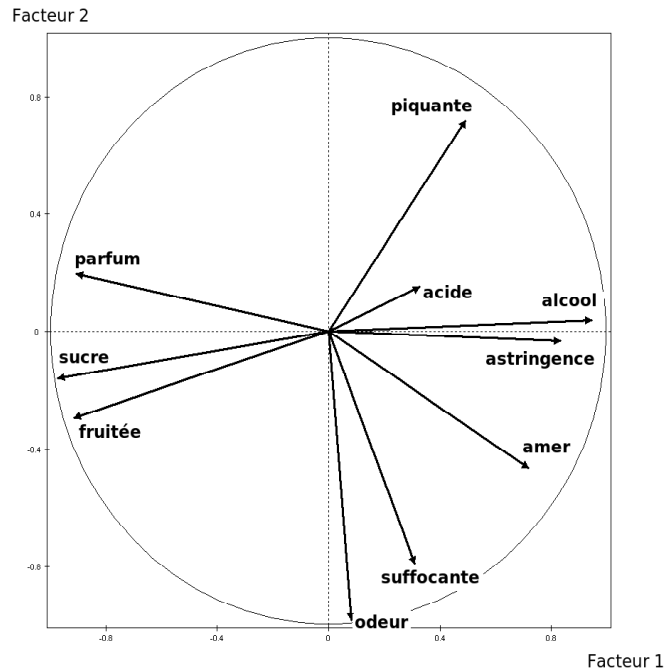


FIG. 3 – Cercle des corrélations pour le premier plan factoriel.

250 variables sont bien représentées dans ce plan factoriel puisque leurs corrélations avec les axes sont relativement importantes (les projections sont proches du cercle de corrélation). L'interprétation que l'on peut faire des deux premiers axes factoriels est la suivante :

- le premier axe factoriel semble opposer le cidre doux (fruité, sucré, parfumé) au cidre brut (plus alcoolisé et astringent),
- 255 – le second axe factoriel semble opposer les cidres ayant une particularité olfactive (forte odeur) aux cidres ayant une certaine particularité gustative (piquance).

Ce plan factoriel étant interprété, nous pouvons y projeter les individus (Figure 4).

Après avoir vérifié que les individus étaient bien représentés sur le plan factoriel en examinant les valeurs des cosinus carrés des angles entre les individus et les axes factoriels (valeurs des cosinus carrés non présentées dans ce document), il semble se dégager 4 groupes de cidres :

- groupe 1 : les cidres 3, 8, 4 et 10 qui sont des cidres doux,
- groupe 2 : les cidres 2, 5, 6 et 7 qui sont des cidres bruts,
- groupe 3 : le cidre 9 qui est un cidre particulièrement odorant et suffocant,
- groupe 4 : le cidre 1 qui est particulièrement piquant.

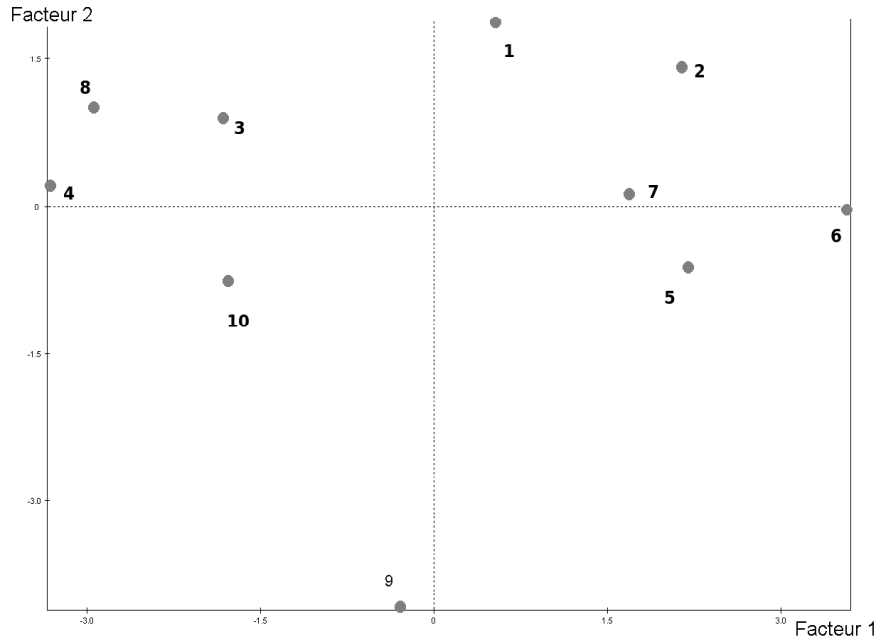


FIG. 4 – Projections des individus dans le premier plan factoriel.

## 265 Annexes

### 5.1 Annexe 1 : démonstration de la formule de l'inertie totale $I_g = \text{tr}(VM)$

$$I_g = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - g)^t M (\mathbf{x}_i - g) = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i^t M \mathbf{y}_i = \frac{1}{n} \sum_{i=1}^n \text{tr}(\mathbf{y}_i^t M \mathbf{y}_i) \quad (28)$$

$$= \text{tr}\left(\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^t M\right) = \text{tr}\left(\frac{1}{n} Y^t Y M\right) \quad (29)$$

$$= \text{tr}(VM) \quad (30)$$

### 270 5.2 Annexe 2 : démonstration du théorème 3.1

Soit  $E_{k+1}$  un espace de dimension  $k + 1$ .

Comme  $\dim E_{k+1} = k + 1$  et  $\dim F_k^\perp = p - k$  on a :

$$\underbrace{\dim(E_{k+1} + F_k^\perp)}_{\leq p} = \underbrace{\dim E_{k+1} + \dim F_k^\perp}_{=p+1 > p} - \dim(E_{k+1} \cap F_k^\perp) \quad (31)$$

d'où nécessairement  $\dim(E_{k+1} \cap F_k^\perp) \geq 1$ .

275 Soit  $\mathbf{v}$  un vecteur de  $E_{k+1} \cap F_k^\perp$ .

Posons  $E_{k+1} = \mathbf{v} \oplus G$  où  $G$  est le supplémentaire orthogonal de  $\mathbf{v}$  dans  $E_{k+1}$ .  $G$  est donc de dimension  $k$ . Soit  $F_{k+1} = F_k \oplus \mathbf{v}$ .

Comme  $\mathbf{v}$  est orthogonal à  $G$  et à  $F_k$ , on a :

$$I_{E_{k+1}} = I_{\mathbf{v}} + I_G, \quad (32)$$

$$I_{F_{k+1}} = I_{F_k} + I_{\mathbf{v}}. \quad (33)$$

<sup>280</sup> Comme  $F_k$  était le sous-espace de dimension  $k$  d'inertie maximale, on a  $I_G \leq I_{F_k}$  donc  $I_{E_{k+1}} \leq I_{F_{k+1}}$ , et ce quel que soit  $E_{k+1}$ .

Le maximum d'inertie est donc réalisé pour l'espace  $F_{k+1} = F_k \oplus \mathbf{v}$  où  $\mathbf{v}$  doit être de sorte que  $I_{\mathbf{v}}$  soit maximale.  $\square$