Journal de la Société Française de Statistique

# Analyses factorielles de données structurées en groupes d'individus

**Title:** Multivariate data analysis of multi-group datasets

Aida Eslami[1], El Mostafa Qannari[2], Achim Kohler[3] et Stéphanie Bougeard[1]

**Résumé :** Après une brève discussion des méthodes d'analyse d'un tableau de données où les individus sont partitionnés en groupes connus *a priori*, l'analyse en composantes principales multi-groupes (Krzanowski, 1984) est plus particulièrement étudiée. Un nouveau critère d'optimisation qui caractérise cette méthode est proposé. Par la suite, la méthode est étendue au cas des données structurées en multi-groupes et multi-tableaux. Le cas particulier où les différents tableaux portent sur les mêmes variables est également considéré. Les démarches d'analyse sont illustrées sur la base d'études de cas.

**Abstract:** An outline of the methods of analysis of a dataset where individuals are partitioned into groups is given. Thereafter, the paper focuses on multi-group principal components analysis (Krzanowski, 1984). A new optimization criterion which characterizes this method is discussed. An extension of the strategy of analysis to the case of multi-block datasets is presented. The particular case where the various blocks pertain to the same variables is also discussed. The methods are illustrated on the basis of case studies.

***Mots-clés :*** analyse en composantes principales multi-groupes, analyse multi-groupes et multi-tableaux
***Keywords:*** multi-group principal components analysis, multi-block multi-group data analysis
***Classification AMS 2000 :*** 62-07

## 1. Introduction

Il arrive souvent que les données à analyser se rapportent à un échantillon d'individus qui sont structurés en groupes. Ceci constitue, en particulier, le cadre général de l'analyse discriminante (Saporta, 2006). Cependant, notre objectif n'est pas de discuter des méthodes de discrimination des groupes d'individus mais plutôt de chercher à identifier une structure commune aux différents groupes en vue d'une représentation factorielle des individus. En d'autres termes, nous nous situons dans un cadre qui s'apparente à l'analyse en composantes principales (ACP) car nous cherchons à restituer l'inertie des tableaux associés aux différents groupes. Ce problème a été étudié dans beaucoup d'articles scientifiques. Nous renvoyons à une publication synthétique de Cazes (2004) à ce sujet. Une étude comparative de différentes méthodes d'ACP multi-groupes a été proposée par Eslami et al. (sous presse). Un bref aperçu de cette étude est présenté ci-après (section 2). Par la suite, nous focalisons sur une méthode appelée 'ACP multi-groupes'

[1] Anses, Département d'épidémiologie animale - Zoopole, BP53, 22440, Ploufragan, France.
E-mail : aida.eslami@anses.fr and E-mail : stephanie.bougeard@anses.fr
[2] LUNAM Université, ONIRIS, Unité Sensométrie et Chimiométrie, Nantes, 44307, France ; INRA, Nantes, 44307, France.
E-mail : elmostafa.qannari@oniris-nantes.fr
[3] Université des sciences de la vie, CIGENE, Ås, Norvège.
E-mail : achim.kohler@umb.no


 ISSN: 2102-6238

qui fut introduite par Krzanowski (1984) (section 3). Cette méthode présente l'avantage d'être directe et simple. En effet, elle revient à effectuer une ACP sur le tableau obtenu en concaténant verticalement les tableaux associés aux différents groupes, ces tableaux ayant été préalablement centrés par groupe.

Nous proposons également une extension de l'ACP multi-groupes au cas des tableaux qui comportent plusieurs blocs de variables (section 4). Ce type de données (tableaux multi-groupes et multi-blocs) est couramment rencontré dans des applications concrètes. En épidémiologie vétérinaire, nous pouvons citer l'exemple de données mesurées sur des abeilles qui sont réparties en colonies (groupes) ; ces données peuvent concerner plusieurs blocs de variables liées à l'environnement des animaux, aux pathologies détectées .... En écologie, des blocs de variables se rapportant respectivement au climat, au paysage et à l'environnement sont régulièrement mesurées sur des sites géographiques qui sont structurés en régions. En analyse sensorielle, différents produits provenant de différents pays ou régions (groupes), sont décrits sur la base de variables structurées en blocs (analyse sensorielle, physico-chimique, données liées à la consommation et aux préférences). Enfin, en chimiométrie, nous pouvons imaginer le cas de données réparties en blocs (mesures physico-chimiques, mesures spectroscopiques, analyse d'images) concernant des produits qui sont eux-mêmes organisés en groupes correspondant à des conditions expérimentales différentes.

Le cas particulier où les blocs de données se rapportent aux mêmes variables présente un intérêt particulier (section 4.3). Cette configuration des données est connue sous l'appellation « données ternaires ». Elle concerne, par exemple, les données évolutives (mesures répétées dans le temps). Un exemple de données ternaires multi-groupes concerne la mesure de variables métabolomiques ou morphologiques sur des individus qui sont répartis en groupes correspondant à des régimes alimentaires spécifiques. Pour l'analyse de ce type de données, nous considérons le même critère d'optimisation que pour le cas de données multi-blocs et multi-groupes mais nous imposons en plus la contrainte que les axes factoriels soient les mêmes d'un bloc à un autre (en plus d'être les mêmes pour les différents groupes). Les stratégies d'analyse de tableaux multi-groupes et multi-blocs que nous proposons sont basées sur des algorithmes itératifs qui nécessitent la détermination de vecteurs propres de matrices. Enfin, nous illustrons les stratégies d'analyse sur la base d'études de cas (section 5).

## 2. Analyse d'un tableau de données structurées en groupes

Nous considérons un tableau de données $X$ constitué par les mesures de $P$ variables quantitatives sur $N$ individus. De plus, nous supposons que $X$ est *a priori* divisé en $M$ sous-tableaux $[(X_1)^T, \ldots, (X_M)^T]^T$. Le sous-tableau $X_m$ est associé à un groupe d'individus de taille $n_m$ ($\sum_{m=1}^{M} n_m = N$). Dans la suite, nous supposons que les tableaux $X_m$ sont centrés. La matrice des variances-covariances associée au groupe $m$ est donnée par $V_m = \frac{1}{n_m}(X_m)^T X_m$.

Le modèle général de l'ACP multi-groupes stipule que les matrices $V_m$ peuvent être approchées par $A\Lambda_m A^T$ avec $A^T A = I$ ($I$ : matrice identité) et $\Lambda_m$, matrice diagonale à éléments positifs. Pratiquement, ce modèle signifie que les différentes ACP effectuées sur des tableaux $(X_1, \ldots, X_M)$ sont supposées conduire aux mêmes axes factoriels. Ceci procure, par conséquent, un modèle parcimonieux qui, en plus de garantir une meilleure stabilité que des ACP séparées effectuées sur les différents groupes (ce qui nécessiterait l'estimation de beaucoup de paramètres), facilite


 ISSN: 2102-6238

l'interprétation des résultats en ce sens que la représentation des variables est commune pour les différents groupes. Les vecteurs constituant la matrice $A$ sont désignés par $a^{(h)}$ pour $h = (1,\ldots,H)$ où $H$ est le rang de la matrice $X$. Les scalaires situés dans la diagonale de la matrice $\Lambda_m$ sont désignés par $\lambda_m^{(h)}$.

La plupart des méthodes d'analyse des données multi-groupes focalisent sur la détermination des vecteurs $a^{(h)}$ $(h = 1,\ldots,H)$ et, par la suite, les scalaires $\lambda_m^{(h)}$ sont déterminés par :

$$\lambda_m^{(h)} = var(X_m a^{(h)}) = (a^{(h)})^T V_m a^{(h)}$$

Ainsi $\lambda_m^{(h)}$ reflète la part d'inertie de $X_m$ restituée par la composante $t_m^{(h)} = X_m a^{(h)}$. Pour quelques méthodes, il est commode d'exhiber, en plus des axes factoriels communs $a^{(h)}$, des axes factoriels associés aux différents groupes définis par $a_m^{(h)} = (X_m)^T t_m^{(h)}$ pour $m = (1,\ldots,M)$. Les composantes globales définies par $t^{(h)} = X a^{(h)}$ $(h = 1,\ldots,H)$. La figure 1 donne un aperçu général des différents éléments et des notations qui seront utiles par la suite.

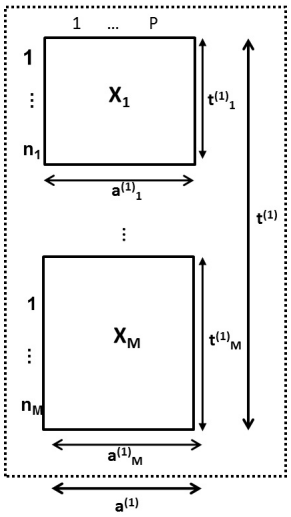



FIGURE 1. *Représentation de la partition d'un tableau X en groupes. Axe factoriel commun* $a^{(1)}$. *Axes factoriels spécifiques aux groupes* $a_1^{(1)},\ldots,a_M^{(1)}$ *correspondant à la première étape de l'analyse. Composantes spécifiques* $t_1^{(1)},\ldots,t_M^{(1)}$. *Composante globale* $t^{(1)}$.

Il convient de noter qu'il existe un parallèle entre les méthodes d'analyse de données multi-groupes et les méthodes d'analyse de tableaux multiples (tableaux de données mesurées sur les mêmes individus mais présentant une structure des variables en plusieurs tableaux). Ainsi plusieurs méthodes de données multi-groupes apparaissent comme des méthodes duales de méthodes d'analyse de tableaux multiples (STATIS duale, AFM duale ...). Le tableau 1 donne un aperçu général des méthodes utilisées pour l'analyse des données structurées en groupes. Pour plus de détails, nous renvoyons le lecteur aux références indiquées dans ce même tableau ainsi qu'à l'article (Eslami et al., sous presse). Une étude comparative des différentes méthodes issue du dernier article cité a indiqué qu'elles peuvent être réparties en deux groupes :

– Méthodes qui s'apparentent à l'ACP : ces méthodes visent à restituer l'inertie des tableaux


 ISSN: 2102-6238

associés aux différents groupes. Parmi ces méthodes nous pouvons citer l'analyse en composantes communes (Flury, 1984), l'ACP multi-groupes (Krzanowski, 1984), STATIS duale (Lavit, 1988; Lavit et al., 1994), AFM duale (Lê et al., 2010), l'analyse de Procrustes généralisée (Eslami et al., sous presse). Sur la base d'exemples concrets, il a été constaté que toutes ces méthodes donnent des résultats qui se recoupent dans une large mesure.

– Méthodes s'apparentant à l'analyse canonique : typiquement, c'est la méthode qui fut proposée par Krzanowski (1979) sous l'appellation 'Between groups comparison'.

TABLEAU 1. *Panorama des méthodes d'analyse des données multi-groupes*

| Type | Objectif général | Méthode (référence) | Démarche |
|---|---|---|---|
| Analyse s'apparentant à l'ACP | Restituer l'inertie des tableaux associés aux différents groupes | Analyse en composantes communes (Flury, 1984) | Les paramètres du modèle ($V_m = A\Lambda_m A^T$) sont estimés par la méthode du maximum de vraisemblance en supposant la multinormalité des données dans les différents groupes. |
| | | ACP multi-groupes (Krzanowski, 1984) | Détaillée ci-après (section 3) |
| | | STATIS duale (Lavit, 1988) | Recherche d'une matrice $V$ qui constitue un compromis des matrices $V_m$ ($m = 1,\ldots,M$). La matrice $A$ (loadings communs) est obtenue par décomposition spectrale de $V$. |
| | | Analyse Factorielle Multiple duale (Lê et al., 2010) | Une ACP est effectuée sur le tableau obtenu en concaténant verticalement les tableaux associés aux groupes, centrés et standardisés. |
| | | Analyse de Procrustes Généralisée duale (Eslami et al., sous presse) | Une analyse de Procrustes Généralisée est appliquée à $(X_m)^T$ conduisant à un tableau compromis $C^T$. La matrice $A$ (loadings communs) est obtenu par décomposition en valeurs singulières de $C$. |
| Analyse s'apparentant à l'analyse canonique | Exhiber des ressemblances entre les profils des individus dans les différents groupes | Comparaison de groupes (Between groups comparison) (Krzanowski, 1979) | Les tableaux $X_m$ sont approchés (via une décomposition en valeurs singulières) par des tableaux $\tilde{X}_m$ de rang inférieur. Une analyse canonique généralisée est appliquée à $(\tilde{X}_m)^T$ ($m = 1,\ldots,M$). |

Dans la suite, nous allons focaliser sur l'ACP multi-groupes car tout en étant conceptuellement très simple, elle a de bonnes propriétés et présente une bonne performance en termes de restitution d'inertie.

## 3. ACP multi-groupes

En partant du modèle stipulant que la matrice $V_m$ ($m = 1,\ldots,M$) peut être approchée par $A\Lambda_m A^T$ ($A^T A = I$), Krzanowski (1984) a remarqué que toute combinaison linéaire des matrices $V_m$, $\sum_{m=1}^{M} \alpha_m V_m$ ($\alpha_m$ : scalaires positifs), peut être approchée par $A\sum_{m=1}^{M} \alpha_m \Lambda_m A^T$. En particulier, si nous considérons la matrice des covariances intra-classes $W = \sum_{m=1}^{M} \frac{n_m}{N} V_m$, nous pouvons l'approcher par $A\sum_{m=1}^{M} \frac{n_m}{N} \Lambda_m A^T$. Ainsi, la stratégie d'analyse consiste à déterminer la matrice $A$ à partir de la décomposition spectrale de la matrice $W$. Par la suite, les matrices $\Lambda_m$ sont déterminées


 ISSN: 2102-6238

par :

$$\Lambda_m = diag(A^T V_m A)$$

Il est clair que l'ACP multi-groupes est intimement liée à l'ACP effectuée sur le tableau $X$, obtenu en concaténant verticalement les matrices $X_m$ ($m = 1,\ldots,M$) car cette ACP conduit également à la diagonalisation de la matrice $W$. L'intérêt de présenter cette analyse en tant qu'analyse multi-groupes plutôt qu'une ACP globale permet d'exhiber des éléments qui peuvent être utiles pour l'interprétation des résultats (figure 1).

Une autre manière de présenter l'ACP multi-groupes consiste en une démarche séquentielle basée sur un critère de maximisation. A la première étape, nous cherchons un vecteur de coefficients (loadings) commun, $a$, ainsi que des coefficients spécifiques associés aux différents groupes : $a_m = (X_m)^T t_m$ ($t_m$ étant un vecteur, à déterminer, de dimension $n_m$) de manière à maximiser le critère :

$$\frac{1}{N}\sum_{m=1}^{M} < (X_m)^T t_m, a >^2 \quad \text{sous les contraintes} \quad \| t_m \| = \| a \| = 1 \tag{1}$$

Il est facile de montrer que pour $a$ fixé, la solution pour $t_m$ est donnée par $t_m = \frac{X_m a}{\|X_m a\|}$. En remplaçant cette valeur dans l'expression à maximiser, il apparaît que nous sommes conduits à maximiser la quantité $\sum_{m=1}^{M} \frac{n_m}{N} var(X_m a)$ (démonstration reportée en Annexe-A). Ceci indique que l'analyse recherchée s'apparente à une 'ACP moyenne' en ce sens que nous cherchons à restituer (en moyenne) le maximum d'inertie avec des composantes qui sont déterminées à l'aide du même vecteur de loadings. Il est facile de vérifier que le vecteur $a$ qui maximise cette inertie moyenne restituée est donné par le vecteur propre de $W$ associé à la plus grande valeur propre. Un deuxième axe factoriel commun pourrait être obtenu en remplaçant le tableau $X_m$ par le tableau $X_m(I - aa^T)$ et en procédant de la même manière que pour la première étape. Cette procédure connue sous l'appellation 'déflation' est, de nos jours, bien connue par les praticiens de l'analyse multivariée des données. Dans le cas présent, nous pouvons vérifier que le deuxième vecteur de loadings est donné par le vecteur propre de $W$ associé à la deuxième plus grande valeur propre. D'autres vecteurs de loadings pourraient être obtenus en réitérant la même procédure. Nous pouvons remarquer que, par construction, les axes factoriels sont orthogonaux. Nous pouvons aussi remarquer que les composantes globales $t^{(h)} = Xa^{(h)}$ ($h = 1,\ldots,H$) sont également orthogonales. En effet, comme nous l'avons indiqué, les mêmes composantes peuvent être déterminées par l'ACP de tableau concaténé $X$.

Il est à noter qu'à partir du critère d'optimisation indiqué ci-dessus, l'analyse multi-groupes apparaît comme étant une analyse duale de l'analyse de co-inertie multiple (Chessel and Hanafi, 1996) qui est, elle-même, intimement liée à l'analyse connue sous l'appellation consensus PCA (Westerhuis et al., 1998).

Comme premiers éléments d'aide à l'interprétation, nous proposons de déterminer les parts d'inertie resituée par les composantes à l'échelle globale et à l'échelle de chaque groupe. Soit $\mathbf{T}_m^{(h)} = [t_m^{(1)},\ldots,t_m^{(h)}]$ la matrice constituée par les $h$ composantes principales associées au groupe $m$ ($m = 1,\ldots,M$) et soit $\hat{X}_m^{(h)}$ la projection de $X_m$ sur $\mathbf{T}_m^{(h)}$. La part d'inertie de $X_m$ restituée par les $h$ premières composantes du groupe $m$ est donnée par :

$$Inertie_m^{(h)} = \frac{trace((\hat{X}_m^{(h)})^T \hat{X}_m^{(h)})}{trace((\mathbf{X}_m)^T \mathbf{X}_m)}$$


 ISSN: 2102-6238

De manière similaire, nous définissions la part d'inertie de $X$ restituée par les $h$ premières composantes globales.

## 4. Analyse d'un tableau de données structurées en multi-blocs et en multi-groupes

### *4.1. Données et notations*

Comme nous l'avons indiqué dans l'introduction, il arrive souvent que les données soient structurées sous forme de tableaux multi-groupes et multi-blocs. Pour l'analyse de ce type de données, nous proposons une extension de l'ACP multi-groupes. Les différents blocs sont notés $X^{(b)}$ $(b = 1,\ldots,B)$. Ces tableaux se rapportent aux mêmes individus mais pas nécessairement aux mêmes variables. Chaque tableau $X^{(b)}$ est lui même structuré en $M$ groupes $X_m^{(b)}$ $(m = 1,\ldots,M)$ que nous supposons, dans la suite, centrés et, si l'utilisateur le juge nécessaire, standardisés.

A l'étape $h$ $(h \geq 1)$, pour laquelle nous omettrons l'indice $(h)$ pour alléger l'écriture, nous cherchons à déterminer pour chaque bloc $X^{(b)}$, un axe factoriel $a^{(b)}$, commun aux différents groupes $X_m^{(b)}$. Nous notons $t$ la composante globale associée à tous les blocs et tous les groupes. Cette composante est obtenue par la juxtaposition verticale des composantes $t_m$ $(m = 1,\ldots,M)$ associés aux différents groupes. Tous ces éléments sont illustrés dans la figure 2.

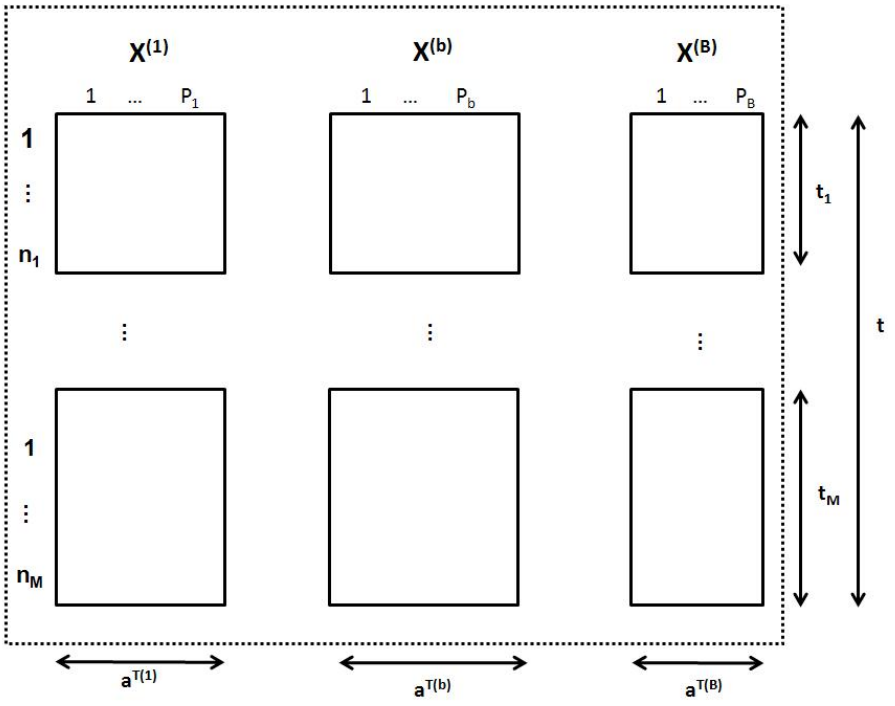



FIGURE 2. *Représentation des données multi-blocs multi-groupes. Pour une dimension donnée, $t$ est la composante globale commune à tous les individus, $(t_1,\ldots,t_M)$ les composantes partielles associées à chaque groupe d'individus, $(a^{(1)},\ldots,a^{(B)})$ les loadings communs à chaque bloc de variables.*


 ISSN: 2102-6238

### *4.2. Analyse proposée*

L'analyse proposée est définie par extension du critère de maximisation (1) qui caractérise l'ACP multi-groupes. Pour cela, nous cherchons à maximiser le critère (2) :

$$\sum_{b=1}^{B}\sum_{m=1}^{M} < (X_m^{(b)})^T t_m, a^{(b)} >^2 \quad \text{sous les contraintes} \quad \|a^{(b)}\| = \|t_m\| = 1 \tag{2}$$

Naturellement, nous retrouvons le critère (1) qui est à la base de l'ACP multi-groupes dans le cas où nous ne disposons que d'un seul bloc ($B = 1$). Le critère (2) peut s'écrire de deux manières différentes :

$$Q_1 = \sum_{b=1}^{B}\sum_{m=1}^{M} (t_m)^T X_m^{(b)} a^{(b)} (a^{(b)})^T (X_m^{(b)})^T t_m$$

$$Q_2 = \sum_{b=1}^{B}\sum_{m=1}^{M} (a^{(b)})^T (X_m^{(b)})^T t_m (t_m)^T X_m^{(b)} a^{(b)}$$

Ceci indique que pour $t_m$ $(m = 1,\ldots,M)$ fixés, le vecteur $a^{(b)}$ $(b = 1,\ldots,B)$ qui maximise le critère $Q_2$ est donné par un vecteur propre normé de $\sum_{m=1}^{M}(X_m^{(b)})^T t_m (t_m)^T X_m^{(b)}$ associé à la plus grande valeur propre. De même, pour $a^{(b)}$ fixés, le vecteur $t_m$ $(m = 1,\ldots,M)$ qui maximise le critère $Q_1$ est donné par un vecteur propre normé de la matrice $\sum_{b=1}^{B} X_m^{(b)} a^{(b)} (a^{(b)})^T (X_m^{(b)})^T$ associé à la plus grande valeur propre. Ainsi, un algorithme de résolution du problème de maximisation (2) est le suivant :

1. Initialisation : choisir des vecteurs $t_m$ $(m = 1,\ldots,M)$ normés
2. Calculer $a^{(b)}$ $(b = 1,\ldots,B)$, vecteur propre de $\sum_{m=1}^{M}(X_m^{(b)})^T t_m (t_m)^T X_m^{(b)}$ associé à la plus grande valeur propre
3. Calculer $t_m$ $(m = 1,\ldots,M)$, vecteur propre de $\sum_{b=1}^{B} X_m^{(b)} a^{(b)} (a^{(b)})^T (X_m^{(b)})^T$ associé à la plus grande valeur propre
4. Réitérer le processus commençant à partir de l'étape 2 jusqu'à convergence ( *i.e.* variation insignifiante, par exemple inférieure à $\varepsilon = 10^{-7}$, du critère (2) entre deux étapes successives).

La convergence de l'algorithme est assurée par le fait qu'à chaque étape, le critère croît. Comme, par ailleurs, ce critère est borné, l'algorithme est, par conséquent, convergent. Cependant, comme c'est souvent cas pour ce type d'algorithmes il peut arriver que la convergence soit réalisée pour un maximum local du critère à maximiser. Le principe de l'algorithme est assez facile à comprendre. En effet, il consiste en une succession d'ACP : pour $a^{(b)}$ fixé, nous considérons la matrice $[X_m^{(1)} a^{(1)},\ldots,X_m^{(B)} a^{(B)}]$ où chaque colonne est la composante spécifique du tableau $X_m^{(b)}$ $(b = 1,\ldots,B)$ associée au vecteur de loadings $a^{(b)}$. La composante globale $t_m$ du groupe $m$ est déterminée comme étant la première composante principale du tableau ainsi constitué. Par la suite, nous constituons le tableau $[(X_1^{(b)})^T t_1,\ldots,(X_M^{(b)})^T t_M]$ correspondant aux vecteurs des loadings associés aux groupes constituant le bloc $b$. Un vecteur de loadings global associé au bloc $b$ est déterminé comme étant la première composante de l'ACP (non centrée) du tableau ainsi formé.


 ISSN: 2102-6238

### 4.3. Cas des données ternaires

Un cas particulier très intéressant de l'analyse de données structurées en plusieurs blocs et en plusieurs groupes concerne les données ternaires multi-groupes. Ici, les blocs de tableaux portent sur les mêmes individus et les mêmes variables. Dans une perspective de rechercher un modèle parcimonieux, nous pourrions considérer un critère similaire à celui indiqué dans (2) mais nous imposons la contrainte que les vecteurs des loadings associés aux différents blocs soient égaux. Ceci conduit au problème d'optimisation suivant (3) :

$$\sum_{b=1}^{B}\sum_{m=1}^{M}\langle (X_m^{(b)})^T t_m, a\rangle^2 \quad \text{sous les contraintes} \quad \| t_m \| = \| a \| = 1 \tag{3}$$

Un algorithme de résolution est donné par :

1. Initialisation : choisir un vecteur $a$ normé
2. Calculer $t_m$ $(m = 1,\ldots,M)$, vecteur propre de $\sum_{b=1}^{B} X_m^{(b)} a a^T (X_m^{(b)})^T$ associé à la plus grande valeur propre
3. Calculer $a$, vecteur propre de $\sum_{b=1}^{B}\sum_{m=1}^{M}(X_m^{(b)})^T t_m (t_m)^T X_m^{(b)}$ associé à la plus grande valeur propre
4. Réitérer le processus commençant à l'étape 2 jusqu'à convergence.

### 4.4. Remarques

Dans la bibliographie, d'autres méthodes pourraient être adaptées et appliquées aux données présentant à la fois une structure multi-blocs et multi-groupes. Parmi celles-ci, deux peuvent être plus particulièrement comparées à la méthode proposée ici. La première, appelée DOuble-Analyse Conjointe de Tableaux (DO-ACT) est proposée par Vivien and Sabatier (2004). Cette méthode cherche l'accord entre deux tableaux présentant respectivement $K$ et $L$ blocs de variables. Pour cela, un critère d'optimisation basé sur une généralisation de la méthode STATIS est proposé et résolu en trois étapes : recherche d'une inter-structure commune à tous les blocs d'un tableau, recherche d'un compromis entre les tableaux, recherche d'une intra-structure commune à tous les tableaux. Nous pensons que l'élaboration d'une version duale d'une telle approche pourrait être intéressante pour le contexte que nous considérons dans cet article. La seconde méthode appelée STATIS-4 (Sabatier and Vivien, 2008) est une généralisation de STATIS et de la méthode DO-ACT. Elle vise à étudier les liens entre plus de deux tableaux, mesurés sur les mêmes observations et présentant une structure multi-blocs. Là aussi, la version duale de STATIS-4 pourrait offrir un cadre très général pour l'étude de données structurées en multi-groupes et multi-blocs.

Toutes les procédures discutées dans cet article ont été implémentées dans l'environnement R. Les programmes sont disponibles auprès des auteurs.

## 5. Illustration

### 5.1. ACP multi-groupes

Dans un premier temps, nous considérons un exemple d'application de l'ACP multi-groupes. Les données concernent des mesures physico-chimiques sur des vins portugais qui sont répartis en


 ISSN: 2102-6238

deux groupes : vins blancs (4898 individus) et vins rouges (1599 individus) (Cortez et al., 2009). Les données ont été centrées par groupe. L'objectif de cette illustration n'est pas de comprendre quelles variables permettent de discriminer les vins blancs et les vins rouges, mais plutôt de mettre en évidence la typologie des vins de chacune des deux catégories. Bien évidemment, l'ACP multi-groupes permet de procurer un modèle plus parcimonieux et des résultats plus faciles à interpréter que des ACP séparées effectuées sur chacun des deux groupes de vins. Ce souci de parcimonie de modèle et de facilité d'interprétation des résultats devient encore plus crucial lorsque le nombre de groupes est relativement important. Le tableau 2 donne la liste des variables mesurées ainsi que les abréviations utilisées pour les représentations graphiques.

TABLEAU 2. *Les listes des variables et abréviations du jeu de données sur les vins portugais.*

| Variables | Abréviations | Variables | Abréviations |
|---|---|---|---|
| Acidité | Acid | Soufre Total | SoufrTot |
| Acidité volatile | AcidVol | Densité | Densite |
| Acide citrique | Citriq | pH | pH |
| Sucres résiduels | Sucre | Sulfates | Sulfat |
| Chlorures | Chlor | Alcool | Alcool |
| Dioxyde de Soufre | DioxSoufr | | |

Le tableau 3 donne les pourcentages d'inertie restituée par les quatre premières composantes principales associées aux groupes ainsi que par les composantes globales.

TABLEAU 3. *ACP multi-groupes : pourcentages d'inertie cumulée restituée par les quatre premières composantes principales.*

| Groupe | Dim1 | Dim2 | Dim3 | Dim4 |
|---|---|---|---|---|
| Vins rouges | 29.19 | 43.27 | 54.48 | 63.76 |
| Vins blancs | 23.36 | 44.94 | 58.40 | 69.52 |
| Global | 27.25 | 43.16 | 54.77 | 64.52 |

La figure 3 donne la représentation des variables sur la base des loadings communs. A l'instar de ce qui est fait en ACP, cette figure pourrait servir pour l'investigation des relations entre les variables et pour interpréter la disposition des individus (non représentés ici). La composante $t^{(1)}$ qui explique 27.25% de l'inertie totale oppose la densité, liée à la mesure des sucres résiduels au degré d'alcool. Ainsi, cette composante reflète la conversion du sucre en alcool lors de la fermentation. La composante $t^{(2)}$ qui représente 15.91% d'inertie totale oppose la mesure de l'acidité et de l'acide citrique au pH.

A titre de comparaison de méthodes, nous avons effectué une ACP sur chacun des groupes de vins séparément. Le tableau 4 donne les pourcentages d'inertie restituée par les composantes principales des deux groupes. Ces résultats se recoupent dans une large mesure avec ceux obtenus pour l'ACP multi-groupes. Ceci montre que, bien qu'apparaissant comme une ACP compromis des deux groupes, l'ACP multi-groupes a un pouvoir de synthèse proche des ACP séparées par groupe.

Pour chacun des groupes, nous avons considéré la configuration des vins sur la base des deux premières composantes principales (ACP séparées) et nous avons calculé son coefficient RV avec la configuration correspondante obtenue sur la base des deux composantes déterminées à l'aide de l'ACP multi-groupes. Ce coefficient permet de mesurer le degré de similitude entre deux


 ISSN: 2102-6238

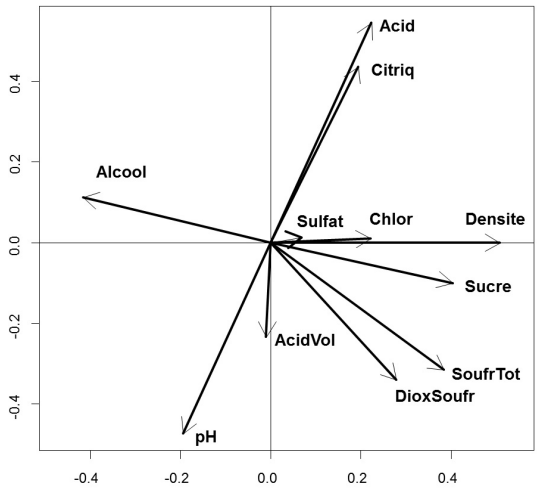



FIGURE 3. *Représentation des variables sur la base des loadings communs.*

TABLEAU 4. *ACP separées : pourcentages d'inertie cumulée restituée par les quatre premières composantes principales des deux groupes.*

| Groupe | Dim1 | Dim2 | Dim3 | Dim4 |
|---|---|---|---|---|
| Vins rouges | 29.29 | 43.61 | 54.72 | 63.98 |
| Vins blancs | 28.17 | 45.68 | 59.78 | 70.81 |

configurations (Robert and Escoufier, 1976). Il ressort que pour les vins rouges, le coefficient RV est égal à 0.93 et, pour les vins blancs, ce coefficient est égal à 0.97. Manifestement, ces coefficients sont très élevés indiquant une forte similitude entre les configurations considérées.

### *5.2. ACP multi-groupes et multi-blocs*

Nous considérons un exemple simple où les données sont structurées en deux blocs et où chaque bloc est lui-même formé de trois groupes d'individus. Il s'agit de mesures effectuées sur des huiles d'olive dont cinq proviennent de Grèce, cinq d'Italie et six d'Espagne. Ces trois pays constituent les groupes. L'objectif est d'explorer les relations entre les variables physico-chimiques et sensorielles en tenant compte de la provenance géographique des trois groupes d'huiles d'olive. Le premier bloc de données est formé par les mesures de six variables sensorielles et le deuxième bloc de données est formé de cinq variables physico-chimiques (tableau 5). Pour plus d'information, nous renvoyons le lecteur à (Massart et al., 1998). Les données sont disponibles dans le package pls de R (Mevik and Wehrens, 2007).

Nous nous situons ici dans une perspective de description des données et non pas de prédiction des données sensorielles à l'aide des données physico-chimiques. L'objectif est d'explorer les relations entre les deux blocs de données en tenant compte de la présence d'une structure de groupes. Pour répondre à cet objectif, nous avons effectué une ACP multi-groupes et multi-blocs sur les données. Pour souci de simplicité, nous nous intéressons seulement aux deux premières composantes principales. Le tableau 6 donne les pourcentages cumulés d'inertie restituée par


 ISSN: 2102-6238

TABLEAU 5. *Les listes des variables sensorielles et physico-chimiques et leurs abréviations.*

| Variables sensorielles | | Variables physico-chimiques | |
|---|---|---|---|
| Variables | Abréviations | Variables | Abréviations |
| Intensité de la couleur jaune | jaune | Acidité | Acid |
| Intensité de la couleur verte | vert | Peroxyde | Perox |
| Intensité de la couleur marron | marron | K232 | K232 |
| Intensité de l'aspect brillant | brillant | K270 | K270 |
| Transparence | transp | DK | DK |
| Aspect sirupeux | sirop | | |

les deux premières composantes principales pour chacun des groupes et chacun des blocs. Les pourcentages d'inertie restituée par les deux premières composantes sont relativement élevés pour les deux blocs indiquant un lien élevé entre eux.

TABLEAU 6. *Pourcentages cumulés d'inertie restituée par les deux premières composantes principales pour chacun des groupes et chacun des blocs.*

| | Données sensorielles | | Données Physico-chimiques | |
|---|---|---|---|---|
| Composantes principales | Dim1 | Dim2 | Dim1 | Dim2 |
| Grèce | 61.83 | 77.73 | 43.15 | 89.04 |
| Italie | 68.09 | 90.41 | 61.43 | 82.28 |
| Espagne | 20.91 | 67.26 | 69.52 | 86.89 |
| Global | 50.28 | 78.47 | 58.03 | 86.07 |

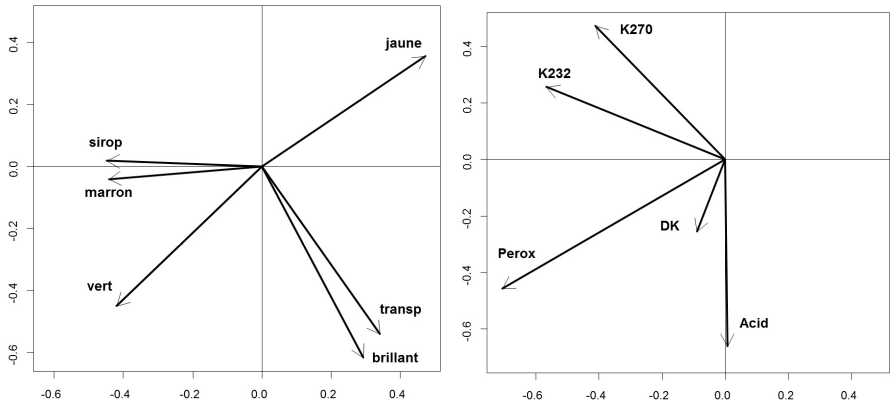
La figure 4 donne la représentation des variables sur la base des loadings communs pour les données sensorielles (bloc 1, figure 4(a)) et pour les données physico-chimiques (bloc 2, figure 4(b)). La figure 5 donne la représentation des huiles d'olive grecques, italiennes et espagnoles. La caractérisation des huiles d'olive pourrait être faite en liaison avec la représentation des variables (figure 4) à l'instar de ce qui est fait couramment en ACP. La représentation des huiles d'olive est donnée séparément pour chaque origine géographique. Une interprétation des différences entre chaque huile peut être donnée au regard des variables dont les liens sont illustrées en figure 4. Par exemple, les huiles d'olive italiennes 2 et 4 ont une couleur plus brune que les autres huiles italiennes.

## 6. Conclusion

L'ACP multi-groupes et l'ACP multi-groupes et multi-blocs peuvent être considérées comme étant, respectivement, des extensions de l'ACP et de l'analyse de co-inertie multiple au cas où les données comportent une structure de groupes. Elles permettent d'opérer une synthèse des données de nature à réaliser une typologie des individus à l'intérieur des différents groupes sur la base d'un modèle parcimonieux. Cette parcimonie du modèle garantit une meilleure stabilité des modèles que des analyses séparées des différents groupes. De plus, elle facilite l'interprétation des résultats.

Par rapport à des méthodes alternatives telles que STATIS duale (Lavit, 1988; Lavit et al., 1994) ou l'Analyse en Composantes Communes (Flury, 1984), l'ACP multi-groupes présente l'avantage d'être facile (algorithme non itératif basé sur la décomposition spectrale de la matrice


 ISSN: 2102-6238




(a) Variables sensorielles (b) Variables physico-chimiques

FIGURE 4. *Représentation des loadings associés aux variables sensorielles (figure 4(a)) et aux variables physico-chimiques (figure 4(b)).*

des covariances intra-classes) tout en présentant de bonnes propriétés et une bonne performance en termes d'inertie restituée dans les différents groupes (Eslami et al., sous presse).

Des investigations supplémentaires sont nécessaires pour confirmer davantage l'apport et la pertinence des méthodes présentées ici et enrichir la panoplie des outils d'interprétation des résultats, notamment graphiques.

Bien que, comme nous l'avons indiqué dans l'introduction, nous n'avons pas abordé ici le problème de discrimination entre les groupes, il nous semble, néanmoins, qu'il y ait une relation intime entre l'objectif de discrimination et l'objectif d'investigation des structures des individus dans les groupes. En effet, le premier aspect se rapporte à la structure 'inter-groupes' alors que le deuxième aspect concerne la structure 'intra-groupes'. Etant donné le lien structurel entre ces deux aspects, il paraît clair qu'une investigation de l'un des deux aspects permet de jeter davantage de lumière sur l'autre aspect. C'est cette direction que nous comptons poursuivre pour des recherches futures en espérant, d'un côté, améliorer les méthodes de discrimination existantes pour le cas d'un seul tableau structuré en groupes et, d'un autre côté, concevoir des méthodes de discrimination pour le cas de tableaux multiples structurés en groupes.


 ISSN: 2102-6238

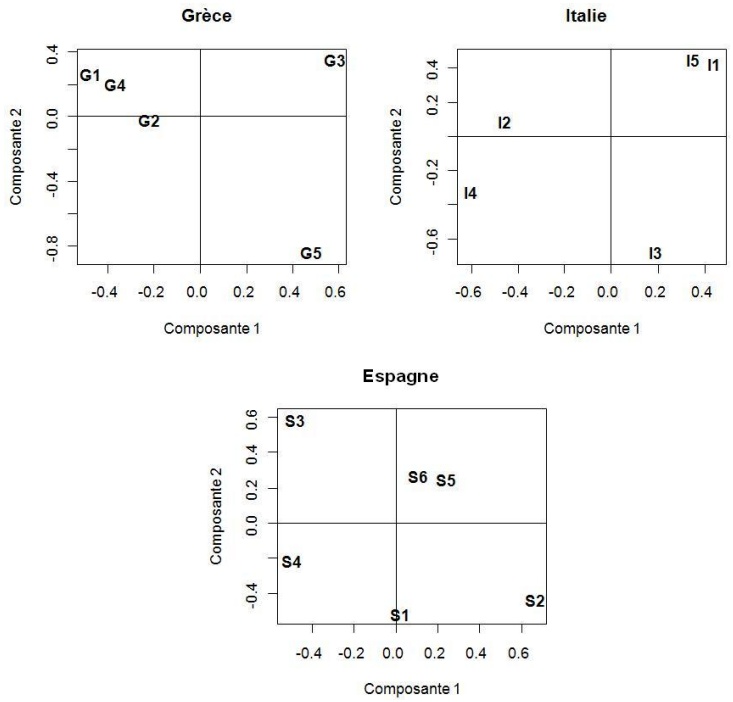



FIGURE 5. *Représentation globale des huiles d'olive grecques, italiennes et espagnoles.*

## Références

Cazes, P. (2004). Quelques méthodes d'analyse factorielle d'une série de tableaux de données. *Revue Modulad*, 31 :1–31.

Chessel, D. and Hanafi, M. (1996). Analyse de la co-inertie de k nuages de points. *Revue de Statistique Appliquée*, 44 :35–60.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, Elsevier*, 47(4) :547–553.

Eslami, A., Qannari, E. M., Kohler, A., and Bougeard, S. (sous presse). General overview of methods of analysis of multi-group datasets. *Revue des Nouvelles Technologies de l'Information*.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, 79 :892–898.

Krzanowski, W. J. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74 :703–707.

Krzanowski, W. J. (1984). Principal component analysis in the presence of group structure. *Applied Statistics*, 33(2) :164–168.

Lavit, C. (1988). *Analyse conjointe de tableaux quantitatifs*. Masson. Masson.

Lavit, C., Escouer, Y., Sabatier, R., and Traissac, P. (1994). The act (statis method). *Computational Statistics & Data Analysis*, 18 :97–117.

Lê, S., Husson, F., and Pagès, J. (2010). Dmfa : Dual multiple factor analysis. *Communication in Statistics-Theory and Methods*, 39(3) :483–492.

Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., de Jong, S., Lewi, P. J., Smeyers-Verbeke, and Pagès, J. (1998). *Handbook of Chemometrics and Qualimetrics*, volume B. Elsevier.

Mevik, B. and Wehrens, R. (2007). The pls package : Principal component and partial least squares regression in r. *Journal of Statistical Software*, 18(2) :1–24.


 ISSN: 2102-6238

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods : The rv-coefficient. *Applied Statistics*, 25(3) :257–265.
Sabatier, R. and Vivien, M. (2008). A new linear method for analyzing four-way multiblock tables : Statis-4. *Journal of Chemometrics*, 22(6) :399–407.
Saporta, G. (2006). *probabilités analyse des données et statistique*. Technip, Paris, 2nd edition.
Vivien, M. and Sabatier, R. (2004). A generalization of statis-act strategy : Do-act for two multiblocks tables. *Computational Statistics & Data Analysis*, 46(1) :155 – 171.
Westerhuis, J. A., Kourti, T., and MacGregor, J. F. (1998). Analysis of multiblock and hierarchical pca and pls model. *Journal of Chemometrics*, 12 :301–321.

## Annexe-A

Dans cette annexe, nous développons l'optimisation du critère $\sum_{m=1}^{M} \left\langle (X_m)^T t_m, a \right\rangle^2$ sous la contrainte $\| t_m \| = 1$ et $\| a \| = 1$. Nous avons :

$$\sum_{m=1}^{M} \left\langle (X_m)^T t_m, a \right\rangle^2 = \sum_{m=1}^{M} \left( (t_m)^T X_m a \right)^2$$

Il est clair que pour $a$ fixé, une solution optimale pour $t_m$ est donnée par :

$$t_m = \frac{X_m a}{\| X_m a \|}$$

En remplaçant cette valeur dans l'expression ci-dessus, il vient

$$\sum_{m=1}^{M} \left( (t_m)^T X_m a \right)^2 = \sum_{m=1}^{M} \left( \frac{a^T (X_m)^T X_m a}{\| X_m a \|} \right)^2$$

$$= \sum_{m=1}^{M} a^T (X_m)^T X_m a = \sum_{m=1}^{M} n_m var(X_m a)$$

De l'avant-dernière expression, nous déduisons que d'optimum est atteint pour $a$, vecteur propre de $\sum_{m=1}^{M} (X_m)^T X_m = N \sum_{m=1}^{M} \frac{n_m}{N} var(X_m) = NW$ associé à la plus grande valeur propre, où, rappelons le, $N$ est le nombre totale d'individus et $W$ est la matrice des covariances intra-classes. En d'autres termes, $a$ est vecteur propre de $W$ associé à la plus grande valeur propre.


 ISSN: 2102-6238