Analyse en Composantes Principales et régression PLS multigroupes.

Application à l'usage du cannabis dans 13 pays Européens.

Stéphanie Bougeard [1], Aida Eslami [2], El Mostafa Qannari [2] & Stéphane Legleye [3]

[1] *Agence de Nationale Sécurité Sanitaire (Anses), BP53, Ploufragan*
[2] *Oniris, Site de la Géraudière, BP 82225, Nantes*
[3] *Institut National d'Études Démographiques (Ined), 133 bd Davout, 75020 Paris*

Résumé. Cet article traite de l'analyse de données multivariées présentant une structure en groupe de leurs individus. Deux méthodes originales sont appliquées, *i.e.*, ACP multigroupe et régression PLS multigroupe, ces méthodes étant focalisées sur la part intra-groupe de la variance ; l'effet du groupe est considéré structurant mais non pertinent dans l'analyse. Elles sont basées sur la maximisation de critères reflétant leurs objectifs, *i.e.*, étude des liens entre variables communs à l'ensemble des groupes. La spécificité de chaque groupe est aussi étudiée au travers la similarité entre groupes et structure commune. Ces méthodes sont illustrées par une enquête internationale relative à la consommation de cannabis d'adolescents scolarisés dans treize pays Européens.

Mots-clés. Données multi-niveaux, ACP multigroupe, PLS multigroupe, consommation de cannabis, sciences sociales.

Abstract. We adress the problem of describing multivariate datasets divided into groups of individuals. We focus herein on multigroup Principal Component Analysis (mgPCA) and multigroup Partial Least Squares (mgPLS), these methods being devoted to the analysis of the within-group part of variance ; the group effect is structuring but not relevant in the analysis. Both these methods are based on criteria to maximize that reflect the objectives to be adressed, *i.e.*, seek common parameters to all the groups in order to study the relationships between variables. To better understand the group specificity in comparison with the common structure, differences and similarities between the groups are also studied. These methods are illustrated on the basis of a questionnaire aiming at studying the cannabis consumption among teenagers of thirteen European countries.

Keywords. Multilevel data, multigroup PCA, multigroup PLS, cannabis, social sciences.

1 Introduction

Cet article traite de l'analyse statistique de données multivariées présentant une structure de leurs individus organisée en plusieurs groupes. Ces données sont associées à

différentes appellations dans la bibliographie, *e.g.*, multi-niveaux, hiérarchiques, segmentées, emboîtées ou multigroupes. Elles sont fréquemment rencontrées en pratique, notamment en biologie. Les individus issus de données multigroupes ne sont pas indépendants les uns des autres, hypothèse fréquemment posée dans les traitements statistiques standards tels que la régression ou l'Analyse en Composantes Principales (ACP). Il convient donc de tenir compte de cette structure forte lors de leur analyse. Le traitement statistique des données multigroupes peut être divisé en deux catégories qui dépendent de l'objectif de celui-ci au regard de la structure en groupe : (i) l'analyse intra-groupe où l'effet du groupe est écarté, et (ii) l'analyse inter-groupe relative à l'analyse discriminante. Par la suite, l'article traite des analyses multigroupes où l'effet du groupe est structurant mais non pertinent dans l'interprétation. Les objectifs des méthodes multigroupes sont de déterminer des paramètres communs à l'ensemble des groupes, *e.g.*, composantes, axes ou coefficients de régression, de façon à tenir compte dans l'analyse de la spécificité des groupes d'individus en comparaison à la structure commune. Dans ce domaine et conformément à ces objectifs, peu de méthodes statistiques multivariées sont proposées. Nous nous focalisons par la suite sur certaines méthodes multigroupes proposées par Eslami (2013c), en particulier sur l'ACP multigroupe (mgPCA) et la régression Partial Least Squares multigroupe (mgPLS). Ces deux méthodes sont illustrées sur la base d'un enquête internationale relative à la consommation de cannabis des adolescents scolarisés de 13 pays Européens.

2 Méthodes multigroupes

2.1 ACP multigroupe

Soit le cas d'un tableau $\mathbf{X}$ constitué de P variables et N individus *a priori* divisés en M groupes $\mathbf{X}_m$, $m = (1, \ldots, M)$. Chaque sous-tableau $\mathbf{X}_m$ de dimension $(N_m \times P)$ est considéré centré. L'objectif de l'ACP multigroupe est d'étudier les différences et similitudes de l'ensemble des individus, ainsi que les liens entre les P variables, dans un espace constitué d'axes communs à tous les groupes a, l'effet du groupe ayant été écarté. De façon à mieux comprendre la spécificité des groupes au regard de la structure commune, les P variables peuvent aussi être vues au travers leurs axes partiels a_m. L'ACP multigroupe consiste à rechercher les vecteurs d'axes communs à tous les groupes a les plus liés aux m vecteurs spécifiques à chaque groupe $(a_1, \ldots, a_M)$ de façon à maximiser le critère (1) (Eslami *et al.*, 2013a).

$$\text{Max.} \quad \sum_{m=1}^{M} < a_m, a >^2 \quad \text{avec} \quad a_m = \mathbf{X}_m' t_m \quad \text{et} \quad \|t_m\| = \|a\| = 1 \tag{1}$$

2.2 Régression PLS multigroupe

Nous nous plaçons à présent dans un contexte de régression où un tableau $\mathbf{Y}$ est expliqué par un tableau explicatif $\mathbf{X}$, ces deux tableaux présentant *a priori* une même structure en M groupes de leurs N individus. L'objectif de la régression PLS multigroupe est de rechercher des axes communs à l'ensemble de ces groupes, *i.e.*, a et b respectivement associés aux tableaux $\mathbf{X}$ et $\mathbf{Y}$. De plus, de façon à mieux comprendre la spécificité des groupes au regard de la structure commune, des axes partiels a_m et b_m, spécifiques à chaque groupe et respectivement associés aux tableaux $\mathbf{X}_m$ et $\mathbf{Y}_m$, sont aussi calculés. La régression PLS multigroupe consiste à rechercher des axes communs à tous les groupes a et b de façon à ce que leurs composantes associées $u_m = Y_m b$ et $t_m = X_m a$ soient les plus liées possible. Ainsi la régression PLS multigroupe est associée au critère à maximiser (2) (Eslami *et al.*, 2013b).

$$\text{Max.} \quad \sum_{m=1}^{M} cov(\mathbf{Y}_m b, \mathbf{X}_m a) \quad \text{avec} \quad \|a\| = \|b\| = 1 \tag{2}$$

Par la suite, les axes spécifiques à chaque groupe sont calculés selon les formules $a_m = \mathbf{X}'_m u_m$ et $b_m = \mathbf{Y}'_m t_m$ sous les contraintes $\|a_m\| = \|b_m\| = 1$.

3 Application

3.1 Données et objectifs

L'intérêt des méthodes multigroupes est illustré sur la base de l'enquête européenne ESPAD menée en 2011. Cette enquête vise à collecter des données de consommation d'alcool et autres drogues suivant un protocole commun à tous les pays (*www.espad.org*). Pour l'année 2011, un module optionnel, *i.e.*, test CAST d'usage du cannabis (=Cannabis Abuse Screening Test), est ajouté dans treize pays Européens (Legleye *et al.*, 2011). Les données sont constituées de 5204 adolescents qui ont déclaré avoir fumé du cannabis lors des douze derniers mois. Ces adolescents proviennent des pays suivants : Belgique (331), Chypre (177), République Tchèque (1013), France (723), Allemagne (365), Italie (617), Kosovo (55), Lettonie (292), Liechtenstein (52), Pologne (1113), Roumanie (93), République Slovaque (246) et Ukraine (127). Le test CAST explore différents aspects de leur consommation de cannabis au travers de six questions : usage non récréationnel (rcast1, rcast2), troubles de la mémoire (rcast3), reproches de la famille ou amis (rcast4), échec lors des tentatives d'arrêt (rcast5) et problèmes associés à la consommation de cannabis (rcast6). L'enquête ESPAD comprend aussi des questions relatives à différents aspects de l'usage de drogue et du contexte de consommation. Neuf questions sont étudiées : consommation de cannabis lors des douze derniers mois (c25b, c25c), âge à la première consommation de cannabis (c26), nombre de cigarettes fumées lors du dernier mois (c09),

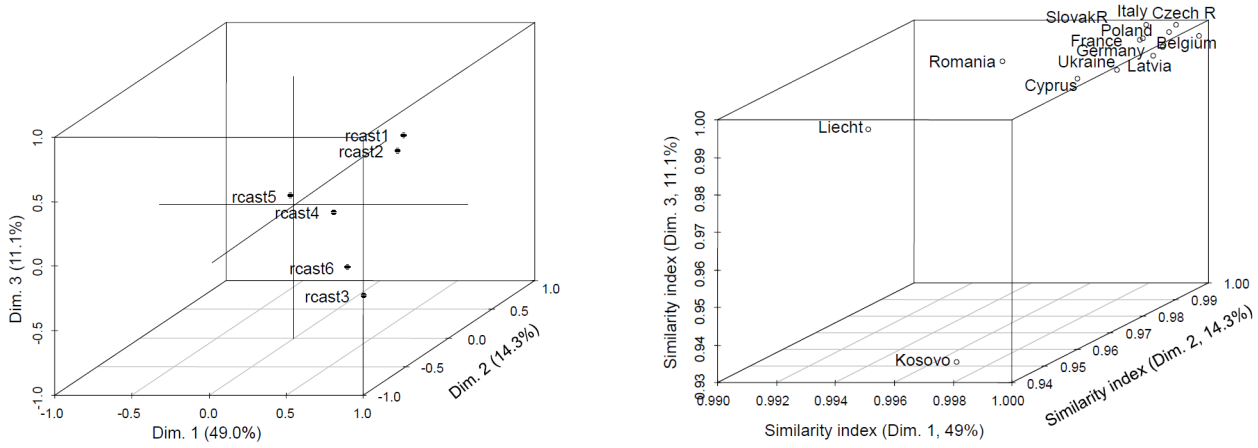
nombre d'ivresses alcooliques dans leur vie et au cours de la dernière année (c19a, c19b), facilité d'achat du cannabis (c24), nombre d'amis consommant du cannabis (c34d) et risque perçu à en consommer (c36h). Le premier objectif est d'étudier les liens entre les six variables du test CAST d'usage du cannabis, ces liens étant communs à l'ensemble des pays. De façon à mieux comprendre la spécificité de chaque pays en lien avec cette structure commune, les différences et similarités entre les treize pays sont aussi étudiées. Le second objectif est d'expliquer le test CAST d'usage du cannabis par les neufs variables décrivant l'usage de drogue et le contexte de consommation. Le principal but est d'étudier ces liens de façon à ce qu'ils soient communs à tous les pays, mais les différences et similarités entre les treize pays ont aussi leur importance.

3.2 ACP multigroupe pour l'analyse du test CAST d'usage du cannabis

L'ACP multigroupe est utilisée pour étudier les liens entre les six variables du test CAST d'usage du cannabis, ces liens étant communs à tous les pays. Comme ces variables ont des variances différentes, celles-ci sont centrées et réduites globalement de façon à avoir la même importance dans l'analyse. L'effet du facteur pays explique 20.9% de la variance totale (part inter-groupe) ; cet effet est retiré de l'analyse mgACP centrée sur l'analyse intra-groupe des données. Les pays ayant aussi des variances différentes pour chacune des six variables étudiées, les données sont de plus centrées et réduites par groupe ; ainsi, chaque pays à le même poids dans l'analyse. Trois dimensions sont retenues pour interpréter la méthode mgPCA ; elles expliquent 74.4% de la variance totale. La Figure 1(a) illustre les axes communs à tous les pays. Il s'ensuit que les deux variables décrivant l'usage non récréationel du cannabis (rcast1, rcast2) sont liées et relativement indépendantes de la variable décrivant les reproches de la famille ou des amis (rcast4) ainsi que des tentatives d'arrêt (rcast5). Puis, la spécificité de chaque pays au regard de cette structure commune est étudiée grâce aux indices de similarité entre les axes communs et partiels, comme illustré par la Figure 1(b). La plupart des pays sont similaires à cette structure commune (*e.g.*, République Tchèque, Pologne) mais le Kosovo et dans une moindre mesure le Liechtenstein présentent des différences, principalement dues au faible usage du cannabis dans ces pays.

3.3 PLS multigroupe pour expliquer le test CAST d'usage du cannabis par les variables de contexte et d'usage

La régression PLS multigroupe est utilisée pour expliquer la consommation de cannabis (tableau à expliquer $\mathbf{Y}$) par les variables qui décrivent l'usage de drogues et le contexte de consommation (tableau explicatif $\mathbf{X}$). Pour les raisons décrites précédemment, l'ensemble des variables est centré et réduit globalement ainsi que par groupe. L'effet du facteur pays explique 11.4% de la variance totale (part inter-groupe) et n'est pas conservé dans



(a) Carte des variables

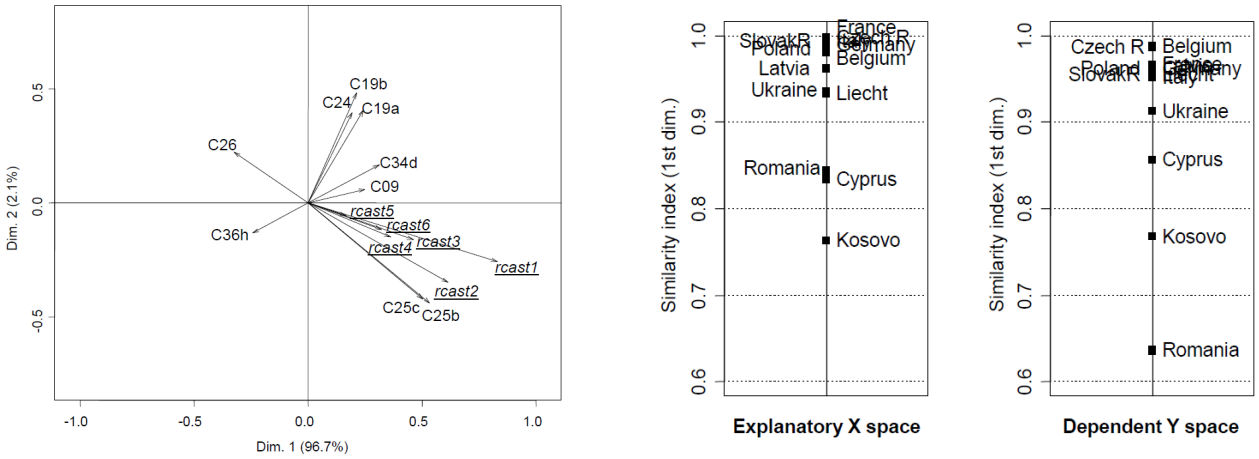
(b) Similarités entre groupes (pays)

FIGURE 1 – Carte des variables de l'ACP multigroupe : (a) Graphe des axes communs et (b) similarités entre les axes par groupe et communs. Illustration sur le test CAST d'usage du cannabis dans treize pays Européens.

la régression PLS multigroupe, centrée sur la part intra-groupe de la variance. Une seule dimension, expliquant 96.7% de la variance totale, est retenue pour l'interprétation. La Figure 2(a) illustre les liens entre variables à expliquer et variables explicatives sur la base des axes communs (liens communs à tous les pays). La principale interprétation est que les deux variables associées à l'usage non-récréationel (rcast1, rcast2) sont associées avec la consommation de cannabis lors de la dernière année ou du dernier mois (c25b, c25c), et anti-corrélées à l'âge de première consommation de celui-ci (c26). Comme pour la méthode mgACP, la particularité de chaque pays est illustrée par l'indice de similarité calculé entre axes communs et spécifiques de chaque groupe. Elle est illustrée par la Figure 2(b). Il s'ensuit que le Kosovo présente la structure la plus différente de la structure commune pour les variables explicatives relatives au contexte et à l'usage, alors que c'est la Roumanie qui présente la structure la plus différente pour les variables à expliquer relatives au test CAST.

4 Conclusion

Dans cet article, nous proposons d'appliquer deux méthodes originales, *i.e.*, ACP multigroupe et régression PLS multigroupe, au traitement statistique de tableaux multivariés présentant une structure en groupe de leurs individus. Les méthodes proposées sont focalisées sur la part intra-groupe de la variance, l'effet du groupe étant structurant mais non pertinent dans l'analyse. Ces deux méthodes sont basées sur la maximisation d'un critère qui reflète les objectifs posés. Un aspect intéressant de ces méthodes est qu'elles recherchent à la fois des axes communs et spécifiques à chaque groupe, tous deux utiles



(a) Carte des variables

(b) Similarités entre groupes (pays)

FIGURE 2 – Carte des variables de la régression PLS multigroupe : (a) Graphe des axes communs (**Y**, italique et souligné, et **X**) et (b) similarités entre les axes par groupe et communs. Illustration sur l'explication du test CAST d'usage du cannabis (**Y**) par les variables relatives à l'utilisation de drogue et au contexte de consommation (**X**) dans treize pays Européens.

aux représentations graphiques et à l'interprétation associée. Les méthodes proposées ainsi que les aides à l'interprétation proviennent de programmes R libres d'accès. Les méthodes mgACP et mgPLS sont des outils puissants pour explorer les données présentant une structure hiérarchisée des individus, comme par exemple les enquêtes internationales.

Bibliographie

[1] A. Eslami, E.M. Qannari, A. Kohler et S. Bougeard (2013a), General overview of methods of analysis of multi-group datasets, Revue des Nouvelles Technologies de l'Information, 25, p 108-123.
[2] A. Eslami, E.M. Qannari, A. Kohler et S. Bougeard (2013b), Multi-group PLS regression. Application to epidemiology, Multi-group PLS regression, In : *New perspectives in Partial Least Squares and Related Methods*, Ed. G. Russolillo, Springer Verlag, p 243-255.
[3] A. Eslami (2013c), Analyses factorielles de données structurées en groupes d'individus. Application en biologie. Thèse de l'Université Rennes 1.
[4] S. Legleye, D. Piontek et L. Kraus (2011), Psychometric Properties of the Cannabis Abuse Screening Test (Cast) in a French Sample of Adolescents, Drug Alcohol Depend, 113(2-3), p 229-235.