Classification de variables : une approche à double critères contrôlés dynamiques

Christian Derquenne

Electricité de France - Recherche et Développement - 7, boulevard Gaspard Monge - 91120 Palaiseau - christian.derquenne@edf.fr

Résumé. La recherche de structures dans les données représente une aide essentielle pour comprendre les phénomènes à analyser. L'apprentissage non supervisé accompagné par des techniques de visualisation en sont les principaux outils. Nous proposons un ensemble de méthodes pour classifier des variables numériques. Celles-ci reposent sur une approche mixte : la corrélation entre les variables initiales et l'unidimensionnalité des groupes obtenus, afin de construire de façon dynamique une typologie en contrôlant le nombre de classes et leur qualité.

Mots-clés. Classification, corrélation, unidimensionnalité, apprentissage non supervisé.

Abstract. The research structures in the data has an essential aid to understanding the phenomena to be analyzed. Unsupervised learning accompanied by visualization techniques are the main tools. We offer a set of methods for clustering numeric variables. These are based on a mixed approach: correlation between the initial variables and one-dimensionality of the resulting groups to dynamically build a typology by controlling the number of classes and quality.

Keywords. Variables clustering, correlation, unidimensionality, unsupervised learning.

1 Contexte - objectif

Que les données soient observées en petit ou en grand nombre, l'exploration, la recherche de structures et la visualisation des données arrivent avant toute exploitation statistique plus poussée (modélisation et/ou prévision, par exemple). La recherche exploratoire de structures dans les données est notamment réalisée à l'aide de la classification automatique d'objets (individus et/ou variables), afin de détecter des groupes homogènes, des agrégats de données. Deux grandes approches de classification non supervisée (analyse exploratoire sans cible a priori) sont disponibles : agrégation par partitionnement (CNH) ou agrégation hiérarchique ascendante (CAH) ou descendante (CDH). Pour la classification hiérarchique, le choix du nombre de classes est généralement réalisé par coupure de l'arbre de classification où la perte d'information entre deux nombres de classes successifs semble "importante". Que ce soit avec une approche par partitionnement ou hiérarchique, le choix du "bon" nombre de classes est généralement exploratoire, donc pas réellement contrôlé statistiquement en fonction de la structure des données. Ce papier se situe dans le cadre de la classification de variables numériques. Plusieurs approches ont été proposées pour résoudre ce problème. Un premier ensemble de méthodes est fondé sur différents critères d'agrégation de la CAH (lien minimum, maximum, Ward, ...) appliqués directement sur la matrice de corrélations entre les variables initiales transformées en matrice de dissimilarités. D'autres méthodes reposent sur la réduction de l'espace factoriel en associant au mieux les variables initiales à de nouvelles composantes (Sarle, 1990, Vigneau et Qannari,

2003, Bühlmann et al., 2013, Chen, 2014). Notre approche est fondée simultanément sur le test statistique de la corrélation linéaire simple de Pearson entre les variables initiales et un test d'unidimensionnalité sur les classes obtenues afin de construire une typologie de façon dynamique au moyen du contrôle du nombre de groupes et de leur qualité. Cette nouvelle méthode s'applique à la CAH et à la CDH, et propose un nombre "optimal" de classes.

2 Une approche à double critères contrôlés dynamiques

2.1 Indépendance *vs* unidimensionnalité

L'approche proposée repose sur l'utilisation conjointe des propriétés d'indépendance et d'unidimensionnalité. En effet, une classe compacte doit être à la fois composée de variables corrélées significativement, mais aussi être unidimensionnelle. En d'autres termes, cette propriété implique forcément la dépendance entre les variables d'un même groupe, par contre la réciproque est fausse. En effet, la dépendance de certaines variables entre elles (effet de chainage, par exemple) n'entraine pas l'unidimensionnalité. Statistiquement, si pour un groupe de variables, l'hypothèse nulle d'indépendance est rejetée, le test d'unidimensionnalité peut l'être également.

2.2 La CAH revisitée

Soit $E_0 = \{X_1, ..., X_q\}$, un ensemble de q variables numériques réelles, l'objectif est de construire une typologie en procédant par agrégation hiérarchique des q classes initiales représentant des singletons. On supposera que les relations entre les $q(q-1)/2$ couples de variables sont soit des relations linéaires, soit présentent une absence de relation (linéaire ou non linéaire).

Première étape de contrôle : Recherche de relations linéaires significatives. Soit $\rho_{jk} = \rho(X_j, X_k)$, le coefficient de corrélation linéaire de Pearson et soit $p_{jk}^{(\rho)}$, la p-valeur associée au test de nullité de ce coefficient. Alors la première classe C_1 est construite comme suit :

$$C_1 = \arg\min_{j,k}(p_{jk}^{(\rho)}) < \alpha_\rho \tag{1}$$

Si aucune p-valeur est inférieure à α_ρ, alors $C_1 = \emptyset$, la typologie sera formée de q classes-singletons et l'algorithme s'arrêtera là, sinon $C_1 = \{X_{(1)}, X_{(2)}\}$. Il y aura donc $q-2$ classes-singletons et une classe de cardinal 2 ($X_{(.)}$ désigne une variable sélectionnée parmi q). $E_1 = E_0 - C_1$ représentera le nouvel ensemble de variables restantes. La classe C_1 sera résumée par la première composante Z_1 de l'ACP. (1) agrège des variables corrélées positivement ou négativement, pour des corrélations seulement positives (1) devient : $C_1 = \arg[(\min_{j,k}(p_{jk}^{(\rho)}) < \alpha_\rho) \wedge (\rho_{jk} > 0)]$. Puis la recherche de relations linéaires significatives se poursuit, seules les corrélations entre les variables de E_1 et Z_1 seront calculées. Enfin (1) sera appliqué sur ces $q-2$ dernières corrélations et sur les $(q-2)(q-3)/2$ de E_1. Trois possibilités de typologie apparaissent. Soit la plus petite p-valeur se situe parmi les nouvelles corrélations calculées et $C_1 = \{X_{(1)}, X_{(2)}, X_{(3)}\}$. Celle-ci sera résumée par la première composante de l'ACP Z_1, mise à jour. Soit la plus faible p-valeur se situe parmi celles de E_1 et dans ce cas une deuxième classe est constituée : $C_2 = \{X_{(3)}, X_{(4)}\}$ générant Z_2. Soit toutes les p-valeurs sont supérieures au seuil α_ρ et dans ces conditions, seule la classe C_1 de deux variables restera avec $q-2$ groupes-singletons.

Seconde étape de contrôle : Unidimensionnalité des groupes. Dès qu'un groupe possède au moins trois variables, il est nécessaire de vérifier son homogénéité ou encore son unidimensionnalité. Un test statistique permet de vérifier cette propriété. On considère que l'espérance des valeurs propres est égale à 1 pour des composantes principales unidimensionnelles. On espère donc que la deuxième valeur propre $\lambda_2 \leq 1$. Le test est $H_0 : \lambda_2 \leq 1$ *vs* $H_1 : \lambda_2 > 1$. La statistique de test (Saporta, 1999) est la suivante :

$$U_{C_m} = \frac{\lambda_2 - 1}{\sqrt{(k_m - 1)/(n-1)}} \tag{2}$$

où k_m et n sont respectivement le nombre de variables dans le groupe C_m et le nombre d'observations. Sous H_0, $U_{C_m} \rightsquigarrow \mathcal{N}(0,1)$. Comme pour (1), nous fixons un seuil : α_U. Dans le cas d'une nouvelle variable $X_{(.)}$ à aggréger, si la p-valeur $p^{(U)}_{C_m} > \alpha_U$ alors C_m sera unidimensionnelle et $X_{(.)}$ restera dans cette classe, sinon elle ne pourra pas y figurer. Dans ces conditions, on sélectionnera la plus petite p-valeur parmi celles qui restent du test (1), si celle-ci est inférieure à α_ρ, alors on appliquera (2). Si aucun des deux tests de contrôle ne passe jusqu'à épuisement des corrélations, alors seul le groupe $C_1 = \{X_{(1)}, X_{(2)}\}$ sera validé avec les $q-2$ classes-singletons. Par conséquent, pour que l'agrégation se poursuive, il faut que $\min_{j,k \in E_m}(p^{(\rho)}_{jk}) < \alpha_\rho$ et que $p^{(U)}_{C_m} > \alpha_U$, ce qui correspond bien aux deux propriétés que doit posséder une classe compacte.

Suite du déroulement du processus et test d'arrêt. Supposons que M classes $C_1, ..., C_M$ aient été formées, où $M < q$, alors à l'étape suivante soit deux classes sont agrégées, soit si l'un des deux tests de contrôle ne passe pas alors le processus d'agrégation de classes s'arrête.

2.3 La CDH revisitée

L'objectif est de construire une typologie en procédant par désagrégation hiérarchique de E_0 en un certain nombre de classes. La validation du découpage de E_0 repose sur le même principe que celui de la CAH au moyen des deux tests de contrôle : corrélation et unidimensionnalité. Mais la constitution des classes est plus complexe. La première étape de la CDH repose sur la notion de distribution des corrélations.

Etape 1 : La distribution mixte. L'objectif est de rechercher la variable la moins corrélée (ou la plus anti-corrélée) avec toutes les autres afin de constituer C_1. Alors pour chaque X_j, la démarche est la suivante. Soit $\phi_j = \sum_{k \neq j} 1_{[p^{(\rho)}_{jk} < \alpha_\rho]}/(q-1)$, la proportion de tests de corrélation significatifs entre X_j et les autres (X_k), tel que $p^{(\rho)}_{jk} < \alpha_\rho$, soit $\psi_j = \sum_{k \neq j} \rho^2_{jk} 1_{[p^{(\rho)}_{jk} < \alpha_\rho]} / \sum_{k \neq j} \rho^2_{jk}$, la part de variabilité expliquée des variables corrélées significativement à X_j et soit enfin U_j la statistique (2) appliquée à l'ensemble des variables corrélées significativement à X_j, alors la première variable qui figurera dans C_1 est sélectionnée comme suit :

$$X_{(1)} = \arg[\min_j \phi_j \wedge \min_j \psi_j \wedge \max_j U_j] \tag{3}$$

où $\wedge$ représente l'utilisation conjointe des trois indicateurs tel que : si pour le min de ϕ_j, il y a plusieurs valeurs de ψ_j alors on en prendra le minimum et si pour ce dernier, plusieurs valeurs

de U_j existent, alors on sélectionnera la plus grande. On aura $C_1 = \{X_{(1)}\}$, puis $X_{(1)}$ sera centrée-réduite et se nommera Z_1. Pour des corrélations seulement positives, les indicatrices de ϕ_j et ψ_j seront $[(p_{jk}^{(\rho)} < \alpha_\rho) \bigcap (\rho_{jk} > 0)]$.

Etape 2 : Choix de la deuxième variable. La démarche de distribution des corrélations est appliquée sur $E_1 = E_0 - C_1$ et permettra d'obtenir $X_{(2)}$. Cette dernière sera incluse dans C_1 si $p_{j=(1),k=(2)}^{(\rho)} < \alpha_\rho$, sinon une deuxième classe sera formée. Dans le premier cas, on désignera par Z_1, la première composante de l'ACP pratiquée sur $\{X_{(1)}, X_{(2)}\}$, sinon la variable $X_{(2)}$ sera centrée-réduite et deviendra Z_2.

Etape 3 : Constitution des classes suivantes. A l'itération s, s variables parmi les q constituent M classes auxquelles sont associés $Z_1, ...Z_m, ..., Z_M$ composantes, alors le processus est le suivant : (i) Application de (3) sur les $q - s$ variables restantes $\rightarrow X_{(s+1)}$. (ii) si $\min_m(p_{s+1,m}^{(\rho)}) < \alpha_\rho$ alors $X_{(s+1)} \rightarrow C_m$, sinon $X_{(s+1)} \rightarrow C_{M+1}$. (iii) (2) est appliqué à C_m, si elle est unidimensionnelle alors $X_{(s+1)}$ reste dedans, sinon (ii) est appliqué sur les Z_m restants, si le test passe alors $X_{(s+1)}$ est inclus dans la classe $C_{m'}$, et (iii) est à nouveau appliqué, etc. L'étape 3 se termine lorsque les q variables sont incluses dans M groupes.

Etape 4 : Amélioration de la typologie. *Phase 4.1 Réallocation des variables dans les classes existantes.* (i) $\forall X_j$, $X_j \rightarrow C_m$ tel que $\max_{m=1,M} \rho^2(X_j, Z_m)$. (ii) Mise à jour de tous les Z_m associés aux classes C_m modifiées. (iii) (2) est appliqué à chaque C_m modifiée si son cardinal est supérieur à 3. (iv) Pour chaque C_m non unidimensionnelle, découpage de celle-ci en deux groupes les plus unidimensionnels. (iv.i) Sélection de 2 noyaux $\theta_m(1)$ et $\theta_m(2)$ (2 variables de C_m les moins corrélées) : $\theta_m(1) = \arg\max_{X_j \in C_m - \{X_k\}} \lambda_1^{(m,k)}$ où $\lambda_1^{(m,k)}$ est la plus grande valeur propre et $\theta_m(2) = \arg\min_{X_j \in C_m - \{\theta_m(1)\}} \rho^2(X_j, \theta_m(1))$. (iv.ii) Attribution des variables à leur noyau le plus proche : $(X_j \in C_m(.)) = \arg\max(\rho^2(X_j, \theta_m(1)), \rho^2(X_j, \theta_m(2)))$. (iv.iii) Calcul de $Z_m(1), Z_m(2)$. A la fin de 4.1, il y aura $M^* \geq M$ classes. *Phase 4.2 Possibilité d'augmenter la qualité* : (i) Recherche de la plus proche voisine de chaque X_j : $X_{k/j} = \arg\max_{k \neq j} \rho^2(X_k, X_j)$. (ii) $\forall X_j$ si $X_{k/j} \notin C_m$ alors X_j exclu de C_m. (iii) Mise à jour des $Z_1, ..., Z_{M^*}$. (iv) Réaffectation à l'un des M^* groupes des X_j exclus, tel que : $(X_j \in C_m) = \arg\max_{m=1,M^*}(\rho^2(X_j, Z_m))$.

2.4 Ni CAH, ni CDH, ni CNH : recherche d'un nombre "optimal" de classes

L'objectif de la méthode NHNP (Non Hiérarchique - Non Partionnement) est de rechercher dès le départ des groupes avec des variables bien corrélées sans a priori sur le nombre de classes. Dans une première étape, M groupes sont détectés à l'aide de l'approche par distribution des corrélations (cf. 2.3, étape 1), puis afin d'améliorer la qualité de la typologie, en termes de compacité, les sous-étapes 4.1 et 4.2 de la CDH sont appliquées. Le critère (3) devient :

$$X_{(1)} = \arg[\min U_{C_1} \wedge \max_j \phi_j \wedge \max_j \psi_j] \tag{4}$$

$X_{(1)}$ sera la première variable qui permettra de constituer C_1, le premier groupe le plus unidimensionnel contenant toutes les variables corrélées significativement à $X_{(1)}$. (4) sera alors à nouveau appliqué sur $E_0 - C_1$ jusqu'à ce qu'il n'y ait plus de variables. Cela pourra entrainer

des classes avec une seule variable. A la fin de ce processus, il y aura M groupes. Ceux qui ne seront pas unidimensionnels seront découpés à l'aide de la démarche de l'étape 4.2 de la CDH. $M^* \geq M$ classes seront ainsi obtenues sur lesquelles l'étape 4 de la CDH sera finalement appliquée. Pour des corrélations seulement positives, ϕ_j et ψ_j sont modifiés comme dans CDH.

2.5 Choix des seuils α_ρ *vs* α_U et critère d'optimisation ?

α_ρ et α_U représentent des paramètres de réglage du nombre de classes de la typologie. En effet, plus α_ρ se rapproche de 0, plus la corrélation en valeur absolue doit être élevée pour que deux variables (initiales ou latentes) constituent une classe, ce qui entraine des typologies avec beaucoup de groupes et inversement. Si $\alpha_\rho = 0$ alors $M = q$; si $\alpha_\rho = 0,5$ alors $M = 1$. De même, plus α_U se rapprochera de 0, plus l'unidimensionnalité sera facile à obtenir et inversement. Si $\alpha_U = 0$ alors $M = 1$; si $\alpha_U = 1$ alors $M = q$. Les trois approches n'ont pas pour objectif d'optimiser un critère comme dans [2 ou 5]. Par exemple dans [4], le critère d'arrêt correspond à l'unidimensionnalité de toutes les classes, alors que pour nos trois approches, c'est la compacité des classes (corrélations significatives entre variables d'un même groupe et unidimensionnalité). De plus, comme indiqué, il est possible d'améliorer la qualité de la typologie, tout en respectant le critère de compacité. Cependant, si l'on désire comparer les résultats issus des différentes méthodes, il est possible d'utiliser : $\pi_M = \sum_{m=1}^{M} \sum_{X_j \in C_m} \rho^2(X_j, Z_m)/q$ qui représentent la proportion d'inertie expliquée par la typologie en M groupes.

3 Application des trois approches et comparaisons

100 jeux de données possèdant 1000 observations et 20 variables ont été simulés. Chacun d'eux est découpé en 9 classes (6 avec une seule variable chacune $X_2, X_3, X_4, X_6, X_7, X_8$, les 3 autres avec, respectivement, 7, 4 et 3 variables), tels que : $X_j = 2X_9 + 2\epsilon_t$ pour $j = 10, ..., 15$ [$X_9 \rightsquigarrow \mathcal{N}(0;1)$] ; $X_j = X_1 + 2\epsilon_t$ pour $j = 16, 17, 18$ [$X_1 \rightsquigarrow \mathcal{N}(0,1)$] ; $X_j = 0,1X_5 + \epsilon_t$ pour $j = 19, 20$ [$X_5 \rightsquigarrow \mathcal{N}(0,1)$] et $\epsilon_t \rightsquigarrow \mathcal{N}(0,1)$. Nous avons $\alpha_\rho = 0,01$ et $\alpha_U = 0,5$. Cette seconde probabilité critique a été choisie car elle permet de comparer les résultats à ceux de VARCLUS. En effet, elle correspond à une deuxième valeur propre inférieure ou égale à 1. La qualité des méthodes est jugée à l'aide des indices d'adéquation de Rand, Jaccard et γ entre la typologie simulée et les classifications obtenues. Ces indices varient entre 0 et 1, plus la valeur obtenue est proche de l'unité, plus l'adéquation est bonne. Le tableau 1 montre que les trois méthodes proposées ont des valeurs médianes d'indices plus élevées que VARCLUS et Ward, même si 9 classes sont imposées. La colonne **BCL** indique le nombre de fois sur 100, où 9 classes ont été détectées. Pour CAH et CDH, plus d'un tiers des typologies estimées en font partie, alors qu'à peine un quart le sont pour NHNP. La colonne **Exact** fournit le nombre de typologies en parfaite adéquation avec la typologie observée : 9 classes et les trois indices égaux à 1. CAH et CDH font bien mieux que VARCLUS et Ward lorsque 9 classes sont imposées. Enfin les trois méthodes proposées obtiennent les trois premières places sur les résultats issus des quatre indicateurs.

Méthodes	**Rand**	**Jaccard**	γ	**BCL**	**Exact**	Pos_{Rand}	$Pos_{Jaccard}$	Pos_{γ}	Pos_{Exact}
CAH	0,9895	0,9431	0,9643	34	15	1	1	1	1
CDH	0,9895	0,9374	0,9610	35	14	1	3	3	2
NHNP	0,9895	0,9411	0,9631	22	10	1	2	2	3
VARCLUS(9)	0,9869	0,9360	0,9583	n.a	5	4	4	4	5
Ward(9)	0,9842	0,9032	0,9400	n.a	9	5	5	5	4
VARCLUS	0,9710	0,8596	0,9089	0	0	6	6	6	6
Ward	0,8263	0,4762	0,6148	0	0	7	7	7	7

Table 1: Comparaison des méthodes

4 Apports, applications et voies futures

Les trois méthodes de classification de variables proposées reposent sur une approche originale utilisant conjointement deux critères de construction des groupes contrôlés par des tests : la corrélation qui permet de capter la force des liens entre les variables et l'unidimentionnalité qui représente un garant de la compacité des classes : l'un ne va pas sans l'autre pour détecter la structure des données. L'approche par distribution permet de mieux détecter le nombre de groupes sans aucun a priori. L'application sur un jeu de données simulées a montré le gain de ces trois méthodes par rapport à celles de VARCLUS et de Ward, en termes de détection du nombre de classes et d'adéquation du contenu des groupes par rapport à la typologie observée. Ces méthodes ont été appliquées à d'autres jeux de données simulées et réelles, et les résultats obtenus ont été du même niveau de qualité. Par la suite, nous comparerons nos trois approches à d'autres. Puis nous généraliserons ces approches sur les axes suivants : non linéarité des variables entre elles (coefficient de Pearson non utilisable), données manquantes, présence de valeurs anormales pouvant biaiser les corrélations, nombre très élevé d'individus ne permettant plus d'utiliser les résultats des tests classiques et nombre important de variables provoquant une augmentation de la complexité pour rechercher une typologie de qualité.

Bibliographie

[1] Bühlmann P., Rütimann P., van de Geer S., and Zhang C-H, (2013): Correlated in regression: Clustering and sparse estimation. *Journal of Stat. Planning and Inference*, **143**(11), 1835-1858.
[2] Chen M., (2014): *Classification de variables autour de variables latentes avec filtrage de l'information : application à des données en grande dimension*, Thèse de doctorat, Université de Nantes, Ecole VENAM.
[3] Saporta G., (1999): Some Simple Rules for interpreting Outputs of Principal Components and Correspondence Analysis, *IXth International Symposium on ASMDA*, Lisbon, Portugal.
[4] Sarle W., (1990): *The VARCLUS Procedure. SAS/STAT 9.2 User's Guide*. Cary, NC: SAS Institute, Inc. **93**, 7453-7484.
[5] Vigneau E. and Qannari E.M., (2003): Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation*, **32**(4), 1131-1150.