Classification de variables avec des relations non linéaires

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. Les méthodes de classification de variables numériques permettent de répondre à cette problématique, mais elles ont été en grande majorité développées pour des variables ayant des liens linéaires. Nous proposons une nouvelle approche pour classifier des variables numériques possédant des relations quelconques au moyen d'un critère d'agrégation hiérarchique. Celle-ci est fondée sur des modèles polynomiaux pour les corrélations et l'ACP non linéaire pour la construction de variables latentes représentantes des groupes.

Mots-clés. Classification, relations non linéaires, apprentissage non supervisé.

Abstract. The research structures in the data has an essential aid to understanding the phenomena to be analyzed. The methods for clustering numeric variables answer to this problem, but the majority has been developed only for linear relationships between variables. We propose a new approach of clustering of variables with non linear relationships by means hierarchical agregating criteria. This ones is based on polynomial models for correlation and non linear PCA to build latent variables representing clusters.

Keywords. Clustering, nonlinear relationship, unsupervised learning.

1 Contexte - objectif

La recherche exploratoire de structures dans les données est essentielle dans de nombreuses applications (biologie, environnement, finance, management de l'énergie, ...) afin de comprendre les comportements des individus, les liens entre les variables, ... Les outils de visualisation, de réduction de dimension, de recherche de patterns permettent de répondre efficacement à ce type de problématiques. Notre objectif est de rechercher des groupes de variables numériques ayant des liaisons quelconques (linéaire ou non linéaire) à l'aide de la classification hiérarchique ascendante. Si plusieurs approches ont été proposées pour classifier des variables liées linéairement, peu de solutions ont été proposées pour le cas non linéaire. Pour le cas linéaire, les principales méthodes reposent sur la réduction de l'espace factoriel en associant au mieux les variables initiales à de nouvelles composantes (Sarle, 1990, Vigneau et al., 2003, Chavent et al., 2011, Bühlmann et al., 2013, Chen M., 2014). Nous avons développé une méthode nommée "double critère contrôlé dynamique" (Derquenne, 2016). Celle-ci est fondée simultanément sur un test d'indépendance linéaire simple entre les variables initiales et/ou des variables latentes (première composante principale de l'ACP) 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 méthode a été développée pour la CAH et la CDH, et propose un nombre "optimal" de classes. Même si les liens linéaires sont présents entre la plupart des phénomènes

de la nature, ce n'est pas le cas dans certains domaines d'application comme la génétique (expression des génes, des protéines), la finance (prix de marché de l'énergie, indicateurs financiers), l'économie, ... Une première approche pour construire une typologie de variables à liens quelconques consiste à appliquer un critère d'agrégation sur une matrice de dissimilarités reposant sur le coefficient de corrélation de Spearman qui préserve à peu près des relations monotones entre variables. Citons une méthode (Chen Y. et al., 2016) utilisant l'information mutuelle pour caractériser et mesurer des interdépendances non linéaires entre les variables. Un modèle de processus de Dirichlet est alors utilisé pour classifier les variables à l'aide de l'information mutuelle. Nous proposons une nouvelle approche fondée sur des transformations polynomiales entre couple de variables initiales et/ou variables latentes (première composante principale issue d'une ACP non linéaire). Cette méthode a seulement été développée pour la CAH et repose sur le principe de construction de typologie (Derquenne, 2016). Cette nouvelle approche sera appliquée sur des données simulées. Les résultats obtenus sont comparés à ceux de méthodes de classification linéaire et à ceux de la classification sur matrice de dissimilarités du coefficient de corrélation de Spearman. La méthode (Chen Y. et al., 2016) n'a pas pu être mise en oeuvre, faute d'un package R disponible. Enfin, les forces et les faiblesses de notre approche seront discutées, ainsi que des améliorations et de nouvelles voies de recherche.

2 Classification de variables avec des relations quelconques

2.1 Principe de l'approche "double critère contrôlé dynamique"

L'approche introduite en 2016 (Derquenne) 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. Nous avions alors notamment développé la CAH sur ce principe.

Soient $X_1, ..., X_q$, q variables numériques dont on suppose que les relations sont linéaires ou absentes, alors la première étape consiste à agréger les deux variables les plus corrélées linéairement pour constituer la première classe. Pour cela, on fixe un seuil critique du test de corrélation (par exemple, $\alpha_\rho = 0,05$), alors si la plus petite p-valeur parmi les $q(q-1)/2$ couples de variables est inférieure à α_ρ, on regroupera ces deux variables. Puis la première composante principale est calculée sur celles-ci, soit Z_1. De nouvelles corrélations sont calculées entre Z_1 et les $q-2$ variables restantes. Trois cas peuvent se présenter : soit une classe de trois variables, soit deux classes de deux variables, soit aucune corrélation significative est trouvée, alors l'algorithme s'arrête. Dans ce dernier cas, il y aura un groupe de deux variables et $q-2$ classes singleton. Si le processus continue, dès qu'un groupe possède au moins trois variables, un test d'unidimensionnalité est pratiqué, tel que $H_0 : \lambda_2 \leq 1$ (Saporta, 1999). Si l'hypothèse nulle d'unidimensionnalité est rejetée, alors on recherche si parmi les p-valeurs restantes issues des tests de corrélations, la plus petite est inférieure au seuil fixé. Si c'est le cas, les trois possibilités indiquées précédemment se

représenteront. Le processus de constitution des classes se poursuit jusqu'à ce que plus aucune p-valeur de corrélation est inférieure à α_ρ et que le test d'unidimensionnalité pour chaque classe est rejeté. A la fin de ce processus, nous obtenons M classes.

2.2 La méthode proposée

L'objectif est de classifier des variables numériques possédant des liaisons quelconques. Soient $X_1, ..., X_q$, q variables numériques et soient $f_{j/k}, f_{k/j}$, les fonctions images quelconques : $X_j = f_{j/k}(X_k)$ et $X_k = f_{k/j}(X_j)$. Si les liaisons entre variables sont seulement linéaires ou absentes alors les fonctions de régression linéaire pour X_j et X_k sont respectivement : $X_j = \beta^{(j/k)} X_k$ et $X_k = \beta^{(k/j)} X_j$, où $\beta^{(j/k)}$ et $\beta^{(k/j)}$ sont les coefficients de régression. Dans ce cas, le coefficient de corrélation de Pearson permet de résumer en une information unique de ces deux fonctions telle que : $\rho_{jk} = \beta^{(j/k)} s_k/s_j = \beta^{(k/j)} s_j/s_k$ où s_j et s_k sont respectivement les écarts-types de X_j et de X_k. Dans le cas de liaisons quelconques, le principe général de construction de la typologie en deux étapes contrôlées dynamiques discuté en 2.1 est également utilisé pour cette nouvelle approche. La première étape, qui consiste à agréger les deux variables les plus dépendantes significativement, passe par la recherche d'une information unique d'un couple de variables et est traité au paragraphe 2.2.1, alors que la seconde (construction de composante synthétique par classe et vérification de l'unidimensionnalité) est développée en 2.2.2.

2.2.1 Recherche de relations quelconques significatives

Dans le cas linéaire, la recherche de la dépendance la plus significative est fournie par la plus petite p-valeur du test de nullité du coefficient de corrélation de Pearson à condition qu'elle soit inférieure à un seuil fixé (cf. 2.1). Comme indiqué en 2.2, la recherche d'une information unique de liaison non linéaire entre deux variables X_j et X_k est asymétrique. Elle dépend des deux fonctions $X_j = f_{j/k}(X_k)$ et $X_k = f_{k/j}(X_j)$. Nous proposons que celles-ci soient de forme polynomiale, alors pour chaque variable du couple (X_j, X_k), nous construisons un modèle de régression polynomial restreint aux degrès statistiquement significatifs :

$$X_j = f_{j/k}(X_k) = \sum_{d \in D_k} \beta_d^{(j/k)} X_k^d + \beta_0^{(j/k)} + \epsilon \tag{1}$$

où D_k est l'ensemble des degrés retenus $d = \{s; 1/s\}$ où $s \in \mathbb{Z}^*$ pour $X_k \in \mathbb{R}$ et $d = \{s\}$ pour $X_k \in \mathbb{R}^{+*}$, $\beta_d^{(j/k)}$ est le coefficient de régression associé au degré d du polynome pour X_k ajustant X_j, et ϵ est une erreur d'espérance nulle et de variance σ^2. Il en est de même pour modéliser X_k à l'aide de X_j :

$$X_k = f_{k/j}(X_j) = \sum_{d \in D_j} \beta_d^{(k/j)} X_j^d + \beta_0^{(k/j)} + \epsilon \tag{2}$$

Pour chaque modèle, les degrés du polynome sont retenus à l'aide d'une méthode de sélection pas à pas (stepwise) des degrés. Le principe revient à fixer tout d'abord deux seuils α_1 pour la proposition d'un nouveau degré du polynôme et α_2 pour la sortie d'un autre qui ne se révèle plus significatif sachant l'apport des autres degrés présents dans l'équation. L'introduction d'un nouveau degré s (nouvelle variable X^s) dans le modèle est possible seulement si sa p-valeur est

la plus petite et si cette dernière est inférieure au seuil α_1. Puis, si une variable déjà entrée possédant la p-valeur du test de Student la plus élevée est supérieure au seuil α_2, alors elle est éliminée. Enfin, nous proposons de choisir le meilleur ajustement d'une des deux variables par l'autre à l'aide du test de Fisher de nullité de l'ensemble des paramètres de chacun des deux modèles, tel que $p_{jk}^{(\rho)} = \min(p_{j/k}^{(\rho)}, p_{k/j}^{(\rho)})$ où $p_{j/k}^{(\rho)}$ et $p_{k/j}^{(\rho)}$ sont respectivement les p-valeurs associées pour ajuster X_j à l'aide de X_k et X_k avec X_j. La première classe C_1 est construite comme suit

$$C_1 = \arg\min_{j,k}(p_{jk}^{(\rho)}) < \alpha_\rho \tag{3}$$

où α_ρ est le seuil au-dessus duquel les X_j et X_k ne seront pas agrégées.

2.2.2 Recherche de composantes synthétiques et unidimensionnalité des groupes

Les deux variables sont résumées sous forme d'une variable latente. La méthode introduite en 2016 (Derquenne) réduisait la dimension des classes au moyen de la première composante principale de l'ACP. Celle-ci est généralisée à l'aide d'une ACP non linéaire dans ce papier. Pour cela, nous utilisons l'approche "Optimal Scaling" (OS) qui recherche des transformations de variables ajustant de façon optimale l'ACP linéaire. OS a l'avantage de préserver beaucoup de propriétés de l'ACP linéaire. En 1959, Guttman a observé que si nous imposons que la régression entre variables transformées de façon monotone est linéaire, alors les transformations sont définies de façon unique. Cependant des approximations sont nécessaires. La fonction de perte de l'ACP-OS est l'inertie résiduelle, comme pour l'ACP linéaire. Mais les transformations appartiennent à une classe restreinte (monotone, polynomiale, spline). Les algorithmes associés sont souvent de type moindres carrés alternés, où les transformations optimales et les approximations des matrices de faibles rangs sont calculées jusqu'à convergence. De nombreuses approches ont été développées notamment dans (Takane et al., 1978) et sont implémentées dans SAS et R.

Nous proposons d'utiliser cette méthode pour obtenir Z_1, la première composante issue de l'ACP non linéaire, celle-ci est intégrée comme nouvelle variable parmi les $q-2$ variables initiales restantes. Puis, les modèles (1) et (2) sont appliqués entre Z_1 et celles-ci. Le processus d'agrégation de variables initiales et/ou de premières composantes principales à l'aide de la recherche de corrélation maximale significative (3) est appliqué, son principe est identique à celui décrit en 2.1. Puis dès qu'un groupe C_m possède au moins trois variables, il est nécessaire de vérifier son unidimensionnalité. Cependant nous ne pouvons pas utiliser le test de Saporta (1999) car les méthodes d'ACP non linéaire fournissent seulement une approximation des valeurs propres. Nous proposons une première solution naturelle (4) qui revient à considérer que si la deuxième valeur propre est inférieure ou égale à 1, alors la classe C_m est unidimensionnelle. Une seconde proposition (5) consiste à construire un intervalle de confiance bootstrap autour de la deuxième valeur propre. Si elle est inférieure à la borne supérieure de cet intervalle, alors la classe C_m est unidimensionnelle. Pour une nouvelle variable $X_{(.)}$ à agréger, nous avons :

$$\text{Si } \lambda_2 \leq 1 \text{ alors } C_m \text{ est unidimensionnelle et } X_{(.)} \in C_m \tag{4}$$

$$\text{Si } \lambda_2 \leq \lambda^*_{1-\alpha/2} \text{ alors } C_m \text{ est unidimensionnelle et } X_{(.)} \in C_m \tag{5}$$

Dans le cas contraire, on sélectionnera la plus petite p-valeur issue de (1) et (2) parmi celles qui restent et la règle (3) sera appliquée, si elle satisfaite alors (4) ou (5) est utilisée à son tour. Si

aucun des deux tests de contrôle ne passe jusqu'à épuisement des corrélations, alors le processus d'agrégation s'arrête pour un nombre M de classes. Au contraire, pour que l'agrégation se poursuive, il faut que (3) et (4) ou (5) soient satisfaits, ce qui correspond bien aux deux propriétés que doit posséder une classe compacte.

3 Application de la CAH non linéaire et comparaisons

Afin d'évaluer la qualité de l'approche proposée et de la comparer à d'autres méthodes, nous avons simulé 40 jeux de données possédant 1000 observations et 18 variables. Chacun d'eux est découpé en 7 classes : $C_1 = (X_5)$, $C_2 = (X_{10})$, $C_3 = (X_6, X_7)$, $C_4 = (X_8, X_9)$, $C_5 = (X_{11}, X_{12}, X_{13})$, $C_6 = (X_1, X_2, X_3, X_4)$ et $C_7 = (X_{14}, X_{15}, X_{16}, X_{17}, X_{18})$. Les relations entre les variables sont les suivantes : $X_1 \rightsquigarrow \mathcal{N}(0;1)$; $X_2 = 2 - 3X_1^2 + 0,5\epsilon_t$; $X_3 = X_1 + 0,5\epsilon_t$; $X_4 = 1,8X_2 + 5 + 0,5\epsilon_t$; $X_5 \rightsquigarrow \mathcal{N}(0;1)$; $X_6 \rightsquigarrow \mathcal{N}(0;1)$; $X_7 = 4\log(X_6) + 1 + 0.5\epsilon_t$; $X_8 \rightsquigarrow \mathcal{N}(0,1)$; $X_9 = -2.5X_8 + 0.5\epsilon_t$; $u \rightsquigarrow \mathcal{U}(0,1)$; $X_{11} = u^2 + 0,3\sin(2\pi u) + 2(\mathcal{U}(0;1) - 0.5)$; $X_{12} = u + 2(\mathcal{U}(0;1) - 0.5)$; $X_{13} = u^3 + u + 1 + 2(\mathcal{U}(0;1) - 0.5)$; $X_{14} \rightsquigarrow \mathcal{N}(0,1)$; X_{15} est une fonction découpée et bruitée de X_{14} ; $X_{16} = 2/X_{15} + 1,5X_{14} + 0.5\epsilon_t$; $X_{17} = -0,8X_{14} + 0.5\epsilon_t$; $X_{18} = X_{14} + X_{15} + X_{16} + X_{17} + 0.5\epsilon_t$ et $\epsilon_t \rightsquigarrow \mathcal{N}(0,1)$. Pour les seuils, nous avons choisi $\alpha_\rho, \alpha_1, \alpha_2 = 0,0001$. 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 l'approche proposée (CAH non linéaire) retrouve 39 fois sur 40 le bon nombre de classes (**BCL**), ainsi que l'attribution des variables à leurs classes réelles (**Exact**). VARCLUS et la méthode CAH linéaire (Derquenne, 2016) obtiennent respectivement 23 et 16 comme bon nombre de classes, mais n'arrivent pas à attribuer correctement les variables dans leurs groupes respectifs. Les valeurs médianes des trois indices d'adéquation sont égales à l'unité pour l'approche proposée. Les résultats des méthodes ClustOfVar, CAH linéaire et VARCLUS sont tout à fait comparables pour ces indices, alors que CLV est plus performante. Enfin le critère d'agrégation du diamètre sur les corrélations de Spearman (Comp-Spearman) est médiocre alors que la non linéarité est tout de même prise en compte. Mais l'application directe de critères d'agrégation sur une matrice de dissimilarités n'est généralement pas très fiable. Les quatre dernières colonnes (tab. 1) fournissent l'ordre d'arrivée de la qualité des méthodes. Ces résultats montrent que l'approche proposée est la plus efficace.

Méthodes	**Rand**	**Jaccard**	γ	**BCL**	**Exact**	Pos_{Rand}	$Pos_{Jaccard}$	Pos_γ	$Pos_{BCL/Exact}$
CAH non linéaire	1	1	1	39	39	1	1	1	1
VARCLUS	0,9171	0,5017	0,6270	23	0	4	4	4	2
CAH linéaire	0,9146	0,5001	0,6207	16	0	5	5	5	3
CLV	0,9618	0,7456	0,8333	0	0	2	2	2	4
ClustOfVar	0,9204	0,5164	0,6373	0	0	3	3	3	4
Comp-Spearman	0,8956	0,2796	0,4691	0	0	6	6	6	4

Table 1: Comparaison des méthodes

4 Apports, applications et voies futures

La méthode de classification proposée repose sur une approche originale utilisant conjointement deux critères de construction des groupes contrôlés par des tests : la corrélation reposant sur des modèles polynomiaux permettant de capter la force des liens non linéaires entre les variables et l'unidimensionnalité des classes au moyen de l'ACP non linéaire qui représente un garant de la compacité des classes. L'application sur un jeu de données simulées a montré le gain de cette nouvelle approche à l'égard des autres méthodes en termes de détection du "bon" nombre de classes et d'adéquation du contenu des groupes vis-à-vis de la typologie observée, même si celles-ci ont été développées pour des relations linéaires entre variables. Seule la CAH avec le critère du lien maximum appliquée à la matrice de dissimilarités des corrélations de Spearman semblait plus adéquate pour traiter des liens non linéaires mais les résultats sont décevants. La nouvelle approche a été appliquée à d'autres jeux de données simulées et réelles, et les résultats obtenus sont du même niveau de qualité. Les voies futures de recherche concernent la prise en compte de données manquantes, la présence de valeurs anormales pouvant biaiser les corrélations (linéaire ou non linéaire), l'adaptation à un nombre très élevé d'individus ne permettant plus d'utiliser les résultats des tests classiques, ainsi qu'à un nombre important de variables provoquant une augmentation de la dimension du problème et donc 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] Chavent M., Kuentz V., Liquet B. et Saracco J., (2011): Classication de variables : le package ClustOfVar, *43ièmes Journées de Statistique*, Tunis, Tunisie.
[3] 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.
[4] Chen Y. and Yang U., (2016): A Novel Information-Theoretic Approach for Variable Clustering and Predictive Modeling Using Dirichlet Process Mixtures, www.nature.com/scientificreports.
[5] Derquenne Ch., (2016): Classification de variables : une approche à double critères contrôlés dynamiques, *48ièmes Journées de Statistique*, Montpellier, France.
[6] Saporta G., (1999): Some Simple Rules for interpreting Outputs of Principal Components and Correspondence Analysis, *IXth International Symposium on ASMDA*, Lisbon, Portugal.
[7] Sarle W., (1990): *The VARCLUS Procedure. SAS/STAT 9.2 User's Guide*. Cary, NC: SAS Institute, Inc. **93**, 7453-7484.
[8] Takane Y., Young F.W. and de Leeuw J., (1978): The principal components of mixed measurement level multivariate data: An alternative least squares method with optimal scaling features *Psychometrika*.
[9] Vigneau E. and Qannari E.M., (2003): Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation*, **32**(4), 1131-1150.