Régression typologique pour données multi-blocs

Ndeye Niang Keita, Gilbert Saporta

REGRESSION TYPOLOGIQUE POUR DONNEES MULTI-BLOCS

Ndèye Niang[1] & Gilbert Saporta[2]

[1,2] CEDRIC CNAM
292, rue Saint Martin, 75141 Paris Cedex 03, France,
[1]ndeye.niang_keita@cnam.fr
[2] gilbert.saporta@cnam.fr

Résumé. Nous nous intéressons à la régression multi-blocs lorsque les individus présentent une structure en groupes inconnus a priori. La régression typologique (ou clusterwise) permet d'apporter une réponse dans le cas d'un seul tableau de variables explicatives. Nous proposons une approche de type hiérarchique à deux niveaux dans laquelle une régression typologique est appliquée à chacun des blocs séparément puis sur les estimations issues du premier niveau. Cette approche générale peut être utilisée dans le cas de la régression linéaire, la régression sur composantes principales ou PLS. La méthode proposée est illustrée sur des données réelles de pollution de l'air intérieur.

Mots-clés. Régression Multi-blocs, régression typologique, classification.

Abstract. Clusterwise linear regression aims at partitioning data sets into clusters characterized by their specific coefficients in a linear regression model. Usually, one dependant variable is related to independent variables which are in a single data table. In this paper we are interested in clusterwise linear regression in the context of multiblock data. We propose a clusterwise multiblock linear regression method with two steps. The first one performs a clusterwise linear regression on each data table yielding a score function. Then in the second step the score functions are combined using again a clusterwise regression. It is a general approach which can be extended to regression on principal components or PLS. The method is illustrated on indoor pollution data.

Keywords. Clusterwise regression, clustering, multiblock regression

1 Introduction

La régression multi-blocs regroupe un ensemble de méthodes de modélisation de plusieurs tableaux de variables mesurées sur les mêmes individus. Il s'agit d'expliquer un ou plusieurs tableaux en fonction d'un ensemble de variables explicatives structurées elles mêmes en blocs. De nombreuses méthodes ont été proposées parmi lesquelles on peut citer la Multiblock PLS (MBPLS), la Hierarchical PLS (HPLS) Westerhuis (1998), la Serial PLS (SPLS) (voir Yaroshchyk (2012) pour une étude comparative de ces méthodes), l'Analyse de co-inertie multiple orthogonale (ACIMO, ACIMOG-PLS) Vivien (2002) ou encore l'analyse canonique généralisée régularisée RGCCA Tenenhaus et Tenenhaus (2011). On trouve une revue de la plupart de ces méthodes dans Vivien (2002). Elles sont largement utilisées en chimiométrie, spectrométrie, et analyse sensorielle entre autres. Ces méthodes ont été développées dans l'objectif de prendre en compte la structuration en blocs des variables, qui apparaît de plus en plus indispensable en pratique.

Dans un grand nombre d'applications en sciences sociales, en environnement ou plus généralement dans le domaine scientifique, un seul ensemble de coefficients de régression pour tous les individus peut parfois ne pas convenir et conduire à des estimations erronées. Cela sera illustré dans l'application en section 4. C'est en particulier le cas lorsqu'il existe une structure de groupes entre les individus. Il existe des méthodes dites multi-group dans lesquelles les classes sont supposées

connues a priori. On peut citer les travaux de Eslami (2013) traitant du cas plutôt non supervisé comme l'ACP multi-group avec des extensions multi tableaux et ceux de Tenenhaus (2011) sur les données multi group et multi-blocs.

Lorsque les groupes sont inconnus, se pose alors le problème de leur détermination comme en classification mais ici les classes doivent être générées selon un modèle de régression linéaire plutôt que selon le critère classique de similarité entre observations.

Dans le cas d'un seul tableau de variables explicatives, différentes méthodes de régression typologique ou clusterwise ont été proposées. Les premiers travaux sont attribués à Spaeth (1979) selon DeSarbo et Cron (1988). Mais on peut aussi citer ceux de Bock (1969) et Diday (1976) puis ceux de Charles (1977). DeSarbo et Cron (1988) propose une méthode de régression linéaire typologique fondée sur un modèle de mélange de gaussienne avec des estimateurs du maximum de vraisemblance et l'algorithme EM.

Plus récemment Preda et Saporta (2005) l'ont utilisée dans le cadre de la régression PLS sur données fonctionnelles. Notons aussi les travaux plus récents de Henin (2002) dans le contexte de la robustesse.

Nous nous intéressons dans cette communication au cas d'une seule variable y à expliquer par un ensemble de tableaux $\mathbf{X}_t$ de variables explicatives, lorsque les individus présentent une structure en groupes inconnus a priori. Nous l'appelons régression typologique multi-tableaux. A notre connaissance elle n'a pas été beaucoup abordée dans la littérature. Nous proposons une approche de type hiérarchique, semblable à la méthode MBPLS, à deux niveaux dans laquelle une régression linéaire typologique est appliquée à chacun des blocs séparément puis sur les estimations issues du premier niveau. Cette approche générale peut être étendue à la régression sur composantes principales ou PLS en cas de multicolinéarité ou si le nombre de variables est très grand comme cela sera discuté dans la section 3. La méthode proposée est illustrée sur des données réelles de pollution de l'air intérieur.

2 La régression linéaire typologique

L'objectif de la régression linéaire typologique est de déterminer une partition d'un ensemble de *n* individus en *K* classes obtenues selon un modèle de régression linéaire reliant une variable $\mathbf{y}$ à un ensemble de variables explicatives $\{\mathbf{x}_j, j = 1,..., p\}$. On note $\mathbf{X}$ la matrice des données associée aux variables explicatives. Cela revient à supposer l'existence d'une variable latente qualitative *C* à *K* modalités telle que $\mathrm{E}(\mathbf{y}|\mathbf{x}) = b_0^k + b_1^k\mathbf{x}_1 + b_2^k\mathbf{x}_2 + \ldots + b_P^k\mathbf{x}_p$ où les b_j^k sont les coefficients de la régression de $\mathbf{y}$ sur les $\mathbf{x}_j$ restreinte aux n_k observations de la classe k décrites par $y^k, \mathbf{X}^k$; avec $n_k \geq p$ pour garantir l'existence d'une solution pour les b_j^k. La régression typologique revient donc à chercher simultanément une partition en *K* classes et le vecteur b^k des coefficients b_j^k correspondant minimisant le critère $\mathrm{Z} = \sum_{k=1}^{K} \left\| \mathrm{X}^k b^k - y^k \right\|^2$.

Diverses méthodes et algorithmes ont été proposés pour l'estimation des coefficients Bock (1969), Diday (1979) et Spaeth (1979). On peut aussi citer les travaux de DeSarbo et Cron (1988) qui utilisent une méthode du maximum de vraisemblance et l'algorithme EM pour estimer les paramètres du modèle. Plus précisément, on suppose que les $(y_i, x_{ij}), i = 1,...,n; j = 1,..., p$ constituent un échantillon d'observations indépendantes issues d'un mélange de lois normales conditionnelles: $y_i \sim \sum_{k=1}^{K} \lambda_k f_{ik}(y_i | X_{ij}, \sigma_k^2, b_j^k)$ où les λ_k sont les proportions inconnues du mélange.

Etant donnés, *K*, **y** et **X**, la méthode fournit les estimations $\hat{\lambda}_k$, $\hat{\sigma}_k^2$ et $\hat{b}_j^k$ des paramètres maximisant la log-vraisemblance des données. Il est alors possible pour une observation *i* de l'affecter à la classe de la partition ayant la probabilité a posteriori estimée $\hat{p}_{ik} = \lambda_k f_{ik}(y_i | X_{ij}, \hat{\sigma}_k^2, \hat{b}_j^k) / \sum_{k=1}^{K} \lambda_k f_{ik}(y_i | X_{ij}, \hat{\sigma}_k^2, \hat{b}_j^k)$ maximale (règle de Bayes). La valeur prédite $\hat{y}_i$ est obtenue en utilisant le modèle associé à la classe d'appartenance de l'observation *i*.

3 Méthode proposée

Nous disposons de *n* observations décrites à l'aide d'une variable **y** à expliquer et un ensemble de variables explicatives structurées en *T* tableaux de p_t variables. On note respectivement **X et** $\mathbf{X}_t$ les matrices des données associées (figure 1).

	$\mathbf{X}_1$	$\mathbf{X}_2$		$\mathbf{X}_T$
X =	*VARIABLES* *1,2, ….j..,* p_1	*VARIABLES* *1,2,..j…,* p_2	……..	*VARIABLES* *1,2,…,j,…* p_T

Figure 1- Matrice des données

Nous formulons le problème de l'extension la régression linéaire typologique au cas multi-tableaux comme la recherche de *K* modèles de régression linéaire multi-tableaux pour chacune des *K* classes d'une variable latente déterminées simultanément à la régression.

Pour en compte explicitement la structuration en tableaux des variables explicatives nous supposons aussi l'existence d'une structure latente des individus propre à chaque tableau. Nous supposons donc cette dernière représentée par une variable C_t à K_t classes ; K_t pouvant être différent de *K* car les tableaux peuvent être porteurs d'une information propre qui pourrait être masquée en imposant à priori la même structure latente à chaque tableau.

Nous proposons d'effectuer une régression typologique sur chaque tableau $\mathbf{X}_t$. On recherche alors simultanément une partition en K_t classes et le vecteur b^{tk} des p_t coefficients b_j^{tk} correspondant minimisant le critère $Z_t = \sum_{k=1}^{K_t} \left\| \mathrm{X}^{tk} b^{tk} - y^k \right\|^2$ .On obtient alors *T* variables $\hat{\mathbf{y}}^t$ contenant les valeurs prédites de **y** localement à chaque classe de la variable C_t. Nous proposons ensuite de combiner ces prédictions à travers une deuxième étape de régression typologique pour obtenir simultanément une partition des observations en *K* classes et les modèles de régression par classe associés. Les *T* coefficients de ces régressions sont notés b_t^{Tk} . Le critère à minimiser est Z alors $= \sum_{k=1}^{K} \left\| \hat{\mathrm{y}}^t b^{Tk} - y^k \right\|^2$.

La méthode est ici proposée dans le cas de la régression linéaire, mais elle peut être naturellement étendue à la régression sur composantes principales ou à la régression PLS en cas de multicolinéarité ou lorsque le nombre de variables est très grand.

Le cas particulier où le grand nombre de variables est supérieur à celui des individus est encore plus fréquent dans la régression typologique puisqu'on applique la régression dans des classes d'effectifs encore plus réduits. La considération des blocs séparément constitue alors une première solution.

4 Application

Nous illustrons la méthode proposée sur des données de pollution de l'air intérieur issues de la campagne logement menée par l'Observatoire de la Qualité de l'Air Intérieur (OQAI) entre 2003 et 2005. Sur un échantillon de 567 logements on dispose de 58 variables structurées en trois blocs thématiques : le premier décrit les caractéristiques des logements avec 32 variables, le deuxième concerne la structure du ménage et contient 5 variables puis le dernier bloc décrit les habitudes des occupants à partir de 21 variables quantitatives. Par ailleurs plusieurs polluants ont été mesurés parmi lesquels le formaldéhyde qui sera la variable explicative dans cette application.
Des études antérieures avaient permis de dégager des typologies des logements selon chaque thématique ainsi qu'une typologie globale justifiant a priori la pertinence d'une approche de type régression typologique pour l'estimation des polluants. Nous avons appliqué une régression linéaire typologique sur les tableaux séparément. La comparaison de leurs performances en terme de R^2, la variance expliquée par les modèles, avec celles de régressions linéaires effectuées sur la totalité de l'échantillon (tableau 1) confirme la pertinence de l'utilisation de la régression typologique.

	LOGT	HAB	MEN
Rlin	0.11	0.037	0.05
RLin_CW	0.99	0.77	0.97

Tableau 1 : R^2 des régressions linéaires (Rlin) et régressions linéaires typologiques (Rlin_CW) .

Le modèle global correspondant à la méthode de régression typologique multi-blocs que nous proposons fournit un R^2 = de 0.999.
Nous avons aussi effectué la régression linéaire et la régression linéaire typologique sur le tableau **X**, sans prise en compte explicite de la structuration en blocs, cela fournit un R^2 de 0.205 et 0.999 respectivement. De nouveau la pertinence de la régression typologique est confirmée. Par ailleurs l'égalité des R^2 à 0.999 semble correspondre à ce qu'on observe dans le cas de la régression PLS multi-blocs (MBPLS). En effet, Westerhuis (1998) montre que les scores obtenus par la MBPLS sont identiques à ceux obtenus avec la PLS sur **X**.
Cependant ici, la régression typologique fournit également des partitions des individus mais les partitions obtenues par les deux méthodes ne sont pas en concordance aussi parfaite, contrairement aux prédictions. En effet, on obtient des indices de Rand variant de 0.72 à 0.8. Une étude approfondie est nécessaire pour la confirmation ces résultats.

Conclusion

Nous avons proposé une méthode de régression typologique multi-tableaux de type hiérarchique à deux niveaux. Les résultats obtenus sur des données réelles de pollution sont encourageants et montrent en particulier la pertinence de la démarche en termes de qualité des prévisions. Cependant, nous n'avons pas explicitement pris en compte les partitions associées aux modèles de chaque bloc, en particulier dans la deuxième étape de notre méthode. Nous envisageons de poursuivre nos travaux dans ce sens dans le but d'obtenir une partition finale plus consensuelle. Des évaluations plus formelles notamment sur des données simulées sont aussi nécessaires.

Bibliographie

[1] WESTERHUIS, J., KOURTI, T.,et MACGREGOR, J. (1998), Analysis of multiblock and hierarchical PCA and PLS models, *Journal of chemometrics*, *12*(5), 301-321.

[2] Yaroshchyk, P., Death, D. L., et Spencer, S. J. (2012), Comparison of principal components regression, partial least squares regression, multi-block partial least squares regression, and serial partial least squares regression algorithms for the analysis of Fe in iron ore using LIBS, *Journal of Analytical Atomic Spectrometry*, *27*(1), 92-98.

[3] Vivien, M. (2002), *Approches PLS linéaires et non linéaires pour la modélisation de multi-tableaux. Théorie et applications*, Thèse de Doctorat, Université Montpellier I.

[4] Eslami A., Kohler A., Qannari E.M., Bougeard S. (2013), General overview of methods of analysis of multi-group datasets, *Revue des Nouvelles Technologies de l'Information,* 108-123.

[5] Tenenhaus, A. et Tenenhaus, M. (2011), Regularized Generalized Canonical Correlation Analysis, *Psychometrika*, *76*(2), 257-284.

[6] Spaeth, H., (1979), Clusterwise linear regression, *Computing* 22, 367–373.

[7] DeSarbo, W.S., Cron, W.L., (1988), A maximum likelihood methodology for clusterwise linear regression, *Journal of classification*, 5, 249–282.

[8] Bock, H.H., (1969), *The equivalence of two extremal problems and its application to the iterative classification of multivariate data,* Lecture Note, Vortragsausarbeitung, Tagung Meolizinische Statistik", Mathematisches Forschungsinstitut Oberwolfach, 1969, pp.10.

[9] Diday, E. (1976), Classification et sélection de paramètres sous contraintes, *Rapport de recherche IRIA-LABORIA*, no 188.

[10] Charles, C., (1977), *Régression typologique et reconnaissance des formes*, Thèse de doctorat, Université Paris IX.

[11] Preda C. et Saporta G. (2005), Clusterwise PLS regression on a stochastic process, *Computational Statistics & Data Analysis*, 49, 99 – 108.

[12] Hennig, C . (2002), Fixed point clusters for linear regression: computation and comparison, *Journal of classification*, 19, 249–276