Prédiction en régression clusterwise PLS

NDèye Niang [1] Stéphanie Bougeard [2] & Gilbert Saporta [1]

[1] *CNAM-CEDRIC, Paris, ndeye.niang_keita@cnam.fr, gilbert.saporta@cnam.fr*
[2] *Anses, Ploufragan, France, stephanie.bougeard@anses.fr*

Résumé. En régression, lorsque les individus présentent une structure en groupes inconnus *a priori*, les méthodes de régression typologique ou clusterwise permettent d'apporter une réponse à travers la recherche simultanée d'une partition des données en un nombre fixé de classes et le modèle de régression local associé à ces classes. La régression clusterwise PLS permet une extension au cas de données de grande dimension et/ou fortement corrélées à travers la détermination de combinaisons linéaires des variables initiales. Se pose alors le problème de la prédiction à partir des modèles locaux en particulier en cas de grande dimension. En effet les composantes PLS sont différentes d'une classe à l'autre et ne peuvent donc être utilisées directement pour déterminer la classe d'appartenance d'un nouvel individu. Nous proposons pour cela d'effectuer une analyse discriminante sur une sélection de composantes principales issues d'une ACP sur les prédicteurs. La méthode proposée est illustrée sur des données simulées.

Mots-clés. Régression PLS, régression clusterwise, classification, discrimination.

Abstract. Clusterwise linear regression aims at partitioning data sets into clusters characterized by their specific coefficients in a linear regression model. High dimensional data and/or the case of multicollinearity can be handled using clusterwise PLS regression based on components which are linear combinations of the initial predictors. The corresponding local PLS models can be used for prediction purpose once the cluster membership of a future observation is determined. The PLS components are related to the clusters and then may differ from one cluster to another. Therefore they cannot be directly used for the cluster membership determination. We propose to use a discriminant analysis on a selected set of principal components from the PCA of the predictors. The method is illustrated on synthetic data.

Keywords. PLS regression, Clustering, Clusterwise regression, Discriminant analysis

1 Introduction

Les méthodes de régression régularisée telles que la régression sur composantes principales, la régression ridge ou la régression PLS permettent de lever les limitations de la régression linéaire classique notamment en cas de multicolinéarité ou lorsque le nombre de variables est inférieur à celui des individus. Tout comme la régression multiple, ces

méthodes fournissent un seul ensemble de coefficients de régression pour tous les individus. Cependant, dans un grand nombre d'applications en sciences sociales, en environnement ou en marketing par exemple, cet unique ensemble de coefficients peut parfois ne pas convenir et conduire à des estimations erronées. C'est en particulier le cas lorsque les individus présentent une structure de groupes sous-jacente mais inconnue. Les méthodes de régression typologique ou clusterwise permettent alors d'apporter une solution à travers la recherche simultanée d'une partition des données en un nombre fixé de classes et le modèle de régression local associé à ces classes. Tout se passe comme en classification mais ici les classes sont générées en minimisant un critère basé sur les résidus de régression plutôt que selon le critère classique utilisant les distances entre observations. On note $\mathbf{X}$ la matrice des données associée aux variables explicatives et $\mathbf{y}$ la variable à expliquer. La régression clusterwise revient à supposer l'existence d'une variable latente qualitative $\mathbf{C}$ à K modalités telle que $E(y|x) = b_0^k + b_1^k x_1 + \ldots + b_P^k x_p$ où les b_j^k sont les coefficients de la régression de $\mathbf{y}$ sur les x_j restreintes aux n_k observations de la classe k décrites par $\mathbf{y}^k$ et $\mathbf{X}^k$ avec $(n_k > p)$ pour garantir l'existence d'une solution pour les b_j^k. Cela revient donc à chercher simultanément une partition en K classes et le vecteur b^k des coefficients b_j^k correspondant minimisant le critère (1).

$$\sum_{k=1}^{K} \|\mathbf{X}^k b^k - \mathbf{y}^k\|^2 \tag{1}$$

Diverses méthodes et algorithmes ont été proposés pour l'estimation des coefficients. Bock (1969), Diday (1976) et Späth (1979) proposent une approche géométrique basée sur un algorithme de type K-moyennes pour minimiser le critère (1). Le nombre de classes est choisi par validation croisée. De Sarbo et Cron (1988) utilisent une méthode du maximum de vraisemblance et l'algorithme EM pour estimer les paramètres du modèle en supposant que les $(y_i, x_{ij}), i = (1, \ldots, n), j = (1, \ldots, p)$ constituent un échantillon d'observations indépendantes issues d'un modèle de mélange gaussien. La détermination du nombre de classes repose sur une utilisation exclusive de l'un des critères BIC ou AIC. La régression clusterwise a été étendue à la régression sur composantes principales par Charles (1977) et à la régression PLS par Esposito Vinzi *et al.* (2005). Plus récemment Preda et Saporta (2005) l'ont utilisée dans le cadre de la régression PLS sur données fonctionnelles. Des extensions à des données multiblocs ont été proposées par Niang et Saporta (2014) puis Niang *et al.* (2015) en particulier dans le cadre de la régression PLS. Ces extensions à la régression sur composantes permettent de lever la contrainte sur la taille des classes $(n_k > p)$, l'étude de données de grandes dimensions et le traitement de la multicolinéarité. Une fois les classes et leur modèle de régression associé obtenus, la phase suivante consiste à utiliser ces modèles locaux pour la prédiction de futures observations après avoir déterminé la classe d'appartenance de l'observation. Notons que les méthodes basées sur le maximum de vraisemblance telles que proposées par exemple dans le package `Flexmix` de `R` ne permettent pas cette prédiction, le calcul des probabilités d'appartenance aux classes nécessitant la connaissance de la valeur observée de $\mathbf{y}$. Ces méthodes mettent l'accent sur

l'ajustement dans l'objectif de trouver le modèle le plus adapté à chaque classe.

D'une manière générale, la phase de prédiction n'est pas souvent traitée explicitement. Généralement, sans donner de détails, les auteurs proposent de déterminer la classe d'appartenance de l'observation à travers une analyse discriminante par exemple et d'utiliser le modèle de régression local associé à cette classe. Nous nous intéressons dans cette communication à cette phase de prédiction dans la régression clusterwise PLS d'une seule variable sur un ensemble de variables explicatives. Les composantes clusterwise PLS issues de régressions PLS locales aux classes sont différentes d'une classe à l'autre : ce sont des combinaisons linéaires différentes des variables initiales. De plus, les nombres $(H_1, H_2, \ldots, H_K)$ de composantes clusterwise PLS peuvent différer d'une classe à l'autre. Se pose alors le problème de leur utilisation dans la prédiction à partir des modèles locaux. Esposito Vinzi *et al.* (2005) proposent une procédure itérative complexe basée sur la recherche d'un modèle compromis à partir d'une régression PLS avec H_{max} composantes clusterwise PLS où H_{max} est le nombre maximal de composantes PLS locales. A notre connaissance ce problème n'a pas été beaucoup abordé dans la littérature. Plusieurs solutions plus simples basées sur des méthodes factorielles de réduction de la dimension de l'espace des prédicteurs sont envisageables pour la recherche d'un espace compromis. Par ailleurs, une approche de type agrégation de modèles pour combiner de façon optimale les prédictions issues des modèles locaux plutôt que l'utilisation d'un seul modèle local pourrait être pertinente. Nous le détaillons en section 2. La section 3 est consacrée à une brève illustration sur des données simulées.

2 Prédiction en régression clusterwise PLS

La régression clusterwise PLS consiste à appliquer une régression PLS plutôt qu'une régression standard dans l'algorithme de la régression clusterwise. Elle revient donc rechercher simultanément une partition des individus en K classes obtenues en minimisant la somme des résidus des régressions PLS locales aux classes de la partition. Le critère à minimiser est donc identique au critère (1) dans lequel $\mathbf{X}^k$ est remplacé par la composante PLS t^k, combinaison linéaire des variables de $\mathbf{X}^k$ ayant la covariance maximale avec $\mathbf{y}^k$, représentant la variable à expliquer restreinte à la classe k. Elle fournit pour chaque classe de la partition optimale un ensemble des composantes PLS et les coefficients de régression associés. Il est alors possible d'obtenir l'équation de régression en fonction des variables initiales pour chaque classe k : $\hat{y^k} = b_0^k + b_1^k x_1^k + \ldots + b_p^k x_p^k$. Ces modèles de régression locaux sont ensuite utilisés pour prédire la valeur de $\mathbf{y}$ pour de futures observations. Mais il est nécessaire au préalable de déterminer la classe d'appartenance des nouvelles observations.

Affectation d'une nouvelle observation aux classes L'affectation d'une nouvelle observation aux classes revient à considérer la variable latente qualitative $\mathbf{C}$ à K modalités représentant la partition des individus comme une variable à prédire à partir des

prédicteurs initiaux $\mathbf{X}$. Nous proposons d'utiliser une analyse discriminante bayésienne paramétrique sous hypothèse de normalité pour obtenir un modèle explicite. Plus précisément on suppose que les K classes sont en proportion $(p_1, p_2, \ldots, p_K)$ dans la population totale et que la distribution d'un vecteur d'observation $x = (x_1, x_2, \ldots, x_p)$ est donnée pour chaque groupe k par une densité $f_k(x)$. Dans le cas classique, cette densité est celle d'une loi normale et les probabilités *a priori* p_k sont estimées par les fréquences empiriques des classes $\mathbf{C}_k$. La probabilité qu'une nouvelle observation x_0 appartienne à la classe k est donnée par la formule de Bayes $P(C_k|x_0) = p_k f_k(x_0) / \sum_{k=1}^{K} p_k f_k(x_0)$. La règle de discrimination bayésienne consiste alors à affecter l'observation x_0 à la classe k_0 ayant la probabilité *a posteriori* maximale : $P(C_{k0}|x_0) = \max_{k=1}^{K} P(C_k|x_0)$. Il est aussi possible d'utiliser une estimation non paramétrique pour les probabilités *a posteriori*. Nous proposons la méthode des k plus proches voisins dans laquelle la matrice de variance intra-classe est utilisée pour le calcul des distances de Mahalanobis entre individus. Dans le cas d'un très grand nombre de prédicteurs ou en cas de multicolinéarité, nous proposons d'effectuer l'analyse discriminante sur une sélection (par une méthode pas à pas par exemple) de composantes principales de l'ACP des prédicteurs $\mathbf{X}$.

Prédiction d'une nouvelle observation Pour une nouvelle observation à prédire, l'analyse discriminante fournit les probabilités d'appartenance aux différentes classes de la partition associée à la régression clusterwise. Deux solutions sont alors possibles : (i) Le modèle correspondant est appliqué à la classe $\mathbf{C}_{k_0}$ la plus probable. La prédiction de la valeur pour l'observation $x_0 = (x_1, x_2, \ldots, x_p)$ est donc $\hat{y_0} = b_0^{k_0} + b_1^{k_0} x_1^{k_0} + \ldots + b_p^{k_0} x_p^{k_0}$ que nous notons $\hat{y}_0^{k_0}$. (ii) Les prédictions des K modèles sont combinées en les pondérant par les probabilités *a posteriori* d'appartenance aux différentes classes : $\hat{y_0} = \sum_{k=1}^{K} P(C_k|x_0) \hat{y}_0^k$.

3 Application

Les données sont simulées pour le cas de $(K = 2)$ classes. Les variables explicatives sont positivement liées à $\mathbf{Y}$ dans la première classe $(b = 1)$ et négativement dans la deuxième $(b = -1)$. Les centres de gravité des deux classes sont bien séparés et les classes ne présentent pas de chevauchement. Les proportions attendues d'individus dans chacune des classes sont égales. Ces données présentent deux profils marqués : (Cas 1) Les données ayant un profil de multicolinéarité comportent $(N = 100)$ observations, $(P = 7)$ variables explicatives dont six sont corrélées les unes avec les autres à la valeur $0,9$ et la septième est totalement redondante avec la sixième, et (Cas 2) Les données ayant un profil où le nombre de variables est plus grand que celui des observations par classe comportent $(N = 50)$ observations, $(P = 30)$ variables explicatives ayant une corrélation les unes avec les autres de $0,7$. De plus, les résidus par classe de ces données sont soit des lois normales, soit des lois uniformes. Par la suite, des données ayant quatre types de profil sont donc étudiées. La méthode de régression PLS1 clusterwise (associée à une seule dimension) est appliquée.

Vingt initiations aléatoires sont appliquées pour éviter les *optima* locaux. Deux méthodes d'affectation, *i.e.*, k plus proches voisins (kNN) et analyse discriminante bayésienne (DA), associées à deux méthodes prédiction, *i.e.*, affectation directe ou pondérée, sont illustrées sur les données simulées. La comparaison des performances des méthodes d'affectation et de prédiction est évaluée par l'erreur de prédiction moyenne ($RMSE.p$) calculée par validation croisée 10-folds. Les performances sont données pour différents nombres de classes : ($K = 1$) soit une régression PLS1 standard, ($K = 2$) associé au vrai nombre de classes simulées et ($K = 5$). Les résultats dont donnés dans la Table 1.

Affectation	Prédiction	Loi résidu	K	Cas 1 : MULTICOL $RMSE.p$	Cas 2 : $p > n_k$ $RMSE.p$
KNN	-	Normale	1	9,62	43,29
KNN	Directe	Normale	2	0,65	0,97
KNN	Pondérée	Normale	2	0,95	1,14
KNN	Directe	Normale	5	0,46	0,40
KNN	Pondérée	Normale	5	0,79	0,52
DA	-	Normale	1	9,62	43,29
DA	Directe	Normale	2	1,88	4,39
DA	Pondérée	Normale	2	1,99	5,10
DA	Directe	Normale	5	1,42	1,33
DA	Pondérée	Normale	5	1,49	0,52
KNN	-	Uniforme	1	8,78	33,28
KNN	Directe	Uniforme	2	0,95	1,55
KNN	Pondérée	Uniforme	2	1,04	1,39
KNN	Directe	Uniforme	5	0,77	1,06
KNN	Pondérée	Uniforme	5	0,94	1,30
DA	-	Uniforme	1	8,78	33,28
DA	Directe	Uniforme	2	1,86	5,83
DA	Pondérée	Uniforme	2	1,79	7,16
DA	Directe	Uniforme	5	1,64	2,23
DA	Pondérée	Uniforme	5	1,63	2,39

TABLE 1 – Erreurs moyennes de prédiction ($RMSE.p$) de la PLS1 clusterwise issues d'une validation croisée 10-folds pour le cas où les données sont fortement multicolinéaires (Cas 1) ou le nombre de variables est plus grand que celui des observations (Cas 2).

Il convient tout d'abord de noter que l'utilisation de la régression PLS clusterwise plutôt que standard améliore très nettement l'erreur de prédiction. Ce résultat est attendu car ($K = 2$) classes distinctes d'individus ont été simulées. On peut aussi noter que l'augmentation du nombre de classes améliore légèrement l'erreur de prédiction dont il conviendra d'étudier la décroissance (plutôt que la minimisation) pour sélectionner le nombre optimal de classes. Par ailleurs, l'affectation de nouveaux individus est plus performante par la méthode des k plus proches voisins (non paramétrique) plutôt que par analyse discriminante, même sur composantes sélectionnées d'ACP. Ce résultat n'est pas surprenant à la vue des profils de données simulés (multicolinéarité ou nombre de variables

plus grand que le nombre d'observations). Les données dont les résidus présentent une loi normale plutôt qu'uniforme présentent de meilleures performances de prédiction pour le cas où le nombre de variables est plus grand que le nombre d'observations par classe ; leurs performances sont comparables pour le cas de données fortement multicolinéaires. La différence de performance de prédiction entre prédiction directe et pondérée semble peu importante et légèrement en faveur de la prédiction directe. Cela est certainement lié à la parfaite séparation des classes.

4 Conclusion

Nous avons proposé une méthode de régression clusterwise PLS adaptée aux données de grande dimension ou multicolinéaires en mettant l'accent sur la prédiction. Les résultats obtenus sur des données simulées montrent la pertinence de la démarche en termes de qualité des prévisions. Cependant, des évaluations plus formelles notamment sur différents cas de données simulées sont nécessaires de même que des applications sur des données réelles. Nous poursuivons nos travaux sur la prédiction avec des approches de type agrégation de modèles et des extensions aux régressions clusterwise multiblocs.

Bibliographie

[1] Bock, H.H. (1969) The equivalence of two extremal problems and its application to the iterative classification of multivariate data, Vortragsausarbeitung, Tagung. *Mathematisches Forschungsinstitut Oberwolfach*
[2] Charles, C. (1977) Régression typologique et reconnaissance des formes. Thèse. Université Paris IX.
[3] DeSarbo, W.S. et Cron, W.L. (1988) A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**, 249-282.
[4] Diday, D. (1976) Classification et sélection de paramètres sous contraintes, *Rapport de recherche INRIA-LABORIA*, 188.
[5] Niang, N. et Saporta, G. (2014) Régression typologique pour données multi-blocs, 46ièmes journées de statistique, juin 2014, Rennes, France.
[6] Niang, N., Bougeard, S., Saporta, G. et Abdi, H. (2015) Clusterwise multiblock PLS, CARME, Septembre 2015, p 58, Naples, Italie.
[7] Preda, G. et Saporta, G. (2005) Clusterwise PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, **49** : 99-108.
[8] Späth, H. (1979) Clusterwise linear regression. *Computing*, **22**, 367-373.
[9] Vinzi, V.E., Lauro, C.N. et Amato, S. (2005) PLS typological regression. In : *New developments in classification and data analysis*, 133-140.