Clusterwise multiblock PLS

Ndèye Niang, Stéphanie Bougeard, Gilbert Saporta

▶ **To cite this version:**

Ndèye Niang, Stéphanie Bougeard, Gilbert Saporta. Clusterwise multiblock PLS. SFC 2018, Sep 2018, Paris, France. hal-02471608

HAL Id: hal-02471608
https://cnam.hal.science/hal-02471608v1

Submitted on 12 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Clusterwise multiblock PLS

Ndèye Niang*, Stéphanie Bougeard**
Gilbert Saporta*

*CNAM-CEDRIC, 292 rue St Martin, 75141 Paris Cedex 03, France
ndeye.niang_keita@cnam.fr, gilbert.saporta@cnam.fr,
** Anses, Département d'Epidémiologie, 22440 Ploufragan, France,
stephanie.bougeard@anses.fr

Résumé. Clusterwise linear regression aims at partitioning a dataset into clusters characterized by their own regression coefficients. To deal with multiblock data, an extension of clusterwise regression to multiblock PLS is proposed. As this method is component-based, it may handle high dimensional data. The interest of the proposed method will be illustrated on the basis of a simulation study.

1 Introduction

Several multiblock regression methods are proposed for the analysis of huge transactional experimental data usually generated as multiblock data. Among them, multiblock Partial Least Squares (mbPLS) aims at exploring and modeling the relationships between several variables to be predicted from several other explanatory ones organized into meaningful blocks. The main idea of this component-based regression method is to find out the best combination of variables within explanatory blocks summarized with components to explain and predict the dependent variables. As an extension of PLS regression it inherits its advantages, that is, it handles situations where the variable number is higher than the observation number as well as multicollinearity.

However in many applications, observations do not come from a single and homogeneous population ; there is an unknown underlying group structure of observations. It follows that the estimation of a single set of regression coefficients for the whole dataset may mask the variables relationships and be misleading. Clusterwise, also called typological, regression is proposed to overcome this drawback in the context of standard regression. Clusterwise regression consists of simultaneously looking for a partition of the observations into clusters and their associated regression model by minimizing the sum of the error sums of squares computed over all the clusters. As in standard linear multiple regression, ordinary least squares or maximum likelihood estimation can be used to estimate clusterwise regression coefficients. Several methods and algorithms have been proposed. For example, in (DeSarbo et al, 1988) a maximum likelihood solution based on finite mixtures of conditional Gaussian distributions is proposed with a EM algorithm to get model parameters estimations. Extensions to principal component regression or PLS have been proposed to deal with multicollinearity and critical situations of small sample size or large variable number.

The problem addressed here is to model and predict multiblock data with the additional constraint of an underlying unknown cluster structure of observations. We propose an new clusterwise multiblock method which extends the typological PLS regression (Vinzi et al, 2005) to multiblock data. Additionally, the associated sequential algorithm guarantee the monotonous decreasing of the criterion to be minimized.

2 Methods

Consider the multiblock setting where a dataset $\mathbf{Y}$ is to be predicted from K explanatory ones $(\mathbf{X}^1, \ldots, \mathbf{X}^K)$. The $\mathbf{Y}$ dataset contains Q variables and each dataset $\mathbf{X}^k$ contains P^k variables $(k = 1, \ldots, K)$. The merged dataset $\mathbf{X} = [\mathbf{X}^1 | \ldots | \mathbf{X}^K]$ related to all the explanatory variables contains $P = \sum_k P^k$ variables. All these quantitative variables are measured on the same N observations and are supposed to be column centred. Furthermore, clusterwise multiblock data suppose that there is an unknown structure of the N observations into G clusters of size $N_1, \ldots, N_G$. In standard multiblock PLS, the relationship between $\mathbf{Y}$ and the K matrices $\mathbf{X}^k$ (stored in $\mathbf{X}$) is first modeled by computing a pair of linear combinations—called *components*–of the columns of, respectively $\mathbf{Y}$ and $\mathbf{X}$ such that these components have maximal covariance. After this first step–equivalent to a standard PLS model–specific components are computed to relate each $\mathbf{X}^k$ to $\mathbf{Y}$. Formally, mbPLS first implements the following optimization problem :

$$\begin{aligned} &\underset{u,t^k,a^k,t}{Argmax} \quad cov^2(u,t) \quad \text{with} \quad u = \mathbf{Y}v, \quad t = \mathbf{X}w = \sum_k a^k t^k \\ &\sum_k \left(a^k\right)^2 = 1, \quad t^k = \mathbf{X}^k w^k \quad \text{and} \quad \|w\| = \|w^k\| = \|v\| = 1 \end{aligned} \tag{1}$$

In a second step, the dependent dataset $\mathbf{Y}$ is predicted (with a standard linear regression) from the component t as $\widehat{\mathbf{Y}} = tc'$.

To improve the $\mathbf{Y}$ prediction, the second order solutions are obtained while deflating the datasets $(\mathbf{X}^1, \ldots, \mathbf{X}^K)$ by means of a regression onto the first global component denoted $t^{(1)}$. The maximization (1) is performed but the original datasets are replaced by the residuals obtained in the deflation step. We denote by O the optimal number of components to keep in the model which is in general estimated by a cross-validation approach. Then the dependent dataset is predicted as : $\widehat{\mathbf{Y}}^{(O)} = \sum_{h=1}^{O} t^{(h)} c^{(h)'} = \sum_{h=1}^{O} \sum_{k=1}^{K} a^{k(h)} t^{k(h)} c^{(h)'}$ with $c^{(h)} = \frac{\mathbf{Y}' t^{(h)}}{\|t^{(h)}\|^2}$ being the vector of the regression coefficients of Y on $t^{(h)}$.

This last regression step corresponds to the following optimization problem :

$$\underset{c}{\text{Argmin}} \quad \|\mathbf{Y} - \sum_{h=1}^{O} \sum_{k=1}^{K} a^{k(h)} t^{k(h)} c^{(h)'}\|^2 \text{ with } t^{k(h)} = \mathbf{X}^{k(h-1)} w^{k(h)} \text{ and } \|w^{k(h)}\| = 1 \tag{2}$$

$\mathbf{X}^{k(h-1)}$ is the residual of the prediction of X^k from $h-1$ previous components $(t^{(1)}, \ldots, t^{(h-1)})$.

mbPLS Criterion As for the standard clusterwise methods, the criterion to minimize is the sum of errors square computed over all the clusters using local multiblock PLS applied to each

cluster instead of standard regression. Let $\mathbf{Y}_g$, $\mathbf{X}_g = [\mathbf{X}_g^1|\ldots|\mathbf{X}_g^K]$ be the parts of the data corresponding to the cluster g respectively for the dependent variables and the explanatory ones. Equation (2) is adapted to the clusterwise context, and so the criterion for clusterwise mbPLS for a given number of clusters G and an optimal number O of components to be included in the model is

$$\underset{c_1,\ldots,c_G}{\text{Argmin}} \quad \sum_{g=1}^{G} \quad \|\mathbf{Y}_g - \sum_{h=1}^{O}\sum_{k=1}^{K} a_g^{k(h)} t_g^{k(h)} c_g^{(h)'}\|^2 \quad \text{with} \quad c_g^{(h)} = \frac{\mathbf{Y}_g' t_g^{(h)}}{\|t_g^{(h)}\|^2} \tag{3}$$

In order to ensure the decreasing monotonicity of the criterion according to the N iterations, we propose to apply a sequential algorithm rather than a standard batch one ; it follows that each observation is assigned to its optimal cluster and the criterion is updated at each step of the algorithm as in standard K-means algorithm. Thereafter, to avoid local optimum, several random initialisations are used. The optimal numbers of clusters and dimensions are unknown and must therefore be determined. Because our method is distribution-free, penalized parametric criteria such as BIC or AIC cannot be used. As the proposed clusterwise multiblock methods can be used to explicitly predict new observations (see below), we propose to select the unknown parameters on the basis of a ten-fold cross-validation procedure where the optimal G and O parameters are selected to minimize the average root mean square error (RMSE) of prediction.

Prediction of new observations Up to our knowledge, in clusterwise framework, the prediction of a new observation is an original question not really addressed. Given a new data point for which only the values of the predictors are available, the prediction of the dependent variables $\mathbf{Y}$ is processed in two sequential steps : (i) the observation is firstly assigned to its nearest cluster and (ii) the relevant cluster regression model is applied to get the predicted $\mathbf{Y}$ value. In the first step, the response variable to predict is the categorical variable whose categories are the cluster labels. The associated predictor variables can be the initial explanatory variables if $\mathbf{X}$ is of full rank. If not, $\mathbf{X}$ can be summarized through components. Then several solutions exist for the assignment of a new observation to the nearest cluster. Geometrical rules based on Mahalanobis distances to the cluster gravity centres can be used or probabilistic rules such as the Bayes rule. In the context of multiblock data with a free distribution setting, we propose to apply a non parametric method using the maximum posterior probability estimated by a K-nearest neighbour method with euclidian distance or the Mahalanobis one ; the k number of neighbours is choosen through a cross-validation procedure.
Once the new observation is assigned to its optimal cluster, this cluster regression model is applied to get the predicted value.

3 Application

The proposed clusterwise method is illustrated in a simulated study through twenty-one situations depending on the general structure and the block structure as well as the cluster structure (Bougeard et al., 2017). Several factors are taken into account : the number of dependent variables, the number of explanatory variables, the number of blocks, the proportion of variables per block, the block weighting scheme when the blocks of variables have different sizes, the within-block correlation, the number of observations, the number of clusters, the

proportion of observations per cluster, and the separation between clusters. For each of the twenty-one case study, twenty datasets were simulated. The first performance criterion is the Root Mean Square Error of prediction evaluated with a ten-fold cross-validation procedure. Not surprisingly, mbPLS always improves the Root Mean Square Error of prediction while taking account the cluster structure of observations. This effect is particularly clear for the simplest simulation cases of well-separated clusters of equal sizes. The second performance criterion is the Adjusted Rand index. Not surprisingly, best performance was achieved for the case of well-separated clusters of equal sizes with an average adjusted Rand of .81 and slightly decreased for the case of well-separated clusters of different sizes with an average adjusted Rand of .77 and for the case of mild-separated clusters of equal sizes with an average Adjusted Rand of .73. The last performance criterion evaluated how well the actual regression coefficients were recovered. The regression coefficients were, most of time, correctly recovered for the case of well-separated clusters of equal sizes but differed slightly from the actual value when the block sizes differed and—but to a lesser extend— when block sizes varied.

4 Conclusion

We propose a new clusterwise multiblock regression method useful for analyzing complex data as found, for example in marketing, biology, or any field dealing with population mixtures. The proposed clusterwise procedure is oriented both towards modeling and prediction and can be applied to any other multiblock component methods.

Références

Bougeard, S., Abdi, H., Saporta, G. et al. Clusterwise analysis for multiblock component methods. Adv Data Anal Classif (2017). https://doi.org/10.1007/s11634-017-0296-8

Bougeard, S., Cariou, V., Saporta, G. et al. Prediction for regularized clusterwise multiblock regression. Applied Stochastic Models in Business and Industry, (2018)

DeSarbo WS, Cron WL, A maximum likelihood methodology for clusterwise linear regression. Journal of Classification, 5, 249-282 (1988)

Vinzi VE, Lauro CN, Amato S, PLS typological regression. In: New developments in classification and data analysis, 133-140 (2005)

Summary

La régression clusterwise vise à partitionner un ensemble de données en classes caractérisées chacune par un modèle de régression linéaire spécifique. Nous proposons une extension à la régression PLS multiblocs. Cette méthode basée sur des composantes permet alors de traiter des données de grande dimension. La méthode proposée est illustrée sur la base d'une étude de simulation.