

Orthogonal Projection to Latent Structures in Models for L-structured Data

Proiezione Ortogonale su Strutture Latenti in Modelli per Dati in Struttura a L

Silvia Squillacciotti^(a,b), Christiane Guinot^(b), Vincenzo Esposito Vinzi^(a)

^(a)Dipartimento di Matematica e Statistica, Università degli Studi di Napoli Federico II,
Via Cintia, 80126, Napoli

^(b)CE.R.I.E.S., Unité de Biométrie et Epidémiologie, 20 rue Victor Noir, 92521
Neuilly-sur-Seine, France

{silvsqui;vincenzo.espositovinzi}@unina.it; christiane.guinot@ceries-lab.com

Riassunto: Nell'ambito della regressione PLS, la selezione di variabili indipendenti viene generalmente effettuata attraverso procedure di cross-validation oppure valutando indici che misurano la capacità predittiva delle variabili. Un approccio diverso, proposto in ambito chemometrico, consiste nel “filtrare” i dati, eliminando la parte di variazione strutturale nelle variabili indipendenti non legata alle variabili dipendenti: le componenti ottenute prendono in conto solo la parte di informazione orientata alla predizione delle variabili dipendenti, permettendo quindi un miglioramento del modello in termini di predittività e interpretabilità. L'obiettivo di questo lavoro è mostrare come tale approccio possa essere esteso anche a strutture di dati più complesse, dette “a L”, in cui la matrice centrale è spiegata da due matrici di descrittori (di linee e di colonne).

Keywords: Partial Least Squares (PLS), Projection to Latent Structures, variable selection, orthogonalisation, L-structures, external information

Introduction

In “L-structures”, a central matrix containing the dependent variables is explained by the interaction between a row-descriptor table and a column-descriptor table. This data structure is frequently encountered in the study of consumer preferences. In this framework, the central matrix ($\mathbf{Y}$), containing the preference scores given by q consumers to n products, is explained by a row-descriptor matrix ($\mathbf{X}$), containing p descriptors of the n products, and by a column-descriptor matrix ($\mathbf{Z}$), containing the m descriptors of the q consumers. The aim of the analysis of such a structure is to study how the interaction between consumer descriptors (such as socio-demographic and/or behavioural characteristics) and product descriptors (such as physico-chemical or sensory variables) influences consumer preferences. Data concerning consumer descriptors, however, is often characterised by a high degree of “noise” which may affect the estimation of the model parameters when studying the effect of these descriptors on dependent variables (in our case consumer preferences). In order to remove, even partially, such noise, one may choose to proceed to variable selection. In PLS Regression (PLSR), variable selection is often performed through cross-validation methods, or by taking into account the VIP (*Variable Importance in Projection*) index (Tenenhaus, 1998). Variables showing a low VIP value (the threshold is generally equal to 1) will be considered as having a weak predictive power, and may be removed from the analysis. This approach may, however, lead to the removal of variables which, although having a low predictivity, are all the same of high importance for the final user or for the research objectives. A different approach has been proposed by Wold and Trygg in the spectroscopic domain, where generally an extremely high

number of predictors, containing a strong “structural variability” is available (Trygg *et al.*, 2002; Trygg, 2002). The principal PLS equations are given in the following equations (1a-c):

$$\hat{\mathbf{Y}} = \mathbf{T}_A \mathbf{C}'_A \quad (1a)$$

$$\mathbf{T}_A = \mathbf{X} \mathbf{W}_A^* \quad (1b)$$

$$\mathbf{X} = \mathbf{T}_A \mathbf{P}'_A \quad (1c)$$

In eq. (1a) $\mathbf{T}_A$ is the matrix containing the A model components (A is chosen by cross-validation) and $\mathbf{C}$ is a matrix of loading vectors. In eq. (1b) $\mathbf{W}^*$ is the matrix of the loading vectors relating the independent variables in $\mathbf{X}$ to the components in $\mathbf{T}_A$. Finally, eq. (1c) shows how PLS components can be used to obtain a “reconstructed” $\mathbf{X}$ matrix according to the PLS model, $\mathbf{X}_{\text{PLS}}$.

PLS components, estimated on the basis of the covariance maximisation criterion, may erroneously take into account a portion of the structural variability, pertinent only to predictors and not related to dependent variables. This may result in a loss in terms of dependent variables’ explained variability and in a lower model interpretability. The technique proposed by Trygg and Wold, *Orthogonal Projection to Latent Structures* (O-PLS), consists of defining loading vectors $\mathbf{w}$ orthogonal to $\mathbf{Y}$ ($\mathbf{w}_{\text{ortho}}$), so as to build score vectors $\mathbf{t}_{\text{ortho}}$ ($\mathbf{t}_{\text{ortho}} = \mathbf{X} \mathbf{w}_{\text{ortho}}$) taking into account only the portion of the predictors’ variability oriented towards the prediction of dependent variables. Through these components it is then possible to reconstruct a “filtered” predictor matrix, $\mathbf{X}_{\text{O-PLS}}$, from which the structural variation, orthogonal to dependent variables, has been removed.

In the present paper an extension of O-PLS to models for data in L-structures will be presented. The first section will be devoted to a description of methodologies for the analysis of L-structures. The discussion will be focused on PLS-based techniques. The second section will be devoted to the methodological aspects of O-PLS and of its extension to L-structure models. Finally, in the conclusions the results from the application of the described methodology will be discussed.

1. PLS Regression with external information on both rows and columns

A number of methods have been proposed in literature to face the analysis of this type of structure (among all: Takane *et al.*, 1991; Amenta *et al.*, 2000; Giordano *et al.*, 1999). In particular, Martens (Martens *et al.*, 2003) proposes an extension of the two-way PLS algorithm based on the Singular Value Decomposition (SVD) of the product matrix ($\mathbf{X}'\mathbf{Y}$) to three-way L-structured data. The adoption of PLS principles is extremely useful in preference studies, where data is usually strongly collinear and more variables than individuals are analysed. Martens’ L-PLSR is based on the SVD of the three-way product matrix ($\mathbf{X}'\mathbf{Y}\mathbf{Z}$), whose meaning, under the point of view of the maximised criterion, is obscure.

In previous works (Esposito Vinzi *et al.*, 2003; Squillacciotti *et al.*, 2003), the authors have proposed a PLS-based alternative approach to L-PLSR, based on a two-step methodology. In step 1, a PLSR of matrix $\mathbf{Y}$ on product descriptor matrix $\mathbf{X}$ is performed. In step 2, the loading vectors $\mathbf{c}_x$ relating the components $\mathbf{t}_x$ to $\mathbf{Y}$ are used as dependent variables in the PLSR on consumer descriptors in $\mathbf{Z}$. Hence in step 2 the influence of consumer descriptors on the relationship between preferences and product descriptors (information contained in matrix $\mathbf{C}_x$) is evaluated. The maximised criterion (PLS) is hence not only homogeneous in the two steps, but also fully logical and interpretable. The final model equation is given in (2):

$$\mathbf{Y} = \mathbf{T}_X \mathbf{C}_Z \mathbf{T}_Z' + \mathbf{Y}_{\text{res}} \quad (2)$$

where $\mathbf{T}_X$ and $\mathbf{T}_Z$ contain, respectively, the score vectors related to matrix $\mathbf{X}$ and $\mathbf{Z}$, $\mathbf{C}_Z$ contains the information relative to the effect of the interaction of the two component matrices on $\mathbf{Y}$ and $\mathbf{Y}_{\text{res}} = \mathbf{T}_X \mathbf{Y}_{\text{EZ}}' + \mathbf{Y}_E$ is the error term ($\mathbf{Y}_E$ and $\mathbf{Y}_{\text{EZ}}$ being, respectively, the error terms from step 1 and step 2).

2. Variable selection in models for L-structured data

Variable selection aiming at improving model predictivity is generally needed in step 2, due to the high degree of noise in consumer descriptors and hence to their scarce predictivity (Martens *et al.*, 2003; Squillacciotti *et al.*, 2003). Two-step PLSR is modified according to O-PLS principles, by introducing the orthogonalisation of loading vectors $\mathbf{w}_Z$ into $\mathbf{w}_{\text{ortho}}$ according to (3):

$$\mathbf{w}_{\text{ortho}} = \mathbf{p} - [\mathbf{t}_w' \mathbf{p} (\mathbf{t}_w' \mathbf{t}_w)^{-1}] \mathbf{t}_w \quad (3)$$

where $\mathbf{t}_w$ are the component scores resuming the relationship between $\mathbf{Z}$ and $\mathbf{C}_X$. The loadings $\mathbf{w}_{\text{ortho}}$ are then used for the computation of the component scores $\mathbf{t}_{\text{ortho}} = \mathbf{Z} \mathbf{w}_{\text{ortho}}$: this assures that the score vectors will take into account only the part of $\mathbf{Z}$'s variability orthogonal to $\mathbf{C}_X$. Afterwards, on the basis of the well known PLS equation allowing the reconstruction of the independent variables' matrix (Tenenhaus, 1998), the orthogonal variation matrix $\mathbf{Z}_{\text{ortho}} = \mathbf{T}_{\text{ortho}} \mathbf{P}_{\text{ortho}}'$ is computed, as well as the filtered data-matrix $\mathbf{Z}_{\text{O-PLS}}$ according to (4):

$$\mathbf{Z}_{\text{O-PLS}} = \mathbf{Z} - \mathbf{T}_{\text{ortho}} \mathbf{P}_{\text{ortho}}' \quad (4)$$

The last step of O-PLS modification in two-step L-PLSR consists of a PLSR of $\mathbf{C}_X$ onto $\mathbf{Z}_{\text{O-PLS}}$. Such PLSR leads to a better result interpretability and to a model with a higher $R^2(\mathbf{C}_X)$ at the price of lower $R^2(\mathbf{Z})$. A lower number of components is hence needed with respect to the model estimated on the unfiltered independent variables (Trygg *et al.*, 2002), and no variable has been wholly removed from the model. Finally, the analysis of the removed orthogonal variation may be performed, for example, through a PCA on $\mathbf{Z}_{\text{ortho}}$.

Conclusions

The described technique has been applied to data concerning consumer preferences on cosmetic products. Matrix $\mathbf{Y}$ (9 x 92) contains the preference scores from 92 consumers on 9 products. Matrix $\mathbf{X}$ (9 x 29) contains 15 physico-chemical and 14 sensory product descriptors. Matrix $\mathbf{Z}$ (92 x 73), finally, contains 73 variables concerning the consumers' age, cosmetic habits, skin characteristics and purchase behaviour.

The global model estimated through two-step L-PLSR (Esposito Vinzi *et al.*, 2003; Squillacciotti *et al.*, 2003) shows that consumer descriptors have a rather low predictive power in step 2

($R^2(C_X)=0.19$ with two significant dimensions). The low predictive power of consumer descriptors on preferences had already been proved by a PLSR of Y' on Z and was, on the basis of the results of a PCA on Z , attributed to the high structural noise in Z (Squillacciotti, 2003). The application of “filtered” two-step L-PLSR has allowed to confirm such conclusions and to find interesting results. Step 2 on the filtered data matrix Z_{O-PLS} led to $R^2(C_X)=0.223$ on two significant dimensions (hence to a slight increase). However, the analysis of the structural variation in Z_{ortho} performed through a PCA led to very interesting results: the principal inertia directions in Z are now immediately visible (2 dimensions account for 80% of the total variability, whereas 29 were needed in PCA on Z). Concluding, the modification of two-step L-PLSR through the inclusion of O-PLS filtering has first of all shown how O-PLS algorithm can be adapted to more complex data-structures for the selection of relevant information. On the particular application discussed here, this procedure has allowed to justify the weak influence of consumer descriptors on the relationship between consumer preferences and product descriptors, and, most of all, to isolate the strong structure pertinent to Z . Analyses on different data sets are on-going. Future research will be focused on the investigation over alternative techniques for variable selection in L-structures based on partial analysis and orthogonal projection.

References

- Amenta P., D'ambra L. (2000), Constrained Principal Component Analysis with External Information, *DMQTE*.
- Esposito Vinzi V., Guinot C., Squillacciotti S. (2003), Two-Step PLS Regression for L-Structured Data: an application in the cosmetics industry. Paper presented at “*SIS 2003: Analisi Statistica Multivariata per le scienze economico-sociali, le scienze naturali e la tecnologia*”, Naples, June 9-11, Invited and Specialized Sessions Volume, RCE Edizioni, Napoli, 157-171.
- Giordano G., Scepti G. (1999), Different Informative Structures for Quality Design, *Journal of the Italian Statistical Society*, 8, 139-149
- Martens H., Anderssen E., Flatberg A., Gidskehaug L.H., Høy M., Westad F., Thybo A., Martens M. (2003), Regression of a data matrix on descriptors of both its rows and of its column via latent variables: L-PLSR, *Computational Statistics and Data Analysis*, (in press).
- Squillacciotti S., Guinot C., Esposito Vinzi V., Tenenhaus M. (2003), PLS-based strategies in consumer preferences and sensory analysis: an application to the cosmetic industry, paper presented at “*PLS-03 – 3rd International Symposium on PLS and Related Methods*”, Lisbon, September 15-17, DECISIA, Montreuil, 311-324.
- Squillacciotti S. (2003), *PLS Modelling of mono- and multi-block L-structures: methodological contributions and applications*, PhD thesis, University Federico II, Naples.
- Takane Y., Shibayama T. (1991), Principal Component Analysis with external information on both subjects and variables, *Psychometrika*, 1, 97-120.
- Tenenhaus M. (1998), *La régression PLS: théorie et pratique*, Technip, Paris.
- Trygg J. (2002), O2-PLS for qualitative and quantitative analysis in multivariate calibration, *Journal of Chemometrics*, 16, 283-293.
- Trygg J., Wold S. (2002), Orthogonal projections to latent structures (O-PLS), *Journal of Chemometrics*, 16, 119-128.