

Three-Way Data Analysis and PLS Path Modelling for Estimating Components of a Temporal Process⁽¹⁾

Analisi di Tabelle a più vie e Modellizzazione basata sul PLS per la Stima delle Componenti di un Processo Temporale

Vincenzo Esposito Vinzi, Germana Scepi

Dipartimento di Matematica e Statistica

Università “Federico II” di Napoli Via Cintia, Monte S. Angelo 80126 Napoli

binci@unina.it, scepi@unina.it

Riassunto: In questo lavoro si propone una strategia che utilizza congiuntamente l’approccio confermativo dei modelli ad equazioni strutturali (secondo il criterio del Partial Least Squares, *PLS Path Modelling*) e tecniche esplorative di analisi dei dati a più vie (Analisi delle Matrici Principali). Tale strategia consente di rappresentare e interpretare strutture di dati costituite da blocchi di variabili osservate sulle stesse unità, identificate da istanti temporali, in occasioni differenti. L’obiettivo finale è quello di arricchire la stima dei coefficienti di un modello temporale con l’interpretazione delle sue diverse componenti.

Keywords: Partial Least Squares (PLS), Structural Equation Modelling, Principal Matrices Analysis, Multi-Block Data Structure, Dynamic Components.

1. Introduction

Within the confirmatory approach to path models, the Partial Least Squares (PLS) algorithm developed by H. Wold (1973) has recently gained popularity because of its flexible estimation options, its “soft” modelling hypotheses and its being feasible for several domains of application. Namely, in a multi-way tables framework, PLS Path Modelling looks for latent variables which summarise one-dimensional blocks of manifest variables taking into account also the causal links of dependence between the blocks. In the following, the PLS approach to path modelling is preferred to maximum likelihood not only because of its robustness in presence of missing data and multicollinearity but also for its prediction-oriented setting.

In the same framework, but with an exploratory objective, several three-way data analysis techniques have been proposed in order to synthesise and graphically represent the relationships *between* and/or *within* multi-way tables (Rizzi, 1989) when no explanatory information is available.

The aim of this paper is to provide a joint strategy of a three-way data analysis technique (Principal Matrices Analysis, PMA, D’Alessio 1989) with PLS path modelling in order to allow for a decomposition and an interpretation of the different

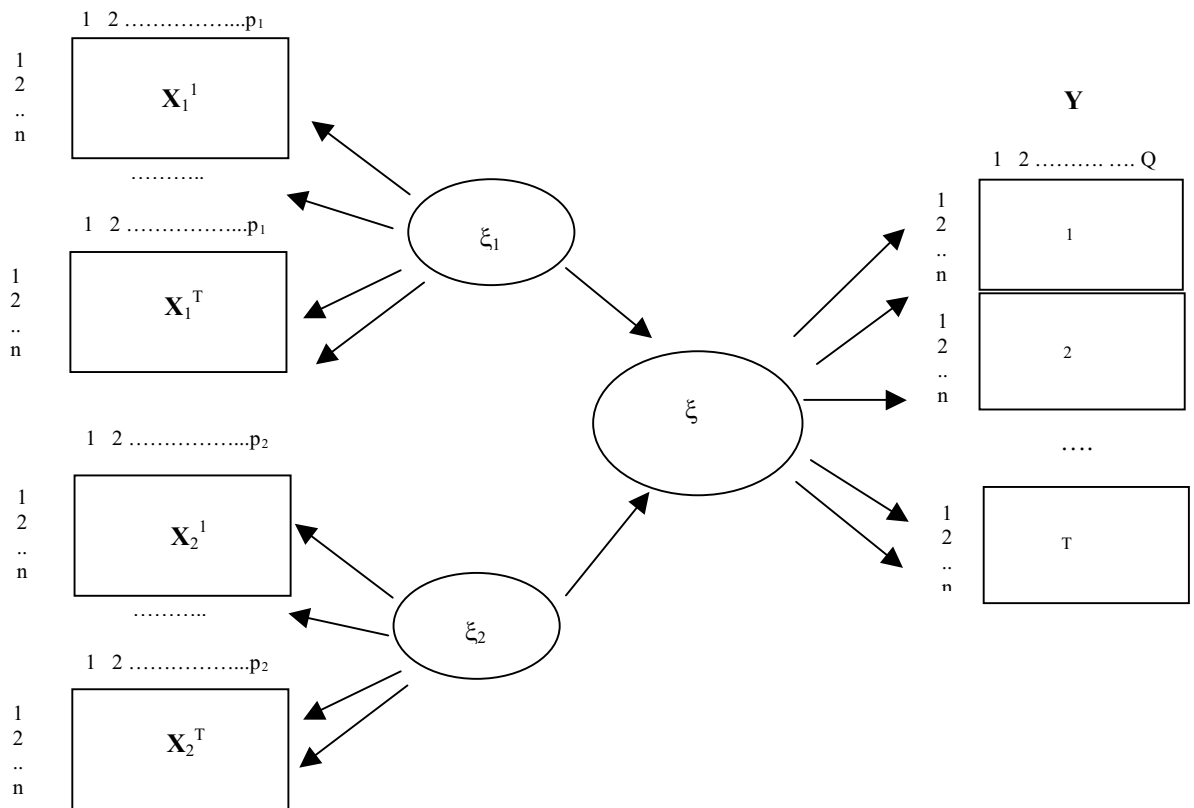
⁽¹⁾ The present paper is financially supported by the PRIN 2000 grant *Data Mining e Analisi Simbolica* (resp: Prof. Carlo Lauro).

components of a temporal process. It will be enhanced how this strategy enriches the results of classical PLS path modelling when the data structure at hand shows a double block-partition, i.e. blocks of variables (by column) and blocks of observational occasions (by row). The proposed strategy will be applied to an industrial production process.

2. Data Structure

The data structure of interest comprises J ($j=1, \dots, J$) groups of manifest variables observed on the same n ($i=1, \dots, n$) time points at T ($t=1, \dots, T$) different occasions. Therefore, for each group of variables, a sequence of T matrices $\mathbf{X}^t(n, p_j)$ is defined. Suppose that each column vector of $\mathbf{X}^t(n, p_j)$ is a realisation of a strictly stationary sequence of random variables. The generic t -th sequence for each group of variables can be thought of as a set of measurements of the same variables in different occasions (e.g. days), while the n rows of each matrix represent sample observations taken at different times in each occasion. Finally, the so-defined structure is constituted by T blocks of J groups of manifest variables.

Figure1: *The Data Structure by blocks*



As in a classical PLS path modelling (Figure 1), we suppose that $(J-1)$ groups of manifest variables play the role of explanatory variables with respect to a set of Q dependent variables which are also manifest. The Q variables are observed on the same n ($i=1, \dots, n$) time points at T ($t=1, \dots, T$) different occasions. Therefore, matrix $\mathbf{Y}$ of

dependent variables has dimensions $(n \times T, Q)$. In Figure 1, the final data structure is represented with $J=3$ variables overall. In particular, we have two exogenous latent variables and one latent variable which is endogenous.

3. PLS Path Model and Estimation

Each group of variables is considered to constitute the observable expression of a latent standardised variable group ξ_j . Each manifest variable x_{jh} ($h=1, \dots, p_j$) in the j -th block is assumed to be connected to its latent variable ξ_j (in a reflective scheme) by the equation:

$$x_{jh} = l_{jh} \xi_j + \varepsilon_{jh}$$

where ε_{jh} is a zero mean random error uncorrelated with the j -th latent variable. Moreover, it is possible to define the following causal relationships between the endogenous latent variable and the exogenous ones:

$$\xi_j = \sum_{j=1}^J \beta_{jj} \xi_j + \zeta_j$$

where ζ_j is a zero mean random error uncorrelated with the J latent variables. The estimation process for the latent variables, based on PLS, comprises two iterative steps: external estimation from the manifest variables and internal estimation from the other latent variables. For each of these steps, different modes exist which provide a flexible way of calculating the coefficients l_{jh} 's and β_{jj} 's (Tenenhaus, 1998). In the following paragraph, we propose a new method for estimating the exogenous latent variables in the case of the peculiar data structure illustrated above. In particular, we aim at obtaining more than one latent variable for each explanatory variable in order to take into account temporal components.

4. Three way data analysis for the decomposition of latent variables

If classical PLS path modelling is directly applied on the data structure shown in Section 2, the results would not explicitly consider the temporal link between the blocks within each group of explanatory variables. In fact, each sequence of explanatory variables would be summarised by a single latent variable as a linear combination of each manifest variable observed on $n \times T$ units. In order to include the temporal aspect in the analysis, a PMA is performed on each sequence of T blocks of the explanatory variables. As a consequence, we may obtain p_j reference matrices for each group of manifest variables which are actually compromises (weighted sums) of the original X 's defined as:

$$CO_a = \sum_{t=1}^T u_{a,t} X^t \quad a \in \{1, \dots, p_j\} \text{ denotes the number of compromises}$$

where $u_{a,t}$ is the t -th element of the a -th eigenvector of the diagonalisation of the interstructure matrix $\mathbf{IS}$ with general element, is_{it} , the RV-correlation coefficient of Escoufier between the variables at T different occasions:

$$is_{it} = \frac{tr(\mathbf{X}_t' \mathbf{X}_{t'})}{\sqrt{Tr(\mathbf{X}_t)^2 Tr(\mathbf{X}_{t'})^2}}$$

In the case of temporal data, the first eigenvector u_1 of matrix $\mathbf{IS}$ can be considered as a component which is not linked to time and is therefore defined as the *core* of the data structure (which usually explains most of the total inertia). On the contrary, the second eigenvector u_2 enhances the main temporal evolution and defines the *trend*, while the third eigenvector u_3 refers to the *cycle* (D'Alessio, 1986).

Following this scheme, we may retain the first three eigenvectors and compute the associated reference matrices $\mathbf{CO}_1$, $\mathbf{CO}_2$, $\mathbf{CO}_3$ which represent the components *core*, *trend* and *cycle* of each group of explanatory variables. These matrices substitute the sequences of explanatory variables on the left hand side of the data structure in Figure 1 and therefore represent the input variables for the external estimation of the latent variables. As a consequence, the generic latent variable ξ_j is decomposed in three components j^{core} , j^{trend} and j^{cycle} . Finally, by means of the structural relationships, such a decomposition is transferred to the endogenous latent variable thus allowing the estimation of its temporal components.

5. An example of application on real data

The proposed methodology is profitably applied to a real data set concerning the continuous production of zinc in an Italian factory which, for confidentiality reasons, is kept anonymous. The manifest explanatory variables measure different set up values of temperature, pressure and other significant environmental variables in the production process. These variables are regularly observed for several days and several times a day and are supposed to influence the amount of zinc and other materials produced by the process. Such a data structure perfectly fits the one defined in Figure 1. For the sake of brevity, the detailed results of the application, and their relevance in terms of interpretation of the latent variables estimates, will be discussed in the extended paper.

References

- D'Alessio G. (1986) L'Analisi di successioni di matrici di dati qualitativi, Tesi di Dottorato di Ricerca- Università La Sapienza –Roma.
- D'Alessio G.(1989) Multistep principal component analysis in the study of panel data, in: Multiway Data Analysis, R.Coppi & S.Bolasco (Eds.), North Holland, 375-381.
- Rizzi A.(1989) On the synthesis of three-way data matrices, in: Multiway Data Analysis, R.Coppi & S.Bolasco (Eds.), North Holland, 143-154.
- Tenenhaus (1998) *La régression PLS, Théorie et Pratique*, Technip, Paris.
- Wold (1973) Nonlinear Iterative Partial Least Squares (NIPALS) Modelling – Some Current Development, in: *Multivariate Analysis*, Krishnajah P.R. (Ed.), Academic Press, New York, 3, 383-407.