CLUSTERING AND STRUCTURAL EQUATION MODELLING

Mario Fordellone[1], Maurizio Vichi[1]

[1] Department of Statistical Sciences, Sapienza University of Rome
(e-mail: mario.fordellone@uniroma1.it, maurizio.vichi@uniroma1.it)

ABSTRACT: The identification of different homogeneous groups of observations and their appropriate analysis in PLS-SEM has become a critical issue in many application fields. Usually, PLS-SEM assumes the homogeneity of all the units on which the model is estimated, and the approaches of segmentation present in literature, consist in estimating separate models for segments of statistical units, which have been obtained by assigning the units to segments *a priori* defined. However, no causal structure among the variables is postulated and this has been considered a limitation. In this paper, a new methodology for simultaneous non-hierarchical clustering and PLS-SEM is proposed. A simulation study and an application on real data are included to evaluate the performance of the proposed methodology.

KEYWORDS: SEM, simultaneous clustering, K-Means, PLS.

1 Introduction

In the last years, Structural Equation Modelling (SEM) has become one of the reference statistical methodologies in the analysis of the statistical relationships between observable (manifest) and non-observable (latent) variables. Structural equation models are used for both to assess unobservable *hidden* constructs (i.e., latent variables) by means of observed variables, and to evaluate the relations between latent constructs. In SEM, variables (manifest or latent) are considered endogenous if they are dependent, i.e., related to a set of variables that explain or predict them. These last are the exogenous variables. SEM has the property to estimate the multiple and interrelated dependence in a single analysis by combining factor analysis and multivariate regression analysis. SEM has been used in many different fields, as in economics and social sciences, in marketing for example to assess customer satisfaction (Squillacciotti, 2010). SEM allows to build latent variables (LVs), such as customer satisfaction, through a network of manifest variables (MVs). As we have noted before, an important research issue in marketing is the measurement of customer satisfaction by using PLS-SEM and the identification of distinctive customer segments has been considered relevant.

Covariance Structure Approach (CSA) (Jöreskog, 1978) and Partial Least Squares (PLS) (Lohmöller, 1989) are the two alternative statistical techniques for estimating such models. The CSA, also referred to as LISREL, uses the ML estimation; thus, has the advantage to allow the researcher to make inference on the results. However, PLS

is considered preferable to CSA in three specific cases: *(i)* when the sample size is small, *(ii)* when the data to be analysed is not multi-normal as required by CSA, and *(iii)* when the complexity of the model to be estimated may lead to improper or non-convergent results (Squillacciotti, 2010).

In this paper, we work to the parsimonious consensus model that identifies the best clustering that best explains the manifest variables reconstructed by a unique common set of measurement/structural relationships. Thus, a new methodology for simultaneous non-hierarchical clustering and PLS-SEM is proposed and named Partial Least Squares *K*-Means (PLS-KM) (Fordellone and Vichi, 2017).

2 Model

Given the $n \times J$ data matrix $\mathbf{X}$, the $n \times K$ membership matrix $\mathbf{U}$, the $K \times J$ centroids matrix $\mathbf{C}$, the $J \times P$ loadings matrix $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_{\mathrm{H}}, \boldsymbol{\Lambda}_{\mathrm{L}}]$, and the errors matrices $\mathbf{Z}$, $\mathbf{E}$, and $\mathbf{D}$, the Partial Least Squares *K*-Means model can be written as follows (Fordellone and Vichi, 2017):

$$\begin{aligned} \mathbf{H} &= \mathbf{H}\mathbf{B}^{\mathrm{T}} + \boldsymbol{\Xi}\boldsymbol{\Gamma}^{\mathrm{T}} + \mathbf{Z}, \\ \mathbf{X} &= \boldsymbol{\Xi}\boldsymbol{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{E}, \\ \mathbf{X} &= \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} = \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}_{\mathrm{H}}\boldsymbol{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}_{\mathrm{L}}\boldsymbol{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{D}, \end{aligned} \tag{1}$$

under constraints: *(i)* $\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda} = \mathbf{I}$; and *(ii)* $\mathbf{U} \in \{0,1\}$, $\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$. Where *H* and *L* are the number of exogenous and endogenous LVs, respectively (i.e., *H*+*L*=*P*). Then, $\mathbf{H}$ is the $n \times L$ matrix of the endogenous LVs with generic element $\eta_{i,l}$, $\boldsymbol{\Xi}$ be the $n \times H$ matrix of the exogenous LVs with generic element $\xi_{i,h}$, $\mathbf{B}$ is the $L \times L$ matrix of the path coefficients $\beta_{l,l}$ associated to the endogenous latent variables, $\boldsymbol{\Gamma}$ is the $L \times H$ matrix of the path coefficients $\gamma_{l,h}$ associated to the exogenous latent variables, $\boldsymbol{\Lambda}_{\mathrm{H}}$ is the $J \times H$ loadings matrix of the exogenous latent constructs with generic element $\lambda_{j,h}$ and $\boldsymbol{\Lambda}_{\mathrm{L}}$ is the $J \times L$ loadings matrix of the endogenous latent constructs with generic element $\lambda_{j,l}$.

Thus, the PLS-KM model includes the PLS-SEM modeling and the clustering equations. The simultaneous estimation of the three sets of equations will produce the estimation of the pre-specified SEM describing relations among variables and the corresponding best partitioning of units.

When applying PLS-KM, the number of groups is unknown and the identification of an appropriate number of *K* clusters is not straightforward. Then, often you need to rely on some statistical criterion. In this paper we use the *gap method* for estimating the number of clusters, i.e., a *pseudo*-F designed to be applicable to virtually any clustering method.

In the preliminary step of the PLS-KM algorithm, the estimation of the PLS-SEM over the entire dataset is carried out; subsequently, the number of the *K* classes is obtained according to the maximum level of the *pseudo*-F function computed on the

estimated latent scores. Then, once chosen the number of clusters, the PLS-KM algorithm optimize the following overall objective function:

$$\underset{\mathbf{UC\Lambda}}{\operatorname{argmin}} \|\mathbf{X}\text{-}\mathbf{UC\Lambda\Lambda}^{\mathrm{T}}\|^{2} \tag{2}$$

Note that because the constraints on **U**, the method can be expected to be rather sensitive to local optima. For this reasons, it is recommended the use of some randomly started runs to find the best solution.

3 Simulation study

Data matrices formed by 100 statistical units and 9 MVs have been simulated for 1000 random generations. The 9 generated variables are split in three blocks related to 3 LVs according the path diagram shown in Figure 1.

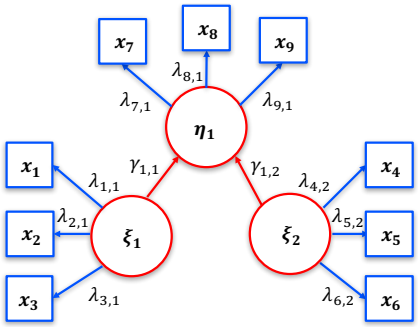

Figure 1 – Path diagram of the SEM model specified by the simulation scheme

Note that, in this simulated model $\mathbf{B} = 1$ (there is only one endogenous LV), whereas the other parameters have been fixed as follows:

$$\mathbf{\Lambda}_{\mathrm{H}}^{\mathrm{T}} = \begin{bmatrix} 0.60 & 0.40 & -0.80 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.85 & -0.85 & 0.50 & 0 & 0 & 0 \end{bmatrix};$$

$$\mathbf{\Lambda}_{\mathrm{L}}^{\mathrm{T}} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0.85 \quad -0.80 \quad -0.50];$$

$$\mathbf{\Gamma}^{\mathrm{T}} = \begin{bmatrix} 0.6 \\ -0.7 \end{bmatrix}.$$

The exogenous latent scores matrix $\mathbf{\Xi}$ has been generated by three different multivariate normal distributions (each with 2 uncorrelated dimensions) to obtain a structure of three groups of units:

$n_1 = 30;\ \boldsymbol{\mu}_1 = [\quad 1 \quad 0]^{\mathrm{T}};\ \boldsymbol{\Sigma}_1 = \mathbf{I}$
$n_2 = 30;\ \boldsymbol{\mu}_2 = [\ -10\ 10]^{\mathrm{T}};\ \boldsymbol{\Sigma}_2 = \mathbf{I}$
$n_3 = 40;\ \boldsymbol{\mu}_3 = [\ 100\ 10]^{\mathrm{T}};\ \boldsymbol{\Sigma}_3 = \mathbf{I}$.

The errors matrix $\mathbf{E}$ has been generated by a multivariate normal distribution (9 uncorrelated dimensions) with means equal to zero (i.e., noise) and standard deviation fixed as: $\sigma = 0.30$ (*low error*), $\sigma = 040$ (*medium error*), $\sigma = 0.50$ (*high error*). Then, we have simulated 1000 random generations of data for each level of error and the performance of the PLS-KM model has been evaluated for each case.

We have split up the analysis of simulation results in two steps. Firstly, we have compared the global quality of the model in the three different error levels using GoF, ARI and AGoF (Fordellone and Vichi, 2017). Secondly, we have analyzed each specific case where ARI is lower than 1 (i.e., when the model identifies a partition that is not the real one) to understand if this cases are local minima or overfitting (results in Tables 1 and 2, respectively).

Table 1 – Means of the indices distributions for the three different error levels

	Mean of the distribution of Communalities mean	Mean of the distribution of R^2 mean	Mean of Goodnes of Fit index	Mean of Adjusted Rand Index	Mean of Adjusted Goodnes of Fit index
σ = 0.30	0.846	0.780	**0.812**	**0.993**	**0.710**
σ = 0.40	0.737	0.614	**0.672**	**0.987**	**0.524**
σ = 0.50	0.705	0.556	**0.626**	**0.984**	**0.470**

Table 2 – Clustering performance of the PLS-KM for the three different error levels

	Found Optimal K (%)	Times model is true (%)	Local minima with 15 random start (%)	Overfitting (%)
σ = 0.30	100.00	99.90	0.00	0.10
σ = 0.40	100.00	88.00	7.40	4.60
σ = 0.50	100.00	72.60	17.20	10.20

4 Application on real data

In this section, an application of the Partial Least Squares *K*-Means model is presented. For this application the European Consumer Satisfaction Index (ECSI) has been used. In particular, we have analyzed the ECSI model in mobile phone industry (Tenenhaus et al., 2005).

4.1 Dataset

Dataset consists in 24 observed variables that represent the answers of 250 consumers of a mobile phone provider. The original items, scaled from 1 to 10, are

transformed into new normalized variables (scaled from 0 to 100). Figure 2 represents the complete ECSI model for the mobile phone industry.

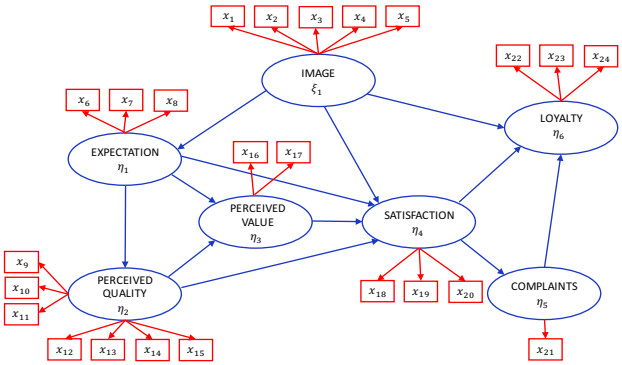

Figure 2 – ECSI model for the mobile phone industry

4.2 Results

In Tables 3 and 4 are shown measurement model and structural model results, respectively.

Table 3 – Loadings matrix Λ obtained by the PLS-KM model

	Image	Expectations	PercQuality	PercValue	Satisfaction	Complaints	Loyalty
X1	0.449	0	0	0	0	0	0
X2	0.398	0	0	0	0	0	0
X3	0.355	0	0	0	0	0	0
X4	0.528	0	0	0	0	0	0
X5	0.486	0	0	0	0	0	0
X6	0	0.615	0	0	0	0	0
X7	0	0.607	0	0	0	0	0
X8	0	0.503	0	0	0	0	0
X9	0	0	0.419	0	0	0	0
X10	0	0	0.284	0	0	0	0
X11	0	0	0.399	0	0	0	0
X12	0	0	0.377	0	0	0	0
X13	0	0	0.375	0	0	0	0
X14	0	0	0.381	0	0	0	0
X15	0	0	0.397	0	0	0	0
X16	0	0	0	0.624	0	0	0
X17	0	0	0	0.781	0	0	0
X18	0	0	0	0	0.558	0	0
X19	0	0	0	0	0.563	0	0
X20	0	0	0	0	0.609	0	0
X21	0	0	0	0	0	1.000	0
X22	0	0	0	0	0	0	0.585
X23	0	0	0	0	0	0	0.099
X24	0	0	0	0	0	0	0.805

Table 4 – Path coefficients matrix Ω obtained by the PLS-KM model

	Image	Expectations	PercQuality	PercValue	Satisfaction	Complaints	Loyalty
Image	0	0.507	0	0	0.177	0	0.201
Expectations	0	0	0.554	0.048	0.071	0	0
PercQuality	0	0	0	0.557	0.509	0	0
PercValue	0	0	0	0	0.191	0	0
Satisfaction	0	0	0	0	0	0.523	0.479
Complaints	0	0	0	0	0	0	0.067
Loyalty	0	0	0	0	0	0	0

Finally, in Table 5 are shown the summary statistics of the three obtained.

Table 5 – Summary statistics of the three groups of mobile phone customers

Group 1

	Image	Expectations	PercQuality	PercValue	Satisfaction	Complaints	Loyalty
Min	0.460	0.180	0.660	0.000	0.537	0.000	0.019
Q1	0.722	0.652	0.775	0.688	0.710	0.778	0.824
Median	0.802	0.773	0.837	0.778	0.787	0.889	0.898
Mean	0.796	0.752	0.840	0.763	0.794	0.832	0.862
Q3	0.861	0.849	0.905	0.878	0.875	1.000	0.956
Max	1.000	1.000	1.000	1.000	1.000	1.000	1.000

N = 92

Group 2

	Image	Expectations	PercQuality	PercValue	Satisfaction	Complaints	Loyalty
Min	0.225	0.145	0.483	0.000	0.273	0.000	0.190
Q1	0.541	0.481	0.594	0.511	0.526	0.556	0.594
Median	0.600	0.584	0.648	0.622	0.599	0.667	0.698
Mean	0.607	0.584	0.643	0.591	0.589	0.638	0.696
Q3	0.681	0.664	0.687	0.667	0.647	0.778	0.804
Max	0.845	1.000	0.831	0.889	1.000	1.000	1.000

N = 112

Group 3

	Image	Expectations	PercQuality	PercValue	Satisfaction	Complaints	Loyalty
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Q1	0.306	0.359	0.284	0.333	0.272	0.333	0.263
Median	0.440	0.497	0.414	0.444	0.353	0.444	0.467
Mean	0.392	0.471	0.398	0.423	0.345	0.447	0.460
Q3	0.494	0.599	0.486	0.556	0.445	0.667	0.626
Max	0.676	0.820	0.704	1.000	0.691	1.000	1.000

N = 46

5 Concluding remarks

In wide range of applications for empirical data analysis, the assumption that data are collected from a single homogeneous population is often unrealistic. In particular, the identification of different groups of observations and their appropriate consideration in PLS-SEM constitutes a critical issue in many fields.

The traditional approach to segmentation in SEM consists in estimating separate models for objects segments which have been obtained either by assigning observations to *a priori* segments. Then, each class has different component scores, structural coefficients, outer weights and loadings.

The PLS-KM approach, instead provides a single SEM guarantying the best partition of objects represented by the best causal relationship in the reduced latent space.

References

Fordellone M. & Vichi M. 2017. Partial Least Squares path modelling and simultaneous clustering.

Jöreskog, K. G. 1978. Structural analysis of covariance and correlation matrices. *Psychometrika*, **43**(4), 443-477.

Lohmoeller, J. B. 1989. Latent variable path analysis with partial least squares. Heidelberg: Physica.

Squillacciotti, S. 2010. Prediction oriented classification in PLS path modelling. *Handbook of Partial Least Squares*, 219-233.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. 2005. PLS path modelling. *Computational statistics & data analysis*, **48**(1), 159-205.