Conjoint use of variables clustering and PLS structural equations modelling

Valentina Stan [1] and Gilbert Saporta [1]

[1] Conservatoire National des Arts et Métiers, 292 Rue Saint Martin, F 75141 Paris Cedex 03, France

Summary

In PLS approach, it is frequently assumed that the blocks of variables satisfy the assumption of unidimensionality. In order to fulfill at best this hypothesis, we use clustering methods of variables. We illustrate the conjoint use of variables clustering and PLS structural equations modelling on data provided by PSA Company (Peugeot Citroën) on customers' satisfaction. The data are satisfaction scores on 32 manifest variables given by 2922 customers.

Keywords: PLS approach, variables clustering

1 Clustering of variables

There are two main methodologies: hierarchical methods and direct partitioning methods. Hierarchical methods are either agglomerative or divisive. Partitioning methods usually require that the number of groups should be defined beforehand and will not be used here. A good partition is such that the variables of the same

class are correlated as much as possible, and two variables belonging to different classes are as uncorrelated as possible. One may distinguish two cases, depending on whether the sign of the correlation coefficient is important or not (i.e. if negative values show a disagreement between variables).

1.1. Agglomerative hierarchical clustering methods

1.1.1 Methods derived from clustering of statistical units(Nakache & Confais; 2005)

Various dissimilarity measures can be used, based on the usual correlation coefficient like:

$1-r_{jj'}$; $1-|r_{jj'}|$ if the sign of the correlation is not important; $s_{jj'}=\cos^{-1}(r_{jj'})$.

Then we use the following strategies of aggregation: single linkage, average linkage, complete linkage, Ward's criteria etc.

1.1.2 The VARHCA method (Vigneau & Qannari, 2003)

Let $C_1, C_2 \ldots C_k$ be k blocks (or clusters) of manifest variables and $Y_1, Y_2 \ldots Y_k$ the standardised latent variables (first principal component) associated respectively with each cluster. Manifest variables are centred, but not necessarily standardized. The following hierarchical procedure aims at locally optimising the criterion T defined by: $T = n\sum_{r=1}^{k}\sum_{j=1}^{p}\delta_{rj}\,\mathrm{cov}^2\left(x_j, Y_r\right)$ *where* $\delta_{rj} = \begin{cases} 1 \text{ if } x_j \in C_r \\ 0 \text{ otherwise} \end{cases}$

- At the first level of the hierarchy, each variable forms a cluster by itself; then, $T_0 = \sum_{j=1}^{p}\mathrm{var}\left(x_j\right)$;
- At level i, one merges the two clusters giving the minimal variation of T: $\Delta T = T_{i-1} - T_i = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A\cup B)}$ where $\lambda_1^{(A)}$, $\lambda_1^{(B)}$, $\lambda_1^{(A\cup B)}$ are the largest eigenvalues of the covariance matrices of the variables in clusters A, B and A∪B.

1.2. Cutting trees

The resulting tree should be cut at a suitable level to get a partition. We use here a criterion of unidimensionality of the groups to obtain this cut. Starting from the root of the tree, we first realize a cut in 2 classes and verify the hypothesis of unidimensionality by using the Cronbach's α or the Dillon-Goldstein's ρ. If these values are close to 1, then the hypothesis of unidimensionality is accepted. Otherwise, we proceed to a cut at the following level of the tree, and so on. We repeat the procedure until we obtain classes satisfying the unidimensionality criteria.

1.3. Divisive methods

SAS VARCLUS procedure is one of the best known. At first step one performs a PCA with all manifest variables. If there is only one principal component with an eigenvalue greater than 1, there is only one cluster.
Otherwise one considers the first two principal components: each manifest variable is associated with the principal component to which it is the closest, in regard to the squared linear correlation coefficient, thus forming two groups of variables. If the second eigenvalue of a group is greater than 1, this group is divided in turn, according to the same method, and so on, until each group has only one principal component.

2 Application to Structural Equation Modeling

Let *p* variables be observed upon *n* units. The *p* variables are partitioned in J subsets or blocks of k_j variables which are presumed to be pertinent for describing the phenomenon. Each of these blocks is designed to describe a theme of the general phenomenon. We shall designate these blocks by X_j and we shall consider them as matrices with dimension (n x k_j) (Tenenhaus § al, 2005). In the following, we shall always suppose that each block is associated with only one latent variable (unidimensionality). Therefore we can identify the blocks by the same name as their latent variable. The latent variable corresponding to the X_j block will be designated by ξ_j.

In order to obtain unidimensional blocks, we propose to use same of the clustering methods, previously presented in §. 1. In the following, we study the specific case where there are no pre-defined causal relationships between the latent variables. We use the blocks obtained by each method to build the causality scheme; at each block one associates a single latent variable.
With the help of experts we propose relationships between latent variables with the aim of explaining the general satisfaction of the customers, and we therefore establish the inner model. To choose the best model from many, we use the global quality criterion developed by Amato et al. (2004):

$$GoF = \sqrt{\overline{communality} \times \overline{R^2}}$$

where $\overline{communality}$ is the average of the communality of each block and measures the quality of the external model. $\overline{R^2}$ is the average of R^2 for each endogenous latent variable. The R^2 measures the quality of the inner model and is calculated for each endogenous variable according to latent variables which explain it.
The software used is the experimental PLSX module of SPAD.

3 Practical application

3.1. The questionnaire

The data obtained are satisfaction scores scaled between 1 and 10 on 32 services for a car. 2922 customers participated. Manifest variables are the followings:

Variable		Variable	
General satisfaction	S01	Radio - CD - rom	S17
General quality	S02	Heating - ventilation	S18
Quality - price ratio	S03	Boot capacity	S19
Absence of small, irritating defects	S04	Security	S20
Absence of noise and vibrations	S05	Braking	S21
General state of the paintwork	S06	Acceleration	S22
Robustness of commands, buttons	S33	Handling	S23
Solidity and robustness	S08	Suspension comfort	S24
Lock, door and window mechanisms	S09	Silence in rolling	S25
Inside space and seat modularity	S34	Maniability	S26
Inside habitability	S11	Direction	S27
Dashboard: quality of materials and finishing	S12	Gears	S28
Insider: quality of mat. and finishing	S13	Mechanic reliability	S29
Front seat comfort	S14	Oil consumption	S30
Driving position	S15	Mechanic's efficiency in solving problems	S31
Visibility from driver's seat	S16	Maintenance cost and repairs	S32

Table 1. Manifest variables

3.2. Clustering variables

We have used $1 - r_{jj'}$ as distance. We have applied 6 clustering methods of variables: single linkage, average linkage, complete linkage, Ward's criterion, VARCLUS and VARHCA. Single linkage and average linkage did not provide well separated clusters, so they are eliminated.

For Ward's criterion, the tree shows that a partition in 8 classes is reasonable and for complete linkage in 6 classes. The partition obtained by cutting VARHCA tree into 7 clusters is here exactly the same as the partition given by VARCLUS. The Cronbach's α coefficients show that the obtained blocks are unidimensional.

In the following, we present the blocks for complete linkage, Ward's criterion VARCLUS and VARHCA:

Ward's criterion						Complete linkage						VARCLUS and VARHCA					
Block	**MV**	**Block**	**MV**	**Block**	**MV**	**Block**	**MV**	**Block**	**MV**	**Block**	**MV**	**Block**	**MV**	**Block**	**MV**	**Block**	**MV**
General satisfaction (Gs)	S01	Solidity (Sd)	S06	Driving quality (Dq)	S20	General satisfaction (Gs)	S01	Comfort (Cf)	S11	Driving quality (Dq)	S20	General satisfaction (Gs)	S01	Solidity (Sd)	S06	Driving quality (Dq)	S20
	S02		S08		S21		S02		S34		S21		S02		S08		S21
	S03		S09		S22		S04		S12		S22		S04		S09		S22
Construct quality (Cq)	S04		S33		S23		S05		S13		S23		S05		S33		S23
	S05	Driving comfort (Dc)	S12		S24	Quality – price ratio (Qpr)	S03		S14		S24	Quality – price (Qp)	S03	Driving comfort (Dc)	S12		S24
Maintenance (Mn)	S31		S13		S25	Maintenance (Mn)	S31		S15		S26		S31		S13		S25
	S32		S14		S26		S32		S16		S27		S32		S14		S26
Interior design (Id)	S11		S15		S27	Solidity (Sd)	S06		S19		S25	Interior design (Id)	S11		S15		S27
	S34		S16		S28		S08		S17		S28		S19		S16		S28
	S19	Interior comfort (Ic)	S17		S29		S09		S18		S29		S34	Interior comfort (Ic)	S17		S29
			S18		S30		S33				S30				S18		S30

Table 2. Blocks of manifest variables after Ward's criterion, complete linkage, VARCLUS and VARHCA

In table 2 we can observed that the blocks "Solidity" and "Driving quality" are identical for all methods. General satisfaction has the same composition for complete linkage, VARCLUS and VARHCA, but partition issued from Ward's criterion is more logical, according to experts. By comparison with the other methods, complete linkage groups in a single block the variables which form the blocks "Interior design", "Driving comfort", "Interior comfort" in Ward's criterion, VARCLUS and VARHCA. For VARCLUS and VARHCA, the variables which are associated to the block "Maintenance" in Ward's criterion and complete linkage, are in the same block with "Quality–price ratio". Complete linkage is the only method which realizes a distinct block for the variable "Quality-price ratio".

3.3. PLS structural models

The clustering techniques provide blocks but not the relationships between them. With the help of experts we then propose relations between blocks, so as to explain the latent variable "General satisfaction". The following figures give the 3 causality schemes:

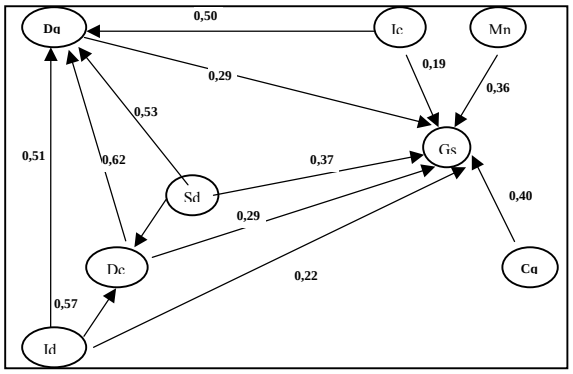

Fig. 1. Causality scheme after Ward's clustering

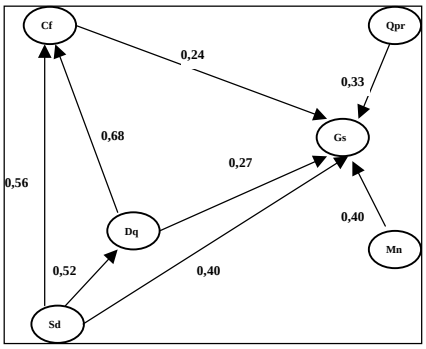

Fig. 2. Causality scheme after complete linkage clustering

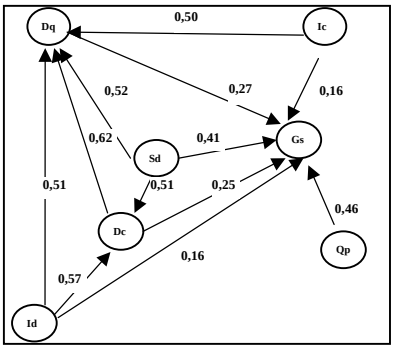

Fig. 3. Causality scheme after VARCLUS or VARCHA clustering

The values of Amato's criterion ($G \circ F$) are:
- for Ward's criterion: $G \circ F = 0{,}48$;
- for complete linkage: $G \circ F = 0{,}42$;
- for VARCLUS: $G \circ F = 0{,}47$.

Ward's clustering gives the best result and will be selected.

3.4. Results and interpretations

The measurement model

After convergence of the PLS algorithm, one obtains the final weights which allow us to link the manifest variables with the latent variables. An example for general satisfaction: $Gsat = 0,22S01 + 0,57S02 + 0,48S03$.

Analyzing the correlations, we observe that all latent variables are well correlated with their own manifest. So, the manifest variables "describe" their latent appropriately and the blocks are therefore validated.
The R^2 coefficients between connected latent variables are:

$$R^2\left(Driving\ comfort; Sd, Id\right) = 0,42$$

$$R^2\left(Driving\ quality; Sd, Id, Dc, Ic\right) = 0,5$$

$$R^2\left(General\ satisfaction; Cq, Mn, Sd, Id, Dc, Ic, Dq\right) = 0,27$$

For "General satisfaction", the R^2 coefficient generated by the other latent variables is 27%, and we consider that as satisfactory because there are 2922 individuals. Analyzing the correlations between the latent variables (fig.1), we can see that to improve "Driving quality", the producer should concentrate on "Driving comfort" (correlation coefficient = 0,62), on the "Solidity" (0,53) and on the "Interior design" (0,51). In order to obtain a good "Driving comfort", the producer could concentrate on "Interior design" (0,57) and on "Solidity" (0,51).

Given the causality scheme, the determination of "general satisfaction" is a complex procedure in which almost all the latent variables are directly involved.
"Construction quality" is the most important variable for the "general satisfaction" (correlation coefficient = 0,40) and the less important is the "Interior comfort" (0,19). Consequently, in order to increase the general satisfaction, the producer should concentrate first on the "construction quality" and then on the "Solidity", "Maintenance", "Driving quality", "Driving comfort" "Interior design" and "Interior comfort".

The equation is as follows:

$$Gs = 0,26Cq + 0,19Mn + 0,15Sd + 0,03Id + 0,10Dc - 0,03Ic + 0,04Dq$$

Conclusions

It must be underlined that this study did not follow the logical sequence of steps of the PLS approach: the construction of a model by experts, the construction of a questionnaire using this model, and the collection of customer data using this questionnaire.
In our case, the process is inverted: we have tried to build a model using data that had already been collected. This fact has obviously effects on the final results which cannot be measured.

With the help of clustering methods of variables we established the external model. According to Amato's criterion, Ward's clustering was choosen as the best technique. But we observe that the values of this criterion for the 3 models are very close.
For the chosen model, a hierarchy of the influence of the latent variables on general satisfaction can be established using the structural model:
I. Construction quality; II. Solidity; III. Maintenance; IV. Driving quality, V. Driving comfort,VI. Interior design, VII. Interior comfort. The results obtained are satisfactory: $R^2 = 27\%$ for a large sample of almost 3000 respondents.

References

Amato, S., Esposito Vinzi, V., Tenenhaus, M. (2004), ‚A global goodness-of-fit index for PLS structural equation modeling', *Oral Communication to PLS Club, HEC School of Management, France, March 24*

Nakache, J.P., Confais J. (2005), ‚Approche pragmatique de la classification', *Ed. Technip, Paris*

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y–M., Lauro, C., (2005), PLS path modeling, *Computational Statistics and Data Analysis,* volume 48, issue 1, 159-205

Vigneau E., Qannari E.M. (2003), ‚Clustering of variables around latent component - application to sensory analysis', *Communications in Statistics, Simulation and Computation* **32**(4), 1131-1150