Evaluating response based segmentation in PLS path modeling

Jacob Eskildsen[1], Rick Edgeman[1], Dan Mønster[2] and John Vestergaard Olesen[1]

[1]ICOA, AU Herning, Aarhus University,
[2]ICOA, Department of Economics, Aarhus University
Fuglesangs Allé 4, 8210 Aarhus V, Denmark
eskildsen@asb.dk

Abstract

Several techniques have emerged for identifying segments in Partial Least Squares. Among these are FIMIX-PLS, PLS-TPM and REBUS-PLS. This study focuses on the latest development, REBUS-PLS, and the performance of this algorithm conditioned on the following five factors: number of segments, root mean squared difference between local models, indicator reliability, structural model specification error and the number of manifest indicators in each block. The study shows that overall the REBUS-PLS algorithm is able to find the solution when the reliability of the indicators is high and the structural model specification error is low. These two factors are thus the ones that truly define the quality of a REBUS-PLS solution, although the other factors also play a role.

Keywords: Market research methodology, simulation study, segmentation, REBUS-PLS

1. Introduction

There are several techniques for identifying segments in Partial Least Squares (PLS) path modeling and the most prevalent ones are Finite Mixture Partial Least Squares (FIMIX-PLS) (Hahn et al., 2002); Partial Least Squares Typological Path Modeling (PLS-TPM) (Squillacciotti, 2007) and REsponse Based Units Segmetation (REBUS-PLS) (Vinzi et al., 2008).

Previous research has shown that both FIMIX-PLS and PLS-TPM are able to detect underlying segments in data and that analyzing data on the aggregate level in the presence of segments can produce ambiguous results (Vinzi et al., 2007). Furthermore this study concluded that *"the PLS-TPM methodology better fits the statistical characteristics of PLS path modeling and has more potentials for extensions to improve model based segmentation"* (Vinzi et al., 2007).

This should bode well for REBUS-PLS since this can be considered as a further development of PLS-TPM that attempts to overcome some of the shortcomings of the latter technique (Vinzi et al., 2007). So far there has however not been enough research done on the performance of the REBUS-PLS algorithm under different data conditions.

The aim of this paper is therefore to test the performance of the REBUS-PLS algorithm with respect to the following five factors:
- Number of segments
- Root mean squared difference between local models
- Indicator reliability
- Structural model specification error
- Number of manifest indicators in each block

2. Background & methodology

The Partial Least Squares model is formulated through three different parts: the inner relations, the outer relations, and the weight relations (Fornell & Cha, 1994, Wold, 1980). The inner relations outline the relationship between the latent variables as shown in (1).

$$(1) \qquad \eta = B\eta + \Gamma\xi + \zeta$$

In (1) η is a vector of the latent endogenous variables and B the corresponding coefficient matrix (Fornell & Cha, 1994, Wold, 1980). ξ is a vector of the latent exogenous variables, Γ the corresponding coefficient matrix and finally an error term, ζ, is included. The second part of the model formulation is the outer relations (Fornell & Cha, 1994, Wold, 1980). This part of the model defines the relationship between the latent variables and the corresponding manifest variables and in contrast to covariance-based methods (such as LISREL) these can both be reflective and formative by nature (Jöreskog & Wold, 1982). Since the analysis performed here is based on reflective outer relations only this situation is mentioned in the following. The general formula for reflective outer relations is shown in (2).

$$(2) \qquad \begin{aligned} y &= \Lambda_y \eta + \varepsilon_y \\ x &= \Lambda_x \xi + \varepsilon_x \end{aligned}$$

In (2) y is a vector of the observed indicators of η and x is a vector of the observed indicators of ξ. Λ_y and Λ_x are matrices that contain the λ_i coefficients which link the latent and the corresponding manifest variables together and ε_x and ε_y are the error of measurement for x and y, respectively (Fornell & Cha, 1994, Wold, 1980).

The weight relations are the final part of the PLS model formulation. In PLS each case value of the latent variables can be estimated through the weight relations shown in (3) as linear aggregates of their empirical indicators. In this case the weights are equal to rescaled loadings from principal components. This is another unique feature of PLS compared to covariance-based methods.

$$(3) \qquad \begin{aligned} \hat{\eta} &= \omega_\eta y \\ \hat{\xi} &= \omega_\xi x \end{aligned}$$

Recently an approach has been developed that allows for response-based segmentation in PLS (Vinzi et al., 2008). The approach has been labeled REBUS-PLS and through the application of this approach one may be able to identify local models that have different structural models and measurement models without making any distributional assumptions. The algorithm of REBUS-PLS is outlined below (Vinzi et al., 2008).

Step 1:   Estimation of global PLS model
Step 2:   Computation of communalities and structural residuals from the global model for all cases
Step 3:   Hierarchical clustering based on the communalities and structural residuals computed in step 2
Step 4:   Choice of number of segments (S) based on the dendogram obtained from step 3
Step 5:   Assignment of cases to each segment according to the cluster analysis
Step 6:   Estimation of the S local models (one for each segment)

Step 7:   Computation of a closeness measure CM for each case with respect to each local model

Step 8:   Assignment of each case to the closest local model

If stability of segment membership THEN step 9 ELSE back to step 6

Step 9:   Description of the obtained segments according to differences among the local models

Previous research has shown that the PLS algorithm is quite robust against anomalies such as (Kristensen & Eskildsen, 2005b, Kristensen & Eskildsen, 2010, Cassel et al., 1999, Westlund et al., 2008):

- Exogenous distribution
- Multicollinearity between latent exogenous variables
- Indicator reliability
- Structural model specification error
- Sample size
- Number of manifest indicators in each block
- Missing values

There is however little evidence as to how the REBUS-PLS algorithm will fare if subjected to the same anomalies. To examine parts of this is the purpose of the following sections.

3. Simulation study

In order to assess the performance of the REBUS-PLS algorithm a simulation study has been constructed. The model applied in the simulation study can be seen in figure 1.
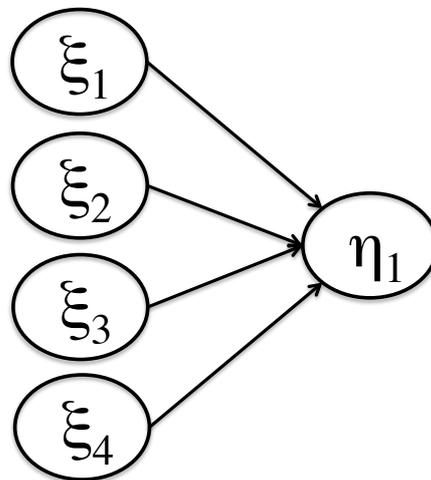

Figure 1: Simulation model

In the study the following factors were included:

- Number of segments [2; 4]
- Root mean squared difference between local models [0.15; 0.25]
- Indicator reliability [1; 5; 10]
- Structural model specification error [1; 5; 10]
- Number of manifest indicators in each block [2; 4; 6]

The sample sizes for the segments were [1000; 500] in the case of 2 segments and [1000; 750; 500; 250] in the case of 4 segments. In bot cases the relationship between the exogenous latent variables has been kept fixed across all segments.

The simulation study is of resolution $2^2 \times 3^3$ and has been carried out as a full factorial with 100 replications of each of the 108 runs. The main dependent variable is the proportion of misclassified cases.

4. Results

In figure 2 below the main effects for indicator reliability and structural model specification error is shown.
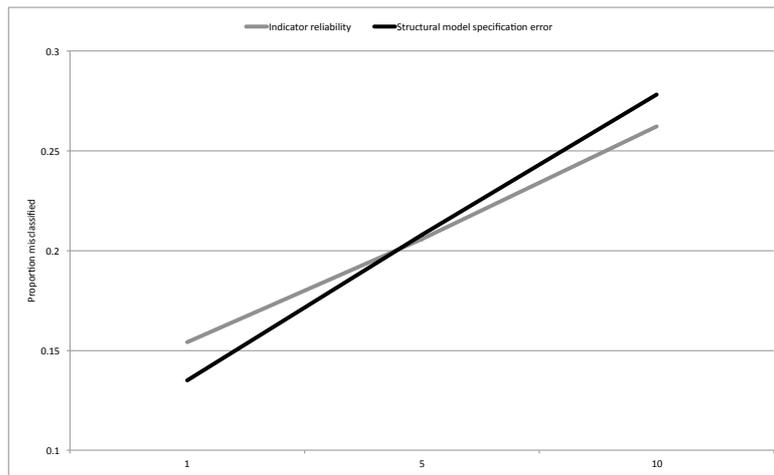

Figure 2: Indicator reliability and structural model specification error

From this figure it is evident that both of these conditions can have dire consequences for the outcome of a REBUS-PLS analysis. If both indicator reliability and is low and the structural model specification error is high then the REBUS-PLS algorithm will be unable to find anything meaningful in the data.

Figure 3 shows the main effects for number of segments, number of indicators and the local model differences.
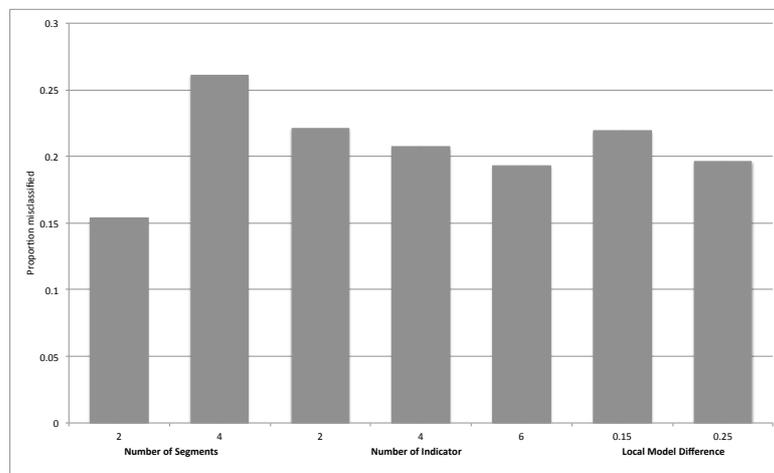

Figure 3: Number of segments, number of indicators and the local model differences.

Figure 3 first of all shows that the REBUS-PLS algorithm has difficulties when the number of true segments increases. Thus one should probably favor solutions with as few segments as possible.

The consistency at large properties of the PLS algorithm (Wold, 1980) also plays a role when it comes to segmentation. REBUS-PLS works better with more indicators and the benefit of increasing the number of indicators does not seem to fade of after 4 indicators as is the case of the PLS algorithm in general (Kristensen &   Eskildsen, 2010).

Finally the differences between the local models also play a role in the ability of REBUS-PLS to find a solution. The more different the models are the better. This can also be used to assess the quality of the final solution of REBUS-PLS although this phenomenon requires more scrutiny before actual guidance for practitioners can be provided.

Overall the REBUS-PLS algorithm is able to find the solution when the reliability of the indicators is high and the structural model specification error is low. These are the properties that truly define the quality of a REBUS-PLS solution.

5. Conclusions

This paper has examined the ability of the REsponse Based Units Segmetation algorithm (REBUS-PLS) (Vinzi et al., 2008) to reproduce predefined segment through the application of a simulation study.

In the study the following factors were included:

- Number of segments [2; 4]
- Root mean squared difference between local models [0.15; 0.25]
- Indicator reliability [1; 5; 10]
- Structural model specification error [1; 5; 10]

The study shows that the reliability of the indicators and the structural model specification error are the properties that truly define the quality of a REBUS-PLS solution although the other factors also play a role.

This study is the first step towards establishing guidelines for practitioners when it comes to evaluating the outcome of a PLS segmentation analysis. More research is needed on the role of a more complicated structure among the exogenous latent variables (kept fixed across segments in this study) as well as the role of the size of the local model differences.

Finally the role of missing values needs to be investigated. Although they have been shown to play a marginal role if handled correctly in the general PLS algorithm (Kristensen &   Eskildsen, 2005a, Kristensen &   Eskildsen, 2010) the role of missing values in REBUS-PLS needs to be examined in greater detail.

References

Cassel, C., Hackl, P. & Westlund, A.H. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics,* **26,** 435-446.

Fornell, C. & Cha, J. (1994). Partial Least Squares. in *Advanced Methods of*

Marketing Research Ed. R.P. Bagozzi. Cambridge: Blackwell.

Hahn, C., Johnson, M.D., Herrmann, A. & Huber, F. (2002). Capturing customer heterogeneity using a finite mixture PLS approach. *Schmalenbach Business Review,* **54,** 243-269.

Jöreskog, K.G. & Wold, H. (1982). The ML and PLS techniques for modeling with latent variables. in *Systems under undirect observations* Ed. H. Wold. New York: North-Holland.

Kristensen, K. & Eskildsen, J.K. (2005a). Missing values, Partial Least Squares, and the estimation of customer satisfaction. in *PLS and Related Methods* Eds. T. Aluja, J. Casanovas, V.E. Vinzi, A. Morineau & M. Tenenhaus. Barcelona: SPAD.

Kristensen, K. & Eskildsen, J.K. (2005b). PLS Structural Equation Modelling for Customer Satisfaction Measurement: Some empirical and theoretical results. in *Handbuch PLS-Pfadmodellierung - Methode, Anwendung, Praxisbeispiele* Eds. F.W. Bliemel, A. Eggert, G. Fassott & J. Henseler. Schäffer-Poeschel.

Kristensen, K. & Eskildsen, J.K. (2010). Design of PLS-based Satisfaction Studies. in *Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields, Springer Handbooks of Computational Statistics* Eds. V.E. Vinzi, W.W. Chin, J. Henseler & H. Wang. Springer Verlag.

Squillacciotti, S. (2007). Prediction-oriented classification in PLS path modeling. in *Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields, Springer Handbooks of Computational Statistics* Eds. V.E. Vinzi, W.W. Chin, J. Henseler & H. Wang. Springer Verlag.

Vinzi, V.E., Ringle, C., Squillacciotti, S. & Trinchera, L. (2007). Capturing and treating unobserved heterogeneityb by response based segmentation in pls path modeling. A comparison of alternative methods by computational experiments. ESSEC RESEARCH CENTER WORKING PAPER DR07019.

Vinzi, V.E., Trinchera, L., Squillacciotti, S. & Tenenhaus, M. (2008). REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling. *Applied Stochastic Models in Business and Industry,* **24,** 439-458.

Westlund, A.H., Kâllstrôm, M. & Parmler, J. (2008). SEM-based customer satisfaction measurement: On multicollinearity and robust PLS estimation. *Total Quality Management & Business Excellence,* **19,** 855-869.

Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce: Theory and application of Partial Least Squares. in *Evaluation of Econometric Models* Ed. J.B. Ramsey. New York: Academic Press.