Assessment of Latent Class Detection in PLS Path Modeling: a Simulation Study to Evaluate the Group Quality Index performance

Laura Trinchera

Assessment of Latent Class Detection in PLS Path Modeling: a Simulation Study to Evaluate the Group Quality Index performance

Laura Trinchera

Abstract Structural Equation Models assume homogeneity across the entire sample. In other words, all the units are supposed to be well represented by a unique model. Not taking into account heterogeneity among units may lead to biased results in terms of model parameters. That is why, nowadays, more attention is focused on techniques able to detect unobserved heterogeneity in Structural Equation Models. However, once unit partition obtained according to the chosen clustering methods, it is important to state if taking into account local models provides better results than using a single model for the whole sample. Here, a new index to assess detected unit partition will be presented: the Group Quality Index. A simulation study involving two different simulation schemes (one simulating the so called null hypothesis of homogeneity among units, and the other taking into account the heterogenous sample case) will be presented.

1 Introduction

Heterogeneity among units is an important issue in statistical analysis. Treating the sample as homogeneous, when it is not, may seriously affect the results [6]. In Structural Equation Models (SEM) [2, 7] all the units are most often supposed to be well described by a unique model. Nevertheless, this hypothesis may often turn to be false. Recently, several techniques able to provide clustering in PLS Path Modeling (PLS-PM) [8, 10] have been presented [5, 6, 9]. However, no matter which method is used to cluster units, once the latent groups are identified, it is important to assess the differences between the detected classes of units and to evaluate the quality of the obtained partition. The first point essentially entails comparing the obtained local models to one another as well as with the global model. In PLS-PM framework only non parametric procedures and resampling methods, such as a bootstrap based

L. Trinchera (✉)
Department of Signal Processing & Electronic Systems, SUPELEC, Gif-sur-Yvette, France
e-mail: laura.trinchera@supelec.fr

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_29,

technique, are available. As regards the second point, i.e. assess the quality of the obtained partition, no specific index or methods have been developed until now. Here we meet this need by presenting a new index to evaluate the quality of the obtained partition: the Group Quality Index (*GQI*).

The remainder of the paper it is organized as follows: first we introduce the *GQI* (cf. 2), then a simulation study to asses the *GQI* properties is presented (cf. 3), to conclude a discussion on the obtained results and of the directions of further research is provided (cf. 4).

2 A New Index to Assess Group Separation in PLS-PM: The Group Quality Index

Assessing the quality of a PLS-PM is a difficult task. It is well known, that PLS-PM is a completely distribution free approach [10]. Thus, standard fit index and inferential process are not yet valid. Moreover, PLS-PM does not seem to optimize a well established global scalar function. Hence, no comparable global goodness of fit criteria are available. Furthermore, it is a variance-based model strongly oriented to prediction. Thus, model validation focuses on the model predictive capability. Following this idea, Amato et al. [1] recently proposed the Goodness of Fit (*GoF*) index. This remains the only available measure to evaluate the global model fitting in a PLS-PM model. Such index has been developed in order to take into account the model performance in both the measurement and the structural model, that is why two different parts compose the index:

$$GoF = \sqrt{\frac{\sum_{q:P_q>1}\sum_{p=1}^{P_q} Cor^2\left(x_{pq}, \hat{\xi}_q\right)}{\sum_{q:P_q>1} P_q} \times \frac{\sum_{j=1}^{J} R^2\left(\hat{\xi}_j, \hat{\xi}_{q:\xi_q\rightarrow\xi_j}\right)}{J}} \qquad (1)$$

where P_q is the number of manifest variables in the q-th block, x_{pq} is the generic manifest variable in the q-th block, $\hat{\xi}_q$ is the generic latent variable score, J is the number of endogenous latent variables in the model and $\hat{\xi}_j$ is the generic endogenous latent variable score.

By looking at Eq. in (1) it is possible to notice that both terms of the product under the square root can be seen as portions of explained variances. As it is well known the R^2 index in a simple regression is an indicator of how well the model fits the data. In fact, the smaller the variability of the residual values around the regression line relative to the overall variability is, the better the prediction obtained by the model is. The residuals play a central role in stating the quality of a model. Following this idea it is possible to rewrite the *GoF* index using residuals as:

$$GoF = \sqrt{\frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \sum_{p=1}^{P_q} \left(1 - \frac{\sum_{i=1}^{N} e_{ipq}^2}{\sum_{i=1}^{N} \left(x_{ipq} - \overline{x}_{pq}\right)^2}\right)} \times \sqrt{\frac{1}{J} \sum_{j=1}^{J} \left(1 - \frac{\sum_{i=1}^{N} f_{ij}^2}{\sum_{i=1}^{N} \left(\hat{\xi}_{ij} - \overline{\hat{\xi}}_j\right)^2}\right)} \tag{2}$$

where e_{ipq} is the measurement model residual for the i-th unit, corresponding to the p-th manifest variable in the q-th block, i.e. the communality residual, and f_{ij} is the structural model residual for the i-th unit, corresponding to the j-th endogenous block. These two kinds of residuals are the same as used in REBUS-PLS algorithm. For further information about how computing these residuals please refers to Trinchera [9] and Esposito Vinzi et al. [5]. In particular, the communality residuals are the residuals of the simple regressions of each manifest variable on the corresponding latent variable, while the structural residuals are the residuals of the OLS simple and multiple regressions of the endogenous latent variables on their exogenous latent variables.

If more than one class is taken into account, i.e. if the N units are split into K classes each one of size n_k, the GoF index as expressed in Eq. (2) can be reformulated leading to the GQI. Therefore, in the case of K classes the GQI can be expressed as:

$$GQI = \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \sum_{p=1}^{P_q} \left(1 - \frac{\sum_{i=1}^{n_k} e_{ipqk}^2}{\sum_{i=1}^{n_k} \left(x_{ipq} - \overline{x}_{pqk}\right)^2}\right)\right]} \times \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{J} \sum_{j=1}^{J} \left(1 - \frac{\sum_{i=1}^{n_k} f_{ijk}^2}{\sum_{i=1}^{n_k} \left(\hat{\xi}_{ijk} - \overline{\hat{\xi}}_{jk}\right)^2}\right)\right]} \tag{3}$$

This index is equal to the GoF in the case of a unique class, i.e. when $K = 1$ and $n_1 = N$. In other words, the GQI computed for the whole sample as a unique class is equal to the GoF index computed for the global model.

If local models performing better than the global model are detected the GQI index will be higher than the GoF value computed for the global model. As a matter of fact, local models performing better than the global model mean working with residuals that are smaller than the ones computed for the global model. And this directly entails obtaining a higher GQI index than the one obtained for the global model. Of course, the GQI can be considered as an average of the class specific GoF index. Nevertheless, expressing the GQI as in Eq. (3), allows us to directly compare

the same index among different partitions of the units (and with the aggregate solution of the global model too).

To assess the quality of the detected partition it is possible to perform a permutation test procedure [3] involving T random replications of the unit partition (keeping constant the group proportions as detected by the chosen clustering method). In this way an empirical distribution of the *GQI* index will be obtained. The *GQI* of the partition obtained by the chosen clustering method will be compared to the empirical distribution in order to assess if the detected partition performs better than a random assignment of the units, and better than the global model.

In the next section a simulation study to investigate the properties of the *GQI* is presented. The use of *GQI* to asses unit partition in a real case application is shown in [4].

3 Simulation Study

3.1 Design of the Numerical Example and Data Simulation

This simulation study aims at testing the *GQI* capability in assessing unit partition in two different situations, i.e. when the simulated data are affected by unobserved heterogeneity and the simulated local models really differ as regards model parameters, and when the simulated data are strictly homogenous, i.e. when the simulated local models do not differ. Here, a simple marketing type model will be used. The postulated model is composed of one latent endogenous variable, *Customer Satisfaction*, and two latent exogenous variables, *Price Fairness* and *Quality* (cf. Fig. 1). Each latent exogenous variable (*Price Fairness* and *Quality*) has five manifest variables (reflective mode), and the latent endogenous variable (*Customer Satisfaction*) is measured by three indicators (reflective mode). Here, we want to assess if in case of heterogenous data, the partition showing the highest *GQI* is the one whit the highest prediction power, i.e. the simulated one. This study intentionally uses a clear cut example of a marketing related path model for data simulation

Fig. 1 Experimental model

purposes. The data generation procedure is based on the LISREL-type approach. In other words, once the model parameters are established, the data are generated according to the implied covariance matrix, using a specific SAS-IML© macro developed by the author. For both the simulation schemes two latent classes, each of 200 units, are supposed to exist. Thus, the data on the aggregate level for each of the numerical examples includes 400 units. Moreover, for each of the postulated simulation scheme 100 sets of simulated data are computed. In total, the analysis involves 200 marketing related numerical examples on different sets of simulated data.

3.1.1 Simulation Scheme for the Heterogeneous Data-Sets

Unobserved heterogeneity involving both the structural and the measurement models directly means working with local models that are different as regards both the path coefficient values and the measurement model parameter values (i.e. the loading and outer weight values). In a simple model, as the one postulated above, heterogeneity in the model implies detecting price sensitive consumers, or those requiring price fairness, and consumers who have the strongest preference for another particular product attribute, e.g. quality. For more details on simulation scheme for heterogenous data-sets please refers to Table 1. 100 data-sets keeping the postulated features have been simulated. For each of these 100 data-set the *GQI* index is computed for both the global model (i.e. by computing the residuals of each unit

Table 1 Simulated values for model parameters

	Heterogenous data-sets		Homogenous data-sets
Model parameters	Class 1	Class 2	Both class 1 and class 2
No. of units	200	200	200
Path Coefficients:			
Price→*Sat*	0.9	0.1	0.8
Quality→*Sat*	0.1	0.9	0.8
Loadings *Price*:			
P_1	0.9	0.9	0.9
P_2	0.9	0.9	0.9
P_3	0.1	0.9	0.9
P_4	0.9	0.9	0.9
P_5	0.9	0.9	0.9
Loadings *Quality*:			
Q_1	0.9	0.9	0.9
Q_2	0.9	0.9	0.9
Q_3	0.9	0.1	0.9
Q_4	0.9	0.9	0.9
Q_5	0.9	0.9	0.9
Loadings *Satisfaction*:			
S_1	0.9	0.9	0.9
S_2	0.9	0.9	0.9
S_3	0.9	0.9	0.9

from the global model regardless of the unit membership to a class) and the simulated local models (i.e. by computing the residuals of each unit from its own local model). Afterwards, for each simulated data-set, 100 random replications of the unit partition in two classes (keeping constant the group proportions as simulated) are computed in order to perform a permutation test. In this way an empirical distribution of the *GQI* index is obtained. The *GQI* obtained for the simulated partition is compared to the empirical distribution in order to assess if the detected partition (in our case the simulated partition) performs better than a random assignment of the units, and better than the global model.

3.1.2 Simulation Scheme for the Homogeneous Data-Sets

In the case of homogenous data-sets all the units are supposed to be well described by a unique model. Two fictitious latent classes showing the same model parameters both in the measurement and in the structural models have been simulated [see Table 1]. 100 data-sets keeping the postulated features have been simulated. For each of these 100 data-sets the *GQI* index is computed for both the global model and the simulated fictitious local models. Once again, for each simulated data-set, 100 random replications of the unit partition in two classes (keeping constant the group proportions as simulated) are computed in order to perform a permutation test. Of course, we expect that the *GQI* indexes for both the global model solution and the (fictitious) partitioned data solution are similar. Moreover, we expect that the *GQI* value computed for the partitioned data solution is not an extreme value of the obtained empirical distribution.

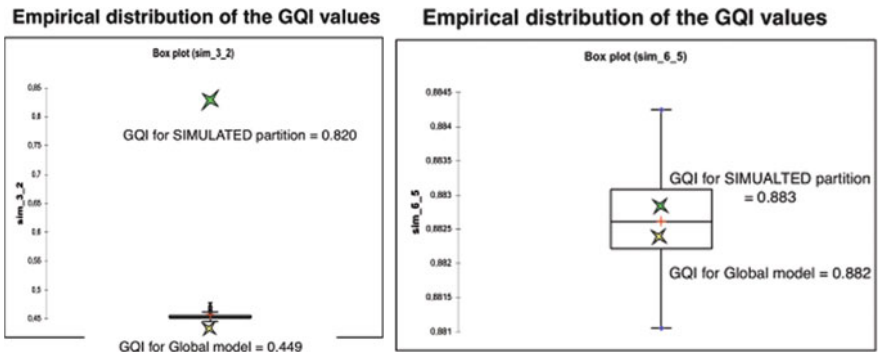
3.2 Simulation Study Results

Following the permutation test approach, each of the 200 data-sets (both homogenous and heterogeneous data) has been randomly divided 100 times into two classes of the same size as the simulated ones. The *GQI* has been computed for each of the random partitions of the units. An empirical distribution of the *GQI* values for a two class partition of the units is therefore obtained for each of the simulated data-sets.

Firstly we present the results obtained for the heterogeneous data-sets. In particular, in Table 2 and in Fig. 2(a) the results obtained as regards one of the 100 simulated heterogenous data-sets are shown. In Fig. 3, instead, the *GQI* distribution for all the 100 heterogeneous data-sets is shown. For each of the simulated heterogenous data-sets, the *GQI* value obtained from the simulated partition of the units, i.e. for real different latent classes, is definitely an extreme value of the distribution (cf. Figs. 2(a) and 3). Moreover, analyzing the box-plot obtained for the empirical distribution of the *GQI* values for a generic heterogenous data-set (cf. Fig. 2(a)), it is possible to notice that the *GQI* computed for the global model (i.e. the *GoF* value computed for the global model) is the smaller value obtained for the *GQI*, except for extreme solutions. This means that a unit partition always surpassed the performance of the global model. In other words, the global model has to be definitely considered as affected by heterogeneity. Moreover, the *GQI* value obtained for the

Table 2 Permutation test results for a generic heterogeneous data-set and a generic homogeneous data-set : simple statistics

Simple statistics	Heterogenous data-set	Homogenous data-set
No. of observations	102	102
Minimum	0.445	0.881
Maximum	0.831	0.884
1^{st} Quartile	0.451	0.882
Median	0.453	0.883
3^{rd} Quartile	0.456	0.883
Mean	0.429	0.883
Lower bound on mean (95%)	0.450	0.883
Upper bound on mean (95%)	0.465	0.882
GQI for SIMULATED partition	0.820	0.882
GQI for the GLOBAL model	0.449	0.883



(a) results for an heterogenous data-set (b) results for an homogenous data-set

Fig. 2 Empirical distribution of the GQI values obtained by permutation test

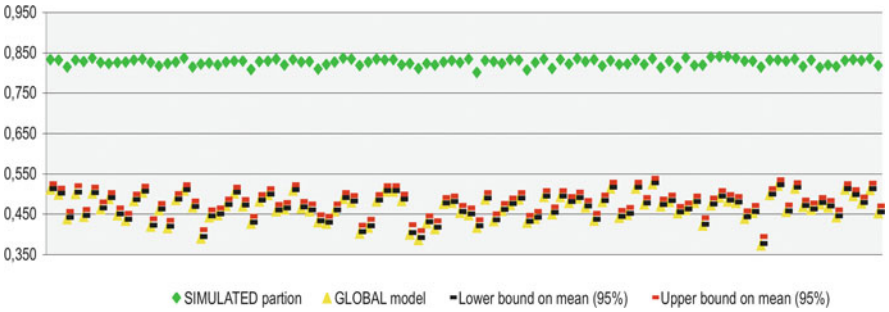


Fig. 3 Permutation test results for all the heterogeneous data-sets

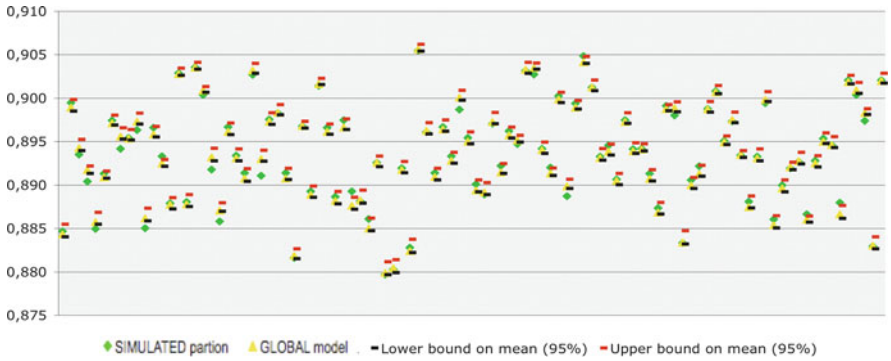


Fig. 4 Permutation test results for all the homogeneous data-sets

simulated partition is the highest obtained value. In Table 2 the simple statistics concerning the empirical distribution of a generic heterogenous data-set are presented. Here we can notice that the *GQI* index computed for the simulated partition is an extreme value as regards the empirical confidence interval ($\alpha = 0.05$) obtaining by permutation test. To conclude, analyzing the Fig. 3, it is possible to notice that similar results are obtained for all the heterogenous simulated data-sets. This allows us to assess that in the case of heterogenous data the simulated partition of the units is better than a random assignment of the units, and is definitely better (in terms of prediction power) than the global model solution.

Results obtained for the homogeneous data-sets are presented in Table 2 and in Figs. 2(b) and 4. Once again results obtained for a generic homogenous data-set are presented in Table 2 and in Fig. 2(b), while the empirical distributions for all the homogeneous data-sets are shown in Fig. 4. Differently form the heterogeneous case, in homogenous data-sets the *GQI* value obtained for the fictitious latent classes is close to the global model ones, as it was obviously expected. As a matter of fact the two latent classes show the same model parameters than the global model. Thus residuals from the local models are similar to residuals computed from the global model. Moreover, random partitions of units in two classes do not improve the predictive power of the models. Following the permutation test approach in the case of homogeneous data-sets no unit partition has to be considered better than the global model solution, i.e. none of the *GQI* values can be considered as an extreme value (cf. Fig. 2(b)). Similar results are obtained for all the homogeneous data-sets. In fact, the empirical confidence interval ($\alpha = 0.05$) for each of the 100 homogeneous data-sets always contains both the global model solution and the simulated one.

4 Discussion and Conclusions

Here, a new index to assess detected unit partition has been presented: the Group Quality Index (*GQI*). This index is a reformulation of the *GoF* index in a multi-group optic. It allows to assess the quality of the obtained unit partition when

performing a clustering method in PLS-PM. This simulation study shows that in the case of homogeneous datasets, the *GQI* computed for a unit partition equals the *GQI* computed for the non partitioned data-set. Instead, in the case of heterogeneous datasets, the *GQI* computed for the *best* unit partition is an extreme value of the *GQI* empirical distribution. Thus, we can conclude that the *GQI* index can be considered as a good indicator to assess if taking into account local models provides better performance (in terms of predictivity power) then using a single model for the whole sample. As future developments are concerned a more complex and more complete simulation study need to be performed so as to consider differences in groups size. Moreover, statistical significance of differences between local parameters needs to be further investigated.

Acknowledgments The Author thanks Vincenzo Esposito Vinzi and Michel Tenenhaus for the invaluable advices.

References

1. Amato, S., Esposito Vinzi, V., Tenenhaus, M.: A global goodness-of-fit index for PLS structural equation modeling. Technical Report, HEC School of Managment, France (2005)
2. Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York, NY (1989)
3. Edgington, E.: Randomization Test. Marcel Dekker Inc., New York, NY (1987)
4. Esposito Vinzi, V., Trinchera L., Amato, S.: PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement. In: Esposito Vinzi, V., Chin, W., Henseler, J., Wang, H. (eds.) Handbook "Partial Least Squares: Concepts, Methods and Applications", Computational Statistics Handbook Series, vol. II. Springer, Europe 47–82 (2010)
5. Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., Tenenhaus, M.: REBUS-PLS: a response-based procedure for detecting unit segments in PLS-PM. Appl. Stoch. Model. Bus. Ind. **24**, 439–458 (2008)
6. Hahn, C., Johnson, M., Herrmann, A., Huber, F.: Capturing customer heterogeneity using a finite mixture PLS approach. Schmalenbach Bus. Rev. **54**, 243–269 (2002)
7. Kaplan, D.: Structural Equation Modeling: Foundations and Extensions. Sage, Thousands Oaks, CA (2000)
8. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, C.: PLS path modeling. Comput. Stat. Data Anal. **48**, 159–205 (2005)
9. Trinchera, L.: Unobserved heterogeneity in structural equation models: a new approach in latent class detection in PLS path modeling. PhD Thesis, DMS, University of Naples (2007)
10. Wold, H.: Modelling in complex situations with soft information. In: Wold, H. (ed.) Third World Congress of Econometric Society, Toronto, Canada (1975)