Understanding PLS path modeling parameters estimates: a study based on Monte Carlo simulation and customer satisfaction surveys

Emmanuel Jakobowicz

CEDRIC-CNAM - 292 rue Saint Martin - 75141 Paris Cedex 03 - France
emmanuel.jakobowicz@free.fr

Summary. PLS path modeling (Partial Least Squares path modeling) has found increased interests since being used in the context of marketing studies. In this paper, we use Monte Carlo simulation and real life datasets to investigate the effects of changes in the model specification or in the data.
We first introduce PLS path modeling and some already well-known properties of this approach; we then conduct studies to estimate robustness of the approach to sample size, blocks size, distribution of the number of manifest variables inside the blocks and modification of the manifest variables values. All these studies are carried out using Monte Carlo simulations and real datasets from customer satisfaction surveys in the French power supply sector. We conclude with some general remarks on a framework to better analyse data and models using PLS path modeling.[1]

Key words: PLS path modeling, parameter estimation, Monte Carlo simulations, marketing data

1 PLS path modeling and properties

PLS path modeling is becoming a major tool in structural equation modeling when working with complex problems or small samples. Its "soft" assumptions[2] gives it serious advantages over covariance structure analysis. It is mainly applied in marketing within the framework of customer satisfaction analysis. In this section, we introduce the approach developed by H. Wold and present basic theoretical properties.

1.1 Methodology

PLS path modeling was introduced in 1982 by H. Wold [6], a good presentation of this approach can be found in [5].

[1] This work was supported by Electricité de France

[2] Soft modeling techniques are opposed to hard modeling techniques because of the soft distributional assumption

PLS path modeling is a confirmatory approach and thus needs a initial conceptual model formed by manifest and latent variables. The model is separated in two sub-models, the measurement or outer model including relations between manifest variables and their associated latent variables and the structural or inner model including relations between latent variables. The approach is based on an algorithm usually called PLS algorithm that estimates latent variables using alternatively the measurement model and the structural model. After convergence, latent variables scores are obtained for each observation and structural relation coefficients are estimated using multiple regressions. Estimation of the latent variables can be done using different schemes:

- For the outer estimation: Mode A (reflective way indicators) or Mode B (formative way indicators)
- For the inner estimation: Centroid, Path or Factorial scheme.

Within the framework of this study, we will focus on the parameterization commonly used in customer satisfaction studies, that is Mode A and centroid scheme.

1.2 Basic properties

H. Wold has demonstrated the convergence of the PLS algorithm for 2 blocks but since nobody succeeded in proving it for the general case but convergence is almost always observed. In this study we are interested in three properties of PLS path modeling estimators:

1. *Inconsistency of latent variables estimates:* Latent variables are estimated as aggregates of manifest variables which involves measurement error.
2. *Bias and inconsistency of loadings and inner structural (path) coefficients:* The coefficients are over or under-estimated. When all correlation coefficients among the K manifest variables are equal to s:

$$bias(\lambda) = \sqrt{\frac{s + \frac{1-s}{K}}{s}} = \frac{1}{\sqrt{\rho_{LV}}} \tag{1}$$

3. *Consistency at large of the model coefficients estimates:* We will deepen this notion in the first study presented bellow.

2 Study of changes in parameters estimation depending on model specification

In this section we first present our methodology for studying how PLS path modeling estimates react to changes in the model specification or in the data. In the whole of this study we try to see if theoretical properties can be empirically observed and if observed properties can be theoretically proved. Two

studies were carried out; in the first one we try to estimate which parameters of the conceptual model and of the data should be improved to obtain better estimates of structural coefficients. The second study, based on customer satisfaction data from the French electricity market, deals with the interpretation of changes in the estimation when manifest variables values change. Until now researchers have focused their studies on PLS path modeling parameterization effects on the latent variables scores or on special parameters of the data [4, 2]. [2] have investigated the robustness of PLS path modeling to skewed data, multicolinearity between manifest variables and misspecification of the model. It appears that bias of PLS estimates is not modified if these parameters change. [4] have focused on effect of changes in the model on the latent variables scores. All through our simulations, we compare our results with the results of these papers.

PLS path modeling is prediction-oriented and research has shown that where possible, covariance structure analysis should be preferred for parameters estimation (because parameters estimates are consistent). However, covariance structure analysis encounters many identification problems, and when identification is not possible, PLS path modeling can be a good solution for parameters estimation. In our studies, our interest is to know how PLS path modeling estimator behaves when models are complex and data have special properties. In this work, we use a slightly modified version of the European Customer Satisfaction Index model [5] (Fig. 1). It has 6 latent variables and covariance structure analysis fails to estimate it. We use this model both with simulated data and real data.

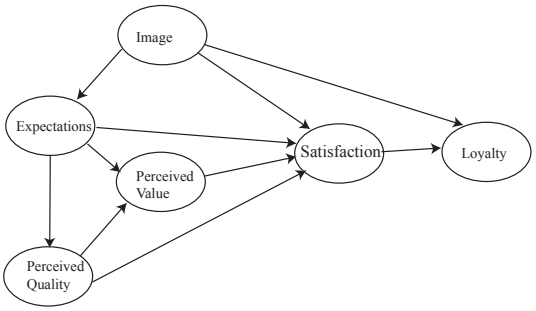

Fig. 1. The ECSI model

2.1 First study

We studied different parameters of both the data and the conceptual model:

- Sample size

- Number of manifest variables for each latent variable
- Unequal distribution of the number of manifest variables in each block.

For each case, we try to present both simulations[3] and theoretical results.

Sample size and number of manifest variables for each latent variable

Both these parameters are linked by this important property:
The PLS approach is consistent "at large".
It has been proven by H. Wold [6]. PLS estimates approach true values as the number of both cases and manifest variables increase. To empirically verify this property, we have simulated datasets with a different number of cases (from 20 to 10′000 cases, 10 replications of 9 datasets) and a different number of manifest variables (from 2 to 10 manifest variables per block, 10 replications of 9 datasets) on the ECSI model. We thus compare structural coefficients, significance of coefficients and manifest variables outer weights. To obtain significance coefficients we used bootstrap methods. We have studied the behaviour of the fit indices that exist in PLS path modeling which are latent variables R^2, Communality, Redundancy and Stone-Geisser's Q^2 but these aspects will not be developed here. In figure 2 we present the mean and standard error of the estimated structural coefficients depending respectively on the number of cases (i) and on the number of manifest variables per block (ii).

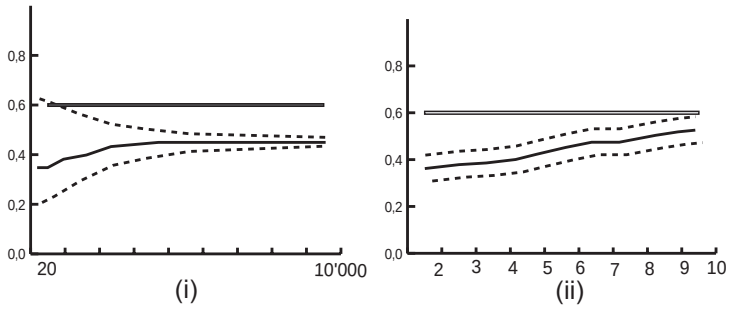

Fig. 2. Evolution of structural coefficients with confidence interval depending on sample size (i) and on the number of manifest variables per block (ii) (the real value of the coefficient is $0,6$)

[3] Our simulation are based on the ECSI model, the exogenous latent variable (Image) and error of the inner and outer models are simulated using fixed distributions ([4] showed that distribution has no impact on model estimation) other latent variables and manifest variables are obtained according to structural relations of the model.

It appears that with a very small sample ($\leq$ 200 cases), bias of parameters estimations is large with large standard errors. For more than 500 cases, there is a clear negative bias that persists without decreasing until 10′000 cases. For the number of manifest variables, results are clearer and bias becomes constantly smaller. Standard errors decrease as sample size increases but do not depend on the number of manifest variables per block. As can be predicted, significance level depends on sample size but not on the number of manifest variables. This property is inherent to the definition of these concepts. The non standardized outer weights also depend on sample size and on the number of manifest variables (as structural coefficient estimates, these estimates have large bias until 200 cases). Our major conclusion is that a certain number of cases is necessary to obtain relevant estimates of model coefficients, but when it is exceeded the sample size does not have an effect on parameters estimates (this complies with [4]'s results). The number of manifest variables per latent variable has a much greater effect on parameters estimates. Increasing the number of manifest variables leads to a reduction of the bias in structural coefficients estimations. When increasing both manifest variables number and sample size, the bias lessens quicker than when increasing them independently. However even with a very large sample (10′000 observations) and a large number of manifest variables (10 per block), structural coefficients are still negatively biased and loadings are still positively biased. It is actually hard to give advice only using simulated data but general advice would be to focus on getting more manifest variables per block than more cases. More particularly, our study shows that for a complex model we need at least 200 observations and 4 manifest variables per latent to obtain robust structural coefficients estimates. Of course, if we have fewer cases or fewer manifest variables, estimates can be analysed but the fact that they are significantly biased should be taken into account. As [2] have shown, latent variables scores are not affected by these changes.

Unequal distribution of the number of manifest variables between blocks

Not much has been studied on the distribution of the number of manifest variables inside the blocks. We developed a simulation process to test the effects of unequal distribution. We used the ECSI model with 24 manifest variables (4 per block) and we modified the number of manifest variables per latent variable (from 2 to 10) focusing on 3 latent variables: the exogenous one (image), the central one (satisfaction) and the one to be explained (loyalty). We obtained 10 datasets (with 20 replicates for each one). We studied the structural coefficients.

Table 1 shows the different bias that appears when modifying this distribution over the structural coefficients. We can see that bias is minimum when working with the classical model (equal number of manifest variables in the 6 blocks). The largest bias on the structural coefficient between satisfaction

Table 1. Bias in the structural coefficients in different distribution of the manifest variables in the blocks (MV: manifest variables)

Case	Structural coefficients bias	
	Image-Loyalty	Satisfaction-Loyalty
Equality	-0,05 (-0,15;0,05)	-0,14 (-0,23;-0,05)
10 MV on Image	-0,05 (-0,16;0,03)	-0,21 (-0,30;-0,14)
10 MV on Satisfaction	-0,10 (-0,19,-0,02)	-0,24 (-0,35;-0,16)
10 MV on Loyalty	-0,10 (-0,18,-0,04)	-0,12 (-0,22,-0,07)
2 MV on Image	-0,11 (-0,17;-0,04)	-0,14 (-0,23;-0,05)
2 MV on Satisfaction	-0,07 (-0,14;0,00)	-0,18 (-0,25;-0,11)
2 MV on Loyalty	-0,12 (-0,20;-0,03)	-0,13 (-0,22;-0,06)

and loyalty appears when working with 10 manifest variables on satisfaction. However, these differences are not significant when studying structural coefficients with confidence intervals. It appears that the fact to have unequal distribution of the manifest variables between blocks has no major effect on parameters estimates. But when working with real data, we face a problem of interpretation. To relate a concept (latent variable) with many facets (with many manifest variables) to a simpler concept (with few manifest variables) can cause badly interpreted results in customer satisfaction analysis (we suppose the internal consistency value remains constant).

2.2 Second study

In this second study, we investigate the effects of manifest variables case values modifications in PLS path modeling. To do so, we use real datasets, simply modifying the manifest variables values. We had three different transformations to study: (i) Translation and homothetic transformation of a manifest variable; (ii) Different scales for the manifest variables; (iii) Changes of manifest variables means and standard errors.

By studying the algorithm, we can see that translation has no effect on parameters estimates and latent variables scores. A homothetic transformation in the case of non standardised data only has effect on outer weights; for a coefficient a with $x' = a * x$, we obtain $w' = \frac{w}{a}$. This transformation effect makes sense when choosing data standardization techniques; if variables are on different scale a standardization of the data should be preferred.
This study shows that scale has no effect when transforming the variables once obtained. If we follow [4]'s procedure, it appears that means of latent variables scores are not transformed and standard errors are smaller when the scale of the answers is larger (when using for example a 5 points scale and a 10 points scale).

It is the last transformation that is of interest; in real life applications, a classic question about structural equation models is: "what happens if the

mean of the answers for this question grows by one point?" Using the method property, we can say that relationships between variables are based on covariance and growth of the mean does not lead to growth in coefficients estimations. But we would like to empirically understand what happens when means and standard errors of manifest variables change. We use 8 real datasets and modify different manifest variables for each one. Each time we try a different modification framework. All manifest variables are based on 10 points scales. Mean and standard error of one variable have been modified for each procedure with a different value of mean and standard error keeping the same scale scheme.

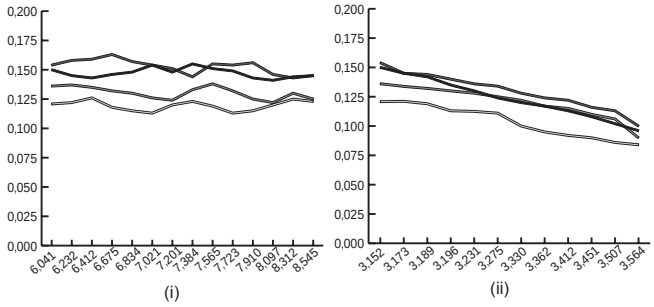

Fig. 3. Example of evolution of parameters estimates of outer weights when increasing the mean of a manifest variable (i) and the standard error (ii)

Figure 3 shows the evolution of outer weight for a designated manifest variable when mean (i) or standard error (ii) are increasing. It appears that outer weights decrease almost linearly when standard errors increase, which means that the relative importance of the manifest variable decreases if answers are not following a narrow distribution. When modifying the mean, no global tendency appears. Relative importance and mean do not seem to be linked.

3 Discussion and conclusions

These two studies reveal some interesting behaviours of PLS path modeling. Variations in model or data specification have major effects on the results and interpretations of this approach which is analytically hard to understand due to the iterative algorithm involved. Clearly our first study gives some clues to parameterizing the model. It confirms [2, 3, 4]s results on bias induced by sample size and by the number of manifest variables per block and shows that large inequalities in the distribution of the number of manifest variables in the blocks will affect mainly model interpretation and to a lesser extent model parameters estimation. The second study is a major help for interpretation

of PLS coefficients. PLS path modeling is widely used in marketing application and frequently results are wrongly interpreted. This study shows that the model parameters estimations are invariant to linear transformation of the data. However we have seen that outer weights which indicate a relative importance of each manifest variable inside a block depend on standard error of the manifest variables. The relationship between outer weights and means are not significant and means should be considered as an information to be used jointly with outer weights. When mean is low and outer weight is high, the values of this manifest variable has to be improved. On the other hand, if outer weight is high and mean is high, increasing the mean is not meaningful in a decision process. The results we obtained cannot easily be generalized. But we can draw a general framework from these two studies:
To get better results or results that can be well interpreted in term of parameters estimates, we should use a sample of at least 200 cases, well explained latent variables (a large number of manifest variables should be preferred to better unidimensionality criterion) and a good distribution of the manifest variables in the different blocks. This sample size can be seen as very large, but compared to covariance structure analysis, it is quite small when looking at the model ([1] advise to use 10 cases for each parameter of the model when data is not normally distributed in covariance structure analysis, that means more than $1'000$ cases). Further research should be done in order to better understand the behaviour of this approach; for example when using formative indicators or ordinal data. Our research is coherent with older works on the subject and confirms that PLS path modeling is a complementary approach to the covariance structure analysis even for parameters estimation.

References

1. P.M. Bentler and C.P. Chou. Practical issues in structural modeling. *Sociological Methods & Research*, 16:78–117, 1987.
2. C. Cassel, P. Hackl, and A.H. Westlund. Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics*, 26:435–446, 1999.
3. W. Chin, B.L. Marcolin, and P.R. Newsted. A partial least squares latent variables modeling approach for measuring interaction effects: Results from a monte carlo simulation study and voice mail emotion/adoption study. In *proceedings of the 17th Int. Conf. on Information Systems*, pages 21–41, 1996.
4. K. Kristensen, J.K. Eskildsen, H.J. Juhl, and P. Ostergaard. PLS structural equation modeling for customer satisfaction: Methodological and application issues. In *13th Int. and 68th Annual American Meeting of the Psychometric Society, Cagliari, Sardinia*, 2003.
5. M. Tenenhaus, V. Esposito Vinzi, Y.-M. Chatelin, and C. Lauro. PLS path modeling. *Computational Statistics and Data Analysis*, 48(1):159–205, 2005.
6. H. Wold. Soft modeling: the basic design and some extensions. In *Systems under Indirect Observation, Part 2*, volume 2, pages 1–54. North-Holland, 1982.