PLS-MGA: A Non-Parametric Approach to Partial Least Squares-based Multi-Group Analysis

Jörg Henseler

Abstract This paper adds to an often applied extension of Partial Least Squares (PLS) path modeling, namely the comparison of PLS estimates across subpopulations, also known as multi-group analysis. Existing PLS-based approaches to multi-group analysis have the shortcoming that they rely on distributional assumptions. This paper develops a non-parametric PLS-based approach to multi-group analysis: PLS-MGA. Both the existing approaches and the new approach are applied to a marketing example of customer switching behavior in a liberalized electricity market. This example provides first evidence of favorable operation characteristics of PLS-MGA.

1 Introduction

For decades, researchers have applied partial least squares path modeling (PLS, see Tenenhaus et al. 2005; Wold 1982) to analyze complex relationships between latent variables. In particular, PLS is appreciated in situations of high complexity and when theoretical explanation is scarce (Chin 1998) – a situation common for many disciplines of business research, such as marketing, strategy, and information systems (Henseler 2010). In many instances, researchers face a heterogeneity of observations, i. e. for different sub-populations, different population parameters hold. For example, institutions releasing national customer satisfaction indices may want to know whether model parameters differ significantly between different industries. Another example would be cross-cultural research, in which the culture or country plays the role of a grouping variable, thereby defining the sub-

J. Henseler (✉)
Institute for Management Research, Radboud University Nijmegen,
Thomas van Aquinostraat 3, 6525 GD Nijmegen, The Netherlands
e-mail: j.henseler@fm.ru.nl

W. Gaul et al. (eds.), *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-24466-7_50,

populations. As these examples show, there is a need for PLS-based approaches to multi-group analysis.

The predominant approach to multi-group analysis was brought foreward by Keil et al. (2000) and Chin (2000). These authors suggest to apply an unpaired samples t-test to the group-specific model parameters using the standard deviations of the estimates resulting from bootstrapping. As Chin (2000) notes, the parametric assumptions of this approach constitute a major shortcoming. As PLS itself is distribution-free, it would be favorable to have a non-parametric PLS-based approach to multi-group analysis.

The main contribution of this paper is to develop a non-parametric PLS-based approach to multi-group analysis in order to overcome the shortcoming of the current approach. The paper is structured as follows. Next to this introductory section, the second section presents the existing approach and elaborates upon its strengths and weaknesses. The third section develops the new approach and describes its characteristics. The fourth section presents an application of both the existing and the new PLS-based approach to multi-group analysis to an example from marketing about the consumer switching behavior in a liberalized electricity market. Finally, the fifth section discusses the findings of this paper and highlights avenues for further research.

2 The Chin/Keil Approach to Multi-Group Analysis

In multi-group analysis, a population parameter θ is hypothesized to differ for two or more subpopulations. At first, we limit our focus on the case of two groups, and will generalize in the discussion.

Typically, multi-group analysis consists of two steps. In a first step, a sample of each subpopulation is analyzed, resulting in groupwise parameter estimates $\tilde{\theta}^g$. In a second step, the significance of the differences between groups is evaluated.

Chin (2000) as well as Keil et al. (2000) propose to use an unpaired samples t-test in order to test whether there is a significant difference between two group-specific parameters. They suggest comparing the parameter estimate of the first group, $\tilde{\theta}^{(1)}$, with the parameter estimate of the second group, $\tilde{\theta}^{(2)}$. The test statistic is as follows (see Chin 2000):

$$t = \frac{\tilde{\theta}^{(1)} - \tilde{\theta}^{(2)}}{\sqrt{\frac{\left(n^{(1)}-1\right)^2}{n^{(1)}+n^{(2)}-2} \cdot se_{\theta^{(1)}} + \frac{\left(n^{(2)}-1\right)^2}{n^{(1)}+n^{(2)}-2} \cdot se_{\theta^{(2)}}} \cdot \sqrt{\frac{1}{n^{(1)}} + \frac{1}{n^{(2)}}}} \tag{1}$$

This statistic follows a t-distribution with $n^{(1)} + n^{(2)} - 2$ degrees of freedom. The subsample-specific parameter estimates are denoted as $\tilde{\theta}^{(g)}$ (with g as a group index), the sizes of the subsamples as $n^{(g)}$, and the standard errors of the parameters as resulting from bootstrapping as $se_{\tilde{\theta}^{(g)}}$. Instead of bootstrapping, sometimes jackknifing is applied (e.g. Keil et al. 2000).

The t-statistic as provided by Equation 1 is known to perform reasonably well if the two empirical bootstrap distributions are not too far away from normal and/or the two variances $n^{(1)} \cdot se^2_{\theta^{(1)}}$ and $n^{(2)} \cdot se^2_{\theta^{(2)}}$ are not too different from one another. If the variances of the empirical bootstrap distributions are assumed different, Chin (2000) proposes to apply a Smith-Satterthwaite test. The modified test statistic becomes (see Nitzl 2010):

$$t = \frac{\tilde{\theta}^{(1)} - \tilde{\theta}^{(2)}}{\sqrt{\frac{n^{(1)}-1}{n^{(1)}} se^2_{\theta^{(1)}} + \frac{n^{(2)}-1}{n^{(2)}} se^2_{\theta^{(2)}}}} \tag{2}$$

Also this statistic follows a t-distribution. The number of the degrees of freedom ν for the t-statistic is determined by means of the Welch-Satterthwaite equation (Satterthwaite 1946; Welch 1947)[1]:

$$\nu(t) = \frac{\left(\frac{n^{(1)}-1}{n^{(1)}} se^2_{\theta^{(1)}} + \frac{n^{(2)}-1}{n^{(2)}} se^2_{\theta^{(2)}}\right)^2}{\frac{n^{(1)}-1}{n^{(1)2}} se^4_{\theta^{(1)}} + \frac{n^{(2)}-1}{n^{(2)2}} se^4_{\theta^{(2)}}} - 2 \tag{3}$$

3 A New PLS-Based Approach to Multi-Group Analysis

It is obvious that the aforementioned approaches to group comparisons with their inherent distributional assumptions do not fit PLS path modeling, which is generally regarded as being distribution-free. Taking into account this criticism against the available approaches, this paper presents an alternative approach to PLS-based group comparisons that does not rely on distributional assumptions. The working principle of the novel PLS multi-group analysis (PLS-MGA) approach is as follows: Just like within the parametric approaches, the data is divided into subsamples according to the level of the grouping variable, and the PLS path model is estimated for each subsample. Moreover, each subsample becomes subject to a separate bootstrap analysis. The novelty of the new approach to PLS-based multi-group analysis lies in the way in which the bootstrap estimates are used to assess the robustness of the subsample estimates. More specifically, instead of relying on distributional assumptions, the new approach evaluates the observed distribution of the bootstrap outcomes. It is the aim of this section to determine the probability of a difference in group-specific population parameters given the group specific estimates and the empirical cumulative distribution functions (CDFs). Let $\tilde{\theta}^{(g)}$ ($g \in \{1, 2\}$) be the group-specific estimates. Without loss of generality, let us assume that $\tilde{\theta}^{(1)} > \tilde{\theta}^{(2)}$. In order to assess the significance of a group effect, we are looking for $P\left(\theta^{(1)} \leq \theta^{(2)} \mid \tilde{\theta}^{(1)}, \tilde{\theta}^{(2)}, \mathrm{CDF}(\theta^{(1)}), \mathrm{CDF}(\theta^{(2)})\right)$.

[1] This notation of the Welch-Satterthwaite equation was derived by Nitzl (2010). Note that the formula proposed by Chin (2000) is incorrect.

Let J be the number of bootstrap samples, and $\tilde{\theta}_j^{(g)*}$ ($j \in \{1, \ldots, J\}$) the bootstrap estimates. In general, the mean of the bootstrap estimates differs from the group-specific estimate, i. e. the empirical distribution of $\theta^{(g)}$ does not have $\tilde{\theta}^{(g)}$ as its central value. In order to overcome this, we can determine the centered bootstrap estimates $\tilde{\theta}_j^{(g)\bar{*}}$ as:

$$\forall g, j: \qquad \tilde{\theta}_j^{(g)\bar{*}} = \tilde{\theta}_j^{(g)*} - \frac{1}{J} \sum_{i=1}^{J} \tilde{\theta}_i^{(g)*} + \tilde{\theta}^{(g)}. \tag{4}$$

Making use of the bootstrap estimates as discrete manifestations of the CDFs we can calculate

$$P\left(\theta^{(1)} \leq \theta^{(2)} \mid \tilde{\theta}^{(1)}, \tilde{\theta}^{(2)}, \mathrm{CDF}(\theta^{(1)}), \mathrm{CDF}(\theta^{(2)})\right) = P\left(\theta_i^{(1)\bar{*}} \leq \theta_j^{(2)\bar{*}}\right) \tag{5}$$

Using the Heaviside step function $H(x)$ as defined by

$$H(x) = \frac{1 + \mathrm{sgn}(x)}{2}, \tag{6}$$

Equation (5) transforms to

$$P\left(\theta^{(1)} \leq \theta^{(2)} \mid \tilde{\theta}^{(1)}, \tilde{\theta}^{(2)}, \mathrm{CDF}(\theta^{(1)}), \mathrm{CDF}(\theta^{(2)})\right) = \frac{1}{J^2} \sum_{i=1}^{J} \sum_{j=1}^{J} H\left(\tilde{\theta}_j^{(2)\bar{*}} - \tilde{\theta}_j^{(1)\bar{*}}\right). \tag{7}$$

Equation (7) is the core of the new PLS-based approach to multi-group analysis. The idea behind it is simple: Each centered bootstrap estimate of the second group is compared with each centered bootstrap estimate of the first group. The number of positive differences divided by the total number of comparisons (i.e., J^2) indicates how probable it is in the population that the parameter of the second group is greater than the parameter of the first group.

4 A Marketing Example

We illustrate the use of both the existing and the new PLS-based approach to multi-group analysis on the basis of a marketing example, namely customer switching behavior in a liberalized energy market. Prior studies and marketing theory (c. f. Jones et al. 2000; de Ruyter et al. 1998) suggest that customers are less likely to switch their current energy provider if they are satisfied or if they perceive high switching costs. From the Elaboration Likelihood Model it can be derived that consumer behavior is contingent on the level of involvement (Bloemer and Kasper 1995; Petty and Cacioppo 1981).

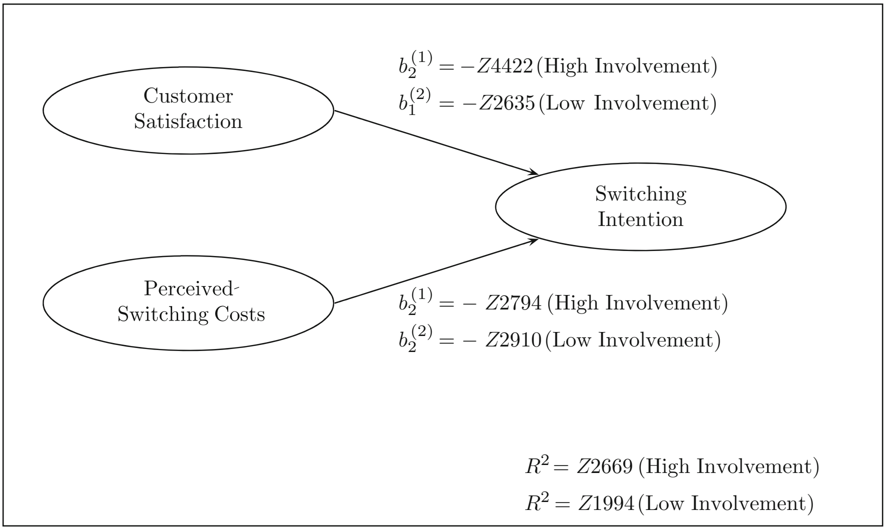

Fig. 1 Structural model with groupwise parameter estimates (standardized PLS path coefficients)

A cross-sectional study among consumers was conducted in order to test the proposed hypotheses. The data at hand stems from computer-assisted telephone interviews with 659 consumers. 334 consumers indicated to be highly involved in buying electricity, while 325 consumers said to have a low involvement. Customer satisfaction, switching costs, and customer switching intention were measured by multiple items using mainly five-point Likert scales.

We create a PLS path model as depicted in Fig. 1. This model captures the two direct effects of customer satisfaction and perceived switching costs on customer loyalty. In order to account for the moderating effect of involvement, we estimate the model separately once for the group of highly involved consumers and once for the group of consumers having low involvement. Figure 1 also reports the standardized path coefficients per group as estimated by means of the PLS software SmartPLS (Ringle et al. 2007).

Moreover, we conduct bootstrap resampling analyses with 500 bootstrap samples per group. Based on the estimates, the bootstrap estimates and their standard deviations, we calculated the p-values for group differences in the effects of customer satisfaction and perceived switching costs on switching intention. Table 1 contrasts the results of the different PLS-based approaches to multi-group analysis, i. e. the parametric test with equal variances assumed (homoskedastic), the parametric test with equal variances not assumed (heteroskedastic), and the non-parametric PLS-MGA.

The different PLS-based approaches to multi-group analysis provide similar results. All approaches find a significant difference in strength of the effect of customer satisfaction on switching intention ($\alpha = .05$). This means, for highly

Table 1 Comparison of statistical tests on group differences

Hypothesis	Statistical test	p-value (one-sided)
Customer satisfaction	Parametric, homoskedastic	.0123
↓	Parametric, heteroskedastic	.0056
Switching intention	PLS-MGA (non-parametric)	.0056
Perceived switching costs	Parametric, homoskedastic	.4308
↓	Parametric, heteroskedastic	.4428
Switching intention	PLS-MGA (non-parametric)	.5528

involved consumers, the level of customer satisfaction is a stronger predictor of switching behavior than for consumers with low involvement. Moreover, all approaches reject a group effect in the impact of perceived switching costs on switching intention. Despite the general convergence of findings, there seem to be notable differences in statistical power between the approaches. For instance, both the parametric test with equal variances not assumed and PLS-MGA are able to detect the group effect on a .01 significance level, whereas the parametric test with equal variances assumed is not.

5 Discussion

It was the aim of this contribution to introduce a non-parametric approach to PLS-based multi-group analysis. The new approach, PLS-MGA, does not require any distributional assumptions. Moreover, it is simple to apply in that it uses the bootstrap outputs that are generated by the prevailing PLS implementations such as SmartPLS (Ringle et al. 2007), PLS-Graph (Soft Modeling, Inc. 1992–2002), PLS-GUI (Li 2005), or SPAD (Test and Go 2006).

Technically, the new approach to PLS-based multi-group analysis, PLS-MGA, is purely derived from bootstrapping in combination with a rank sum test, which makes it conceptually sound. Still, its use has only been illustrated by means of one numerical example. Future research should conduct Monte Carlo simulations on PLS-MGA in order to obtain a better understanding of its characteristics, such as for instance its statistical power under various levels of sample size, effect size, construct reliability, and error distributions.

Further research is also needed to extend PLS-MGA to analyze more than two groups at a time. As a quick solution, multiple tests with a Bonferroni correction could be performed. Alternatively, an adaptation of the Kruskal-Wallis test (Kruskal and Wallis 1952) to PLS-based multi-group analysis might be promising.

Finally, PLS-based multi-group analysis has been limited to the evaluation of the structural model so far, including this article. However, PLS path modeling does not put any constraints on the measurement model so that measurement variance could be an alternative explanation for group differences. Up to now, no PLS-based approaches for examining measurement invariance across groups have been proposed yet. Given its ease and robustness, PLS-MGA may also be the point of departure for the examination of group differences in measurement models.

References

Bloemer JMM, Kasper HDP (1995) The complex relationship between customer satisfaction and brand loyalty. J Econ Psychol 16:311–329

Chin WW (1998) The partial least squares approach to structural equation modeling. In: Marcoulides GA (ed) Modern methods for business research. Lawrence Erlbaum Associates, Mahwah, NJ, pp 295–336

Chin WW (2000) Frequently Asked Questions – Partial Least Squares & PLS-Graph. URL http://disc-nt.cba.uh.edu/chin/plsfac/plsfac.htm

Henseler J (2010) On the convergence of the partial least squares path modeling algorithm. Comput Stat 25(1):107–120

Jones MA, Mothersbaugh DL, Beatty SE (2000) Switching barriers and repurchase intentions in services. J Retail 76(2):259–274

Keil M, Tan BCY, Wei KK, Saarinen T, Tuunainen V, Wassenaar A (2000) A cross-cultural study on escalation of commitment behavior in software projects. Manag Inform Syst Q 24(2):299–325

Kruskal W, Wallis W (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47(260):583–621

Li Y (2005) PLS-GUI - Graphic User Interface for Partial Least Squares (PLS-PC 1.8) Version 2.0.1 beta. University of South Carolina, Columbia, SC

Nitzl C (2010) Eine anwenderorientierte Einführung in die Partial Least Square (PLS)-Methode. Universität Hamburg, Institut für Industrielles Management, Hamburg

Petty R, Cacioppo J (1981) Attitudes and persuasion: classic and contemporary approaches. W.C. Brown, Dubuque, Iowa

Ringle CM, Wende S, Will A (2007) SmartPLS 2.0 M3. University of Hamburg, Hamburg, Germany, URL http://www.smartpls.de

de Ruyter K, Wetzels M, Bloemer JMM (1998) On the relationship between perceived service quality, service loyalty and switching costs. Int J Serv Ind Manag 9(5):436–453

Satterthwaite F (1946) An approximate distribution of estimates of variance components. Biometrics Bull 2(6):110–114

Soft Modeling, Inc (1992–2002) PLS-Graph version 3.0. URL http://www.plsgraph.com

Tenenhaus M, Vinzi VE, Chatelin YM, Lauro C (2005) PLS path modeling. Comput Stat Data Anal 48(1):159–205

Test & Go (2006) SPAD version 6.0.0. Paris, France

Welch B (1947) The generalization of 'Student's' problem when several different population variances are involved. Biometrika 34(1/2):28–35

Wold HOA (1982) Soft modelling: the basic design and some extensions. In: Jöreskog KG, Wold HOA (eds) Systems under indirect observation. Causality, structure, prediction, vol II. North-Holland, Amsterdam, New York, Oxford, pp 1–54