DETERMINING THE NUMBER OF SEGMENTS IN FIMIX-PLS

Marko Sarstedt, Ludwig-Maximilians-University Munich, Germany
Manfred Schwaiger, Ludwig-Maximilians-University Munich, Germany
Christian M. Ringle, University of Hamburg, Germany

ABSTRACT

Response-based segmentation using the finite mixture partial least squares approach (FIMIX-PLS) has recently received increasing research attention in the marketing and management disciplines. Different approaches for latent class detection have been proposed which generalize, for example, tree-like structure (Sánchez and Aluja 2006), finite mixture (Hahn et al. 2002), fuzzy regression (Palumbo et al. 2008), distance-based (Esposito Vinzi et al. 2007) and genetic algorithm approaches (Ringle et al. 2009a) to PLS path modeling. Sarstedt's (2008a) review of PLS segmentation techniques characterizes FIMIX-PLS as the primary choice for latent class detection and segmentation tasks within a PLS context. FIMIX-PLS was introduced by Hahn et al. (2002) and allows for a simultaneous estimation of model parameters and segment affiliations of observations. The approach has performed favourably in different simulation studies (Ringle et al. 2009b) as well as empirical applications (e.g. Hahn et al. 2002; Ringle et al. 2009c) and has been made available to researchers by the software application SmartPLS (Ringle et al. 2005). An unresolved problem in the application of FIMIX-PLS concerns the issue of model selection, i.e. the determination of the number of segments underlying the data. Unlike other latent class procedures, the FIMIX-PLS algorithm allows for the computation of several statistical model selection criteria (Sarstedt 2008b). These are well-known from finite mixture model literature and their performance has been discussed in several simulation studies in different context. However, despite the growing popularity of FIMIX-PLS and the importance of an accurate model selection, up to date, no such simulation study exists for this procedure. Consequently, the scope of this paper is to examine the performance of several model selection criteria for determining the number of segments in FIMIX-PLS by conducting a Monte Carlo simulation study.

EXPERIMENTAL SET-UP

The simulation study considers ten major information and classification criteria that have frequently been used in marketing applications (i.e. AIC, AIC_3, BIC, CAIC, MDL_2, MDL_5, HQ, NEC, CLC, and ICL-BIC). Seven data characteristics (factors) are manipulated. The choice of the factors and their levels is influenced by studies conducted by Hawkins et al. (2001), Williams et al. (2002) and Andrews and Currim (2003a, 2003b): Number of observations [50, 100 or 300], number of segments [2,3], size of the smallest segment [15%, 30%], distance between path coefficients [0.2, 0.8], model complexity [low, high], measurement model of the exogenous latent variables [reflective, formative], measurement error of the manifest variables / disturbance term of the endogenous latent variables [15%, 35%]. The study builds on a data generation procedure which has been described and applied in previous PLS path modeling based simulation studies (e.g., Henseler and Chin 2009). For each possible combination of factor levels, data sets are generated. The six factors with two levels and the one factor with three levels result in $2^6 \cdot 3^1 = 192$ possible factor level combinations that are considered in this study. Each simulation run is conducted four times for s=1,…,5 segments and the percentages of data sets in which every criterion identified the expected number of segments (i.e. the success rate) is computed.

RESULTS

The analysis results show that three criteria - the BIC, the CAIC and the ICL-BIC - exhibit comparably high success rates (BIC: 45%, CAIC: 46%, ICL-BIC: 42%) and relatively low overfitting quotas (<20%). The CAIC has the highest overall success rate and performs favorably in situations where small random samples or formative measures are used. A joint consideration of the CAIC and ICL-BIC takes also the entropy into account, which indicates the degree of separation between the segments to assess whether the analysis provides well-separated clusters. This kind of information criterion combination offers the best FIMIX-PLS analysis results in our study. In contrast, the NEC and the MDL_5 criterion show the worst performance in that they uncover the expected model only in approximately every fifth case. FIMIX-PLS evaluation of results represent a requirement for PLS path modeling (Ringle et al. 2009c). As a result, the identification of suitable model selection criteria improves the applicability of FIMIX-PLS and allows for a more prudent use in PLS applications. The findings of this study provide researchers and practitioners with a specific decision-making support in their PLS path modeling analyses.

References Are Available on Request