ANALYZING MULTIGROUP DATA WITH STRUCTURAL EQUATION MODELS

N. Görz, L. Hildebrandt, D. Annacker

Institut für Marketing II,
Humboldt-Universität zu Berlin, D-10178 Berlin, Germany

Abstract: In empirical applications of structural equation modeling researchers often assume that the sample under investigation is homogenous unless observed charateristics allow for a division of the sample into mutual exclusive homogenous subgroups. If such information is not available, unobserved heterogeneity can be taken into account by a finite-mixture approach (Arminger et al. (1998); Jedidi et al. (1997)). The simulation study presented in this paper reveals that this approach clearly outperforms a sequential procedure combining cluster and multigroup analysis.

1 Introduction

Empirical analyses of structural equation models (SEM) in marketing often assume homogenous samples or form "homogenous" groups a priori based on theoretical considerations regarded as sufficient to identify the main causes of differences in the means of the relevant variables and/or the relationships among them (Hildebrandt, Homburg (1998)). If the sources of heterogeneity can be easily observed or if appropriate indicators for latent causes exist (observed heterogeneity), the proposed model can be estimated by separate analyses of the subsamples or by using the multigroup option available in programs like LISREL (see e.g. Jöreskog, Sörbom (1996)). If, however, indicators for an a-priori segmentation are missing, this unobserved heterogeneity poses serious problems, e.g. biased parameter estimates (Jedidi et al. (1997)). In this article a more sophisticated methodology based on finite mixtures of mean- and covariance structures will be described, which solves many of these shortcomings.

2 SEM and Unobserved Heterogeneity

2.1 The traditional approach

Traditionally, unobserved heterogeneity in structural equation modeling was taken into account by a sequential two-step procedure (Jedidi et al. (1997)): (1) "homogenous" groups are formed by performing cluster analysis on the

data set (e.g., K-means clustering), and (2) a multigroup SEM based on separate covariance matrices for each segment is estimated. A major shortcoming of this approach is that it ignores important information about the data. Although the researcher frequently has specific hypotheses about the relationships between the analyzed variables, traditional cluster analysis assumes independence among the variables. In the case of highly-correlated variables data-reduction techniques (e.g., principal component analysis) are proposed, but also in this case the conceptual flaws are accompanied by serious statistical problems (Jedidi et al. (1997)).

2.2 Finite Mixtures of Conditional Mean- and Covariance Structure Models

2.2.1 Mixtures of Conditional Normal Densities

Analyses of mixed distributions suppose that the data are a composition of two or more populations mixed in different proportions (McLachlan, Basford (1988)). A conditional finite-mixture model for observed continuous dependent random variables y_i and continuous and/or dummy regressors x_i is defined as follows (Arminger, Stein (1997)):

$$h(y_i, x_i) = f(y_i|x_i)g(x_i) \tag{1}$$

where $h(y_i, x_i)$ is a multivariate density function for the identically and independently distributed data, $g(x_i)$ is an arbitrary marginal density of the regressor variables and the conditional density function $f(y_i|x_i)$ is specified as

$$f(y_i|x_i) = \sum_{g=1}^{G} \omega_g f_g(y_i|x_i). \tag{2}$$

The mixing components are denoted as g, where the number of components is usually unknown. For the mixing probabilities ω_g the following restrictions apply:

$$\sum_{g=1}^{G} \omega_g = 1 \tag{3}$$

$$\omega_g \geq 0 \qquad (g = 1, ..., G). \tag{4}$$

A multivariate normal density $f_g(y_i|x_i) = \Phi(y_i; \mu_{ig}, \Sigma_g)$ with expectation μ_{ig} and covariance matrix Σ_g is assumed for the group-specific conditional densities. The conditional mean for the different groups is parameterized as a reduced form regression model

$$E(y_i|x_i, g) = \mu_{ig} = \gamma_g + \Pi_g x_i. \tag{5}$$

The vector γ_g including regression constants, the matrix Π_g of regression coefficients and the covariance matrix Σ_g are parameterized in a vector $\vartheta \in \Theta \subset \mathcal{R}^d$, which leads to specific structures for the moments.

2.2.2 Mean- and Covariance Structure Models and Conditional Mixtures

A number of articles contributed to the development of finite-mixture models in SEM (DeSarbo and Cron (1988); Jones and McLachlan (1992); Yung (1997); Jedidi et al. (1997); Arminger et al. (1998)). Because the model by Arminger et al. is the most general we restrict our presentation to this approach.

Finite mixtures of conditional mean- and covariance structures extend the general theory on mixed distributions in two directions: (1) the set of observed variables is divided into dependent variables y and independent regressors x for which no distributional assumptions are made, and (2) the group specific expectations and covariances can be arbitrarily parameterized, e.g. as factor-analytic models or SEM with and without latent variables. A general conditional LISREL model for a latent variable vector η_i can be specified as (Arminger et al. (1998)):

$$\eta_i|(x_i, g) = B_g\eta_i + \Gamma_g x_i + \zeta_i^{(g)} \tag{6}$$

where $\zeta_i^{(g)} \sim N(0, \Psi_g)$. Assuming a specific measurement model for the vector of dependent variables y_i results in a conditional expected value

$$E(y_i|x_i, g) = \nu_g + \Lambda_g(I - B_g)^{-1}\Gamma_g x_i = \gamma_g + \Pi_g x_i, \tag{7}$$

and a conditional covariance matrix

$$V(y_i|x_i, g) = \Lambda_g(I - B_g)^{-1}\Psi_g(I - B_g)^{-1'}\Lambda_g' + \Theta_g = \Sigma_g. \tag{8}$$

An example for the general model is depicted in figure 1. As can be seen, two sources of heterogeneity are distinguished: the regressor variables x represent known causes of differences between the groups (dashed arrows). Unobserved latent causes of heterogeneity are partly modeled by γ_g (dotted arrows). In many cases, however, the sources of heterogeneity are unknown, and the general model reduces to an unconditional finite-mixture model. This special case of the model by Arminger et al. (1998) corresponds with the finite-mixture LISREL-type model described in Jedidi et al. (1997).

2.2.3 Identifiability and Estimation

Suppose a specified model is identifiable for a single group, or identifiability can be established by parameter restrictions across groups, then the finite-mixture SEM is identified if the variables follow a multivariate normal distribution in the unknown groups (for a proof see Jedidi et al. (1997); see also Titterington et al. (1985); McLachlan, Basford (1988)).

Arminger et al. (1998) present three different procedures for the parameter estimation: (1) a two-stage estimation procedure, (2) a direct algorithm,

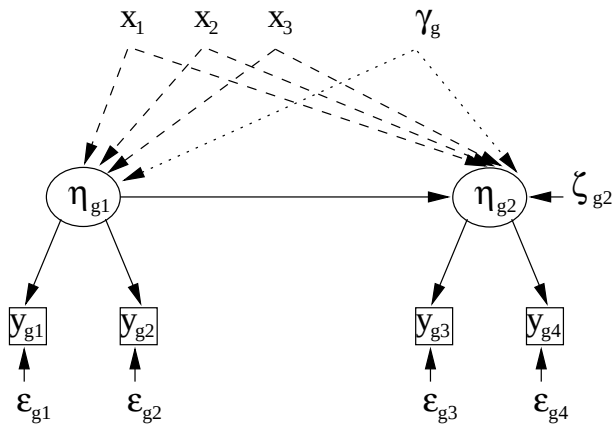

Figure 1: Representation of a Finite Mixture SEM

and (3) a gradient algorithm. The two-stage procedure starts with an EM estimation of the reduced-form parameters γ_g, Π_g and Σ_g and their asymptotic covariance matrix (Jones, McLachlan (1992)). In the second stage, the parameters of the structural model are estimated using minimum distance estimation (Arminger (1995)). In contrast to the two-stage approach, in the direct estimation procedure the model restrictions are taken into account at the level of the fuzzy group formation. Because reduced-form parameters (e.g., the covariance matrix Σ_g) are not provided, model specification only relies on theoretical considerations. Also, computation time is considerably higher but can be reduced by applying the gradient EM algorithm, where only one iteration is performed in each M-step (Becker et al. (1997)).

3 A Simulation Study

3.1 The Concept of the Simulation Study

The aim of this simulation study is to assess the performance of both the traditional approach and the finite-mixture methodology when the sample analyzed is heterogeneous. A heterogenous sample ($n = 20,000$) has been created by merging simulated data for two different groups, each of sample size 10,000. Both data sets have been generated by the software TETRAD II (Scheines et al. (1994)). The data for the two groups were simulated according to the path diagrams in figure 2. Both groups differ in their causal structures as well as in the structural parameters but have the same measurement models (parameter sets conform to typical values in empirical research). In addition, particular group means were specified (see table 1). All variables follow a multivariate normal distribution within the groups. We assume that no information about the sources of heterogeneity is available.

The data set has been analyzed using different approaches under specific

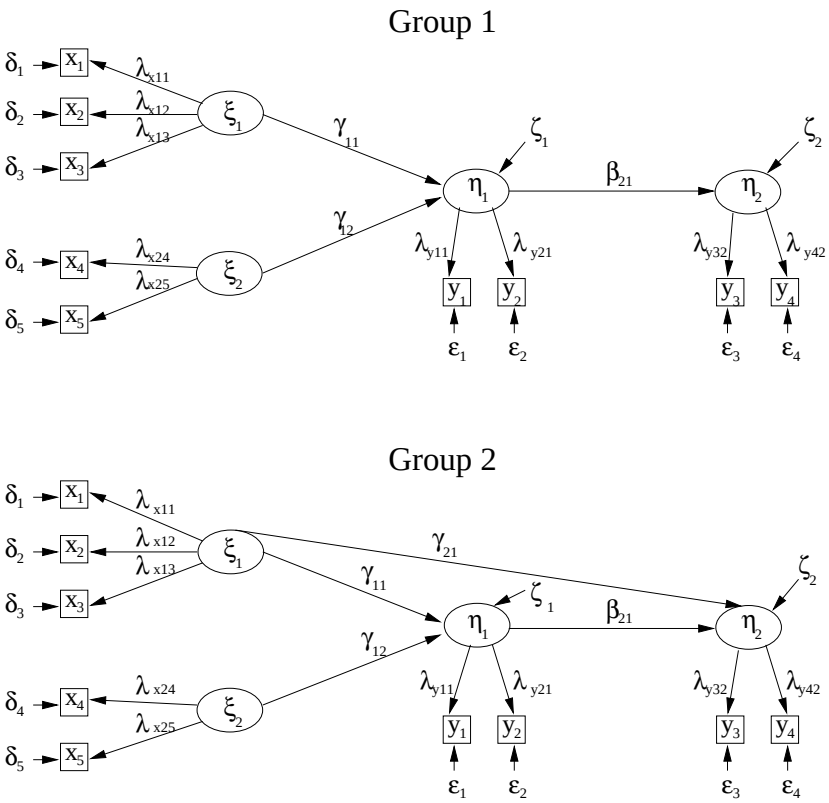

Figure 2: Path Diagrams for the Simulated Data Sets

Parameters	Group 1	Group 2	Means	Group 1	Group 2
γ_{11}	0.70	0.20	$\mu_{\xi 1}$	0.00	1.00
γ_{12}	0.60	-0.30	$\mu_{\xi 2}$	0.00	1.00
γ_{21}	0.00	0.80	$\mu_{\eta 1}$	0.00	1.00
β_{21}	0.60	0.30	$\mu_{\eta 2}$	0.00	1.00
λ_{x1}	1.00	1.00	μ_{x1}	0.00	1.00
λ_{x2}	0.90	0.90	μ_{x2}	0.00	1.00
λ_{x3}	0.80	0.80	μ_{x3}	0.00	1.00
λ_{x4}	1.00	1.00	μ_{x4}	0.00	1.00
λ_{x5}	0.90	0.90	μ_{x5}	0.00	1.00
λ_{y1}	1.00	1.00	μ_{y1}	0.00	1.00
λ_{y2}	0.90	0.90	μ_{y2}	0.00	1.00
λ_{y3}	1.00	1.00	μ_{y3}	0.00	1.00
λ_{y4}	0.80	0.80	μ_{y4}	0.00	1.00

Table 1: Parameters for the Simulated Models

heterogeneity assumptions (see table 2). As a benchmark for both the finite-mixture and the traditional approach we estimated a multigroup SEM based

on the covariance matrices for the original groups. In addition, a common SEM was estimated under the assumption of homogeneity.

Multigroup SEM is performed by the standard SEM software AMOS 3.6 (Arbuckle (1997)). The program MECOSA 3.01 (Arminger et al. (1996)) was applied for the finite-mixture model. K-means cluster analysis in SPSS 8.0 was used to form groups according to the traditional approach.

Approach	Assumption
SEM	Homogeneity
Multigroup SEM	Observed heterogenity
Traditional Approach Finite-mixture SEM	Unobserved heterogeneity

Table 2: Heterogeneity Assumptions of the Applied Approaches

3.2 Results of the Simulation Study

In order to assess the performance of the different procedures, both the recovery of the parameters and measures of overall fit (χ^2, AIC, RMSEA, AGFI) were used (see table 3). The χ^2 test statistic and the AIC indicate an acceptable fit for the multigroup and the finite-mixture SEM, whereas the model for the traditional approach is just rejected, as is the common SEM under a homogeneity assumption. With respect to RMSEA and AGFI the picture is more ambiguous. Although both measures show that the multigroup SEM and the finite-mixture model are preferable, the two other models are not strictly rejected. The low RMSEA for the traditional approach even indicates a good fit (Browne, Cudeck (1993)). These results support the view that some of the widely-used fit measures are not sensitive enough to detect heterogeneity problems and that information criteria like AIC, CAIC, or BIC should be used instead (Jedidi et al. (1997)).

Because the original measurement models did not differ across groups, we obtained similar results for all procedures and factor loadings were well reproduced (we therefore do not report these results, however, they are available upon request). With respect to the structural parameters, the most biased estimates resulted for the common SEM as one would expect (see especially the parameter estimates for γ_{12} and γ_{21}). Even the traditional approach produced poor results compared to the finite-mixture model. For group 1 a significant parameter γ_{21} was estimated, even though the original value is zero. The estimation of the finite-mixture SEM led to parameters that were similar to those from the multigroup SEM, with only negligible deviations from the original values.

	SEM	Multigroup SEM	Traditional Approach	Finite Mixture SEM
Model fit				
AIC				
- estimated model	1,434.06	126.20	837.79	125.40
- saturated model	90.00	180.00	180.00	180.00
- independence model	144,245.45	86,623.66	73,350.96	86,656.23
χ^2	1,388.06	44.20	755.59	43.40
- df	22	49	49	49
- p	0.00	0.67	0.00	0.70
RMSEA	0.06	0.00	0.03	0.00
AGFI	0.97	1.00	0.98	1.00
Parameter estimates				
group proportions				
- ω_1			0.3684	0.5011
- ω_2			0.6316	0.4989
structural parameter group 1				
- γ_{11}	0.56	0.69	0.64	0.69
- γ_{12}	0.30	0.59	0.54	0.59
- γ_{21}	0.72	-0.01 (n.s.)	-0.08	-0.01 (n.s.)
- β_{21}	0.55	0.59	0.42	0.59
structural parameter group 2				
- γ_{11}		0.19	0.27	0.18
- γ_{12}		-0.31	-0.06	-0.31
- γ_{21}		0.79	0.85	0.79
- β_{21}		0.30	0.39	0.31

Table 3: Simulation Results

4 Summary

The results of the simulation study clearly show that the finite-mixture approach outperforms the traditional sequential procedure in the case of unobserved heterogeneity. Therefore, finite-mixture SEM offer a promising methodology for segmentation tasks in market research. In practice, however, data requirements are formidable. In order to get consistent estimates, relatively large data sets are necessary (at least 1000 observations for moderately complex models; see Jedidi et al. (1997)). With the exception of consumer panel data, survey samples are normally relatively small. Further simulation studies should examine the properties of the finite-mixture approach in small samples. In addition, deviations from the normality assumption should be analyzed, with and without exogenous regressors.

References

ARBUCKLE, J. (1997): AMOS User's Guide, Version 3.6. SmallWaters Corporation.

ARMINGER, G. (1995): Specification and Estimation of Mean Structures: Regression Models. In: Arminger, G., Clog, C.C., and Sobel, M.E. (Eds.), Handbook of Statistical Modeling for the Social and Behavioral Sciences, Plenum, New York, 77–183.

ARMINGER, G. and STEIN, P. (1997): Finite Mixtures of Covariance Structure Models with Regressors. Technical report, Bergische Universität Wuppertal and Gerhard Mercator Universität Duisburg.

ARMINGER, G., STEIN, P. and WITTENBERG, J. (1998): Mixtures of Conditional Mean- and Covariance Structure Models. Technical report, Bergische Universität Wuppertal and Gerhard Mercator Universität Duisburg.

ARMINGER, G., WITTENBERG, J. and SCHEPERS, A. (1996): MECOSA 3 User Guide. ADDITIVE GmbH.

BECKER, M., YANG, I. and LANGE, K. (1997): EM Algorithms without Missing Data. *Statistical Methods in Medical Research, 6, 37–53.*

BROWNE, M.W. and CUDECK, R. (1993): Alternative Ways of Assessing Model Fit. In: Bollen, K.A. and Long, J.S. (Eds.), Testing Structural Equation Models, Sage, Newbury Park, 136–162.

DESARBO, W.S. and CRON, W.L. (1988): A Maximum Likelihood Methodology for Clusterwise Linear Regression. *Journal of Classification, 5, 249–282.*

HILDEBRANDT, L. and HOMBURG, C. (Eds.) (1998): Die Kausalanalyse: Instrument der empirischen betriebswirtschaftlichen Forschung. Schäffer-Poeschel, Stuttgart.

JEDIDI, K., JAGPAL, H. and DESARBO, W. (1997): Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. *Marketing Science, 16, 39-59.*

JÖRESKOG, K. and SÖRBOM, D. (1996): LISREL 8: User's Reference Guide. Scientific Software International, Inc.

JONES, P. and MCLACHLAN, G. (1992): Fitting Finite Mixture Models in a Regression Context. *Australian Journal of Statistics, 32, 233-240.*

MCLACHLAN, G. and BASFORD, K. (1988): Mixture Models: Inference and Applications to Clustering. New York, Marcel Dekker.

SCHEINES, R., SPIRTES, P., GLYMOUR, C. and MEEK, C. (1994): TETRAD II: Tools for Causal Modelling. Hillsdale, Lawrence Erlbaum.

TITTERINGTON, D., SMITH, A. and MAKOV, U. (1985): Statistical Analysis of Finite Mixture Distributions. Wiley, Chichester.

YUNG, Y.F. (1997): Finite Mixtures in Confirmatory Factor-Analysis Models. *Psychometrika, 62, 297-330.*