WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Graphical Markov Models as an alternative to SEM

EMAC 2013
Modelling and Forecasting

Kathrin Gruber, Radoslaw Karpienko, Thomas Reutterer
June 6, 2013

Contents

- Overview of SEM in Marketing Research
- Introduction to Graphical Markov Models and their relation to SEM (and PLS)
- An empirical application: the CRAN Motivation Survey

The Role of SEM in Marketing Research (1)

Structural Equation Modeling (SEM) is a very popular tool for theory testing in marketing and behavioral sciences (Steenkamp & Baumgartner, 2000; Baumgartner & Homburg, 1996; Hulland et al., 1996)

- SEM manage the inclusion of multiple endogenous / exogenous constructs
- SEM account for measurement error in the latent constructs
- Sound theoretical assumptions are confronted with their "fit" with directly observable data (both structural and measurement model)
- Two SEM-philosophies: Covariance-based SEM vs. Variance-based partial least squares SEM (cf. Hair et al., 2012)

The Role of SEM in Marketing Research (2)

- Underlying theoretical justification of SEM models are not always so sound (sometimes they are very weak)
- SEM practice often degenerates to an "exploratory device" for identifying "best" model fitting empirical data (in particular when it comes to "adjust" the measurement models)
- Formal assumptions:

	Cov.-based **SEM**	Var.-based SEM (**PLS**)
Assumptions	multivariate Norm.	no distributional assumptions (appl. for nom., ord. & cont.)
Estimation	ML (or GLS/WLS)	Componentwise
Models	multivariate	uni- & multivariate
Meas. Mod.	factor scores	factor scores
Error Corr.	Bootstrap	Bootstrap
R-packages	lavaan, sem, ...	semPLS, pls, plspm, ...

Graphical Markov models (1)

The approach based on graphical Markov models (in particular DAG) imposes less rigorous assumptions.

- ... are multivariate statistical models, where a graph G ($G = (V, E)$) describes independence statements in the joint distribution
  - **V**: r.v. are denoted by nodes (discrete or continuous)
  - **E**: cond. association parameters in the distribution are represented by edges

A Markovian model is equivalent to a recursive model in SEM.

Example:

$$Z \rightarrow Y \leftarrow X$$

cond. independence rel. $X \perp Z | Y$

$$f(x, y, z) = f(x) \underbrace{f(y|x) f(z|y)}_{linear\ \ regressions}$$

Graphical Markov models (2)

- An initial sequence of the r.v. is defined by (e.g.) research hypotheses (**dependence chain** or joint response chain graph $\mathcal{G} = \{V(1), \ldots, V(p)\}$, i.e., an ordered disjoint partitioning of V)

Example:

$V(1)$		$V(2)$		$V(3)$		$V(4)$
A	$\leftarrow$	B, C	$\leftarrow$	D, E	$\leftarrow$	F, G, H

- All variables are assigned to a higher order component of G are considered conditionally on the prior components
- The density factorizes to

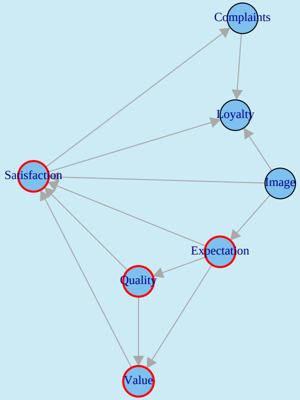
$$f_V = f_{V_p|V_{p-1}\ldots V_1} \cdot f_{V_{p-1}|V_{p-2}\ldots V_1} \cdot \ldots \cdot f_{V_2|V_1} \cdot f_{V_1}$$

Graphical Markov models (3)

- If the variables can be ordered in a recursive response structure we call the **graph a directed acyclic graph** $\mathcal{G}_{dag}$ (variables are neither explanatory nor responses for itself).
- The density f admits a recursive factorization according to $\mathcal{G}$ (y is a random vector)
$$f(y) = \prod_{p \in V} f(y_p | y_{pr(p)})$$
- Models of this type can be constructed via a set of **univariate cond. models**.

The ECSI (European Customer Satisfaction Index) Study (1)

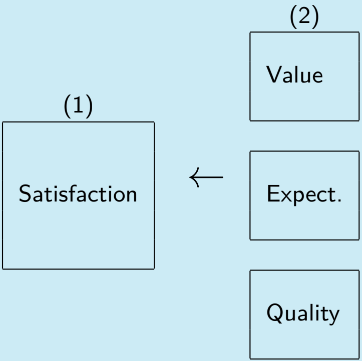
Ex.: Structural model describing causes and consequences of customer satisfaction.



(cf. Tenenhaus et al., 2005)

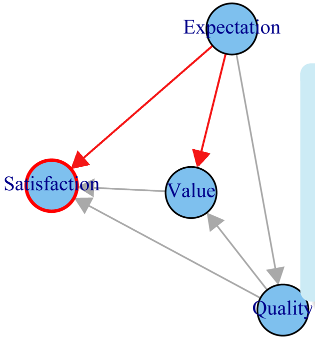
The ECSI (European Customer Satisfaction Index) Study (2)

dependence chain



- First block: all variables in the second block are regressed as independent ones on satisfaction
- Second block: all variables are alternatingly considered as dependent and independent

The European Customer Satisfaction Index (3)



- SEM & PLS: same results (cf. Tenenhaus et al., 2005)
- DAG: same results $f(s,e,v,q) = f(s|v,q)f(v|q)f(e|q)f(q|v,e)$
- Automatic model selection only in DAG (Coefficients for Expectation are very small, $p > 0.05$)

DAG advantages:

- Equation parameters are regression coefficients (interpretation of structure in terms of independencies).
- No overparametrization and consequently no problems of identification.
- General results are available to read directly off the graph of the model.

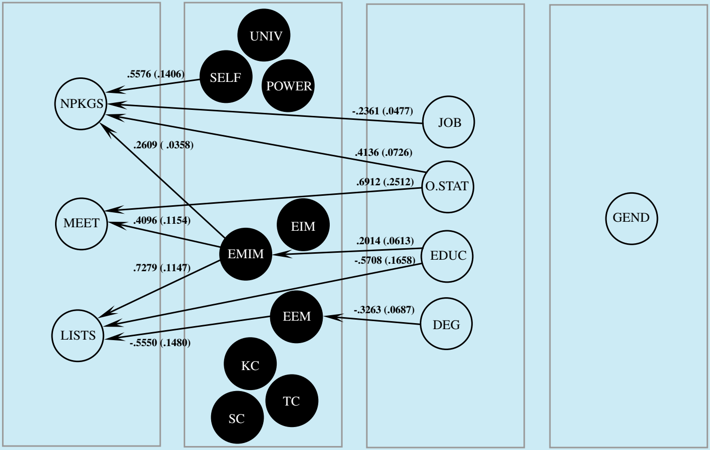
The CRAN Motivation Survey

- Conducted in April/May 2010 to study software developers' motivation to contribute to the OSS project the R-project for statistical computing.
- 4,274 authors of R packages were contacted via Email (CRAN, R-Forge, Bioconductor).
- 782 contributors completed the questionnaire consisted of 120 items (incl. different forms of individual participation and potential motivational drivers).

Dependence chain: Initial ordering of the response (1), intermediate (2 and 3) and purely explanatory (4) variables

(1)		(2)		(3)		(4)
NPKGS MEET (1 = Yes) LISTS (1 = Yes)	$\leftarrow$	UNIV POWER SELF EEM EMIM EIM TC KC SC	$\leftarrow$	JOB (1 = academics) O.STAT (1 = full) EDUC (1 = statistics) DEG (1 = postdoc)	$\leftarrow$	GEND

Model graph of the final graphical chain model. Binary variables are depicted as circles and numerical variables as dots (NPKGS is a discrete poisson distributed variable). Arrows indicate significant relationships.



Summary & Discussion

- The presented kind of Markovian models is particularly interesting for social and behavioral sciences (observational studies, complex multivariate dependencies, existing substantive knowledge)
- Combination of graphical Markov model technique, model building and methods for scaling provide a useful alternative to SEM.
- Only an ordinal structure behind the model has to be specified (no theoretical restrictions on the form of the conditional distributions)
- Variable of mixed measurement scale types can be modeled both within and between levels.

Literature

- Cook, J., & Stefanski, L. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association, 89*(428), 1314-1328.
- Cox, D., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation* (Vol. 67). Chapmann & Hall/CRC.
- De Ayala, R. J. (2008). *The theory and practice of item response theory.* The Guilford Press.
- Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge University Press.
- Steenkamp, J., & Baumgartner, H. (2000). On the use of structural equation models for marketing modeling. *International Journal of Research in Marketing, 17*, 195-102.
- Tenenhaus, M, & Vinzia, V., E. & Laurob, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis, 48*, 159–205.

Appendix - Model Selection (1)

- By the model selection algorithm of Cox & Wermuth (1996) (heuristic based on backward and forward selection)
- At each step a variable is regressed on all variables belonging to a chain component with a lower number (**univariate conditional regression**)
- Performed for every (potential) response variable to break up complex structures into tractable subcomponents

Non-recursive linear models in SEM are equivalent to block recursive regression models (Lauritzen & Wermuth, 1990)

Appendix - Model Selection (2)

Step 1

Screening for interactions and nonlinear relations by forward selection (full model)

$\Downarrow$

Regression based on main effects (+ nonlinear terms and/or interaction)
using backward selection strategy leads to a reduced model

$\Downarrow$

Step 2

Check for interactions and nonlinear relations based on the reduced model, again backward selection leads to a even more reduced model

$\Downarrow$

Check for interactions and nonlinear terms, again backward selection leads to the finally selected model

Appendix - The CRAN Motivation Survey Measurement Model

- Latent variables are estimated by the use of **item response theory**.
  - Each subject is mapped on a latent dimension by estimating a person parameter (corresponds to factor scores from FA).
  - Different methods are possible (for an overview see de Ayala, 2008)
- additionally: variables are corrected for additive measurement error by the use of **simulation extrapolation method** (SIMEX, Cook & Stefanski, 1994)