Graphical Markov Models as an alternative to SEM

Abstract

We propose the use of directed acyclic graphs (a subclass of graphical Markov Models) as an alternative to structural equation models. The method imposes less rigorous assumptions on the data and the underlying theoretical framework The relationships between the latent and structural variables can be incorporated in a multivariate model, by specifying only an initial ordering. Thus, they appear to be particularly suitable for situations, when the theoretical foundations are weak or ambiguous. We demonstrate the potential capabilities of the methodology using data collected from an exploratory study on the motivational drivers of voluntary contributors to the R-project.

Keywords: *Structural equation models, graphical models, directed acyclic graphs, marketing modeling*

Track: *Modeling and Forecasting*

1 Introduction

In the past decades, the application of structural equation models (SEMs) has been steadily growing in marketing and the behavioral sciences (Baumgartner & Homburg, 1996, Steenkamp & Baumgartner, 2000). SEMs are confirmatory modeling techniques for simultaneously assessing and testing hypothesized interrelationships among latent variables and their corresponding measurement models. In the SEM-framework, theory serves as a conceptual framework for developing a model which implies a covariance structure and is able to detect latent factors in theoretical constructs (Bollen, 1989).

However, SEMs require all posited relationships (even the structure of observed variables) to be well specified before the model can be estimated. They are typically also very sensitive to violations of the underlying assumptions (e.g., if there are nonlinear relationships) and can produce misleading results even if the assumed interrelationship structure is only partially false. Thus, the use of SEMs is limited to applications with a very strong and substantial theoretical foundation. Another issue with SEMs is that in empirical research practice, an initial model is often estimated based on theory, but later on modified in order to improve model fit with the observed data. In fact, most of the commonly used software packages offer some kind of "modification indices" to support the model user in this respect (for critics of this practice see, e.g., MacCallum, Roznowski & Necowitz 1992; Spirtes, Glymour & Scheines, 2000).

This paper introduces an approach, which imposes less rigorous assumptions on both the data and the underlying theoretical framework. The proposed modeling technique is based on directed acyclic graphs (a special case of graphical Markov models). The approach is combined with the features of item response theory (IRT) for measuring latent variables. Similar to confirmatory factor analysis, the IRT framework allows to interpret the responses to the measurement items as observable manifestations of hypothesized constructs. To avoid measurement error the model variables are error corrected by the use of simulation and extrapolation (SIMEX). The combination of these modeling components can be used to explain interdependencies when the theoretical foundations of the underlying study area are weak or ambiguous.

The next section introduces the building blocks of the proposed methodology. To empirically illustrate the capabilities of the modeling technique, we present results from a study, which aims to detect motivational patterns for software developers to contribute to the Open Source Software (OSS) project R.

2 Graphical modeling

Graphical Markov models are multivariate statistical models where a graph describes independence statements in the joint distribution. In the setting of graphical Markov Models p nodes $V = \{1, \dots, p\}$ in a graph denote random variables $Y_1, \dots, Y_p$ and there is at most one edge i, j between each pair of nodes i and j. The edges represent conditional association parameters in the distribution of Y_V. The variables in this so-called independence graph may be categorical and modeled by discrete random variables, or numerical and modeled by continuous variables.

Research hypotheses (or any other substantive knowledge) are used to specify an ordered sequence of the variables starting with purely explanatory (independent) variables and ending with one ore more responses. The variables are arranged in a joint response chain graph $\mathcal{G}$ into subsets (typically displayed by boxes) from left to the right, starting with responses of primary interest.

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

Equation 1 is an ordered partition of the set V of all p nodes into subsets as $(a, b, c, \dots)$ to a dependence chain. If it is possible to order the variables in a way that some are responses to others in a recursive response structure this type of graphical models is called a directed acycilc graph $\mathcal{G}_{dag}^{V}$ (Cox & Wermuth, 1996; Wermuth & Cox, 2004; Pearl, 2000). The variables are neither explanatory nor responses for itself and thus all edges in the independence graph $\mathcal{G}_{dag}^{V}$ are directed and there is no direction-preserving path from a node back to itself. This graph defines Y_p to be independent of remaining ancestors. The density f of the distribution P admits a recursive factorization according to $\mathcal{G}$ such that

$$f(y) = \prod_{p \in V} f(y_p | y_{pr(p)} \tag{2}$$

where y is a random vector. In this way the joint distributions are decomposed recursively into conditional joint distributions and simplified by conditional independencies. Additionally this greatly reduces the amount of computation when calculating the density $f(y)$ of a $\mathcal{G}_{dag}^{V}$. The next subsections explains how the relationships of the graphical model can be estimated.

2.1 Structural part

As mentioned, the variables are assigned into disjoint blocks based on substantive knowledge and ordered such that all variables in a later chain component are considered conditional on the prior components. Now an appropriate model selection technique is employed to identify a subset $J \subseteq K, K = \{1, \dots k\}$ such that all coefficients $\beta_j, j \in J$ are different from zero and the remaining $\beta_i, i \in K \setminus J$ are equal to zero.

In this paper we use a heuristic strategy introduced by Cox & Wermuth (1996), which is based on the calculation of univariate regression models as a sequence of conditional distributions. This strategy can roughly be divided into two steps where at each step a variable $Y_p \in V$ is regressed on all other variables belonging to a lower-level chain component. In the first step of the algorithm there is a screening for two-way interactions (in all possible trivariate models) and nonlinearities (quadratic terms) are applied to all univariate generalized linear regressions. Only significant effects beyond a specified threshold (e.g., $p < 0.01$) are included in the respective models. In a second step, a backward selection procedure for regressions with main effects, nonlinear terms and/or interaction terms (different selection criterions are possible, e.g., AIC, BIC, deviance, Wald-test) is used to reduce the set of variables. Finally, there is another check for interactions and nonlinearities based on the reduced model, i.e., including all two-way interactions and quadratic terms for those variables and effects that have been selected in the previous step. Again backward selection leads to the final model. This procedure is performed for every (potential) response variable to reduce complex structures into tractable subcomponents.

This type of graphical Markov models has been successfully applied in many domains (Cox & Wermuth, 1996; Edwards, 1995; Lauritzen, 1996; Oliver & Simth, 1990; Sprites et al., 1993), but it requires the accommodation of latent variables to make it fully usable as an alternative to SEM. This can be accomplished by integrating IRT and a method for bias correction of the estimates to cover and supplement the features of confirmatory factor analysis within the framework of SEM. Additionally, it is also possible to keep or drop relationships that should or should not be included in the model in the estimation process of the chain graph.

2.2 Latent part

Similar to an SEM-setting the variables used in the graphical Markov model can be conceptualized as latent variables by the use of IRT (see, e.g. De Ayala, 2008, for an overview). Latent variables are not directly observed and must be inferred from manifest responses. A person parameter is estimated for each subject, which maps the subject on a latent dimension. These parameters correspond to factor scores as known from factor analysis. The person parameters as well as their standard errors are used as input for the graphical modeling procedure.

Because of the latent variable approach and our inability to directly observe the variables of interest, the regression framework we use could result in seriously biased parameter estimates. To account for measurement error we use the SIMEX method (Cook & Stefanski, 1994), which allows us to correct the effect estimates in the presence of additive measurement error. This method is especially helpful for complex models with a simple measurement error structure.

The SIMEX-method uses the relationship between the size of the measurement error σ_u^2 and the bias of the effect estimator when ignoring the measurement error, that defines the function

$$\sigma_u^2 \longrightarrow \beta^\star(\sigma_u^2) := \mathcal{F}(\sigma_u^2) \tag{3}$$

where $\beta^\star$ is the limit to which the naive estimator converges as the sample size $n \to \infty$. Although there is measurement error in the data, the SIMEX method approximates the function $\mathcal{F}(\sigma_u^2)$ by a parametric approach $\mathcal{F}(\sigma^2, \Gamma)$.

The combination of all of these statistical methods for scaling and model building can provide a useful and powerful alternative to SEM. It can be used to identify interdependency structures among (ordered) subsets of variables even in the extreme case of absence of any plain underlying theory. The next section illustrates this capacity in an empirical application.

3 Empirical application

Studying software developers' motivations to contribute to OSS projects is not an easy and straightforward task. There are many external factors and hybrid forms of motivation that might potentially play a role and have to be taken into account when one wishes to explain OSS contributions. Thus, previous empirical findings in this research area are rather limited and partially ambiguous (Roberts et al., 2006). In this study, we demonstrate the modeling approach presented above using data collected from the CRAN survey, which was conducted to examine the motivation of voluntary contributors to the R project. The survey was conveyed on the popular platforms 'CRAN', 'R-Forge', and 'Bioconductor'.

The directly observable responses (dependent variables) of primary interest are different forms of individual participation in the project. In addition, a wide range of potential motivational drivers (multi-item scales) was compiled based on an exhaustive literature review (Morgeson & Humphrey, 2006; Schwartz, 1992; Reinholt, 2006). The final questionnaire consisted of 120 items and also included some socio-demographic variables. The survey was conducted in April/May 2010. 4,274 authors of R packages (OSS contributors) were contacted via Email and asked to participate in the online survey. By the end of May 2010, 782 OSS contributors completed the full questionnaire.

Our literature review on potentially relevant explanatory variables of OSS participation suggests a first ordering of the studied variables as depicted in Tab. 1. There are three response variables of primary interest: The number of published packages (NPKGS), participation to mailing lists (LISTS) and to meetings (MEET). Several variables serve as "intermediates", because they can be considered as potentially explanatory for the higher-ordered variables (NPKGS, LISTS, and MEET) and as a response to others lower in the ordered block structure. These intermediates are: Task characteristics (TC), knowledge characteristics (KC), social characteristics (SC), extreme extrinsic motivation (EEM), well internalized extrinsic motivation (EMIM), extreme intrinsic motivation (EIM), universalism (UNIV), power (POWER) and self direction (SELF). The intermediates are all scaled by using a two-parametric logistic model (2PL; Birnbaum, 1968) and define the latent motivational dimension. The set of purely explanatory variables, i.e., occupational status (O.STAT), field of education (EDUC), degree (DEG), job category (JOB) and gender (GEND), is represented at the far right end of the chain. The effects on the response variables are analyzed for all variables listed in the boxes to the right.

We fitted the chain graph to the data using the model selection procedure suggested by Cox & Wermuth (1996). The Wald-test criterion was used in the backward selection and only

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

Table 1: Initial ordering of the response (block 1), intermediate (blocks 2 and 3) and purely explanatory (block 4) variables as dependence chain.

strong effects ($p < 0.01$) are included in the respective models. The procedure was performed for every response variable. The effects of the resulting univariate generalized linear regression models on the response variables are represented as edges in the final model graph (see Fig. 1). For sake of model simplicity block regression was used where associations within the groups of pure responses, as well as pure explanatories have been omitted (cf. Wermuth, 1998).

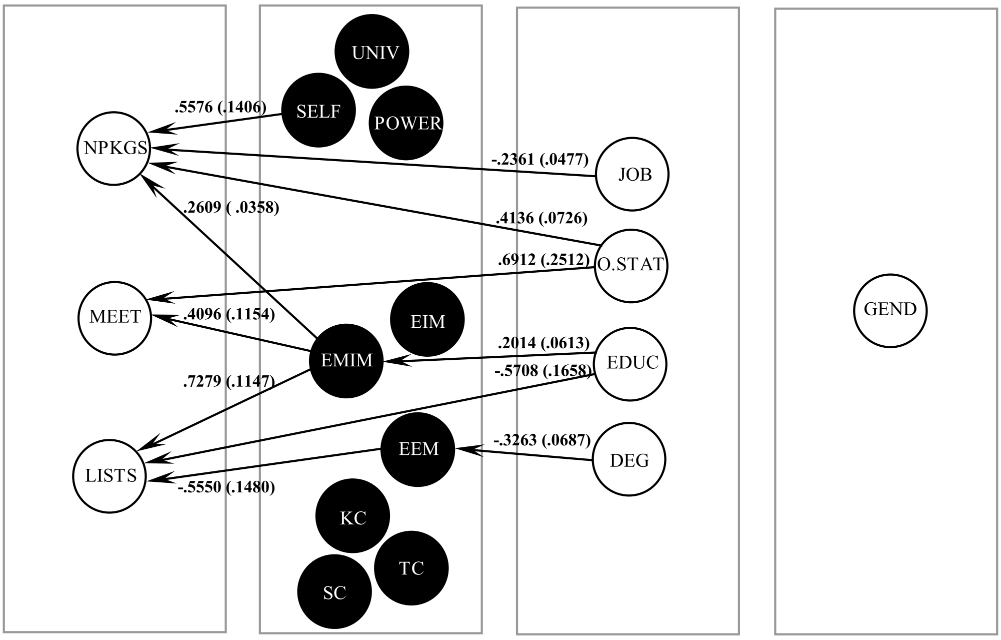

Figure 1: Model graph of the final graphical chain model. Binary variables are depicted as circles and numerical variables as dots (NPKGS is a discrete poisson distributed variable). Arrows indicate significant relationships, the effect estimates are displayed above (plus standard errors in brackets).

4 Discussion and Summary

In the present application of the proposed method based on directed acyclic graphs, the final graphical model depicts main drivers of contributing to the R project. The effects from the second on the first block of variables can be directly interpreted. The effects of nominal variables in the third block can be interpreted as a difference in means, whereas negative values indicate higher means for the one-encoded group. The effects of EMIM on all three response variables imply that hybrid forms of motivation (well internalized extrinsic / intrinsic motivation) is a main driver of voluntary contributing to OSS projects. Also SELF shows a positive effect on NPKGS. Thus, the more self directed software developers are, the more packages they tend to release. The significant effect of DEG on EEM provides evidence that prae-doc's are more likely to be external regulated and post-doc's to be more introjected. The direct effect of JOB on NPKGS indicates that academics are more likely to contribute than others (Henkel, 2006). Additionally, the effect of EDUC on LISTS shows that statisticians subscribe to more mailing lists. The effects of O.STAT on NPKGS and MEET indicate that contributors who work part time may develop more packages and attend more meetings. This supports the finding from literature that contributors to OSS are typically hobbyists who contribute in their free time (Shah, 2006). In summary, the final graphical model captures many results which are consistent with previous findings reported in the relevant literature.

References

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, *13*(2), 139 - 161.

Birnbaum, A. (1968). Some latent trait models and Their Use in Inferring an Examinee's Ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores.* London: Addison Wesley.

Bollen, K. (1989). *Structural equations with latent variables.* Wiley New York.

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). Chapman & Hall/CRC.

De Ayala, R. J. (2008). *The theory and practice of item response theory.* The Guilford Press.

Edwards, D. (1995). *Introduction to graphical modelling.* Springer Verlag, New York.

Henkel, J. (2006). Selective revealing in open innovation processes: the case of embedded linux. *Research Policy*, *35*(7), 953–969.

Lauritzen, S. (1996). *Graphical models.* Oxford University Press.

Lederer, W., & Küchenhoff, H. (2006). A short introduction to the simex and mcsimex. *The Newsletter of the R Project Volume 6/4, October 2006*, 26.

MacCallum, R., Roznowski, M., & Necowitz, L. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin; Psychological Bulletin*, *111*(3), 490.

Morgeson, F., & Humphrey, S. (2006). The work design questionnaire (wdq): Developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology*, *91*(6), 1321–1321.

Oliver, R., & Simth, J. (1990). *Influence diagramms, belief nets and decision analysis.* Wiley, London.

Pearl, J. (1998). Graphs, causality and structural equation models. *Sociological Methods and Research*, *27*, 226–284.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge University Press.

Reinholt, M. (2006). *No more polarization, please! Towards a more nuanced prespective on motivation in organizations)* (Tech. Rep.). Center for Strategic Managment Working Paper Series: Copenhagen Business School, Copenhagen, Denmark.

Rizopoulos, D. (2006). ltm: An r package for latent variable modelling and item response theory analysis. *Journal of Statistical Software*, *17*(5), 984–999.

Roberts, J., Il-Horn, H., & Sandra, A. (2006). Understanding the motivations, participations and performance of open source software developers: A longitudinal study of the apache projects. *Management Science*, *52*(7), 984–999.

Schwartz, S. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology*(25), 1–65.

Shah, S. (2006). Motivation governance and the viability of hybrid forms in open source software development. *Management Science*, *52*(7), 1000–1014.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search. adaptive computation and machine learning.* MIT Press, Cambridge.

Sprites, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction and search.* Springer-Verlag, New-York.

Steenkamp, J., & Baumgartner, H. (2000). On the use of structural equation models for marketing modeling. , *17*, 195–102.

Stefanski, L., & Cook, J. (1995). Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, *90*(432), 1247–1256.

Wermuth, N. (1998). Graphical markov models. *Encyclopedia of Statistical Science*, *2*.

Wermuth, N., & Cox, D. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(3), 687–717.