Social science application of graphical models on mobility data

Summary of the PhD dissertation of

Renáta Németh

Supervisor:	Doctoral Program in Sociology
Tamás Rudas, DSc	Eötvös Loránd University
Department of Statistics	Faculty of Social Sciences

Budapest, 2009

Motto:

> *"There has been a revolution in the study of social mobility: the once dominant Blau-Duncan paradigm has been overthrown by log-linear modeling. [...] The log-linear revolution was a noble experiment, and at first seemed to offer a bright new future beyond Blau and Duncan. [...] In the end, I believe Goldthorpe's experiment failed. A decade and more has passed. Endless models have been fitted; legions of design matrices passed in review; chi-squares marshaled. But we have learnt little new of substance. Frequencies in a crude cross-classification of father's occupation by son's occupation are described in loving detail, with some ad-hoc interpretation but little theory. [...] This compares poorly with the statistically simpler but conceptually more sophisticated results of the Blau-Duncan paradigm."*
>
> Jonathan Kelly: The failure of a paradigm (1990)

Research focus

Graphical models are models based on graphs in which nodes represent random variables, and the (lack of) edges represent conditional independence assumptions. Hence they provide a compact representation of joint probability distributions. Three graph types are distinguished which differ with regard to their edges. A graph with only undirected edges (lines) is called an **undirected graph** (UG**.** Variables in this case are considered on equal footing. Applications of UGs (also called Markov random fields) include models for spatial statistics and image analysis. In some cases the role of variables is not symmetrical, that is *X* may affect *Y* but *Y* may not affect *X*. In such situations the use of UGs would be unnatural; which motivated the introduction of **directed acyclic graph**s (DAGs). DAGs have directed edges (arrows), but contain no directed cycles. DAG models are also called Bayesian networks or belief networks and occur in artificial intelligence, genetics and many other fields. This work was extended to models permitting both directed and undirected edges, such graphs are called **chain graphs** (CGs). CGs are used when undirected associations are also allowed, e.g. if it is not known, whether (1) *Y* affects *X* or *X* affects *Y*, or (2) the association between *X* and *Y* is caused by an unmeasured third variable.

The study of these models is an active research area, with many questions still open. Some of them are answered in my dissertation, which deals with **graphical models on categorical data**. Continuous data with normality assumed and categorical data turned out to be quite different from the perspective of graphical modeling. The main difference is that the pairwise independence statements $A \perp B$ and $A \perp C$ imply $A \perp BC$ only in the case of normally distributed variables but not in the case of categorical data. That is why it is meaningful to introduce many different rules (Markov properties) in the categorical case to read-off independences from the graph. E.g. if there is a missing edge between *A* and *B* and another between *A* and *C*, then the pairwise Markov property implies only $A \perp B$ and $A \perp C$, while the local Markov property implies $A \perp BC$ as well. Under the assumption of normality the two rules do not differ.

Basic question of my dissertation is **how to parameterize the contingency table** in order to easily define graphical models on it. This is an essentially important issue, since (1) a graphical model could be given by restricting the parameters (preferably in an easy way, setting some of them to zero). Additionally, (2) it is very often the case that one moves from a poorly fitting model to a slightly less restrictive model in the hope of a better model fit, by drawing new edges in the graph, step by step. Or on the contrary, the initial model is a well fitting one, and one searches for a slightly more restrictive model in the hope of a still good model fit, by removing edges in the graph. One can obtain the possible new model by omitting some (or introducing new) parameter restrictions. Finally, (3) interpretation of a model is usually based on estimated parameter values. In either case, the parameters the researcher uses to describe the models will have a great impact on the result. By using improper parameterization it can occur that (1) it is not clear how to restrict the parameters in order to define the model, (2) in some stage of the model selection procedure it is not clear how to translate the new (removed) edge into the language of parameter restrictions.

These problems can be avoided by defining graphical model with marginal loglinear parameters. As it is readily implied by the results of Bergsma and Rudas (2002), parameterizations that meet some simple combinatorial assumptions assures that missing edges correspond to some of the parameters being zero, and the parameters of the model can be interpreted in a straightforward way. These parameterizations have further important properties concerning existence, estimation and testing of models.

Sociological application of graphical models may arise primary in the field of social mobility, concerning models of status attainment process. These approaches typically describe social dynamics with casual process models, using categorical data (social position, educational level etc). Returning to the motto cited from Kelly: using graphical models both the approach of Goldthorpe and Erikson (defining occupational status as class position by using categorical variables) and the aim of Blau and Duncan (modeling mobility pathways via direct and indirect effects) can be achieved.

New scientific achievements

1.

Graphical models based on directed acyclic graphs can be parameterized with marginal loglinear parameters in a way that (1) the parameters are smooth and variationally independent from each other (consequently, they may be interpreted individually, without reference to the other parameters), (2) linear restrictions on the values of these parameters do not lead to contradictions, and (3) the models have standard asymptotic behavior.

(Publication: Rudas, Bergsma, Németh, 2006. In the dissertation: Chapter 4.5)

2.

Seven classes[1] of CG models do not have a hierarchical marginal loglinear parameterization. These models correspond to the light cells of the table below.

(Chapter 4.6)

Table 1 Types of CG models, in case of categorical data and positive distribution. Dark background: each models are smooth, light: some of them are non-smooth, white: not known, =: equivalence, ⇒: implication.

Type	*Markov property*			
1 (LWF, most conditional)	Global =	Block-recursive =	Local =	Pairwise
2 (AMP)	Global =	Block-recursive ⇒	Local =	Pairwise
3		Block-recursive ⇒		Pairwise
4 (most marginal)		Block-recursive ⇒		Pairwise

[1] The 4X4 table below would result in sixteen models, but the literature introduces only twelve of them. Empty cells of the table correspond to the non-introduced models. Five out of the twelve are equivalent with some of the remaining seven models.

3.

If one tries to parameterize the other four model classes (white and dark cells of the table above) in a straightforward way, by parameterizing the implied $S_1, S_2, ..., S_n$ conditional independence statements within the corresponding $M_1, M_2, ..., M_n$ marginals, then the parameterization obtained

- is not hierarchical in general (parameters may be incompatible: the same effect should be set to zero within two different marginals, or a non-zero effect pertaining to a smaller marginal should be set to zero within a greater marginal)
- is not ordered decomposable in general (the example is based on a Type-2 CG).

(Chapters 4.3 and 4.6)

4.

The problems mentioned in the previous point can be solved in some cases with the application of conditional independence properties (C1)-(C4) (see Lauritzen, 1996, pp. 29). But the general applicability of the procedure is not guaranteed. Conditional independence as a logical system is non-complete, so (C1)-(C4) are not always enough to prove the equivalence of two equivalent statements. As it is implied by the definition of non-completeness, this problem cannot be solved, neither by expanding (C1)-(C4) nor by replacing them with other axioms.

(Chapters 4.3)

5.

Goodman introduced categorical path models called modified path models (1973) based on DAGs. The parameterization Goodman proposed is problematic, since

- factorization of probabilities cannot always be obtained directly because of the partial ordering of the vertex,
- construction of restricted parameters is not unique, and not all the possible constructions are such that the model can be defined by setting some of the parameters to zero,
- not all the possible constructions are such that the independence statements defining the model can be directly related to the restricted parameters,
- important properties of parameterizations (e.g. ordered decomposability) are not guaranteed to hold.

(Chapter 4.1)

6.

On the other hand, modified path models can be parameterized with marginal loglinear models, in a way that all the desirable properties listed in point 1 hold.

(Publication: Rudas, Bergsma, Németh, 2006. In the dissertation with more general results: Chapter 4.7)

7.

Definitions for CGs given by different authors are not standardized, moreover, are not compatible. Basically two definitions can be distinguished: Type 1 (Frydenberg, 1990, Lauritzen, 1996, Andersson et al., 2001, Lauritzen, Richardson, 2002, Drton, 2008) and Type 2 (Whittaker, 1990, Cox, Wermuth, 1996). DAGs and UGs are special cases of Type-1 CGs. There are graphs belonging to Type-2 graphs but not to Type-1 graphs. These are not just technical but also subject matter differences: there are models, which can be given by Type-2 graphs but cannot be given by Type-1 graphs.

(Chapters 3.2, 3.6.1, 3.6.2)

8.

Choosing between alternative statistical tools for modeling social mobility cannot be reduced to a statistical question. The answer of the question whether graphical models are more appropriate then its alternatives depends on the given theoretical conception and research question. The main aspects influencing the answer are:

- continuous or categorical variables are more appropriate to operationalize the social indicators? (in case of graphical models the latter)
- micro or macro level approach is followed? (in case of graphical models the former)
- is a multistage casual process to be modeled, by distinguishing direct and indirect effects? (graphical models and structural equations models are both appropriate)
- are the effects to be decomposed numerically into direct and indirect effects? (structural equations models could be appropriate but graphical models are not)
- is the joint distribution described by conditional independence statements? (graphical models are appropriate but structural equations models are not)

(Chapter 6.2)

9.

In the dissertation it is presented how the meaning of the most known mobility models (Blau-Duncan, Treiman, Wisconsin model) and their variants found in the literature would be modified if they were interpreted as graphical models. It is also shown, that the modification of the hypothesis can be easily followed by changing the graph (replacing arrows by lines and vice versa, drawing new edges, removing edges etc). An example is presented in point 11, where the modification of the Treiman-model based on Boguszak et al. (1990) can be found.

(Chapter 3.8)

10.

Testing Treiman's modernization hypothesis in Hungary during the transition period based on the model below:

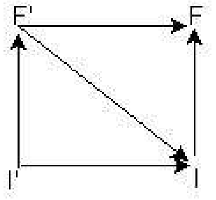

Figure 1 The model (I: education, F: occupation, I': father's education, F': father's occupation)

Social mobility data are analyzed for Hungary from surveys conducted by the Hungarian Central Statistical Office in 1983, 1992 and 2000. The main focus is put on testing Treiman's modernization hypothesis that was posed in 1970 and is still widely cited today in the context of transition. The fitted models are graphical models based on directed acyclic graphs and the values of marginal log-linear parameters are used to gain insight into the strengths of associations. The main findings include that the process of status-attainment seems to be basically unchanged for women, but some of the Treiman-like associations move toward greater social closure. That is, our findings do not support the hypothesis of a trend toward increasing social fluidity in Hungary between the early 1980s and 2000.

(Publication: Németh 2006b, 2007. In the dissertation Chapter 7.1)

11.

Comparing the Treiman-model and its modified version based on Boguszak et al. (1990):

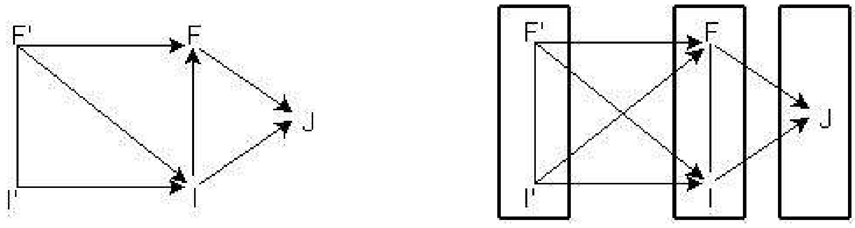

Figure 2 The Treiman-model (Treiman, 1970, originally Duncan et al., 1968) and its modification proposed by Boguszak et al. (1990). I: education, F: occupation, I': father's education, F': father's occupation, J: income

According to the theory of compensation of previous discrimination proposed by Boguszak et al., in the socialist Czechoslovakia (1) occupation has an effect on education as well, and (2) father's education has a direct effect on his son's/daughter's occupation. The former is a consequence of the negative educational discrimination. Discriminated children with highly educated fathers nevertheless inherited the cultural capital, and they found other ways to achievement. The latter is a result of the positive educational discrimination, which was exercised in favor of the politically reliable blue-collar workers who lacked the qualification needed for managerial positions.

Based on Hungarian, Czechoslovakian and American ISSP data from 1992 the original Treiman-model fits the data better than its variant proposed by Boguszak et al. That is, the data don't support the theory of compensation of previous discrimination

(Publication: Németh 2006a, Németh, Rudas, Bergsma, 2006. In the dissertation Chapter 7.2)

References

Andersson, S. A., Madigan, D., Perlman, M. D. (2001): Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics*, 28, 33-85.

Bergsma, W., Rudas, T. (2002): Marginal models for categorical data. *The Annals of Statistics*, (30/1), 140-159.

Cox, D.R., Wermuth, N. (1996). *Multivariate Dependencies – Models, Analysis and Interpretation*. London: Chapman & Hall.

Drton, M. (2008). Discrete chain graph models. *Bernoulli*, accepted.

Frydenberg, M. (1990): The chain graph Markov property. *Scandinavian Journal of Statistics*. 17, 333-353.

Goodman, L.A. (1973): The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika*, 60, 179-192.

Kelley, J. (1990): The failure of a paradigm: Log linear models of social mobility. In , Modgil, C. (szerk.): *John Goldthorpe: Consensus and Controversy* London, England: Falmer Press, 319-46.

Lauritzen, S. L. (1996): *Graphical Models*. Oxford: Clarendon Press.

Lauritzen, S. L., Richardson, T. S. (2002): Chain graph models and their causal interpretation (with discussion). *Journal of the Royal Statistical Society, Series B,* 64 , 321 - 361.

Németh, R. (2004): An application of marginal log-linear models to examine changes in social mobility in Hungary during the transition period. In *Recent Developments and Applications In Social Research Methodology. Proceedings of the RC33 Sixth International Conference on Social Science Methodology*, (RC33), Amsterdam.

Németh, R. (2006a): Graphical models on categorical data – with social science application. In: Némedi, D., Somlai, P., Szabari, V., Szikra, D. (eds.) *Kötő-jelek.* Annual Book of the Sociology PhD Program of Eötvös Loránd University, Budapest. (Hungarian)

Németh, R. (2006b): Changes in social mobility in Hungary during the transition period. *Szociológiai Szemle*, 2006/4, 19-35. (Hungarian)

English version of the above paper: Németh, R.: (2007) Changes in social mobility in Hungary during the transition period. *Review of Sociology*, 13/1, 49-66.

Németh, R., Rudas, T., Bergsma, W. (2006): *Analyzing categorical data with graphical models - a social science application.* SMABS-EAM (Society for Multivariate Analysis in the Behavioural Sciences, European Association of Methodology) konferencia, Budapest.

Rudas, T., Bergsma, W., Németh, R. (2006): Parameterization and estimation of path models for categorical data. In: Rizzi, A., Vich, M. (eds.) *COMPSTAT 2006 Proceedings in Computational Statistics*, Physica-Verlag, 383-394.

Rudas, T., Bergsma, W., Németh, R. (2009): Markov marginal models for categorical data. (manuscript)

Whittaker, J. (1990): *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons.

Content of the dissertation