Chaotic Modeling and Simulation (CMSIM) 3: 175–184, 2020

Smooth Chain Graph Model of type II: a learning procedure

Federica Nicolussi

Department Economics and Quantitative Methods, via Conservatorio, 7, Universit degli Studi di Milano, Italy.
(E-mail: federica.nicolussi@unimi.it)

Abstract. Chain Graph Models (CGs) are a widely used tool to describe the conditional independence relationships among a set of variables. One of the advantages lies in the possible use undirected and directed arcs to link vertices representing variables in the graph. There are four ways to read off the conditional independencies from a chain graph. Each way differs from the other in the way of interpret the missing (un)directed arcs, (see Drton 2009). Different problems can be address with different CGs, however often it is not clear which type of CGs is the best in order to describe the multivariate system of relationships underlying the selected variables. In this work, we propose a learning algorithm, based on a Monte Carlo procedure, that consider the system of independencies underlying all four CGs and select the type and the graph which optimize a score function. When we handle with categorical variables, we take advantage of the marginal models (Bergsma and Rudas, 2002) to parametrize the joint and marginal probability distribution of the variables. Unlikely, Bergsma and Rudas, 2002 showed that particular combinations of conditional independences have no a smooth parametrization. Nicolussi and Colombi, 2013 and 2017, provide the condition according to (any type of) CG admits a smooth parametrization. In the learning procedure we consider only the smooth CGs, that is they admit a smooth parametrization. This approach is implemented to study the poverty status and particularly how this one can be affected from a group of selected variables. We took advantage of the cross-section data sets of Hungarian Household. This analysis highlighted a strong effect of the considered social variables on the poverty status.
Keywords: Chain Graph models; categorical data; learning procedure; poverty status.

1 Introduction

In social studies, there is increasing attention to multivariate models that can capture and describe multiple aspects in a simple way. In particular, it is worthwhile to observe how one or more study variables are affected by other factors. It is also plausible to think that different relationships link the variables studied (i.e. symmetrical, asymmetrical, or causal). Chain Graph models

Received: 29 June 2020 / Accepted: 30 July 2020

ISSN 2241-0503

well represent complex conditional independence assumptions through a particular graph, so-called Chain Graph (CG). Moreover, they well shape both direct and indirect associations. One of the most advantages of these models lies in their immediacy of the representativeness. Indeed, in the graph, each vertex depicts a variable, and the arcs depict association between the two variables represented by the nodes. In particular, each undirected arc pretends for the symmetric association, and each directed arc denotes an effect according to the direction of the arrow. In this work, we consider the four types of Chain Graph models presented in the literature, each of which able to depict peculiar connections, [11]). Section 1.1 is addressed to introduce this topic. In literature, it is widespread to shape Undirected Graph models for discrete variables with log-linear parameters to expound the associations among these variables, [14]. Unfortunately, log-linear parameters can not describe different kinds of relationships in the same models, and these parameters are left in favor of the more general Marginal models, [4]. Bergsma and Rudas (2002) introduced these models to model several 4dependences among a set of discrete variables. For this reason, the Chain Graph Models for discrete variables use marginal models, adding the visual tool to describe the system of relationship between variables, [17], [18], [26], [21]. The potential of these models on social and economic studies is shown by Nemeth and Rudas 2013 [19], [20], [22]. This work aims to highlight whether and how other selected variables can cause income inequality and, more specifically, the poverty status. We take advantage of the Household Monitor survey of TARKI for the Hungarian study. In literature, many works study these two data-sets under the poverty issue. Most of these work on cross-section data use among other classical log-linear models to describe the multivariate system of relationship, see, for instance, [13], [16], [6], and [23]. In this work, we also replied to some results known in the literature. The paper follows this structure. In Sections 1.1 and 1.2, we give a brief introduction to Chain Graph models and Marginal models. In Section 2, we propose a learning procedure for the final graph. Finally, in Section 3, we expound the study data sets and the used methodology step by step, and we show the results of the analysis of the data set are explained. Furthermore, in Section 4, we added a brief conclusion to summarize the output of the analysis.

1.1 Chain Graph Models

Different multivariate analysis to model the relationships among a set of variables exist in literature. Graphical models take advantage of the visual impact that does easily interpretable complex associations. A CG is a graph that includes both directed and undirected arcs while excluding any direct or semi-directed cycle. A CG can be decomposed into so-called chain components, ordered according to the direction of the arrow. Each component is an undirected sub-graphs that contains only undirected arcs, while the vertices in different components are linked to each other by directed arcs. CG Models use chain graphs to represent a system of conditional independencies in a collection of variables. Each variable would be represented as a vertex. In contrast, arcs would represent symmetrical or asymmetrical relationships between them

concerning whether the arc is directed or not so that the lack of an arc represents conditional independence. There are four types of CG Models available in the literature for data analysis, which differ in the way to explain the independence statements (see [11]). However, only three of these have suitable features to describe some problems. In this work, we consider only these three. The CG models proposed by [15] and [12], hereafter LWF CGMs, unifies the directed and undirected graphs approach. The CG models proposed by [1] (AMP CGMs) describe the dependence structures among regression residuals. These two models interpret the lack of an undirect arc conditionally to the other variables in the same component. On the other hand, the CG models proposed by [8] and [25] (MR CGMs) marginalize over these last variables and are suitable to describe multivariate regression systems. Further, LWF CGMs interpret the lack of a direct arc conditionally to the other variables in the same component, while AMP CGMs and MR CGMs marginalize over these last variables. For more profound dissertations about these models and their application, see [19].

1.2 Marginal log-linear parameterization

Log-linear parameters are a useful tool to handle with categorical variables but they are not able to depict conditional independence restrictions involving subsets of variables. Since often the inherent independencies of a CG model concern subsets of variables, we need a most flexible tool, such as marginal log-linear parameters. Marginal log-linear parameters are standard log-linear parameters defined within subsets of contingency tables obtained by marginalizing over one or more variables, [4]. Bergsma and Rudas (2002) show that by building the parameters according to two specific properties (of hierarchy and completeness) the asymptotic properties of parameters hold.
Let consider for instance a set of two variables A and B collected in a contingency table of dimension naXnb with probability π_{ij} where $i = 1, \ldots, n_A$, $j = 1, \ldots, n_B$. Let furthermore consider $\{A; AB\}$ as marginal sets. Then the marginal log-linear parameters are given by:

$$\begin{aligned} \eta_A^A &= \left\{ \log\left(\frac{\pi_{i+}}{\pi_{1+}}\right) \right\}_{i=2,\ldots,n_A} \\ \eta_B^{AB} &= \left\{ \log\left(\frac{\pi_{1j}}{\pi_{11}}\right) \right\}_{j=2,\ldots,n_B} \\ \eta_{AB}^{AB} &= \left\{ \log\left(\frac{\pi_{11}\pi_{ij}}{\pi_{i1}\pi_{1j}}\right) \right\}_{i=2,\ldots,n_A;j=2,\ldots,n_B} \end{aligned} \tag{1}$$

where $\eta_\bullet^\star$ denotes the vector of log-linear parameters concerning the variables $\bullet$ in the marginal distribution $\star$. The symbol $+$ in the probability π denotes the marginalization over the variables in that position.
There are many ways to aggregate the probabilities in the log-linear parameters, the widely diffuse is the baseline criterion, such as in the formula 1, that compares each probability with the probability of the so-called "reference" category, in our case the first one.
However, a more meaningful criterion to describe ordinal variables is the so-called global criterion that compares the cumulative probabilities with the

retro-cumulative probabilities. For instance, the logits of an ordinal variables A evaluated in the marginal A is

$$\eta_A^A = \left\{ \log \left(\frac{\pi_{(A>a_j)}}{\pi_{(A\leq a_j)}} \right) \right\}_{j=\ldots,n_A-1} \tag{2}$$

where n_A is the level number of the variable A. For more details see [2]. System of independencies can be easily represented by setting to zero specific parameters defined in particular marginal distributions. In this way, each missed arc (directed or undirected) in the chain graph corresponds to a set of marginal log-linear parameters constrained to zero. In particular, given three variables A, B and C, to describe the sentence A is independent by B given C (denoted with $A \perp B|C$) the parameters η_{AB}^{ABC} and η_{ABC}^{ABC} must be constrained to zero. For more detail, see [4]. The definition of the marginal sets is crucial for representing different independencies at the same time. Rudas et al. 2010 showed how to define the marginal sets corresponding to the LWF CGMs and MR CGMs, [26]. Nicolussi and Colombi, 2017 showed how to define the set of marginals corresponding to a subset of AMP CGMs, [21].

1.3 Learning procedure

In order to select the CG models (the system of conditional relationships) best performing the data, we take advantage of a Bayesian learning algorithm, that is a variant of the posterior distribution over graphical models. The algorithm requires the evaluation of the marginal likelihood, which can be approximated through a maximum likelihood estimation of the Bayesian Information Criterion score (BIC), and the assignment of a prior probability to the graph. We carry out three parallel learning procedures one for any assumption of underlying CGM. At the end we chose the best fitting model among the resulting models from the three procedures, according to the BIC.
The used procedure is based on the algorithm proposed by [5] and it is described in Algorithm 1. Once chosen one graphical model among the ones described above, we set G_0 equal to the graph without missing arcs.

2 Poverty study

2.1 Data and methods

The results presented in this work are gained from the cross-sections Household Monitor survey carried out by TARKI Social Research Center (Monitor-TARKI) during 2012. It counts 4838 statistical units, each of which has a weight that takes into account the gender, the age, the highest education level of the subject and the reference person of the household, the settlement type, and the number of the household members. The survey considers each family member as a statistical unit. Furthermore, we computed the household equivalence income as the sum of household income weighed by the number of household members. The final contingency table with the collected data has

Algorithm 1 Learning procedure

$G_t = G_0$
while the number of times we choose, consecutively, the graph G0 is less than two times the number of possible edges of the model or graph has been tested against all the other possible graphs (less than an edge). **do**
  Randomly select one edge $(\gamma, \delta) \in (V \backslash E)$
  if if the edge is present in G_t **then**
    remove it
  else
    add it
  end if
  calculate the score of G_t: $score(G_t)$
  calculate the probability $P = min[(score(G_t) - score(G_0)); 1]$
  set $G_0 = G_t$ with probability P.
  return G_3
end while

59 on 192 empty cells.
This work aims to describe the system of relationships among factors that we use as an indicator of wealth and social inequality. In particular, the main factor in analyzing is poverty status (P). This factor refers to the household, which is defined poor whether the equivalent income of a household is less than 60% of median national income. The Employment (E) - evaluated as work intensity- the Status of the Flat (F) and the Type of household (T) were considered as social factors. Finally, the Gender (G) of the subject was considered. Below, we list the variables with their categories

P : Poverty [No, Yes];
E : Employment [0; 0.01-0.49; 0.50-0.99;1];
F : Status in the Flat [owner; rent; other];
T : Type of household [One person; Couple or other without children; Lonely parent with children; couple or other with children];
G : Gender [Male, Female].

Within the cells of the contingency table, we collect the personal weights (W) instead of the classical frequencies. In order to model the variables with the CG models we consider three groups of variables (three chain components): Anagraphical (G), Social (E, F and T) and Wealth (P), and we investigate which model is suitable for well describing the relationships among these factors. The choice of the grouping the variables supposes symmetric relationship between the variables within the same component and asymmetric, causal, relationship between variables in different components.

We test all these models by constraints to zero specific log-linear parameters in selected marginal distributions. According to most of CG models taken into account, it is sufficient to use (G); (E,F,T,G),(P,E,F,T,G) as a hierarchical partial ordered list of marginal sets. However, some independencies require addition marginal sets such as (P,E,F,T), (E,G), (F,G) and (T,G), see for more detail [26], [21]. To describe the dependence relationships between the

factors we chosen the baseline logit for the categorical variables and global logit for the ordinal variable (E). In order to select the best fitting model we adopt the procedure displayed in Algorithm 1. All analysis are carry out with the statistical software R ([24]) with the help of packages `hmmm`, ([7]), `igraph` ([9]) and `gRbase` ([10]).

2.2 Results

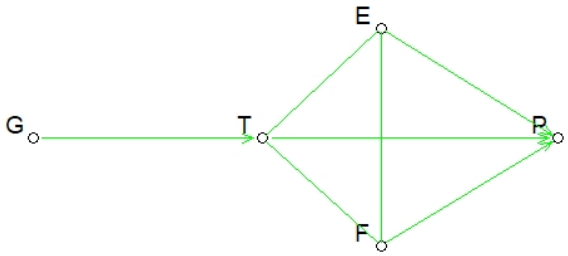

Fig. 1. Chain graph representing the best fitting model

Figure 1 shows the chain graph model which best represents the structure of independence among the factors. Indeed, the three learning procedures lead to the same graph implying the same independence statements. In particular, the graphical model represented in Figure 1 presumes that the gender does not affect the status of poverty given the three social factors considered - $P \perp G|TEF$- and it does not affect even the work intensity and the status of the flat of the household, given by the type of household -$EF \perp G|T$-. In Table 2.2, we reported the parameterization associated to the CG in Figure 1. Here, in the first row, we listed the constrained parameters, in correspondence of the marginal distribution where they are defined. Instead, in the second row, we reported the free parameters. The chosen model presents a likelihood ratio statistic of 145, 7348 which leads to an acceptable p-value of 0, 070 if we consider as the degree of freedom the 122 constrained parameters.
The following tables report the estimate free parameters concerning the two-order effects, conferred to the arcs of the graph in Figure 1. Higher absolute values of these parameters denote strong association. With the Wald test, we evaluated whether the parameters are singularly significant, different from zero. The symbols $*$, $**$, or $***$ denote the significance at the 0.05, 0.01, and 0.001, respectively. Table 2 reports the parameters η_{GT}^{GTFE} concerning the only arc (directed) starting from the gender (G). Each parameter is compared with the reference category. The influence of gender (G) on the type of household (T) is not statistically significant, but in the whole model we can not omit this link even if it is weak.

Table 3, 4 and 5 report the parameters describing the component of social variables TFE. The association between T and F is described in Table 3. Except for the parameter associated to the modalities ?rent? of F and ?couple or other

Marginal	G	GTFE	TFEP	GTFEP
Effect set to zero		GE; GF; GTE; GFE; GTF; GTFE	FEP; TFP; TFEP	GP;GTP;GFP; GEP; GTFP; GTEP; GFEP; GTFEP
Free effect	G	T; E; TF; TE; GT; FE;	P; TP; FP; EP; TEp	

Table 1. Likelihood Ratio Statistic: 145,7348; df 122 (63) p-value 0,070397 (1.669027e-08)

GT	*t2*	*t3*	*t4*
F	-0,35	0,25	-0,37

Table 2. Monitor-TARKI survey: η_{GT}^{GTFE} based on baseline logits for both variables. The reference category is *t1=single with no children* for the type of household (T) and *Male* for the gender (G). The other categories are: F=*Female*, *t2=Couple or other without children*, *t3=lonely parents with children*, *t4=couple or other with children*.

without children? of T, the other parameters are not statistically significant. This parameter denotes that the couples without children with a propensity to a rental house are $e^{?1.88} = 0.153$ times the lonely subjects without children with a propensity to a rental house. However, the connection T−F is stronger than the G→T (the absolute values of parameters in Table 3 are greather than the ones in Table 2).

F-T	*t2*	*t3*	*t4*
f2	-1,88*	-23,62	1,06
f3	-19,71	0,18	-0,27

Table 3. Monitor-TARKI survey: η_{TF}^{GTFE} based on baseline logits for both variables. The reference category is ?single with no children? for the type of household (T) and ?owner? for the flat (F). f2=?renter?, f3=?other?, t2=?couple or other without children?, t3=?lonely parents with children?, t4=?couple or other with children?.

In Table4 are listed the parameters concerning the association between the work intensity of the household (E) and the type of household (T). This association is statistically significant and the parameters grow to the increasing of work intensity. The first parameter denotes that the propensity to work (work intensity greater than zero) in the couples without children is about $e^{0.92} = 2.51$ times the same in the single without children. This ratio grows when the hours of work grow except in the case of couples with children where the trend is opposite. The connection is always positive but the modality ?couple or other with children? which presents a negative trend.

The parameters in Table 5 describe the arc between the variables F and E. The connection between these two variables is weak and mainly negative in the last modalities. In the component of social variables the most substantial connection lies between the work intensity and the type of household (E ? T).

E-TT	*t2*	*t3*	*t4*
> 0	0,92***	0,51*	-0,35
$> 0,49$	1,34***	0,65*	-0,85*
$> 0,99$	2,31***	0,71***	-0,94****

Table 4. TARKI survey: η_{TE}^{GTFE} based on baseline logits for T and global logit for E. The reference category is t1: ?single with no children? for the type of household (T). The other categories are t2=?couple or other without children?, t3=?lonely parents with children?, t4=?couple or other with children?.

f2	-1,03	0,43	-1,44
f3	-0,28	-1,19	-0,07

Table 5. Monitor-TARKI survey:η_{FE}^{GTFE} based on baseline logits for F and global logit for E. The reference category is ?owner? for the type of Flat (F). f2=?renter?, t3=?other?.

The last three tables in (6) refer to the directed arcs from the social variables T, F, and E to the poverty indicator P. The variable that strongly affects the poverty index is reasonably the work intensity. In particular, the propensity to be poor of employed people ($E > 0$) is $e^{-1.7} = 0.18$ times the propensity to be poor of unemployed people ($E \leq 0$). This gap increases by growing the work intensity. Indeed, the last parameter means that the the propensity to be poor in subjects with full-time job (is about $e^{-2.81} = 0.06$ times the subject having work intensity at most equal to 0.99. Even the type of household has a strong and significant influence on poverty. These parameters suggest that the propensity to be poor for a couple or other without children (T=t2) is $e^{?1.12} = 0.33$ times with respect to the single without children. This trend changes when we consider a family with children. For instance, the couples or other with children (T=t4) have about 3.49 times more possibility than a single without children to be poor. Finally, there is a lack of statistical evidence that the type of contract flat (F) affects the poverty index.

TP	Yes
t2	-1,12 ***
t3	0,37
t4	1,25***

FP	Yes
f2	0,81
f3	1,06

EP	Yes
> 0	-1,7 ***
$> 0,49$	-2,07
$> 0,99$	-2,81***

Table 6. Monitor-TARKI survey: η_{TP}^{TFEP}, η_{FP}^{TFEP} and η_{EP}^{TFEP} based on baseline logits for both T, F and P and global logit for E. The reference category is *owner* for the type of Flat (F), *single without children* for (T) and *not poor* for the poverty index (P). The other categories are f2=*rent*, t3=*other*, t2=*Couple or other without children*, t3=*lonely parents with children*, t4=*couple or other with children*.

3 Conclusion

The analysis of the Hungarian study from TARKI survey (2012), Hungarian shows that the gender of the subject (G) does not affect the poverty (P) fixed the type of household (T), the work intensity (E), and the status of flat (F). Further, all the social variables (respectively, work intensity, the status of flat and type of household) affect the poverty status. In detail, there is an effect of gender only on the type of household and the effects of all the social variables E, F, and T on the poverty status (P). The work intensity shows the strongest link with the poverty and highlights a trade-off between poverty and work intensity. The other significant connection is between the type of household and poverty. In this case, the estimated model shows that singles without children are more likely to be poor than a couple without children but are less likely to be poor than lonely parents or couples with children.

4 Acknowledgments

The research leading to these results has received support under the European Commission's 7th Framework Programme (FP7/2013- 2017) under grant agreement n. 312691, InGRID - Inclusive Growth Research Infrastructure Diffusion. The responsibility for all conclusions drawn from the data lies entirely with the authors.

References

1. Andersson, S. A., Madigan, D., and Perlman, M. D. (2001). Alternative Markov properties for chain graphs. Scandinavian journal of statistics, 28(1), 33-85.
2. F. Bartolucci, R. Colombi, A. Forcina, An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints, Statistica Sinica 17(2): 691.(2007)
3. J. Janssen and R. Manca. *Semi-Markov risk models for finance, insurance and reliability*, Springer, New York, 2007.
4. W. P. Bergsma, T. Rudas. Marginal models for categorical data. The Annals of Statistics 30.1: 140-159.(2002)
5. R. Breen, (2008). Statistical models of association for comparing cross-classifications. Sociological Methods & Research, 36(4), 442-461
6. Y. Chzhen and J. Bradshaw(2012). Lone parents, poverty and policy in the European Union. Journal of European Social Policy, 22(5), 487-506.
7. R. Colombi, S. Giordano, M. Cazzaro, and the R Development Core Team. hmmm: Hierarchical Multinomial Marginal Models. R package version 1.0-1 (2013)
8. Cox, D. R and Wermuth, N. (1993) Linear dependencies represented by chain graphs, *Statistical Science*, **8**(3) 204—218.
9. G. Csardi, T. Nepusz. The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org
10. C. Dethlefsen, S. Hjsgaard (2005). A Common Platform for Graphical Models in R: The gRbase Package. Journal of Statistical Software, 14(17), 1-12. URL http://www.jstatsoft.org/v14/i17/.
11. M. Drton, Discrete chain graph models, Bernoulli 15(3): 736- 753.(2009)

12. Frydenberg, M. (1990). The chain graph Markov property. Scandinavian Journal of Statistics, 333-353.
13. M. de Graaf-Zijl and B. Nolan, B. (2011). Household joblessness and its impact on poverty and deprivation in Europe. Journal of European Social Policy, 21(5), 413-431.
14. S. L. Lauritzen. Graphical Models. Oxford University Press, 1996.
15. Lauritzen, S. L., and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. The annals of Statistics, 31-57.
16. H. Lohmann(2011). Comparability of EU-SILC survey and register data: The relationship among employment, earnings and poverty. Journal of European social policy, 21(1), 37-54.
17. A. Forcina, M. Lupparelli and G.M. Marchetti (2010). Marginal parameterizations of discrete models defined by a set of conditional independencies. Journal of Multivariate Analysis, 101(10), 2519-2527.
18. G.M. Marchetti, M. Lupparelli(2011). Chain graph models of multivariate regression type for categorical data. Bernoulli. 17: 827-844.
19. R. Nemeth, T. Rudas, (2013a). On the application of discrete marginal graphical models. Sociological Methodology, 43(1), 70- 100.
20. R. Nemeth, T. Rudas(2013). Discrete Graphical Models in Social Mobility Research-A Comparative Analysis of American, Czechoslovakian and Hungarian Mobility before the Collapse of State Socialism. Bulletin of Sociological Methodology, 118(1), 5-21.
21. F. Nicolussi, R. Colombi (2017). "Graphical Model of type II: a smooth subclass". Bernoulli, 23(2), 863-883.
22. F. Nicolussi, and F. Mecatti (2014b). A smooth subclass of graphical models for chain graph: towards measuring gender gaps. Quality & Quantity, 1-15.
23. V. Polin and M. Raitano(2014). Poverty Transitions and Trigger Events across EU Groups of Countries: Evidence from EUSILC. Journal of Social Policy, 43(04), 745-772.
24. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
25. Richardson, T., and Spirtes, P. (2002). Ancestral graph Markov models. The Annals of Statistics, 30(4), 962-1030.
26. T. Rudas, W. P. Bergsma, and R. Nemeth(2010). Marginal log-linear parameterization of conditional independence models. Biometrika 97.4 : 1006-1012.