Proceedings

5th Stochastic Modeling Techniques and Data Analysis International Conference with Demographics Workshop

SMTDA2018

Editor

Christos H Skiadas

12 - 15 June 2018

Cultural Centre of Chania, Crete, Greece

Imprint
Proceedings of the 5th Stochastic Modeling Techniques and Data Analysis International Conference with Demographics Workshop
Chania, Crete, Greece: 12-15 June, 2018

Published by: ISAST: International Society for the Advancement of Science and Technology.
Editor: Christos H Skiadas

Social application of multivariate regression chain graph models

Federica Nicolussi and Manuela Cazzaro

Department of Statistics and Quantitative Methods, via Bicocca degli Arcimboldi 8,
University of Milano Bicocca, Milano, Italy
(E-mail: federica.nicolussi@unimib.it; manuela.cazzaro@unimib.it)

Abstract. In this work we focus on the study of relationships among a set of categorical variables. Beyond the conditional and marginal relationships we also consider the context-specific independencies that are particular conditional independencies holding only for certain values of the conditioning variables. At this aim we use an improve of chain graphical models combined with the hierarchical multinomial marginal models. A social application on the life satisfaction is provide.
.

Keywords: Chain Regression Model, Multivariate Regression Model, Context-specific independence.

1 Introduction

This work studies how the satisfaction of the interviewees' life can be affected by individual characteristics and personal achievement and, at the same time, how the personal aspects can affect the educational level and the working position. We propose to describe this kind of relationships through a multivariate logistic regression model based on the Chain Graph model. By following the approach of Marchetti and Lupparelli, [4], in fact, we take advantage of a particular case of Chain Graph model, called "of type IV", in order to express variables as *purely explicative*, *purely response* or *mixed* variables. In addition, we also study the relationships under the context-specific independence point of view. This means that we study if there are conditional independencies that hold only for a subset of categories of the conditioning variables. Formally, a context-specific independence (CSI) has the form $A \perp B|C = i_C$ where A, B and C are three sets of variables and i_C is the vector of certain values of the variables in C. Nyman et al. [6] handle with the context-specific independencies in graphical models, through the so-called *strata* added to the graphs. We improved their approach by implementing the *strata* also in the Chain Graph models, see [5]. This work is finalized in showing the multiple aspects that it is possible to highlight by implementing these models, in both graphical and parametric point of views.
This work follows this structure. In Section 2 is presented the graphical models

5^{th}*SMTDA Conference Proceedings, 12-15 June 2018, Chania, Creete, Greece*

and the parametrization that we adopted. In Section 3 we analysed the ISTAT dataset on the "*aspects of everyday life*", [3]. Interesting results were showed.

2 Methodology

Graphical models take advantage of graphs to represent system of (conditional and/or marginal) independencies among a set of variables. In a graphical model, each vertex of the graph represents a variable and any missing arc is a symptom of independence. In this work we consider chain graph where the arcs between two vertices can be both directed or undirected. In a Chain Regression Graph Models (CRGM), variables linked by undirected arcs have a symmetric relationship (as two covariates or two response variables for instance). Each directed arc links a covariate to its dependent variable. The rules to extract a list of independencies from a graph are called Markov properties. In this work we take advantage of the so-called multivariate regression Markov properties, see [4]. In order to consider also the CSIs we improve the graphical models through labelled arcs. The label on the arcs reports the list of categories of the conditioning variables according to the arc vanishes. These labels are exactly the values of the conditioning variables for which the CSI holds. We refer to this new graphical model as Stratified Chain Regression Graph Model (SCRGM). Figure 1 reports examples of CRGM and SCRGM. Note that the independencies represented by these kinds of graphical models

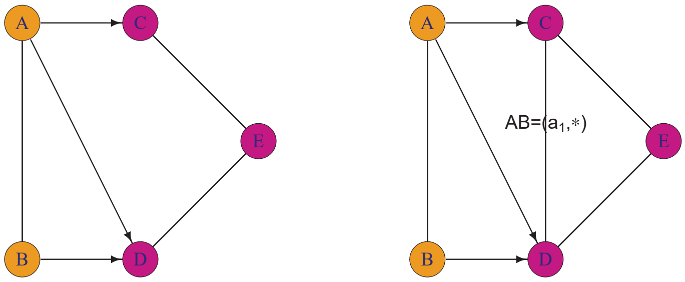

Fig. 1. On the left, a CRGM representing $C \perp D|AB$, $B \perp C|A$ and $AB \perp E$. On the right, a SCRGM representing $C \perp D|AB = (a_1, *)$, $B \perp C|A$ and $AB \perp E$. Note that the response variables are displayed in purple and the covariates in orange.

are to be understood marginally with respect to the other response variables. As mentioned before, we can split the variables in "responses" and "covariates". We adopt the Hierarchical Multinomial Marginal (HMM) parameters, denoted with the symbol η, see [1], to model the dependence among these variables. The HMM parameters are sum of logarithms of probabilities defined on marginal and joint distributions according to certain properties of hierarchy and completeness. For this reason, each parameter is denoted by the marginal distribution where it is evaluated $\mathcal{M}$, by the set of the variables involved by the parameter $\mathcal{L}$ and by the values of variables in $\mathcal{L}$ which it refers $i_{\mathcal{L}}$: $\eta_{\mathcal{L}}^{\mathcal{M}}(i_{\mathcal{L}})$. Note that, when $\mathcal{L}$ is composed of only one variable, the HMM parameter is a logit. In correspondence of each subset of response variables we can build a regression model. For instance, by considering the SCRGM in Figure 1 (right side), with regard to the dependent variable D, we have:

$$\eta_D^{ABD}(i_D|i_{AB}) = \beta_{\emptyset}^D + \beta_A^D(i_A) + \beta_B^D(i_B) + \beta_{AB}^D(i_{AB}) \tag{1}$$

Similar regression models can be built for the other response variables or for combination of these.

Each marginal or conditional independence on the data corresponds to certain zero constraints on these parameters. Quite similar situation is obtained when we consider the CSIs, see [5]. In particular, by looking to the SCRGM in Figure 1 (right side), in addition to the model in formula (1), we have that the model with response variable C, due to the independence $B \perp C|A$ becomes:

$$\eta_C^{ABC}(i_C|i_{AB}) = \beta_{\emptyset}^C + \beta_A^C(i_A).$$

The response variables C and D together can be represented by the following model, by taking into account the CSI $C \perp D|AB = (a_1, *)$, when $AB \neq (a_1, *)$:

$$\eta_{CD}^{ABCD}(i_{CD}|i_{AB} \neq (a_1, *)) = \beta_{\emptyset}^{CD} + \beta_A^{CD}(i_A) + \beta_B^{CD}(i_B) + \beta_{AB}^{CD}(i_{AB})$$

and it is equal to zero when $AB = (a_1, *)$. Note that the other parameters $\eta_E^{ABE}(i_E|i_{AB})$, $\eta_{CE}^{ABCE}(i_{CE}|i_{AB})$, $\eta_{DE}^{ABDE}(i_{DE}|i_{AB})$ and $\eta_{CDE}^{ABCDE}(i_{CDE}|i_{AB})$ are all equal to zero according to the independence $AB \perp E$. Thus, given a graph, we have a list of independencies among the variables. These independencies correspond to certain constraints on the HMM parameters. The unconstrained parameters describe the dependence.

3 Application

In this section we present an application on a real dataset in order to highlight the relationships that are among a set of variables. At first we identify the best fitting SCRGM. Each model is tested by evaluating the likelihood ratio test G^2. The selection is done through a three step algorithm explained below.

Step 1: We test all the plausible pairwise independencies (involving only two variables at time) in the complete graph. From this step we select the models with a *p-value* greater than 0.01.

Step 2: We further investigate all possible CSIs concerning the independence discarded in step 1, by applying, to the graph labelled arcs (one at time) with all possible labels. We take advantage also of mosaic plots. In this step we take into account the models with a *p-value* greater than 0.1.

Step 3: From all admissible models selected in the previous two steps, we test all possible combinations of marginal, conditional and CSIs and we maintain

the one with lower AIC (Akaike Information Criterion) between the models with a *p-value* higher than 0.05.
All the analysis are carried out with the statistical software R [7], and the `package hmmm`, [2].

3.1 Survey on multiple aims analysis

From the survey on the every day aspects of life, [3], we select 5 variables: Gender (G) (*male*=1, *female*=2), Age (A) ($25-34$=1, $35-44$=2, $45-54$=3, $55-59$=4, $60-64$=5), Educational level (E) (*less than high school*=1, *high school, no college* =2, *bachelor degree* =3, *doctoral degree*= 4), Working condition (W) (*looking for a job*=1, *unemployed*=2, *employed*=3) and Life satisfaction (S) ($0-4$: low satisfaction=1, $5-7$: medium satisfaction=2, $8-10$: high satisfaction=3). The survey covers 23880 interviewers, collected in a contingency table of 360 cells of which only 8 cells are null.
The aim of the analysis is to investigate how gender and age affect the educational level and the working condition; at the same time we want to study how all these variables affect the life satisfaction. Thus, we select different marginal distributions where to study the dependencies in order to satisfy these aim. First of all, we consider the smallest (G, A) for studying the symmetrical relationship between the *personal* characteristics. Then we add, one at time and then together, the *achievement* variables (E and W) in order to investigate how the *personal* characteristics affect these dependent variables. Finally, we consider the joint distribution for the study of the life satisfaction. Thus, the class of marginal distributions is $\{(G, A); (G, A, E); (G, A, W); (G, A, E, W); (G, A, E, W, S)\}$.
With the selected marginal distributions, the list of all possible pairwise marginal/ conditional independencies is (a) $G \perp A$, (b) $G \perp E|A$, (c) $A \perp E|G$, (d) $G \perp W|A$, (e) $A \perp W|G$, (f) $E \perp W|AG$, (g) $G \perp S|AEW$, (h) $A \perp S|GEW$, (i) $E \perp S|GAW$, (j) $W \perp S|GAE$. Among this list of independencies only the first present the empirical evidence. Thus, only the undirected arc between the variables G and A should be missed from the CRGM ($G^2 = 10.74$, `df=4, p-value=` 0.03). Getting to the second step at the procedure, the study of the CSI leads to several plausible models. In figure 2 are depicted the mosaic plots concerning the variables W and S in two different conditional distributions in order to have an idea of the independencies that hold only in particular conditioning distributions. In particular, in the plot on the left, where the squares of the mosaic form a quite regular grid, there is high evidence of independence between W and S. On the other hand, the plot on the right shows evidence of strong dependence among the two variables because their ”irregular” lines. By testing all possible combinations of the plausible models we select the one reported in Figure 3. The SCRGM in Figure 3 represents a system of relationships where the gender G and the age A are marginally independent (missing undirected arc between G and A); where the gender G does not affect the life satisfaction S when the age A is among $45-54$, the educational level is doctoral degree and the working condition W is employed (labelled directed arc between G and S); where the educational level E does not affect S for male among 35

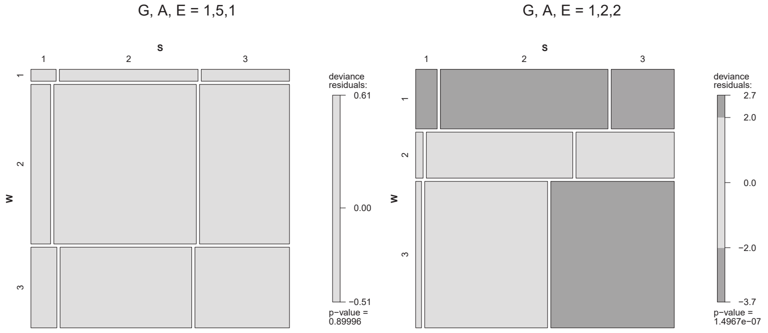

Fig. 2. Mosaic plots concerning the variables W and S in two conditional distributions. On the left, G *male*, A=60 − 64 and E=*less than high school*. On the right, G= *male*, A=35 − 44 and E=*high school*.

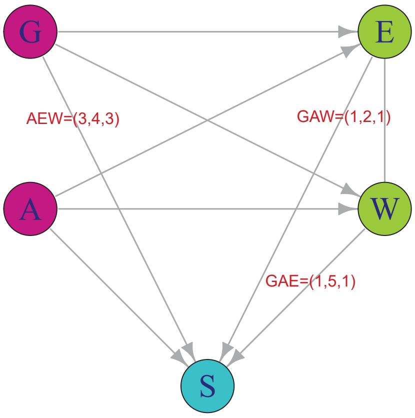

Fig. 3. Best fitting SCRGM. $G^2 = 25.68$, `df=`16, `p-value=`0.06, `AIC=`−658.32.

and 44 years old that are looking for a job; where the working condition W does not affect S for old male with the lowest educational level. The undirected arc with no label shows the strongest dependence because any independence holds among all possible conditioning distributions.

The dependence structure is described by the regression models. For brevity we report only the model with S as dependent variables in Tables 1. In Table 1 the

Table 1. Regression parameters $\eta_S^{GAEWS}(i_S|i_{GAEW})$ of dependent variable S for all possible values of covariates G, A, E and W

GAEW	S=2	S=3	GAEW	S=2	S=3	GAEW	S=2	S=3	GAEW	S=2	S=3
(1111)	1,204	-0,405	(2441)	34,897	33,981	(1342)	1,792	0,693	(2313)	2,159	1,099
(2111)	24,569	23,653	(1541)	23,288	23,47	(2342)	1,705	1,253	(1413)	2,241	1,569
(1211)	1,256	0,34	(2541)	0,693	-0,693	(1442)	1,386	0	(2413)	1,312	0,887
(2211)	0,511	-0,693	(1112)	0,693	0,693	(2442)	1,872	1,946	(1513)	1,72	1,085
(1311)	0,93	-1,099	(2112)	2,303	2,197	(1542)	1,658	1,658	(2513)	2,639	1,099
(2311)	0,754	-0,288	(1212)	0,511	-30,445	(2542)	2,457	2,485	(1123)	2,163	1,887
(1411)	0,262	-0,916	(2212)	1,856	0,875	(1113)	1,846	0,693	(2123)	2,52	2,165
(2411)	35,234	34,674	(1312)	-0,47	-1,674	(2113)	25,039	24,058	(1223)	2,117	1,638
(1511)	1,72	1,085	(2312)	0,981	-0,105	(1213)	3,02	1,504	(2223)	2,09	1,771
(2511)	1,386	0,693	(1412)	1,386	0,223	(2213)	1,386	0,511	(1323)	2,359	1,899
(1121)	1,285	0,28	(2412)	1,353	0,502	(1313)	1,992	1,269	(2323)	2,318	1,836
(2121)	2,022	1,022	(1512)	1,72	1,085	(2313)	2,159	1,099	(1423)	2,206	1,593
(1221)	1,256	0,34	(2512)	1,483	0,853	(1413)	2,241	1,569	(2423)	2,106	1,692
(2221)	1,418	0,853	(1122)	0,134	-0,847	(2413)	1,312	0,887	(1523)	2,09	1,705
(1321)	0,991	-0,089	(2122)	1,492	0,894	(1513)	1,72	1,085	(2523)	2,327	2,225
(2321)	1,187	0,159	(1222)	0,747	-0,811	(2513)	2,639	1,099	(1133)	2,632	2,367
(1421)	0,719	0,1	(2222)	1,638	1,172	(1123)	2,163	1,887	(2133)	2,355	2,001
(2421)	0,847	0	(1322)	1,073	0,074	(2123)	2,52	2,165	(1233)	2,42	2,197
(1521)	0,802	-0,619	(2322)	1,776	1,105	(1223)	2,117	1,638	(2233)	2,592	2,285
(2521)	1,735	1,099	(1422)	2,132	1,488	(2223)	2,09	1,771	(1333)	2,38	2,126
(1131)	1,996	0,986	(2422)	2,227	1,514	(1323)	2,359	1,899	(2333)	2,363	1,842
(2131)	1,738	0,838	(1522)	2,639	2,303	(2532)	1,979	1,52	(1433)	2,398	1,756
(1231)	1,256	0,34	(2522)	1,648	1,304	(1142)	3,689	3,497	(2433)	1,98	1,445
(2231)	1,471	0,83	(1132)	1,405	0,731	(2142)	2,983	2,584	(1533)	2,655	2,147
(1331)	1,292	-0,223	(2132)	2,865	2,325	(1242)	0,405	0	(2533)	1,624	1,293
(2331)	1,306	0,353	(1232)	-0,486	-0,773	(2242)	2,159	2,12	(1143)	2,803	2,733
(1431)	0,871	-0,492	(2232)	2,215	1,843	(1342)	1,792	0,693	(2143)	3,049	3,054
(2431)	0,452	0	(1332)	1,299	0,606	(2342)	1,705	1,253	(1243)	3,016	2,958
(1531)	1,012	-0,134	(2332)	2,075	1,504	(1442)	1,386	0	(2243)	2,862	2,875
(2531)	1,705	1,253	(1432)	2,277	1,386	(2442)	1,872	1,946	(1343)	2,413	2,244
(1141)	1,73	0,857	(2432)	2,534	2,104	(1542)	1,658	1,658	(2343)	2,413	2,244
(2141)	2,054	1,076	(1532)	1,67	1,792	(2542)	2,457	2,485	(1443)	2,28	2,335
(1241)	1,256	0,34	(2532)	1,979	1,52	(1113)	1,846	0,693	(2443)	3,517	2,979
(2241)	2,979	2,457	(1142)	3,689	3,497	(2113)	25,039	24,058	(1543)	4,007	4,159
(1341)	1,658	0,811	(2142)	2,983	2,584	(1213)	3,02	1,504	(2543)	2,833	2,639
(2341)	2,12	1,386	(1242)	0,405	0	(2213)	1,386	0,511			
(1441)	24,611	0	(2242)	2,159	2,12	(1313)	1,992	1,269			

positive parameters refers to situations where the frequency of subjects with life satisfaction S neutral or high is greater than the frequency of low life satisfaction. From the table it is possible to deduce that, only for male among 45 and 54 years old with the lowest educational level and unemployed, the frequency of unsatisfied is greater than the neutral category ($\eta_S^{GAEW}(2|1312) = -0.47$). In the same categories, also the frequency of high satisfied people is lower than the unsatisfied ($\eta_S^{GAEW}(3|1312) = -1.674$). We can find a similar trend among the categories male between 35 and 44 years old with a bachelor degree and unemployed ($\eta_S^{GAEW}(2|1232) = -0.486$ and $\eta_S^{GAEW}(3|1232) = -0.773$). On the other hand, in correspondence of female among 55 and 59 years old with lower educational level that are looking for a job we have very few number of unsatisfied with respect to the other categories ($\eta_S^{GAEW}(2|2411) = 35.234$ and $\eta_S^{GAEW}(3|2411) = 34.674$).

4 Conclusion

The SCRGMs presented in this work are a useful tool to explore and represent the system of relationships among a set of categorical variables. In particular, the labelled arcs in the graph suggest which dependence relationships are weak. The regression parameters, in addition, quantifies the dependence relationships. These results are presented through an application to a life satisfaction. Here, for brevity, few comments and partial results are presented. However, it is possible to deep the analysis and the study of unconstrained parameters.

References

1. Bartolucci, Francesco, Colombi, Roberto, & Forcina, Antonio. *An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints.* Statistica Sinica, **17**, 691-711 (2007).
2. Colombi, Roberto, Giordano, Sabrina & Cazzaro, Manuela. *hmmm: An R Package for Hierarchical Multinomial Marginal Models.* Journal of Statistical Software, **59**(11), 1-25, (2014).
3. Istat. Multiscopo ISTAT - *Aspetti della vita quotidiana. UniData* - Bicocca Data Archive, Milano. Codice indagine SN167. Versione del file di dati 1.0 (2015)
4. Marchetti, Giovanni M., & Lupparelli, Monia. *Chain graph models of multivariate regression type for categorical data.* Bernoulli, **17**(3), 827-844, (2011).
5. Nicolussi, Federica & Cazzaro, Manuela. *Context-specific independencies for ordinal variables in chain regression models.* arXiv:1712.05229, (2017).
6. Nyman, Henrik, Pensar, Johan, Koski, Timo, & Corander, Jukka. *Context specific independence in graphical log-linear models.* Computational Statistics, **31**(4), 1493-1512, 2016.
7. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014.