Context-specific independence in innovation study

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. The study of (in)dependence relationships among a set of categorical variables collected in a contingency table is an amply topic. In this work we want to focus on the so called context-specific independence where the conditional independence holds only in a subspace of the outcome space. The main aspects that we introduce concern the definition in the same model of marginal, conditional and context-specific independencies, through the marginal models. Furthermore, we investigate how it is possible to test these context-specific independencies when there are ordinal variables. Finally, we propose a graphical representation of all the considered independencies taking advantages from the chain graph model. We show the results on an application on ”The Italian Innovation Survey” of Istat (2012).
Keywords: Context-specific independence, ordinal variables, graphical models, innovation.

1 Introduction

In the field of the categorical variables, with the term context-specific (CS) independence we refer to the particular conditional independence that holds only for some modalities of the variable(s) in the conditioning set, but not for all. That is, given three variables X_1, X_2 and X_3 we describe this situation as $X_1 \perp X_2|X_3 = c_3$, where c_3 is a subset of all possible values of X_3. Among other, Højsgaard (2004) [12] and Nyman (2016), [11] deepen this topic. In this paper we want to improve the main results of these works by dealing with CS independencies concerning subsets of all the considered (also ordinal) variables. At this aim we use the Hierarchical Multinomial Marginal Models (HMMMs), see Bartolucci, Colombi and Forcina, 2007 [1]; Cazzaro and Colombi, 2014 [3]. The need of this parametrization chases the will of consider a model where we want to test simultaneously marginal and conditional independencies. In addition, it uses also local logits evaluated on different marginal contingency tables in order to consider the ordered modalities of the CS conditioning variables.

The paper is organized as follows. In Section 2 we introduce the constraints to impose on the HMMM in order to represent also CS independencies. The proposed model is also represented through a Stratified Chain Graph Model (SCGM), an extension of Stratified Graphical Model proposed by Nyman (2016) [11], that uses a Chain Graph Model (CGM) to represent the classical conditional independencies and labelled arcs in the graph to denote CS independencies. The details are explained in Section 3.

Finally we analyze a real dataset, “The Italian Innovation Survey” of Istat

(2012) [5], in order to investigate the effect of the innovation in different aspects of small and medium Italian enterprises on the grown in revenue terms. The procedure and the results are showed in Section 4. In Section 5, we summarize the main results of this work and future research.

2 Parametrization for context specific independencies

Let us consider q categorical variables $(X_1, \ldots, X_q)$ taking values $(i_1, \ldots, i_q)$ in the contingency table $\mathcal{I} = (n_1 \times \cdots \times n_q)$, where the modalities of the generic variable X_j, i_j takes value in $\mathcal{I}_j$. A parametrization of a model able to capture marginal and conditional independencies among non ordinal variables comes through the marginal model, see Bergsma and Rudas, 2002 [2], which defines the classical log-linear parameters on marginal distributions by respecting certain properties of completeness and hierarchy. The marginal parameters are $\eta^{\mathcal{M}}_{\mathcal{L}}(i_{\mathcal{L}})$ where $\mathcal{M}$ refers to the marginal set, $\mathcal{L}$ denotes the subset of variables which the parameter pertains and $i_{\mathcal{L}}$, in parenthesis, the modalities of the variable selected in $\mathcal{L}$ (when the parenthesis are omitted means that the parameters refer to each $i_{\mathcal{L}} \in \mathcal{I}_{\mathcal{L}}$). The following example shows how to define the marginal parameters in order to describe a conditional independence.

Example 1 Let us consider a set of four variables, say X_1, X_2, X_3 and X_4 and suppose we are interested in describing the independence $X_1 \perp X_2 | X_3$. At this aim, we have to define the marginal sets $\{(1,2,3),(1,2,3,4)\}$ where $(1,2,3,4)$ is a shortcut for $(X_1X_2X_3X_4)$. Then, we define the classical log-linear parameters on the contingency table $\mathcal{I}_{1,2,3}$ restricted to $(1,2,3)$ and the remaining parameters on the unrestricted contingency table $\mathcal{I}$. Finally, we have to constrain to zero the parameters associated to the statement of independence $\eta^{1,2,3}_{1,2}$ and $\eta^{1,2,3}_{1,2,3}$.

Now, let us collect 4 subsets of variables, supposing A, B, C and D. As we mentioned, our aim is to find a parametrization able to describe, beyond the classical statements of conditional independencies, the following statement of CS independence, formally:

$$A \perp B | (C = i_C, D), \qquad i_C \in \mathcal{K} \tag{1}$$

where i_c is the vector of certain modalities of variables in C which take values in $\mathcal{K}$ that is a subset of the modalities of C $(\mathcal{I}_C)$ for which the conditional independence holds. The independence in formula (1) holds if the marginal log-linear parameters satisfy the following constraints

$$\sum_{\substack{v \in \mathcal{V} \\ c \in \mathcal{P}(C)}} \eta^{\mathcal{M}}_{vc}(i_v i_c) = 0 \qquad i_v \in \mathcal{I}_v \quad i_c \in \mathcal{K} \tag{2}$$

where $\mathcal{P}(\cdot)$ denotes the power set, $\mathcal{V} = \{(\mathcal{P}(A) \setminus \emptyset) \cup (\mathcal{P}(B) \setminus \emptyset) \cup \mathcal{P}(D)\}$ and $\mathcal{K}$ is a subset of the modalities of C $(\mathcal{I}_C)$ for which the CS independence holds.

Example 2 (recall Example 1) Let suppose that we want to define through marginal model the CS independence $X_1 \perp X_2 | X_3 X_4 = i_4$, with $i_4 \in \mathcal{K}$ where $\mathcal{K} \subseteq I_4$ is a subset of the modalities i_4 of X_4 for which the conditional independence holds. The constraints on the marginal parameters will be in this case

$$\eta_{1,2}^{1,2,3,4}(i_1 i_2) + \eta_{1,2,3}^{1,2,3,4}(i_1 i_2 i_3) + \eta_{1,2,4}^{1,2,3,4}(i_1 i_2 i_4) + \eta_{1,2,3,4}^{1,2,3,4}(i_1 i_2 i_3 i_4) = 0$$

$$i_1 \in \mathcal{I}_1, \quad i_2 \in \mathcal{I}_2, \quad i_3 \in \mathcal{I}_3, \quad i_4 \in \mathcal{K}.$$

Now, we consider the case where we have at least an ordinal variable. In this unexplored case we move in the HMMM framework, see Bartolucci, Colombi and Forcina, 2007 [1] and Cazzaro and Colombi, 2014 [3]. In the HMMMs, beyond the baseline parameters, we can use parameters η coded with different criteria in order to consider the possible proper order of the modalities. In this work we take advantage from the local logits that compare the probability of a cell π_i with the previous one, for instance, referring to variable X_1 we have $\eta_1^1(i_1) = log(\frac{\pi_{i_1}}{\pi_{i_1 - 1}})$.

The independence in formula (1) holds if the parameters of HMMM, coded with local logits, satisfy the following constraints

$$\sum_{\substack{v \in \mathcal{V} \\ c \in \mathcal{P}(C)}} \sum_{i_c^* \leq i_c} \eta^{vc}(i_v i_c) = 0 \qquad i_v \in \mathcal{I}_v \quad i_c \in \mathcal{K} \tag{3}$$

where $\mathcal{P}(\cdot)$ denotes the power set, $\mathcal{V} = \{(\mathcal{P}(A) \setminus \emptyset) \cup (\mathcal{P}(B) \setminus \emptyset) \cup \mathcal{P}(D)\}$ and $\mathcal{K}$ is a subset of the modalities of C ($\mathcal{I}_C$) for which the CS independence holds, see [10].

Example 3 By considering the CS independence in Example 2, by adopting local logit for coding the conditioning variable, the constraints in formula (3) become

$$\begin{aligned} &\eta_{1,2}^{1,2,3,4}(i_1, i_2) + \eta_{1,2,3}^{1,2,3,4}(i_1 i_2 i_3) + \\ &+ \textstyle\sum_{i_4^*=1}^{i_4} \eta_{1,2,4}^{1,2,3,4}(i_1 i_2 i_4^*) + \sum_{i_4^*=1}^{i_4} \eta_{1,2,3,4}^{1,2,3,4}(i_1 i_2 i_3 i_4^*) = 0 \end{aligned} \tag{4}$$

with $i_1 \in \mathcal{I}_1$, $i_2 \in \mathcal{I}_2$, $i_3 \in \mathcal{I}_3$ and $i_4 \in \mathcal{K}$. It is worthwhile to note that the constraints in formula (3), when we deal with local logit, correspond to the CS independence $X_1 \perp X_2 | X_3 X_4 \leq i_4$, $i_4 \in \mathcal{K}$.

3 Stratified Chain Graph models

A *Chain Graph* is a graph with both directed and undirected arcs and without any directed or semi-directed cycle. The vertices of a chain graph are decomposable in so-called *Chain Components*, denoted by $T_1, ..., T_s$. Within these chain components there are only undirected arcs and between vertices

belonging to different components there are only directed arcs, all head toward the same direction. Trivially, the Chain Graph Models (CGM) are graphical models which take advantages from chain graphs to describe a system of independencies. There are different types of CGM, see Drton, 2009 [4], that interpret in different way the presence/absence of directed/indirected arcs. In this work we use the CGM of type I, see Lauritzen and Wermuth, 1989 [7] and Frydenberg, 1990, [6], as natural generalization of classical graphical models. CGMs are used when the variables to analyze are of different nature, such that they can be naturally collected in different components. Furthermore, it is reasonable to suppose that between variables within the same component there is a kind of dependence relationship that differs from the relationship between variables collected in different components. Therefore, it is possible to define an explicative order between the variables collected in different components.
As it is shown in Rudas, Bergsma and Németh, 2010, [13] and in Nicolussi, 2013, [9], the marginal log-linear models and the HMMMs give a suitable parameterization for the CGM of type I. Now, the improvement in CGMs necessary to represent the CS independencies closely follows the Nyman's approach (Nyman, 2016 [11]) for undirected graphs. Thus we introduce the Stratified Chain Graph Models (SCGM) as extension of stratified graphical models, [11]. A stratified chain graph has, in addition to the previous graphs, labeled arcs. These identify the "stratum" of the models, that is the modality(ies) of the variable(s) in the conditional set according to the context-specific independence.

Example 4 Let us consider 5 variables X_1, X_2, X_3, X_4 and X_5. Suppose that, according to the nature of the variables we can split them in two components such that variables X_1 and X_2 can be considered explicative for X_3, X_4 and X_5. The SCGM represented by the graph in Figure 1 is one possible situation that can occur. In this case we have the conditional independencies $X_3 \perp X_2|X_1$ and $X_5 \perp X_1X_2|(X_3, X_4)$ and the CS independence $X_3 \perp X_4|(X_1 = i_1^*, X_2, X_5 = i_5^*)$.

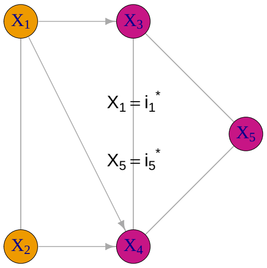


Fig. 1. SCGM with the labelled arc $X_3 - X_4$ referring to modality i_1^* of X_1 and modality i_5^* of X_5.

4 Application on real data

In this section we investigate the potential of a model that simultaneously, consider marginal, conditional and CS indpendencies on a set of (ordinal) categorical variables. Our aim is to study the effect of innovation in small and medium Italian enterprises, during the 2009-2012, in the revenue growth. With the term "innovation" we refer to any improvement in product, services, productive line, logistic system, organization and investment in Research and Development (R&D) area. We used the "Italian innovation survey on SM enterprises" [5].

Thus we considered the *revenue growth in 2012*, **G** (Yes, No) henceforth denoted as variable **1**, as the pure response variable. Then, we took into account the innovation through three dichotomous variables referring to the period 2009-2012: *innovation in products or services or production line or investment in R&D*, **IPSP** (Yes, No), *innovation in organization system*, **IORG** (Yes, No) and *innovation in marketing strategies*, **IMAR** (Yes, No), henceforth denoted as variables **2, 3** and **4** respectively. Finally, other variables concerning the firm's featuring in 2009-2012 were collected: the *main market (in revenue terms)*, **MARK** (A= Regional, B= National, C= International), the *percentage of graduate employers*, **DEG** (1= 0% $\vdash$ 10%, 2= 10% $\vdash$ 50%, 3=50% $\vdash$ 100%) and the *enterprise size*, **TYP** (1= Small, 2= Medium), henceforth denoted as variables **5, 6** and **7** respectively.

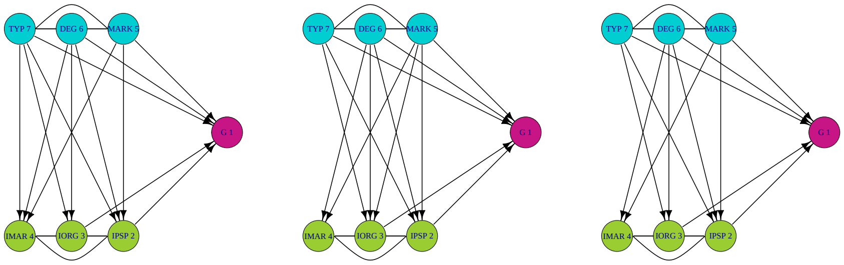
In order to analyze this dataset, we build a chain graph with three components according to the nature of the variables, so in the first component we collect the firm's features (**MARK 5**, **DEG 6**, **TYP 7**), in the second component the innovations variables (**IPSP 2**, **IORG 3**, **IMAR 4**) and in the third component the revenue growth **G 1**. Then, starting from the complete chain graph, where there are all possible edges, corresponding to the saturated HMMM, we tested all chain graph models of type I with only one missing edge, in order to investigate, one by one, which pairwise relationship is plausible. The test was lead with the maximum likelihood ratio test, by comparing the likelihood of unconstrained HMMM, with the likelihood of the corresponding constrained model. In the HMMM, the parameters of dummy variables were codified with baseline logits while the parameters referring to the ordinal **MARK** and **DEG** were codified with local logits.

We removed from the complete chain graph all the edges which given positive results in the previous tests, obtained in this way a reduced CGM. Subsequently, we test the reduced CGM adding one by one all the edges previously removed. Table 1 shows the statistic test, the degree of freedom and the p-value of the HMMM for the main significant models. The numbers involved in the independencies represent the variables in the order of presentation. The CGMs associated to these three HMMMs were depicted in Figure 2.

It is clear (i.e. it is common to all models), that the growth (**1**) is independent by the innovation in the marketing strategies (**4**) given by the remaining variables (**2, 3, 5, 6, 7**). In model A we have that the innovation in the organization system (**3**) is independent on the market where the enterprise works (**5**) given the other variables concerning the innovation and the firm's features

Name	Independencies	G^2	df	p-value
A	$1 \perp 4 \mid 2,3,5,6,7$ $3 \perp 5 \mid 2,4,6,7$	100.88	84	0.1012
B	$1 \perp 4 \mid 2,3,5,6,7$ $4 \perp 7 \mid 2,3,5,6$	91.87	81	0.1921
C	$1 \perp 4 \mid 2,3,5,6,7$ $3 \perp 5 \mid 2,4,6,7$ $4 \perp 7 \mid 2,3,5,6$	112.02	93	0.0872

Table 1. Values of likelihood ratio test G^2 of HMMM associated to CG models.



(a) (b) (c)

Fig. 2. CG models.

(**2, 4, 6, 7**). On the contrary, in model B we have that the innovation in marketing strategies (**4**) is independent on the enterprise's size (**7**) given the other variables concerning the innovation and the firm's features (**2, 3, 5, 6**). Model C is the union of the independencies in model A and in model B. As we can see from Table 1 by choosing a reference level of the first type error α equal to 0.1 we reject the null hypothesis, thus we have no enough evidence to choose the model C. Thus, we considered the three independencies characterizing model C like CS independencies and we test all possible alternatives. The more interesting models were reported in Table 2. The preferable model, according to the parsimonious principle, is C4. The difference between models C and C4 is the independence concerning the organization system (**3**) and the market where the enterprise works (**5**). In fact, in C4 this independence holds only when the conditioning variable percentage of graduated employers (**6**) is lower than 10% or greater than 50% that we can assume as indicator of unspecialized or high specialized firms. This means that only when the percentage of graduated employers is between 10% $-$ 50% the market affects the innovation in the organization system.

The stratified chain graph associated to the model C4 is depicted in Figure 3. In this graph the labeled arc between the node **MARK** and **IORG** reports the modalities of the variables **DEG** according to the arc is removed. That is, only when the variable **DEG** assume the first or the third modality,

Name	Independencies	G^2	df	p-value
C1	$1 \perp 4 \mid 2,3,5,6,7$ $3 \perp 5 \mid 2,4,(6=1),7$ $4 \perp 7 \mid 2,3,5,6$	94.75	85	0.22002
C2	$1 \perp 4 \mid 2,3,5,6,7$ $3 \perp 5 \mid 2,4,(6=2),7$ $4 \perp 7 \mid 2,3,5,6$	102.77	85	0.09205
C3	$1 \perp 4 \mid 2,3,5,6,7$ $3 \perp 5 \mid 2,4,(6=3),7$ $4 \perp 7 \mid 2,3,5,6$	101.08	85	0.1125
C4	$1 \perp 4 \mid 2,3,5,6,7$ $3 \perp 5 \mid 2,4,(6=1,3),7$ $4 \perp 7 \mid 2,3,5,6$	105.09	89	0.1171

Table 2. Values of likelihood ratio test G^2 test of HMMM

there is **MARK** independent by **IORG** given by **ISPS**, **IMAR**, **DEG** and **TYP**.

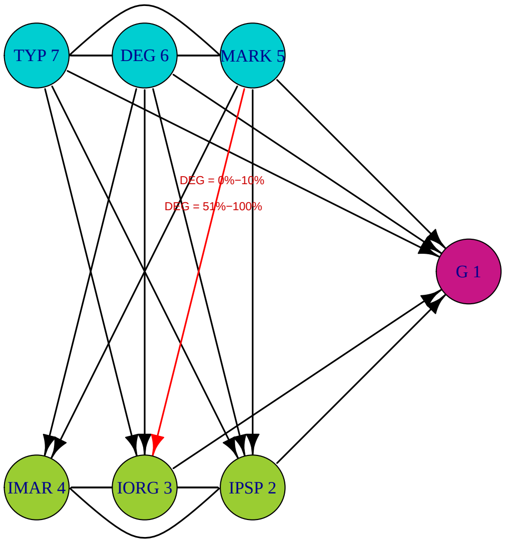


Fig. 3. SCG model C4.

Finally, in Table 3 we report the values of the second order marginal log-linear parameters (referring to paired variables) of model C4. At first we remind that these are defined in the first marginal distribution where they occur. In this case, the marginal subsets associated to the CG models in Figure 2 and to the SCG model in Figure 3 are $\{(5,6,7),(2,3,4,5,6,7),(1,2,3,4,5,6,7)\}$. Furthermore, we remind that in order to define the conditional (marginal) independencies in model C4 we have to constraint to zero the parameter $\eta_{1,4}^{1,2,3,4,5,6,7}$ and all the higher order parameters, defined in the marginal set $(1,2,3,4,5,6,7)$, containing the paired variables $(1,4)$ and also the parameter $\eta_{4,7}^{2,3,4,5,6,7}$ and all the higher order parameters, defined in the marginal set $(2,3,4,5,6,7)$, containing the paired variables $(4,7)$. Finally, in order to define the CS independence, according to the formula (3), we have to constrain to zero the sum of parameters $\eta_{3,5}^{2,3,4,5,6,7}$ and all the higher order parameters, defined in the marginal $(2,3,4,5,6,7)$, containing the paired variables $(3,5)$ but where the variable 6 assumes value 1 or 3. Note that in Table 3, the parameters $\eta_{3,5}^{2,3,4,5,6,7}$ are free and assume value zero. This reveals the lack of relationship between the variables **MARK** and **IORG** at least concerning the parameters of third or higher order.

		G 1	IPSP 2	IORG 3	IMAR 4	MARK 5		DEG 6	
Variable	Modalities	*Yes*	*Yes*	*Yes*	*Yes*	*National*	*Internat.*	*10%-50%*	*≥ 50%*
ISPS 2	*Yes*	0.1927 (0.0793)							
IORG 3	*Yes*	0.1023 (0.0709)	1.8221 (0.0827)						
IMAR 4	*Yes*	0 (0.0000)	1.4848 (0.0907)	1.9967 (0.0764)					
MARK 5	*National*	0.0980 (0.0688)	0.6378 (0.0960)	0 (0.0000)	0.3005 (0.0928)				
	Internat.	0.4668 (0.1486)	0.1517 (0.1815)	0 (0.0000)	-0.2096 (0.1912)				
DEG 6	*10%-50%*	0.0332 (0.0821)	0.5020 (0.10400)	0.4372 (0.0988)	0.4323 (0.0927)	0.6902 (0.0567)	0.2547 (0.0856)		
	≥50%	-0.1333 (0.1436)	-0.0422 (0.2070)	0.5048 (0.1624)	0.3451 (0.1746)	0.1758 (0.1024)	-0.1186 (0.1493)		
TYP 7	*Medium*	0.3700 (0.0790)	0.6447 (0.1064)	0.5687 (0.0868)	0 (0.0000)	0.9878 (0.0497)	0.7591 (0.0775)	1.1702 (0.0543)	-0.3302 (0.0899)

Table 3. Second order marginal log-linear parameters.

From Table 3 we can see that between the three innovation variables there is a strong (positive) second order association: (**IPSP**, **IORG**) with log odds ratio of 1.82, (**IPSP**, **IMAR**) with log odds ratio of 1.49 and (**IMAR**, **IORG**) with log odds ratio of 2. In the graph they correspond to the undirected arcs between the nodes **2** and **3**. This means that is more likely to have firms that improve innovations in different levels. Another strong association is between the firm's dimension and the main market. In particular it seems, reasonably, that bigger is the firm bigger is the market where it works. It is also worthwhile to focus on the parameters concerning the variable **DEG**, which discriminates between a conditional and a CS independence in model C4. In particular from Table 3 it came to light that there is a reverse direction between the parameters (all positive) referring to the $10\% - 50\%$ modality and the one referring to the $\geq 50\%$ which are more than half negative. This means that moving from the unspecialized firm (less than 10% graduate) to a medium specialized firm ($10\% - 50\%$ graduate), we have a positive association with all the other variables. On the other hand, by considering the highly specialized firm ($\geq 50\%$ graduate) with respect to the medium specialized firm, we can see that there is a negative trend with the revenue growth. The same trend occurs also with the innovation in product, services, product line and R&D (**IPSP**), the main market (**MARK**) and the firm's size (**TYP**). This change probably would been unobserved by codifying the parameters with baseline logits. Furthermore, by accepting the conditional independence $3 \perp 5|2, 4, 6, 7$ we would not focus on the variable 6.

5 Conclusion

In this work we showed how to represent CS independencies in HMMMs when we treat with ordinal variable and we are interested in representing also marginal and conditional independencies. We also provide a graphical representation based on chain graph in order to give visual simplification of the relationships among the variables.

The final SCGM have been chosen following a two steps procedure to identify the best CGM and then by watching the problem at hand to find the "strata" of the graph, but further research will be dedicated to implement the procedure able to test all possible models (testing all hypothesis of independence). Furthermore, other research involves the definition of constraints for parameters coded with "global" or "continuation" logits. It should be interenting also to study the definition of SCGM by considering the Chain Graph Models of type 4, see Drton (2009) [4], with the parameterization explained by Marchetti and Lupparelli (2011) [8].

References

1. Bartolucci, F., Colombi, R., and Forcina, A. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, **17**, 691 – 711, 2007.
2. Bergsma, W. P., and Rudas, T. Marginal models for categorical data. *Annals of Statistics*, **30(1)**, 140 – 159, 2002.
3. Cazzaro, M., and Colombi, R. Marginal nested interactions for contingency tables. *Communications in Statistics - Theory and Methods*, **43(13)**, 2799 – 2814, 2014.
4. Drton, M. Discrete chain graph models. *Bernoulli*, **15(3)**, 736 – 753, 2009.
5. Istat. Italian innovation survey on SM enterprises, 2012.
6. Frydenberg, M. The chain graph Markov property. *Scandinavian Journal of Statistics*, **17(4)**, 333-353, 1990.
7. Lauritzen, S. L., and Wermuth, N.. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, **17(1)**, 31-57, 1989.
8. Marchetti, G. M., and Lupparelli, M.. Chain graph models of multivariate regression type for categorical data. *Bernoulli*, **17(3)**, 827 – 844, 2011.

9. Nicolussi, F. Marginal parameterizations for conditional independence models and graphical models for categorical data. *PhD Thesis.* University of Milano Bicocca. 2013.
10. Nicolussi, F., Cazzaro, M..Context-specific independencies for ordinal variables in chain regression models. arXiv preprint arXiv:1712.05229 2017.
11. Nyman, H., Pensar, J., Koski, T., and Corander, J. Context-specific independence in graphical log-linear models. *Computational Statistics*, **31(4)**, 1493 – 1512, 2016.
12. Højsgaaard, S. Statistical inference in context specific interaction models for contingency tables. *Scandinavian journal of statistics*, **31(1)**, 143-158, 2004.
13. Rudas, T., Bergsma, W. P., and Németh, R.. Marginal log-linear parameterization of conditional independence models. *Biometrika*, **97(4)**, 1006-1012, (2010).