Bernoulli **27**(3), 2021, 2091–2116
https://doi.org/10.3150/20-BEJ1302

Context-specific independencies in stratified chain regression graphical models

FEDERICA NICOLUSSI[1] and MANUELA CAZZARO[2]

[1]*University of Milan, Via Conservatorio, 7, 20122 Milano MI, Italy. E-mail: federica.nicolussi@unimi.it*
[2]*University of Milan Bicocca, Via Bicocca Degli Arcimboldi 8, 20126 Milano, MI, Italy. E-mail: manuela.cazzaro@unimib.it*

Graphical models are a useful tool with increasing diffusion. In the categorical variable framework, they provide important visual support to understand the relationships among the considered variables. Besides, particular chain graphical models are suitable to represent multivariate regression models. However, the associated parameterization, such as marginal log-linear models, is often difficult to interpret when the number of variables increases because of a large number of parameters involved. On the contrary, conditional and marginal independencies reduce the number of parameters needed to represent the joint probability distribution of the variables. In compliance with the parsimonious principle, it is worthwhile to consider also the so-called context-specific independencies, which are conditional independencies holding for particular values of the variables in the conditioning set. In this work, we propose a particular chain graphical model able to represent these context-specific independencies through labeled arcs. We provide also the Markov properties able to describe marginal, conditional, and context-specific independencies from this new chain graph. Finally, we show the results in an application to a real data set.

Keywords: Graphical models; stratified Markov properties; categorical variables; multivariate regression models; marginal models

1. Introduction

Different statistical models can study the relationships among a set of categorical variables collected in a contingency table depending on the focus of the analysis. When the considered variables have different nature, and the causal interpretation can explain their link, a multivariate regression system could be a suitable tool. In this case, we can suppose that each variable can assume a different role. It can be a response variable, hereafter purely response; it can be a covariate, hereafter purely covariate; or finally, it can be explanatory for some variables and response for others, hereafter a mixed variable.

For this purpose, we take advantage of graphical models. Graphical models rotate around a system of independencies among variables that they easily represent through graphs. However, the relevance of the considered models, as we will see, is also because these models always have a smooth likelihood function, which in general is false.

Different graphical models exist in literature. See, for instance, Lauritzen [15], Wermuth and Cox [32], and Whittaker [33] for an overview. Here, we take advantage of a particular type of chain graphical models known as chain regression graphical models (CRGMs) (see Cox and Wermuth [8], Richardson and Spirtes [26] and Drton [9]. Indeed, these models describe simplified multivariate regression models, where each dependent variable cannot be a covariate for another dependent variable considered at the same level of the first one, Marchetti and Lupparelli [17]).

Marginal and conditional (in)dependencies among a set of variables are deeply studied through graphical models, though the representation of context-specific independencies (CSIs) is not widely spread. With the term CSI, we mean conditional independence that holds only in a subspace of the

outcome of the variables in the conditioning set. The study of these CSIs is dual. First, they allow to focus on categories of certain variables that truly discriminate between the independence of two sets of variables. Additionally, and no less important, the CSIs permit the reduction of the number of parameters needed to represent the joint probability distribution.

Different authors faced the problem of the graphical representation of CSIs by proposing several possible solutions. The manifold suggestions are clues of the difficulty and non-triviality of this topic. In particular, Boutilier et al. [4] take advantage of a Bayesian network and represent each conditional probability distribution through a tree. Likewise, Højsgaard [10,11] proposed the "split model", where the same graph represents the relationships among a set of variables in correspondence to the different values of a chosen subset of variables. Furthermore, La Rocca and Roverato [13] and Roverato [27] depict the CSIs in graphs by considering some categories as variables. Moreover, Sadeghi [29] consider the CSIs in the discrete 'determinantal' point processes represented through undirected and bidirected graphical models. Finally, Pensar et al. [24], Nyman et al. [23] and [22] generalize the graphical model (for undirected and directed acyclic graphs) with the so-called stratified graphical model that summarizes the CSIs in only one graph.

In this work, we propose the stratified chain regression graphical model (SCRGM) as a generalization of the CRGMs, representing CSIs. We use a parametrization based on the hierarchical multinomial marginal models (HMMMs) to include also CSIs in the regression models (see for details Bartolucci et al. [2], Cazzaro and Colombi [6], Nicolussi and Cazzaro [19]).

The work follows this structure. In Section 2, we give the state of the art in CRGMs suitable for multivariate regression frameworks. Sections 3 and Section 4 are reserved to the original results of this paper. Indeed, in Section 3, we present a new SCRGM, that extends the CRGMs described in Section 2 by considering also the CSIs. Here, new suitable Markov properties, the rules to extract a list of independencies from a graph, were proposed by considering either the *global* or the *pairwise* approach. Furthermore, we prove the equivalence of these two approaches. We also state the list of conditional independencies and CSIs compatible with an SCRGM. In Section 4, we introduce a parameterization suitable for the SCRGM by taking advantage of the HMMMs. In particular, we use two approaches for coding the variables in the parameters, the *baseline* and the *local* code (see Bartolucci et al. [2] and Cazzaro and Colombi [5]), in order to have more meaningful parameters in the case of ordinal variables. We highlight the connection between the SCRGMs and the HMMMs through constraints on suitable parameters. In Section 5, we provide some applications to a real data set. In Section 6, we report the conclusion. Appendix A.1 contains a dissertation on the existence of the maximum likelihood estimation. All the proofs of the theorems are listed in Appendix A.2 to improve the readability of the paper.

2. Chain regression graphical model

A CRGM is a particular chain graph model known as CGM of type IV, see Drton [9]. A summary of the chain graph model used to represent multivariate regression models follows, such as explained in Marchetti and Lupparelli [17]. The SCRGM proposed in this work is a generalization of this CRGM.

Formally, a *graph* $\mathcal{G} = \{V, E\}$ is a collection of two sets, the one of vertices or nodes (V) and the one of edges or arcs (E). The admitted edges can be *directed*, represented by an arrow ($\rightarrow$), or *bidirected*, represented by a double headed arrow ($\leftrightarrow$). In the literature, the double headed arrow can be replaced with dashed segments, as in Marchetti and Lupparelli [17]. We prefer the double headed arrow representation in order to use an homogeneous notation with the larger class of *mixed-graph*, see Lauritzen and Sadeghi [14].

Two vertices linked by a bidirected arc are called *adjacent* ($\gamma \leftrightarrow \delta$). Two vertices linked by an arrow $\gamma \rightarrow \delta$ are *parent* (γ) and *child* (δ), respectively. A *path* between two nodes γ and δ is a sequence

of non repeated nodes linked by directed or bidirected arcs. A *collider* in a path is a node with two arrowheads pointing to it, such as $\rightarrow \gamma \leftarrow$, or $\leftrightarrow \gamma \leftrightarrow$, or $\leftrightarrow \gamma \leftarrow$. The *anterior* set of the node γ, denoted with $\text{ant}(\gamma)$, is the set of nodes linked by a path, containing only directed arcs, pointing to γ. A set of vertices is *connected* if any pair of nodes in A is linked by a *path* belonging in A. Otherwise, the set is said *non-connected*. A *chain graph* is a graph $\mathcal{G} = \{V, E\}$ with both directed and bidirected arcs in E and without either directed or semi-directed cycle. That is, by following the direction of the arrows, no path starts and ends in the same vertex. Given a set A of vertices, the *parent* set of A, $\text{pa}_{\mathcal{G}}(A)$ is composed of all vertices that are parents of at least one vertex in A. The so-called *chain components*, denoted by $T_1, \ldots, T_s$, make a partition of the vertices of the *chain graph* according to the following conventions: adjacent vertices must belong to the same component. In contrast, vertices linked by directed arcs must belong to different components. With the term *parent component*, $\text{pa}_T(T_h)$, we refer to the set of components, from which at least one directed arc starts, pointing to the component T_h. The subscript T refers to the common name used for components, and it is needed to discriminate it from the parents of a set A, $\text{pa}_{\mathcal{G}}(A)$ mentioned above. We consider the components $T_1, \ldots, T_s$ of a chain graph as partially ordered, such that if $h < l$, then $T_l \notin \text{pa}_T(T_h)$. Finally, we define the set of *predecessors* of a component T_h, $\text{pre}(T_h)$, as the union of the components coming before in the chosen order of the components.

Graphical models take advantage of graphs by representing the connection among a set of variables. In this work, we focus on a vector of $|V|$ categorical variables, $X_V = (X_j)_{j \in V}$, taking values $\boldsymbol{i}_V = (i_1, \ldots, i_j, \ldots, i_{|V|})$ in the contingency table $\mathcal{I} = (I_1 \times \cdots \times I_{|V|})$, with joint probability distribution P.

A probabilistic independence model $\mathcal{J}(P)$ is a list of independence statements $\langle A, B|C\rangle$, interpreted as "A is independent of B given C", induced by a joint probability distribution P.

Sadeghi and Lauritzen [30] summarized some properties of a generic independence model $\mathcal{J}$:

Definition 2.1. Given A, B, C, and D disjoint subsets of X_V,

S1 $\langle A, B|C\rangle \in \mathcal{J}$ if and only if $\langle B, A|C\rangle \in \mathcal{J}$ *(symmetry)*;
S2 if $\langle A, B \cup D|C\rangle \in \mathcal{J}$, then $\langle A, B|C\rangle \in \mathcal{J}$ and $\langle A, D|C\rangle \in \mathcal{J}$ *(decomposition)*;
S3 if $\langle A, B \cup D|C\rangle \in \mathcal{J}$, then $\langle A, B|C \cup D\rangle \in \mathcal{J}$ and $\langle A, D|C \cup B\rangle \in \mathcal{J}$ *(weak union)*;
S4 $\langle A, B|C \cup D\rangle \in \mathcal{J}$ and $\langle A, D|C\rangle \in \mathcal{J}$ if and only if $\langle A, B \cup D|C\rangle \in \mathcal{J}$ *(contraction)*;
S5 if $\langle A, B|C \cup D\rangle \in \mathcal{J}$ and $\langle A, D|C \cup B\rangle \in \mathcal{J}$ then $\langle A, B \cup D|C\rangle \in \mathcal{J}$ *(intersection)*;
S6 if $\langle A, B|C\rangle \in \mathcal{J}$ and $\langle A, D|C\rangle \in \mathcal{J}$ then $\langle A, B \cup D|C\rangle \in \mathcal{J}$ *(composition)*;

A graphical model is a representation of the probabilistic independence model $\mathcal{J}(P)$ of the collection of variables X_V. In general, each vertex γ in the graph represents one variable X_γ. Any directed arc from γ to δ stands for asymmetric dependence between X_γ, and X_δ, which is the variable X_γ affects X_δ and not the reverse. Finally, any bidirected arc between two vertices γ and δ stands for the symmetric dependence between the corresponding two variables. As a consequence, each missing arc (directed or bidirected) denotes an independence relationship.

This ability to depict different relationships makes the CRGM a suitable graphical tool to represent multivariate regression models. Indeed, in a CRGM, the analyzed variables follow an inherent explanatory order where some variables are covariate of other ones, which can be, in turn, covariate of the other ones. Thus, the partition of the vertices in components comes naturally according to the variables represented by the vertices. Furthermore, from this "classification" of the variables, the CRGMs are useful to represent both conditional and marginal independencies.

As shown by Drton [9], given a chain graph, there are different criteria to read off a list of independent restrictions among variables. These rules are called Markov properties, and they characterize

four types of chain graph models. In this work, we consider the approach of Cox and Wermuth [8] and Richardson and Spirtes [26].

By following Richardson and Spirtes [26], we need to introduce the m-separation criterion in order to define the list of independence statements.

Thus, given a chain graph $\mathcal{G}$, a path in $\mathcal{G}$ is an m-connecting path given the subset of vertices C if all colliders are in $C \cup \text{ant}(C)$ and all its non-colliders are outside C. Given two disjoint subsets of vertices A and B, they are m-separates given C if there is no m-connecting paths between A and B.

Definition 2.2. Given a *chain graph* $\mathcal{G}$, an independence model $\mathcal{J}$ defined over X_V satisfies the global Markov property w.r.t. $\mathcal{G}$ if, for A, B and C disjoint subsets of X_V, it holds that

$$\text{if A and B are } m\text{-separates given C } \rightarrow \langle A, B|C\rangle \in \mathcal{J}. \tag{1}$$

The resulting independence model is said faithful to the graph and it is denoted with $\mathcal{J}(\mathcal{G})$.

Lauritzen and Sadeghi [14] proved that any independence model faithful to a graph $\mathcal{J}(\mathcal{G})$ satisfies all the properties in the Definition 2.1.

Example 2.1. Let us consider the chain graph $\mathcal{G}$ in Figure 1(a). The paths linking the nodes 3 and 4 are three. In the path $3 \leftrightarrow 5 \leftrightarrow 4$, the node 5 is a collider, thus 3 and 4 are m-connected given $C = (5)$. In the path $3 \leftarrow 1 \rightarrow 4$ the node 1 is not a collider, thus the nodes 3 and 4 are m-separated given $C = (1)$. Finally, in the path $3 \leftarrow 1 \leftrightarrow 2 \rightarrow 4$ the nodes 1 and 2 are not a collider, thus the nodes 3 and 4 are m-separated given $C = (1, 2)$. In the same way, the node 5 is m-separated from the nodes $(1, 2)$ given the empty set (all paths between the two sets of nodes contains *colliders*). Further, for example, the nodes 2 and 3 are m-separated given the node 1. By applying the Markov property in Definition 2.2, we get, among other, the following list of independence statements: (i) $\langle X_3, X_4|X_1\rangle$, (ii) $\langle X_3, X_4|X_{12}\rangle$, (iii) $\langle X_2, X_3|X_1\rangle$, (iv) $\langle X_5, X_{12}|\varnothing\rangle$, and (v) $\langle X_{35}, X_2|X_1\rangle$. Note that, not all the statements are necessary, for instance applying the *contraction* property (S4) to (ii) and (iii) we obtain (vi) $\langle X_3, X_{24}|X_1\rangle$. Now, from this last, in force of the *decomposition* (S2) also (i) $\langle X_3, X_4|X_1\rangle$ holds. Thus, it is unnecessary to include also the (i) in the list.

3. Stratified chain regression graphical model

For a triplet of disjoint sets $A, B, C \subseteq V$, we say that X_A and X_B are independent given X_C if and only if $P(i_A, i_B|i_C) = P(i_A|i_C)P(i_B|i_C)$ for all $i_C \in \mathcal{I}_C$. More generally, we say that a context-specific independence holds if there exists a nonempty subclass $\mathcal{K}_C$ of $\mathcal{I}_C$ such that the factorization above holds for all $i_C \in \mathcal{K}_C$, formally

$$X_A \perp\!\!\!\perp X_B|(X_C = \boldsymbol{i}_C), \quad \boldsymbol{i}_C \in \mathcal{K}_C. \tag{2}$$

In general, in formula (2) it may occur that the above factorization is satisfied for all the values of a subset of X_C in combination with a selection of values of the remaining variables in the conditioning set. In this case, it is possible to partition the conditioning set C as $l \cup (C \backslash l)$, and the class $\mathcal{K}_C$ of values $\boldsymbol{i}_C$ can be obtained as $\mathcal{K}_C = \mathcal{K}_l \times \mathcal{I}_{C\backslash l}$. This means that the CSI holds for all values $\boldsymbol{i}_{C\backslash l} \in \mathcal{I}_{C\backslash l}$, combining with the values $\boldsymbol{i}_l \in \mathcal{K}_l$.

By considering an independence model $\mathcal{J}(P)$ faithful to a probability distribution P of the X_V variables, we interpret the statement $\langle A, B|C; \mathcal{K}_C\rangle$ as "X_A is independent of X_B given X_C when

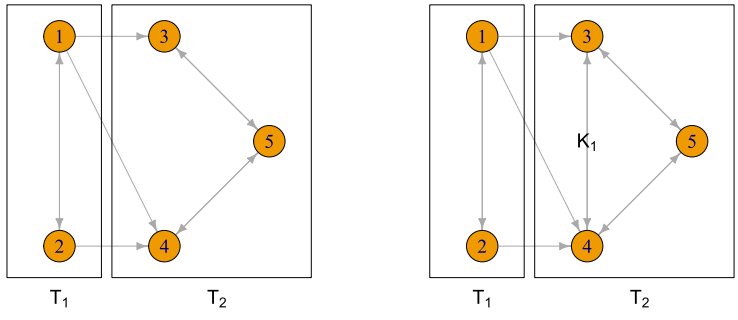

Figure 1. A chain graph (a) and a stratified chain graph (b), both with components $T_1 = (1, 2)$ and $T_2 = (3, 4, 5)$. In figure (b) there is a labeled arc representing the *stratum* between the vertices 3 and 4, with the label $\mathcal{K}_1$.

X_C is equal to $\boldsymbol{i}_C$ taking values in the class $\mathcal{K}_C$". Trivially, when $\mathcal{K}_C = \mathcal{I}_C$, the CSI is conditional independence, and the previous independence statement becomes $\langle A, B|C\rangle$.

The following lemma highlights that not all the CSI statements are admissible.

Lemma 1. *Let consider the probabilistic independence model $\mathcal{J}(P)$ faithful to the joint probability distribution P of the vector of variables X_V. If $\langle A, B|C\rangle$ belongs to $\mathcal{J}(P)$ then $\langle A, C|B; \mathcal{K}_B\rangle$, with $\mathcal{K}_B \neq \mathcal{I}_B$ is not representable in the same independence model $\mathcal{J}(P)$.*

To represent also the CSIs in a graphical model, we propose the SCRGM based on a stratified chain graph (SCG) obtained by adding strata to the chain graph presented in the previous section. Thus, likewise to stratified graphical models (SGMs), proposed by Nyman et al. [23], we represent the CSIs through labeled arcs, also called *strata*. Trivially, if there is a stratum between the nodes γ and δ, the label reports the class of category(ies), $\mathcal{K}_C$, according to the arc is missing. Thus, we say that the variables X_γ and X_δ, associated to the end-point nodes of the labeled arc, are independent given X_C, taking values in $\mathcal{K}_C$.

In order to make less messy the SCG, usually in the label, we specify the class $\mathcal{K}$ only for a subset of variables $l \subseteq C$ such that $\langle X_\gamma, X_\delta|C; \mathcal{K}_l \times \mathcal{I}_{C\setminus(l)}\rangle \in \mathcal{J}$. Thus, the variables in the conditioning set, but not quoted in the label are suppose to assume all possible values.

Example 3.1 shows briefly how to interpret the *strata* in the SCG. The formal definition of the SCG will follow.

Example 3.1. In Figure 1(b), the label on the arc between the nodes 3 and 4 reports the value $\mathcal{K}_1$, referring to the variable X_1. This means that, when the variable X_1 takes values in $\mathcal{K}_1$ the arc is missing and the two nodes are not linked anymore but they are m-separated given $C = (1)$ and $C = (1, 2)$ (see Example 2.1 for details). Thus, the labeled arc stands for the CSI statements $\langle X_3, X_4|X_1; \mathcal{K}_1\rangle$ and $\langle X_3, X_4|X_{12}; \mathcal{K}_1 \times \mathcal{I}_2\rangle$.

Definition 3.1. A stratified chain graph $\mathcal{G} = \{V, E, \mathcal{S}\}$ is a collection of two sets -the one of vertices V, the one of arcs E- and a class of strata $\{\mathcal{K}_l\}_{l\in\mathcal{S}}$, where $\mathcal{S}$ is the class of sets of variables quoted in all the strata and it is closed under countable unions.

Since any stratum links two vertices at time, it is natural to formulate the CSIs in a *pairwise* approach, which explains the relationship between two paired variables. The pairwise Markov properties defining an SCRGM are described in the following definition.

Definition 3.2. Given a SCG $\mathcal{G}$, the induced independence model $\mathcal{J}(\mathcal{G})$ obtained by applying the pairwise stratified Markov properties is composed of the following independence statements:

- For any missing arc between γ and δ in $\mathcal{G}$:
  pM1. $\langle X_\gamma, X_\delta | X_{\text{pre}(T_h)} \rangle \in \mathcal{J}(\mathcal{G})$, when $\gamma, \delta \in T_h$;
  pM2. $\langle X_\gamma, X_\delta | X_{\text{pre}(T_h)\setminus\delta} \rangle \in \mathcal{J}(\mathcal{G})$, when $\gamma \in T_h$ and $\delta \in \text{pre}(T_h)$.
- For any labeled arc between γ and δ with label $\mathcal{K}_l$ in $\mathcal{G}$:
  pS1. $\langle X_\gamma, X_\delta | X_{\text{pre}(T_h)}; \mathcal{K}_l \times \mathcal{I}_{\text{pre}(T_h)\setminus l} \rangle \in \mathcal{J}(\mathcal{G})$, when $\gamma, \delta \in T_h$;
  pS2. $\langle X_\gamma, X_\delta | X_{\text{pre}(T_h)\setminus\delta}; \mathcal{K}_l \times \mathcal{I}_{\text{pre}(T_h)\setminus(l\cup\delta)} \rangle \in \mathcal{J}(\mathcal{G})$, when $\gamma \in T_h, \delta \in \text{pa}_T(T_h)$.

The first two properties (*pM1* and *pM2*) are the *pairwise* Markov properties for the CRGM (Marchetti and Lupparelli [16], Sadeghi and Wermuth [31]). The last two (*pS1* and *pS2*) are the equivalent rules for the *strata*. Note that, moving from the Markov property *pM2* to the Markov property *pS2*, the belonging set of δ is reduced from $\text{pre}(T_h)$ to the set $\text{pa}_T(T_h)$. Indeed, if absurdly, there could be a labeled arc between a vertex in T_h and one in $\text{pre}(T_h)\setminus\text{pa}_T(T_h)$, this last vertex, by definition, should be a parent of T_h.

Example 3.2. The SCG $\mathcal{G}$ in Figure 2(a) has 4 variables, 3 directed arcs and two strata on bidirected arcs. All strata in $\mathcal{G}$ refer to the variable X_1 thus the class $\mathcal{S}$ is equal to $\{(1)\}$. However, the two strata have different labels: $\mathcal{K}_1 = \{2\}$ on the arc $2 \leftrightarrow 3$ and $\mathcal{K}_1 = \{2; 3\}$ on the arc $3 \leftrightarrow 4$. By applying the pairwise Markov properties in Definition 3.2, according to *pS1*, we get that, when $X_1 = 2$, the three variables X_2, X_3, and X_4 are mutually independent given X_1: $\langle X_2, X_3, X_4 | X_1; \mathcal{K}_1 = \{2\} \rangle$. Besides, when $X_1 = 3$, X_3 and X_4 are independent given X_1: $\langle X_3, X_4 | X_1; \mathcal{K}_1 = \{3\} \rangle$. Finally, according to *pM1*, the conditional independence $\langle X_2, X_4 | X_1 \rangle$ holds.

Example 3.3. The SCG $\mathcal{G}$ in Figure 3(a), has two *strata*: $\mathcal{K}_2 = \{2; 3\}$ on the arc $1 \to 4$ and $\mathcal{K}_2 = \{2\}$ on the arc $3 \to 4$. By applying *pS2*, we have $\langle X_1, X_4 | X_{23}; \{2; 3\} \times \mathcal{I}_3 \rangle$. Further, we also have $\langle X_3, X_4 | X_{12}; \mathcal{I}_1 \times \{(2)\} \rangle$. Let us suppose that the variable X_1, as well as the variable X_3, is fully described by three categories labeled (1), (2), and (3). Then the first independence holds for the categories $\boldsymbol{i}_{23}$ belonging to $\{(2, 1); (3, 1); (2, 2); (3, 2); (2, 3); (3, 3)\}$, and the second independence holds for the categories $\boldsymbol{i}_{12}$ belonging to $\{(1, 2); (2, 2); (3, 2)\}$.

In order to provide the *global* Markov property for an SCG, we have to introduce two new graph definitions.

Definition 3.3. Given a SCG $\mathcal{G}$, the chain graph obtained by replacing all labeled arcs with unlabeled arcs is called full graph, $\mathcal{G}^F$.

Definition 3.4. Let us consider a SCG $\mathcal{G}$, with the class of strata $\{\mathcal{K}_l\}_{l\in\mathcal{S}}$. For any value $\boldsymbol{i}_l \in \mathcal{K}_l$ and $l \in \mathcal{S}$, the associated reduced chain graph $\mathcal{G}^R(\boldsymbol{i}_l)$ is obtained:

- by deleting all the labeled arcs having the particular value $\boldsymbol{i}_l$ in the label $\mathcal{K}_l$;
- by replacing the remaining labeled arcs with unlabeled arcs.

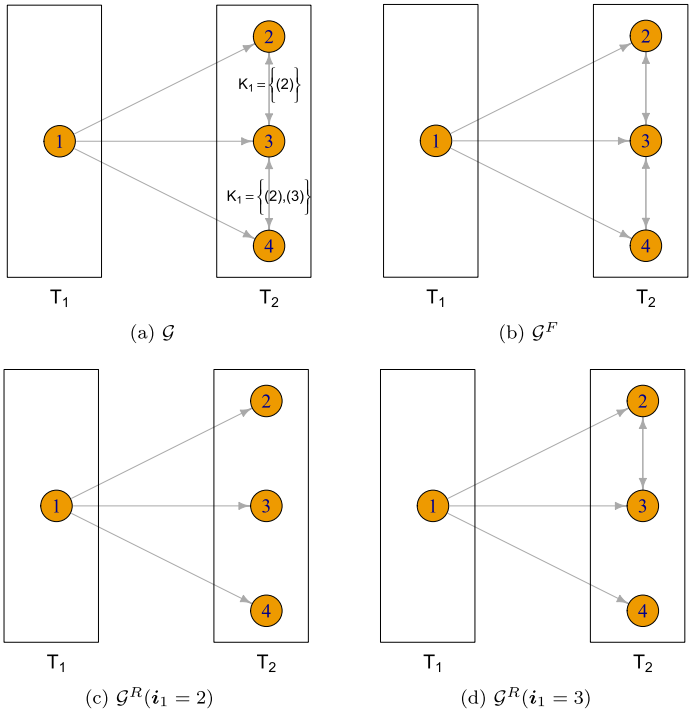

Figure 2. (a) Stratified chain graph $\mathcal{G}$. (b) Full chain graph $\mathcal{G}^F$ of $\mathcal{G}$. (c) Reduced chain graph in correspondence with $\boldsymbol{i}_1 = 2$, $\mathcal{G}^R$ $(\boldsymbol{i}_1 = 2)$. (d) Reduced chain graph in correspondence with $\boldsymbol{i}_1 = 3$, $\mathcal{G}^R$ $(\boldsymbol{i}_1 = 3)$.

Note that, there is only one full graph $\mathcal{G}^F$ associated with one SCG $\mathcal{G}$, while there are many possible reduced graphs $\mathcal{G}^R$, one for each different cell $\boldsymbol{i}_l \in \mathcal{K}_l$ and $l \in \mathcal{S}$.

Now we have all the elements to define the *global Markov* property defining a SCRGM.

Definition 3.5. Given an SCG $\mathcal{G}$, the induced independence model $\mathcal{J}(\mathcal{G})$ defined over X_V satisfies the global stratified Markov property w.r.t. $\mathcal{G}$ if

gM in the full chain graph,

$$\text{if A and B are } m\text{-separates given C} \longrightarrow \langle A, B|C\rangle \in \mathcal{J}(\mathcal{G})$$

gS in any reduced chain graph $\mathcal{G}^R(\boldsymbol{i}_l)$,

$$\text{if A and B are } m\text{-separates given C} \longrightarrow \langle A, B|C; \boldsymbol{i}_l \times \mathcal{I}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G})$$

To avoid unnecessary statements in the independence model $\mathcal{J}$ resulting from Definition 3.5, we apply the following rules.

Definition 3.6. Given an SCG $\mathcal{G}$, let be $\mathcal{J}(\mathcal{G}^F)$ the independence model resulting from the full graph $\mathcal{G}^F$ and $\mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ the independence model resulting from any reduced graph $\mathcal{G}^R(\boldsymbol{i}_l)$ for any $\boldsymbol{i}_l \in \mathcal{K}_l$ and for any $l \in \mathcal{S}$. Then

R1 if $\langle A, B|C\rangle \in \mathcal{J}(\mathcal{G}^F)$ and $\langle A, B|C; \mathcal{K}_l \times \mathcal{I}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(i_l))$ then $\langle A, B|C\rangle \in \mathcal{J}(\mathcal{G})$ (i.e., it is unnecessary to include also the second statements because the first one implies it);

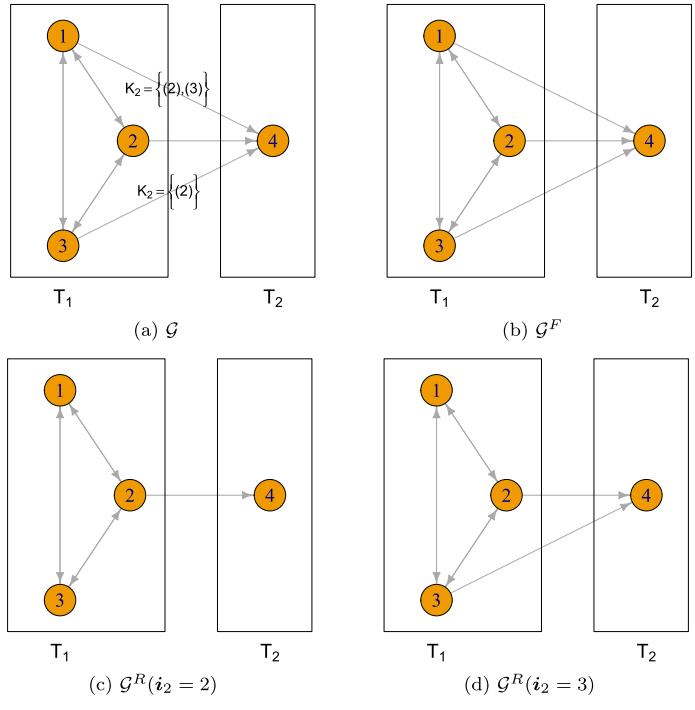

Figure 3. (a) Stratified chain graph $\mathcal{G}$. (b) Full chain graph $\mathcal{G}^F$ of $\mathcal{G}$. (c) Reduced chain graph in correspondence with $\boldsymbol{i}_2 = 2$, $\mathcal{G}^R(\boldsymbol{i}_2 = 2)$. (d) Reduced chain graph with correspondence with $\boldsymbol{i}_2 = 3$, $\mathcal{G}^R(\boldsymbol{i}_2 = 3)$.

R2 if $\langle A, B|C; i_l \times \mathcal{I}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and $\langle A, D|E; \boldsymbol{j}_m \times \mathcal{I}_{C\setminus m}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{j}_m))$, with $l \cap m = \varnothing$ or $\boldsymbol{i}_{l\cap m} = \boldsymbol{j}_{m\cap l}$, then $\langle A, F|G; \mathcal{K}_{l\cup m} \times \mathcal{I}_{G\setminus(l\cup m)}\rangle \in \mathcal{J}(\mathcal{G})$ where $F = B \cup D$, $G = C \cup E\setminus(B \cup D)$ and $\mathcal{K}_{l\cup m} = \{(\boldsymbol{i}_l, \boldsymbol{j}_{m\setminus l})\}$.

Example 3.4 (Continuation of Example 3.2). In Figure 2(b), the full graph $\mathcal{G}^F$ associated to the SCG $\mathcal{G}$ in Figure 2(a) is depicted. In $\mathcal{G}^F$ the nodes 2 and 3 are m-separated given 1 because 1 is not a collider in the path liking the nodes 2 and 4. Thus, $\langle X_2, X_4|X_1\rangle \in \mathcal{J}(\mathcal{G})$. In Figure 2(c), the reduced graph $\mathcal{G}^R$ $(\boldsymbol{i}_1 = 2)$ is depicted. This graph represents the independence model when the variable X_1 is equal to 2. In $\mathcal{G}^R$ $(\boldsymbol{i}_1 = 2)$ the nodes 2, 3, and 4 are mutually m-separated given 1, thus $\langle X_2, X_3, X_4|X_1; \{(2)\}\rangle \in \mathcal{J}(\mathcal{G})$. However, we already stated that, for all values of X_1, $\langle X_2, X_4|X_1\rangle \in \mathcal{J}(\mathcal{G})$, holds. Therefore, it is enough to consider $\langle X_{24}, X_3|X_1; \{(2)\}\rangle \in \mathcal{J}(\mathcal{G})$. Finally, in Figure 2(d) there is the reduced graph $\mathcal{G}^R$ $(\boldsymbol{i}_1 = 3)$ representing the independence model when the variable X_1 is equal to 3. Here, we have that the nodes $(2, 3)$ and 4 are m-separates given 1, thus $\langle X_{23}, X_4|X_1; \{(3)\}\rangle \in \mathcal{J}(\mathcal{G})$.

Example 3.5 (Continuation of Example 3.3). In Figure 3(b) the full graph is depicted. Since this graph has no missing edges, it represents a model where all variables affect each others. When the variable X_2 is equal to 2, the independence model is represented by the reduced graph $\mathcal{G}^R(\boldsymbol{i}_2 = 2)$ in Figure 3(c). By applying the global Markov properties in Definition 3.5, we get the independence statement $\langle X_{13}, X_4|X_2; \{(2)\}\rangle \in \mathcal{J}(\mathcal{G})$. On the other hand, the reduced graph $\mathcal{G}^R(\boldsymbol{i}_2 = 3)$ in Figure 3(d), represents the independence model when the variable X_2 is equal to 3. Here the statement $\langle X_1, X_4|X_{23}; \{(3)\} \times \mathcal{I}_3\rangle \in \mathcal{J}(\mathcal{G})$ holds.

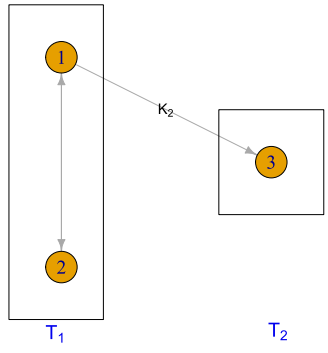

Figure 4. SCRG with components $T_1 = (1, 2)$ and $T_2 = (3)$ with a non-representable *stratum*.

Theorem 3.1. *Any independence model $\mathcal{J}(\mathcal{G})$ faithful to a SCG $\mathcal{G}$, satisfies all the properties in Definition 2.1.*

Corollary 3.1. *Given a SCG $\mathcal{G}$, the induced independence model $\mathcal{J}(\mathcal{G})$ obtained by applying the pairwise Markov properties in Definition 3.2 is equivalent to the one obtained by applying the global Markov property in Definition 3.5.*

A *stratum* can be represented both by a bidirected or a directed labeled arc. However, in Lemma 1, there are restrictions on the independence statements belonging to the same independence model. In Lemma 2 we use the result of Lemma 1 to define the admissible *strata* in the SCRGM. The following example shows the logic of the previous assertion.

Example 3.6. Let us consider the chain graph in Figure 4. According to the gM in Definition 3.5 the conditional independence $\langle X_2, X_3|X_1\rangle$ holds, but at the same time, according to gS in Definition 3.5 the CSI $\langle X_1, X_3|X_2; \mathcal{K}_2\rangle$ holds too. According to the conditional independence $P(X_3|X_{12}) = P(X_3|X_1)$, the variables X_2 does not affect the conditional distribution of X_3 given X_1. However, according to the CSI, when the values assumed by X_2 belongs to the class $\mathcal{K}_2$, $P(X_3|X_{12}) = P(X_3)$ and when X_2 assumes values do not belonging to $\mathcal{K}_2$, the previous probability becomes $P(X_3|X_{12}) = P(X_3|X_1)$. This means that the values of X_2 effectively affects the probability of X_3. The only compatible situation with the previous statement is that $P(X_3|X_1) = P(X_3)$, but this holds if $\langle X_1, X_3|X_2\rangle$.

The problem arises if there is at least one variable contained in l, $l \in \mathcal{S}$, that does not point to γ and δ, the endpoints of the labeled arc. Nyman et al. [23] dealt with this situation, and in their Theorem 2, they give the condition for the existence of a stratum in a bidirected graphical model. This result is generalized to the SCRGM in Lemma 2.

Lemma 2. *Given an SCRGM, all the vertices in l referring to the stratum with label $\mathcal{K}_l$ must be*

- *parents of both the endpoints of the labeled arc between γ and δ, $(\gamma \leftrightarrow \delta)$;*
- *adjacent to δ and parent of γ when $\gamma \in T_h$ and $\delta \in \mathrm{pa}_T(T_h)\backslash\mathrm{pa}_{\mathcal{G}}(\gamma)$, $(\gamma \rightarrow \delta)$.*

Example 3.7. Figure 2 satisfies the two conditions in Lemma 2, trivially. In Figure 3, in all strata, the set l is equal to vertex 2 only. In both strata, the conditions of Lemma 2 hold because the vertex 2 is adjacent to the other vertices in $\mathrm{pa}_T(T_h)$.

4. Parameterization for stratified chain regression graph models

In Sections 2 and 3, we faced with independence models with different types of independence statements. In this section, we consider a parameterization of the probability distribution to represent these independence models easily, to model through linear constraints on the parameters, and to shape the dependence relationships through regression models. In particular, we take advantage of the HMMMs and a re-parameterization of these last. Indeed, in the literature, the marginal models proposed by Bergsma and Rudas [3] are widely used to parameterize graphical models for categorical variables, see among other Rudas et al. [28], Marchetti and Lupparelli [16], Marchetti and Lupparelli [17], Nicolussi [18]. The HMMMs generalize these by admitting a different form of the parameters.

4.1. Parameterization for categorical variables

Briefly, HMMMs are a generalization of the classical log-linear models where the parameters, henceforth HMM parameters, are evaluated in opportune marginal distributions $P_{\mathcal{M}}$, with $\mathcal{M} \subseteq V$, and they are specified by assigning a logit type to each variable of the marginal distribution. Let $\pi_{\mathcal{M}}(\boldsymbol{i}_{\mathcal{M}})$ be the probability of the cell $\boldsymbol{i}_{\mathcal{M}}$ of the contingency table $\mathcal{I}_{\mathcal{M}}$. In the vector $\pi_{\mathcal{M}}$, we collect all the probabilities $\pi_{\mathcal{M}}(\boldsymbol{i}_{\mathcal{M}})$ following the lexicographical order. As in the log-linear models when the parameter refers to a single variable, it becomes a logit; when it refers to more than one variable, it becomes a contrasts of logits. As simplification, in this section we consider the parameters based on *baseline* logits. Thus, given a marginal set $\mathcal{M}$ and an interaction set $\mathcal{L}$, that is, the set of variables that the parameter refers, the HMM parameter, evaluated in the categories $\boldsymbol{i}_{\mathcal{L}}$, is:

$$\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = \sum_{\mathcal{J} \subseteq \mathcal{L}} (-1)^{|\mathcal{L} \setminus \mathcal{J}|} \log \pi_{\mathcal{M}}(\boldsymbol{i}_{\mathcal{J}}, \mathbf{1}_{\mathcal{M} \setminus \mathcal{J}}). \tag{3}$$

In the formula, $\mathbf{1}_{\mathcal{M} \setminus \mathcal{J}}$ is the vector of coordinates equal to the first category for each variable X_j, such that $j \in \mathcal{M} \setminus \mathcal{L}$. It is worthwhile to consider that, if there is at least one $j \in \mathcal{L}$, such that $i_j = 1$ the parameter in formula (3) becomes zero.

In general, we define a class of partially ordered marginal sets $\mathcal{H} = \{\mathcal{M}_j\}$ by respecting the inclusion such that if $i < j$, then $\mathcal{M}_j \not\subset \mathcal{M}_i$. In the HMMMs, the definition of the parameters within the marginal distributions must satisfy the properties of *completeness* and *hierarchy*, that is, for a subset of variables $\mathcal{L}$ must be only one $\eta_{\mathcal{L}}^{\mathcal{M}_j}$ where $\mathcal{M}_j$ is the first marginal set in $\mathcal{H}$ such that $\mathcal{L} \subseteq \mathcal{M}_j$, see Bergsma and Rudas [3], Bartolucci et al. [2].

Example 4.1. By considering two variables $V = \{1, 2\}$, both with three categories $(1, 2, 3)$ and the hierarchical class of marginal sets $\mathcal{H} = \{(1), (12)\}$, the vector of parameters $\boldsymbol{\eta}$ is

$$\boldsymbol{\eta} = \left[\eta_1^1(2), \eta_1^1(3), \eta_2^{12}(2), \eta_2^{12}(3), \eta_{12}^{12}(2,2), \eta_{12}^{12}(2,3), \eta_{12}^{12}(3,2), \eta_{12}^{12}(3,3)\right]'.$$

These parameters are

$$\eta_1^1(2) = \log\left(\frac{\pi(2+)}{\pi(1+)}\right), \qquad \eta_2^{12}(2) = \log\left(\frac{\pi(12)}{\pi(11)}\right),$$

$$\eta_{12}^{12}(2,2) = \log\left(\frac{\pi(11)\pi(22)}{\pi(21)\pi(12)}\right), \qquad \eta_{12}^{12}(3,2) = \log\left(\frac{\pi(11)\pi(32)}{\pi(31)\pi(12)}\right)$$

$$\eta_1^1(3) = \log\left(\frac{\pi(3+)}{\pi(1+)}\right), \qquad \eta_2^{12}(3) = \log\left(\frac{\pi(13)}{\pi(11)}\right),$$

$$\eta_{12}^{12}(2,3) = \log\left(\frac{\pi(11)\pi(23)}{\pi(21)\pi(13)}\right), \qquad \eta_{12}^{12}(3,3) = \log\left(\frac{\pi(11)\pi(33)}{\pi(31)\pi(13)}\right)$$

where the symbol $\pi(i+)$ refers to the marginal probability of X_1 and where the commas within the parentheses are omitted for short.

As mentioned in Sections 2 and 3, the approach of the (S)CRGMs seems natural when we want to explain the effect of some variables (covariates) on a set of dependent variables that can be, in turn, covariates for other dependent variables. In this section, we present an opportune reparametrization of the HMM parameters able to capture this regression form. Here, we want to improve the CRGMs, as presented by Marchetti and Lupparelli [16], by simplifying the regression equations given the CSIs. Thus, first, we define the appropriate hierarchical class $\mathcal{H}$ of marginal sets, likewise in Nicolussi [18], as the partially ordered marginal sets in $\mathcal{H}_1$ and $\mathcal{H}_2$, where

$$\begin{aligned} \mathcal{H}_1 &= \big\{\big(\mathrm{pa}_T(T_h) \cup A\big), h = 1, \ldots, s;\, A \subseteq T_h\big\} \quad \text{and} \\ \mathcal{H}_2 &= \big\{\big(\mathrm{pre}(T_h) \cup T_h\big), h = 1, \ldots, s\big\}. \end{aligned} \tag{4}$$

The elements of $\mathcal{H}_1$ and $\mathcal{H}_2$ form $\mathcal{H}$, where the elements respect the inclusion property, as specified in the definition of $\mathcal{H}$. Then, focusing on each group of dependent variables, we define the HMM parameters (3) evaluated in each conditional distribution identified by the levels of the covariates. This means that, for each subset of dependent variables $A \subseteq T_h$, we define the parameters $\eta_A^{A \cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A | \boldsymbol{i}_{\mathrm{pa}_T(T_h)})$ evaluated in each values $\boldsymbol{i}_{\mathrm{pa}_T(T_h)} \in \mathcal{I}_{\mathrm{pa}_T(T_h)}$ of the covariates $\mathrm{pa}_T(T_h)$. All these parameters can be expressed as a combination of regression parameters as described in the formula (5). Hereafter, to make the notation more readable, in the subscript and in the superscript we cancel the union symbol, thus $A \cup B$ becomes AB.

Definition 4.1. Given an (S)CRGM, for any subset A of the set of response variables T_h, we have the following regression model:

$$\eta_A^{\mathcal{M}}(\boldsymbol{i}_A | \boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = \sum_{t \subseteq \mathrm{pa}_T(T_h)} \beta_t^A(\boldsymbol{i}_t), \quad \forall h = 1, \ldots, s A \subseteq T_h, \tag{5}$$

where $\mathcal{M} = A \cup \mathrm{pa}_T(T_h)$.

Theorem 4.1. *The regression parameters $\beta_t^A(\boldsymbol{i}_t)$ in the regression model (5), are the HMM parameters*

$$\beta_t^A(\boldsymbol{i}_t) = \eta_{tA}^{\mathcal{M}}(\boldsymbol{i}_{tA}), \quad \forall t \subseteq \mathrm{pa}_T(T_h) \neq \varnothing, \tag{6}$$

where $\mathcal{M} = A \cup \mathrm{pa}_T(T_h)$.

Example 4.2. Let us consider the SCRGM represented by the SCG $\mathcal{G}$ in Figure 1(b) where there are two components. The vertices in the first represent the covariates X_{12}, while the vertices in the second component represent the purely dependent variables X_{345}. Thus, according to formula (4), the class of marginal sets is composed of {(12), (123), (124), (125), (1234), (1235), (1245), (12345)}. By focusing only on the dependent variable X_4, we can express the regression model that explains the variable X_4 as function of X_{12} as follows:

$$\eta_4^{124}(\boldsymbol{i}_4 | \boldsymbol{i}_{12}) = \beta_\varnothing^4 + \beta_1^4(\boldsymbol{i}_1) + \beta_2^4(\boldsymbol{i}_2) + \beta_{12}^4(\boldsymbol{i}_{12})$$

$\forall \boldsymbol{i}_{12} \in \mathcal{I}_{12}$ and $i_4 \in \mathcal{I}_4$. Remember that, when $i_4 = 1$, the HMM parameters involving X_4 are zero by definition thus, also $\eta_4^{124}(1|\boldsymbol{i}_{12})$ is equal to zero. By applying formula (6), we see that the β parameters are

$$\begin{aligned}\beta_{\varnothing}^4 &= \eta_4^{124}(\boldsymbol{i}_4),\\ \beta_1^4(\boldsymbol{i}_1) &= \eta_{14}^{124}(\boldsymbol{i}_{14}),\\ \beta_2^4(\boldsymbol{i}_2) &= \eta_{24}^{124}(\boldsymbol{i}_{24}),\\ \beta_{12}^4(\boldsymbol{i}_{12}) &= \eta_{124}^{124}(\boldsymbol{i}_{124}).\end{aligned}$$

Then by applying formula (5) to each subset of the dependent variables, we should have also the regression models $\eta_3^{123}(\boldsymbol{i}_3|\boldsymbol{i}_{12})$, $\eta_5^{125}(\boldsymbol{i}_5|\boldsymbol{i}_{12})$, $\eta_{34}^{1234}(\boldsymbol{i}_{34}|\boldsymbol{i}_{12})$, $\eta_{35}^{1235}(\boldsymbol{i}_{35}|\boldsymbol{i}_{12})$, and $\eta_{345}^{12345}(\boldsymbol{i}_{345}|\boldsymbol{i}_{12})$ for all values $\boldsymbol{i}_{12} \in \mathcal{I}_{12}$.

The regression models in formula (5) describes the relationships between the dependent variables in $A \subseteq T_h$ and the covariates in $\mathrm{pa}_T(T_h)$ for each $h = 1, \ldots, s$. However, the first components have no parents and the $\mathrm{pa}_T(T_h)$ is empty, thus the variables in T_h have no covariates to explain them. In this case we use the HMM parameters:

$$\eta_A^A(\boldsymbol{i}_A) \quad \forall A \subseteq T_h \text{ such that } \mathrm{pre}(T_h) = \varnothing. \tag{7}$$

Theorem 4.2. *The regression parameters in formula* (5) *and the HMM parameters in formula* (7) *are a* 1:1 *function* (*a re-parametrization*) *of the HMM parameters* $\boldsymbol{\eta}_{\mathcal{L}}^{\mathcal{M}}$, $\forall \mathcal{L} \in \mathcal{P}(V)$ *and* $\forall \mathcal{M}_j \in \mathcal{H}$, *where* $\mathcal{P}(\cdot)$ *denotes the power set.*

As a consequence, a parameterization based on the parameters in formulas (5) and (7) is smooth, since the HMM parameters define a *smooth* parameterization of the set of all strictly positive probability distributions P. On the other hand, the parameters are not variation independent because the marginal sets in $\mathcal{H}$ are not ordered decomposable (unless the number of vertices for any component is at most two).

An SRCGM is represented by linear constraints on the regression parameters in formula (5). In general, is not necessarily true that, a parameterization is able to represent all the statements in an independence model faithful to a CG, $\mathcal{J}(\mathcal{G})$, see Drton [9], Nicolussi and Colombi [21]. Theorem 4.3 explains how to constrain the parameters according to any missing arc and any labeled arc of the SCRGM.

Theorem 4.3. *Given an SCG* $\mathcal{G}$, *the induced independence model* $\mathcal{J}(\mathcal{G})$ *is equivalent to the regression models in formula* (5), *coded with the baseline aggregation criterion of the categories*, *where for any component* T_h, *with* $h = 1, \ldots, s$, *the following constraints holds.*

In the full chain graph $\mathcal{G}^F$, *for any subset* A *of* T_h:

i. *if* A *is non connected*, $\forall \boldsymbol{i}_{\mathrm{pa}_T(T_h)} \in \mathcal{I}_{\mathrm{pa}_T(T_h)}$

$$\eta_A^{A \cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = 0; \tag{8}$$

ii. *if* A *is connected*, $\forall \boldsymbol{i}_{\mathrm{pa}_T(T_h)\setminus \mathrm{pa}_{\mathcal{G}}(A)} \in \mathcal{I}_{\mathrm{pa}_T(T_h)\setminus \mathrm{pa}_{\mathcal{G}}(A)}$,

$$\eta_A^{A \cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = \sum_{t \subseteq \mathrm{pa}_{\mathcal{G}}(A)} \beta_t^A(\boldsymbol{i}_t). \tag{9}$$

In any reduced chain graph $\mathcal{G}^R(\boldsymbol{i}_l)$, *for any subset* A *of* T_h:

iii. *if* A *is non connected*, $\forall \boldsymbol{i}_{\mathrm{pa}_T(T_h)} \in (\mathcal{K}_l \times \mathcal{I}_{\mathrm{pa}_T(T_h)\setminus l})$,

$$\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = 0; \tag{10}$$

iv. *if* A *is connected*, $\forall \boldsymbol{i}_{\mathrm{pa}_T(T_h)\setminus \mathrm{pa}_{\mathcal{G}^R(\boldsymbol{i}_l)}(A)} \in \{\mathcal{K}_l \times \mathcal{I}_{\mathrm{pa}_T(T_h)\setminus(\mathrm{pa}_{\mathcal{G}^R(\boldsymbol{i}_l)}(A)\cup l)}\}$,

$$\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = \sum_{t\subseteq \mathrm{pa}_{\mathcal{G}^R(\boldsymbol{i}_l)}(A)} \beta_t^A(\boldsymbol{i}_t). \tag{11}$$

Note that, according to (i) and (iii), the regression models involving two or more dependent variables, that are each other conditionally independent, are set to zero. Instead, according to the *(ii)* and *(iv)*, the dependent variables are explained only by their parents.

The constraints (iii) and (iv) come from the linear constraints on HMM parameters that satisfy the CSIs (see La Rocca and Roverato [13], Nicolussi and Cazzaro [19]). Indeed, given a CSI as $\langle A., B|C; \mathcal{K}_l \times \mathcal{I}_{C\setminus l}\rangle$, if the variables in the vector X_C are coded with the *baseline* criterion, then the constraints on the HMM parameters are:

$$\sum_{c\subseteq C} \eta_{vc}^{\mathcal{M}}(\boldsymbol{i}_v, \boldsymbol{i}'_c) = 0, \quad \boldsymbol{i}_v \in \mathcal{I}_v, \boldsymbol{i}'_c \in \mathcal{I}_c \cap (\mathcal{K}_l \cap \mathcal{I}_{C\setminus l}), \tag{12}$$

$\forall v \in \mathcal{V} = \{v \subseteq (A \cup B) : A \cap v \neq \varnothing, B \cap v \neq \varnothing\}$ and for a marginal set defined by the variables $v \cup C \subseteq \mathcal{M} \subseteq A \cup B \cup C$.

Note that, zero is an admissible value for the parameters in formula (12), without implying any other CSI, except for when $c = \varnothing$. On the contrary, this framework would fail the hierarchy property. Besides, according to the parsimonious principle, simpler models can be achieved setting to zero the parameters with order higher than two.

Example 4.3 (Continuation of Example 4.2). Let us consider the SCRGM represented b y the SCG $\mathcal{G}$ in Figure 1(b). Then the following parameters entirely describe the relationships among all the variables;

$$\begin{aligned}
\eta_3^{123}(\boldsymbol{i}_3|\boldsymbol{i}_{12}) &= \beta_\varnothing^3 + \beta_1^3(\boldsymbol{i}_1), \quad \forall \boldsymbol{i}_3 \in \mathcal{I}_3, \forall \boldsymbol{i}_{12} \in \mathcal{I}_{12}\\
\eta_4^{124}(\boldsymbol{i}_4|\boldsymbol{i}_{12}) &= \beta_\varnothing^4 + \beta_1^4(\boldsymbol{i}_1) + \beta_2^4(\boldsymbol{i}_2) + \beta_{12}^4(\boldsymbol{i}_{12}), \quad \forall \boldsymbol{i}_4 \in \mathcal{I}_4, \forall \boldsymbol{i}_{12} \in \mathcal{I}_{12}\\
\eta_5^{125}(\boldsymbol{i}_5|\boldsymbol{i}_{12}) &= \beta_\varnothing^5, \quad \forall \boldsymbol{i}_{12} \in \mathcal{I}_{12}\\
\eta_{34}^{1234}(\boldsymbol{i}_{34}|\boldsymbol{i}_{12}) &= \begin{cases} 0, & \forall \boldsymbol{i}_{34} \in \mathcal{I}_{34}, \forall \boldsymbol{i}_{12} \in \mathcal{K}_1 \times \mathcal{I}_2,\\ \beta_\varnothing^{34} + \beta_1^{34}(\boldsymbol{i}_1) + \beta_2^{34}(\boldsymbol{i}_2) + \beta_{12}^{34}(\boldsymbol{i}_{12}), & \forall \boldsymbol{i}_{34} \in \mathcal{I}_{34}, \forall \boldsymbol{i}_{12} \notin \mathcal{K}_1 \times \mathcal{I}_2, \end{cases}\\
\eta_{35}^{1235}(\boldsymbol{i}_{35}|\boldsymbol{i}_{12}) &= \beta_\varnothing^{35} + \beta_1^{35}(\boldsymbol{i}_1), \quad \forall \boldsymbol{i}_{35} \in \mathcal{I}_{35}, \forall \boldsymbol{i}_{12} \in \mathcal{I}_{12}\\
\eta_{45}^{1245}(\boldsymbol{i}_{45}|\boldsymbol{i}_{12}) &= \beta_\varnothing^{45} + \beta_1^{45}(\boldsymbol{i}_1) + \beta_2^{45}(\boldsymbol{i}_2) + \beta_{12}^{45}(\boldsymbol{i}_{12}), \quad \forall \boldsymbol{i}_{45} \in \mathcal{I}_{45}, \forall \boldsymbol{i}_{12} \in \mathcal{I}_{12}\\
\eta_{345}^{12345}(\boldsymbol{i}_{345}|\boldsymbol{i}_{12}) &= \beta_\varnothing^{345} + \beta_1^{345}(\boldsymbol{i}_1) + \beta_2^{345}(\boldsymbol{i}_2) + \beta_{12}^{345}(\boldsymbol{i}_{12}), \quad \forall \boldsymbol{i}_{345} \in \mathcal{I}_{345}, \forall \boldsymbol{i}_{12} \in \mathcal{I}_{12}.
\end{aligned}$$

An SCRGM is represented by a set of linear constraints on the HMM parameters. This implies that, it belongs to the curved exponential family, see Bergsma and Rudas [3], Bartolucci et al. [2]. Besides,

whatever is the sampling scheme we assume in the data collection, the log-likelihood is proportional to

$$l(\boldsymbol{\theta};\boldsymbol{n}) \propto \sum_{\boldsymbol{i}\in\mathcal{I}} n(\boldsymbol{i})\theta(\boldsymbol{i}) - \sum_{\boldsymbol{i}\in\mathcal{I}} \log\big(\theta(\boldsymbol{i})\big), \tag{13}$$

where $n(\boldsymbol{i})$ is the observed frequency of the cell $\boldsymbol{i} \in \mathcal{I}$ and $\theta(\boldsymbol{i})$ is the logarithm of the expected value of the cell $\boldsymbol{i}$. The maximization of this function, under the linear constraints of the Theorem 4.3, can be achieved through iterative algorithm, for more details, see Aitchison and Silvey [1], Bartolucci et al. [2], Marchetti and Lupparelli [17]. Further considerations about the convergence of the iterative algorithm are postponed in Appendix A.1.

Example 4.4. Let us consider the SCRGM represented by the SCG $\mathcal{G}$ in Figure 2(a) where the dependent variables are represented by the vertices 2, 3, and 4 in the component T_2, while the single covariate is represented by the vertex 1 in the component T_1. In compliance with Definition 4.1, we define the class of marginal sets such as {(1), (12), (13), (14), (123), (124), (134), (1234)}. In Example 3.4, we showed that, according to the global Markov properties in Definition 3.5, the independence model $\mathcal{J}(\mathcal{G})$ is composed of the following list of independencies:

(a) $\langle X_2, X_4 | X_1 \rangle$;
(b) $\langle X_{24}, X_3 | X_1; \{(2)\}\rangle$;
(c) $\langle X_{23}, X_4 | X_1; \{(3)\}\rangle$.

Now by applying Theorem 4.3, we obtain the following constraints on the parameters.

According to (a), we get $\eta^{124}_{24}(\boldsymbol{i}_{24}|\boldsymbol{i}_1) = 0$ for any $\boldsymbol{i}_1 \in \mathcal{I}_1$ and $\boldsymbol{i}_{24} \in \mathcal{I}_{24}$. According to (b), we get that $\eta^{1234}_{23}(\boldsymbol{i}_{23}|\boldsymbol{i}_1 = 2) = \eta^{1234}_{34}(\boldsymbol{i}_{34}|\boldsymbol{i}_1 = 2) = \eta^{1234}_{234}(\boldsymbol{i}_{234}|\boldsymbol{i}_1 = 2) = 0$ for all $\boldsymbol{i}_{23} \in \mathcal{I}_{23}$, $\boldsymbol{i}_{34} \in \mathcal{I}_{34}$, and $\boldsymbol{i}_{234} \in \mathcal{I}_{234}$. Finally, according to (c), we get that $\eta^{1234}_{24}(\boldsymbol{i}_{24}|\boldsymbol{i}_1 = 3) = \eta^{1234}_{34}(\boldsymbol{i}_{34}|\boldsymbol{i}_1 = 3) = \eta^{1234}_{234}(\boldsymbol{i}_{234}|\boldsymbol{i}_1 = 3) = 0$, for all $\boldsymbol{i}_{24} \in \mathcal{I}_{24}$, $\boldsymbol{i}_{34} \in \mathcal{I}_{34}$, and $\boldsymbol{i}_{234} \in \mathcal{I}_{234}$.

Example 4.5. Let us consider the SCRGM represented by the SCG $\mathcal{G}$ in Figure 3(a), where the dependent variable is represented by the vertex 4 in the component T_2 and the covariates are the vertices 1, 2, and 3 in the component T_1. We build the class of marginal sets such as {(123), (1234)}. From the SCRGM, we can extract the following list of independencies, see the Example 3.5:

(a) $\langle X_{13}, X_4 | X_2; \{(2)\}\rangle$ from the Figure 3(c);
(b) $\langle X_1, X_4 | X_{23}; \{(3)\} \times \mathcal{I}_3\rangle$, from the Figure 3(d).

According to (a), when $i_2 = 2$, we get that

$$\eta^{1234}_4(\boldsymbol{i}_4|\boldsymbol{i}_{123}) = \beta^4_{\varnothing} + \beta^4_2(\boldsymbol{i}_2).$$

Besides, when $i_2 = 3$, the regression parameters becomes

$$\eta^{1234}_4(\boldsymbol{i}_4|\boldsymbol{i}_{123}) = \beta^4_{\varnothing} + \beta^4_2(\boldsymbol{i}_2) + \beta^4_3(\boldsymbol{i}_3) + \beta^4_{23}(\boldsymbol{i}_{23}).$$

Thus, the following formula summarizes all the regression parameters:

$$\begin{aligned}\eta^{1234}_4(\boldsymbol{i}_4|\boldsymbol{i}_{123}) = \beta^4_{\varnothing} + \beta^4_2(\boldsymbol{i}_2) + \mathbb{1}_{i_2\notin(2,3)}\big(\beta^4_1(\boldsymbol{i}_1) + \beta^4_{12}(\boldsymbol{i}_{12})\big) \\ + \mathbb{1}_{i_2\notin(2)}\big(\beta^4_3(\boldsymbol{i}_3) + \beta^4_{13}(\boldsymbol{i}_{13}) + \beta^4_{23}(\boldsymbol{i}_{23}) + \beta^4_{123}(\boldsymbol{i}_{123})\big)\end{aligned}$$

In the previous equations, the symbol $\mathbb{1}$ is equal to 1 when the condition in the subscript is satisfied and zero otherwise.

The independence model faithful to a CRGM is always representable trough constraints on HMM parameters via an opportune choice of the marginal sets in $\mathcal{H}$, as proved in Nicolussi [18], Marchetti and Lupparelli [16]. Note that this result is not necessarily right for chain graphs satisfying other Markov properties, as shown in Nicolussi and Colombi [21]. In the case of SCRGM, the constraints needed for the CSIs involve the same parameters needed for the conditional independencies, thus the only limits on the compatible independence statements are listed in Lemma 1. In the following example, we show that the same limits comes by also using the HMM parameters.

Example 4.6. Let us consider the SCG in Figure 4. As discussed in Example 3.6, the independence model $\mathcal{J}(\mathcal{G})$ should contain the statements $\langle X_2, X_3|X_1\rangle$ and $\langle X_1, X_3|X_2; \mathcal{K}_2\rangle$. According to the first statement we have the constraints $\eta^{123}_{23}(\boldsymbol{i}_{23}) = \eta^{123}_{123} = 0$ for all $\boldsymbol{i}_{23} \in \mathcal{I}_{23}$ and $\boldsymbol{i}_{123} \in \mathcal{I}_{123}$. According to the second statement we have the constraints $\eta^{123}_{13}(\boldsymbol{i}_{13}) + \eta^{123}_{123} = 0$ for all $\boldsymbol{i}_{13} \in \mathcal{I}_{13}$ and $\boldsymbol{i}_2 \in \mathcal{K}_2$. By considering together the constraints, since $\eta^{123}_{123} = 0$ the second constraints becomes $\eta^{123}_{13}(\boldsymbol{i}_{13}) = 0$. But, these two last constraints hold if and only if the statement $\langle X_1, X_3|X_2\rangle$ holds.

4.2. Parameterization for ordinal variables

The considerations performed in Section 3 and especially in Section 4, are typically for unordered (nominal) variables. To contemplate the order of the categories of the variables, we use parameters based on a different aggregation criterion, such as the *local* logit. This aggregation criterion consists in replacing the cells $\mathbf{1}_{\mathcal{M}\setminus\mathcal{J}}$ in formula (3) with the coordinates $((\boldsymbol{i}_{\mathcal{L}\setminus\mathcal{J}} - 1), \mathbf{1}_{\mathcal{M}\setminus\mathcal{L}})$, where $(\boldsymbol{i}_{\mathcal{L}\setminus\mathcal{J}} - 1)$ denotes the level $(i_j - 1)$ for all $j \in \mathcal{L}\setminus\mathcal{J}$. Example 4.7 shows the form of the HMM parameters when we use parameters based on *local* logits.

Example 4.7. Let us consider two variables X_V, $V = \{1, 2\}$, both with three ordered values $(1, 2, 3)$ and the class of marginal sets $\mathcal{H} = \{(1), (12)\}$. Then, when the variables concerning the whole set of variables are based on the *local* logits, each parameter in the vector $\boldsymbol{\eta}$ is

$$\eta^{1}_{1}(2) = \log\left(\frac{\pi(2+)}{\pi(1+)}\right), \qquad \eta^{12}_{2}(2) = \log\left(\frac{\pi(12)}{\pi(11)}\right),$$
$$\eta^{12}_{12}(2,2) = \log\left(\frac{\pi(11)\pi(22)}{\pi(21)\pi(12)}\right), \qquad \eta^{12}_{12}(3,2) = \log\left(\frac{\pi(21)\pi(32)}{\pi(31)\pi(22)}\right),$$
$$\eta^{1}_{1}(3) = \log\left(\frac{\pi(3+)}{\pi(2+)}\right), \qquad \eta^{12}_{2}(3) = \log\left(\frac{\pi(13)}{\pi(12)}\right),$$
$$\eta^{12}_{12}(2,3) = \log\left(\frac{\pi(12)\pi(23)}{\pi(22)\pi(13)}\right), \qquad \eta^{12}_{12}(3,3) = \log\left(\frac{\pi(22)\pi(33)}{\pi(32)\pi(23)}\right).$$

The multivariate system of regression models based on the HMM parameters, such as described in Theorem 4.1, holds, whatever is the type of logits considered. As a consequence, the results in Theorems 4.1 and 4.2 still hold also when we use the *local* criterion.

Formula (2), in Section 3, presents the definition of a CSI as a conditional independence that holds only when the variables in the conditioning set assume (are equal to) certain categories $\boldsymbol{i}_C$ in $\mathcal{K}_C$. Here, by considering the order of the categories, we define the CSIs like a conditional independence that holds only when the variables in the conditioning set are equal or lower than a certain value $\boldsymbol{i}'_C$.

Hence, we take into account a subset of CSIs in formula (2) where the class $\mathcal{K}_C$ is composed of all cells of coordinates $\bigcap_{j\in C}(i_j \leq i'_j)$.

Formally, we say that X_A and X_B are independent given X_C when the variables in X_C assume values lower or equal to the threshold $\boldsymbol{i}'_C$ and we write it as

$$\langle X_A, X_B | X_C; \leq \boldsymbol{i}'_C \rangle. \tag{14}$$

According to this new definition, the labels in the SCRGM become inequalities.

This new formulation leads to a new constraint on the regression parameters. In particular, by following Nicolussi and Cazzaro [19], we have that whatever the chosen aggregation criterion, the CSIs in formula (14) are represented by the following constraints on the HMM parameters in formula (3):

$$\eta^{\mathcal{M}}_{vc}(\boldsymbol{i}_v, \boldsymbol{i}_c) = 0 \quad \forall \boldsymbol{i}_c \leq \boldsymbol{i}'_c, \boldsymbol{i}'_c \in (\mathcal{K} \cap \mathcal{I}_c), \boldsymbol{i}_v \in \mathcal{I}_v \tag{15}$$

$\forall v \in \mathcal{V} = \{v \subseteq (A \cup B) : A \cap v \neq \varnothing, B \cap v \neq \varnothing\}$ and $\forall c \subseteq C$. Note that, unlike the constraints in formula (12), here, we set to zero certain parameters and not the sum of them.

By applying these new constraints to the regression parameters, the following result is reached out.

Theorem 4.4. *Given an SCRGM, the global stratified Markov properties with inequality labels in Definition 3.5, are equivalent to the constraints in formula* (5)*,when the parameters concerning the whole set of variables are based on local logits.*

Note that, if we want to focus on a CSI like this:

$$\langle X_A, X_B | X_C; \geq \boldsymbol{i}'_C \rangle, \tag{16}$$

it is enough to reverse the order of the modalities.

5. Application

In this section, we present an application on a real data set. In particular, we select a set of variables intending to investigate the relationships among these by supposing that some groups of variables can affect others unilaterally. Thus, we suggest an SCRGM, learned from the data, able to describe these associations. At first, we collect the variables in chain components, splitting the pure response variables from the pure covariates and the mixed variables.

In compliance with these chain components, we define the class of marginal sets, according to formula (4). Finally, we should test all possible conditional, marginal, and context-specific independencies and choose the best-fitting model, but this procedure is computationally expensive. However, several procedures lead to the choice of a model. Undoubtedly, the aim of the analysis plays an important role. If, for instance, we want to find the system of different independence relationships among the variables, we can overlook to the nature of the variables (if ordinal or nominal) and use the parameters concerning the whole set of variables *based* on the baseline logits and consider the formulation of the CSIs in formula (2). In the research of the best-fitting model, the preference is for the plausible models with the lowest number of free parameters (according to the parsimonious principle). Analysis involving a large number of variables where the description of each dependence relationship is unnecessary and of difficult realization, usually adopt this approach. This setting could create a model with a lot of CSIs holding for many different categories of conditioning variables that are arduous, even to list. On the other hand, when the aim of the analysis is the interpretation of the dependence relationships

Algorithm 1 Learning procedure

$G_1 = G_0$
for $\gamma, \delta \in V$ and $C \subseteq \text{pre}(\gamma \cup \delta)$ **do**
  test the statement $\langle X_\gamma, X_\delta | X_C \rangle$ with the likelihood ratio test
  if there is the evidence of $\langle X_\gamma, X_\delta | X_C \rangle$ for at least one C **then**
    remove the arc $\gamma \to \delta$ or $\gamma \leftrightarrow \delta$ from G_1
  end if
end for
$G_2 = G_1$
for $(\gamma, \delta) \in E_1$ **do**
  set $C = \text{pre}(T_h)$
  for $\boldsymbol{i}'_C \in \mathcal{I}_C$ **do**
    test the statement $\langle X_\gamma, X_\delta | C; \geq \boldsymbol{i}'_C \rangle$ with the likelihood ratio test
    if there is the evidence of $\langle X_\gamma, X_\delta | C; \geq \boldsymbol{i}'_C \rangle$ for at least one $\boldsymbol{i}'_C$ **then**
      replace the arc $\gamma \to \delta$ or $\gamma \leftrightarrow \delta$ with one labeled arc in G_2
    end if
  end for
end for
$G_3 = G_2$
test the independence model G_3
while there is no evidence for the independence model faithful to G_3 **do**
  Select one edge $(\gamma, \delta) \in (V \backslash E_2)$, set $C = \text{pre}(T_h)$ and select one index $\boldsymbol{i}'_C \in \mathcal{I}_C$. Add to G_3 the labeled arc γ, δ with the label $\boldsymbol{i}'_C$ and test the corresponding. model
  return G_3
end while

among a selected set of variables, it is worthwhile to reduce the possible CSIs to the one defined in formula (14) and/or in formula (16), expressed in terms of inequalities. This approach could not lead to the best-fitting model, compared with the previous approach; however, the resulting model is more meaningful.

In this application, we are inclined to adopt the second point of view, even if a final comparison with an SCRGM with only CSIs expressed through equality terms is mentioned.

To obtain the best-fitting model in this framework, we first split the variables X_V into groups (corresponding to chain components), and we confer an established direction of the arrow among the components. We call $\mathcal{G}_0$ the starting chain graph with that partition of the variables and with all possible arcs. Then we adopt the procedure listed in Algorithm 1. Note that, in the algorithm the symbol E_1 stays for the set of edges concerning the graph G_1.

Innovation study survey 2010–2012. The section aims to build a chain regression model that studies the effect of innovation in some aspects of the enterprise's life on revenue growth without omitting the main features. Thus, we collect the following variables from the survey on the innovation status of small and medium Italian enterprises from 2010 to 2012, ISTAT [12]. At first, as pure response, we consider the *revenue growth* variable in 2012, *GROW* (Yes = 1, No = 0), henceforth denoted as variable X_1. Then as mixed variables, we consider the innovation through three dichotomous variables referring to the period 2010–2012: *innovation in products or services or production line or investment in R&D*, *IPR* (Yes = 1, No = 0), *innovation in organization system*; *IOR* (Yes = 1, No = 0); and *innovation in marketing strategies*, *IMAR* (Yes = 1, No = 0), henceforth denoted as variables X_2, X_3, and X_4,

respectively. Finally, the role of purely covariates is entrusted to the variables concerning the firm's featuring in 2010–2012: the *main market (in revenue terms)*, *MRKT* (1 = Regional, 2 = National, 3 = International); the *percentage of graduate employers*, *DEG* ($0\% \vdash 10\% = 1$, $10\% \vdash 50\% = 2$, $50\% \vdash 100\% = 3$); and the *enterprise size*, *DIM* (Small = 1, Medium = 2), henceforth denoted as variables X_5, X_6, and X_7, respectively. The survey covers 18697 firms, collected in a $2 \times 2 \times 2 \times 2 \times 3 \times 3 \times 2$ contingency table.

To analyze this data set, we build a chain graph with three components according to the nature of the variables; thus, in the first component, we collect the firm's feature variables (X_{567}); in the second component, the innovations variables (X_{234}); and in the third component, the revenue growth variable (X_1). According to formula (4), we consider the following marginal sets: {(5, 6, 7); (2, 5, 6, 7); (3, 5, 6, 7); (4, 5, 6, 7); (2, 3, 5, 6, 7); (2, 4, 5, 6, 7); (3, 4, 5, 6, 7); (2, 3, 4, 5, 6, 7); (1, 2, 3, 4, 5, 6, 7)}. The parameters associated with the dichotomous variables were based on *baseline* logits, while the parameters concerning the variables with three levels are based on the *local* one.

According to the learning procedure in Algorithm 1, the best-fitting model is represented in Figure 5. In correspondence to this model, the likelihood ratio test produced the following results `Gsq` $= 155.79$, `df` $= 132$, `p-val` $= 0.08$, and `AIC` $= -108.21$. By applying the global stratified Markov properties to the graph, we obtain:

- $X_1 \perp\!\!\!\perp X_{46} | X_{2357}$,
- $X_1 \perp\!\!\!\perp X_{234} | (X_{567} \geq (2, 3, 1))$,
- $X_1 \perp\!\!\!\perp X_{246} | (X_{357} \geq (0, 2, 1))$,
- $X_3 \perp\!\!\!\perp X_5 | (X_{67} \geq (3, 2))$.

In this model, the innovation in marketing strategies, X_4, and the percentage of graduated employers, X_6, do not affect the revenue growth of the firms, X_1, conditioning on the other variables. Furthermore, there are other weak relationships defined by the CSIs represented by the strata. Firstly, the revenue growth, X_1, is independent of all the types of innovation, X_{234}, when the primary market where the firm works is national or international ($X_5 \geq 2$) and the degree of the graduated employer is high ($X_6 = 3$), whatever the firm dimension ($X_7 \geq 1$). Second, we have that the revenue growth, X_1, is independent of the innovation in product and services, X_2, and in marketing strategies, X_4, and from the percentages of graduated employers, X_6, whatever is the innovation in the organization system ($X_3 \geq 0$) and the firm dimension ($X_7 \geq 1$), and when the firm does not work in the regional market. Finally, the organization system's innovation, X_3, is independent of the market where the firm works, X_5, when the firm is medium ($X_7 = 2$), and when the percentage of graduated employers is the highest ($X_6 = 3$).

These independencies correspond to simplifications on the regression models represented in the SCRGM in Figure 5. In fact, according to the conditional independence $X_1 \perp\!\!\!\perp X_4 | X_{23567}$, we get that all the covariate $\beta^1_{4t}(\boldsymbol{i}_{4t})$ of the regression model $\eta^V_1(\boldsymbol{i}_1 | \boldsymbol{i}_{V\setminus 1})$ are null for all the subsets t of $(2, 3, 5, 6, 7)$. Furthermore, $\beta^1_{3,t}(i_{3t})$ is null for the subsets t of $(2, 4, 5, 6, 7)$ and for the categories $\boldsymbol{i}_t = \boldsymbol{i}_{24567} \cap \mathcal{I}_t$, where $\boldsymbol{i}_{24567} \geq (1, 1, 1, 3, 1)$. Similarly, $\beta^1_{2,t}(i_{2t})$ is null for the subsets t of $(3, 4, 5, 6, 7)$ and for the categories $\boldsymbol{i}_t = \boldsymbol{i}_{34567} \cap \mathcal{I}_t$, where $\boldsymbol{i}_{34567} \geq (0, 0, 3, 1, 1)$. Finally, $\beta^1_{6,t}(i_{6t})$ is null for the subsets t of $(2, 3, 4, 5, 7)$ and for the categories $\boldsymbol{i}_t = \boldsymbol{i}_{23457} \cap \mathcal{I}_t$, where $\boldsymbol{i}_{23457} \geq (0, 0, 0, 2, 1)$. The values of the regression model that explain the dependent variable X_1 as a function of the remaining variables in X_V are displayed in Table 2 in Nicolussi and Cazzaro [20].

To support the learning procedure by considering only the CSIs expressed through inequality terms, such as in formula (14), we report also the best-fitting SCRGM obtained by considering all the possible CSIs in equality terms in Step 2 of the procedure. Figure 6 displays the graph representing the independencies $X_1 \perp\!\!\!\perp X_4 | X_{23567}$, $X_1 \perp\!\!\!\perp X_3 | X_{24567} = \boldsymbol{i}'_{24567}$ for all $\boldsymbol{i}'_{24567} \in \mathcal{K}_{24567}$,

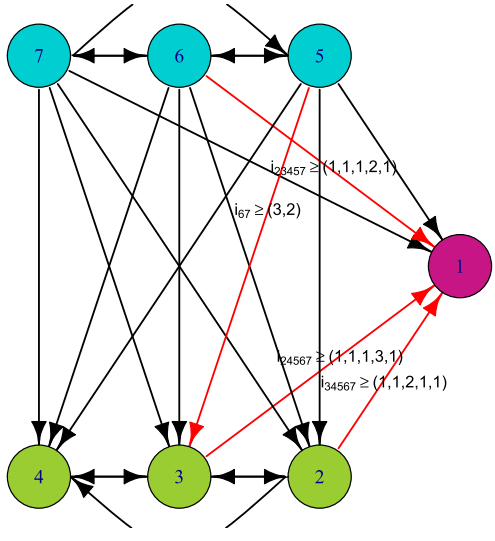

Figure 5. SCRGM with components $T_1 = (5, 6, 7)$, $T_2 = (2, 3, 4)$ and $T_3 = (1)$, with inequality constraints.

$X_1 \perp\!\!\!\perp X_6 | X_{23457} = \boldsymbol{i}'_{23457}$ for all $\boldsymbol{i}'_{23457} \in \mathcal{K}_{23457}$ $X_2 \perp\!\!\!\perp X_7 | X_{56} = (3, 3)$. Note that the graph is not exhaustive to explicit the independencies because the list of the categories in the strata is too big to be displayed. Indeed, the full list is in Table 4 in Nicolussi and Cazzaro [20].

Finally, the output of this application shows a little aspect of what we can derive from the application of this model. For instance, once fitted, the model can be used to forecast the values of some dependent variables given the covariate, or again, looking at the regression parameters, it is possible to define a

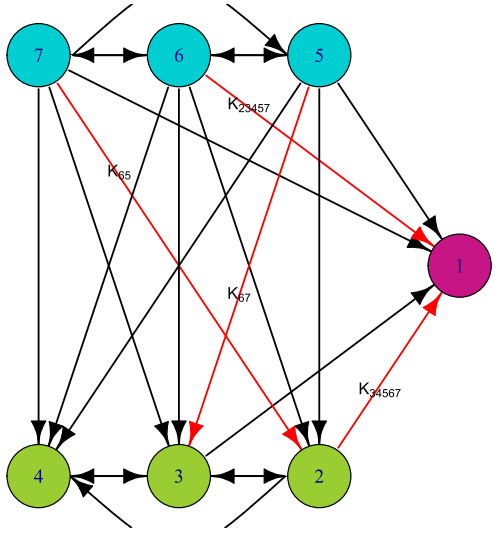

Figure 6. SCRGM with components $T_1 = (5, 6, 7)$, $T_2 = (2, 3, 4)$, and $T_3 = (1)$, with equality constraints.

strategy where to invest. The possibilities are several. It depends on the aim of the analysis. The HMM parameters (that are not listed here) can be used to study the relationships among the variables.

All the analysis were carried out with the statistical software R, (R Core Team [25]), with the help of the package `hmmm`, Colombi et al. [7], for testing the HMM models.

6. Conclusions

In this work, we generalized the concept of the chain regression graphical models through the CS independencies. We provided original methodological results such as the pairwise and global stratified Markov properties needed to read the independencies from an SCRGM. Further, we listed the conditions for admissible labeled arcs in the SCGs. We proposed a system of regression models faithful to the SCRGM through opportune constraints. Furthermore, we considered a subset of the SCRGM with inequalities constraints in the CSIs. These subclasses of the models have different advantages, especially when we want to deepen the meaning of the parameters instead of focusing only on the parsimonious principle. The application showed these two different ways to face the SCRGMs. Besides, we suggested to take advantage of a parameterization based on the HMMMs. This class of models is widely studied and offer several advantages. Indeed, the chosen parameterization is smooth, but its parameters are not variation independent. Undoubtedly, investigating further properties of this parameterization can be considered an interesting challenge for future works. Assuredly, additional context-specific independence statements make, the already huge model space, larger, and there is no optimal solution for exploring this vast model space. In this work, we proposed a possible way to investigate this space well aware of limiting the research to a subspace. A deepen research of algorithms for the learning procedure, with a study on performance, is a topic for future research.

Appendices

A.1. Further results

Note on the maximum-likelihood estimations. As Bartolucci et al. [2], we put all the HMM parameters in the vector $\boldsymbol{\eta}$ by following the lexicographical order. Thus, the linear constraint on the HMM parameters such as in formula (12) can be expressed as $E\boldsymbol{\eta} = \mathbf{0}$ for a suitable constraints matrix. All the considerations carried out about the existence of maximum likelihood estimations, and the convergence of the iterative algorithm hold if E is a full rank matrix. The row number of E is the number of linear constraints, while the column number is the number of parameters. In order to have a full-rank matrix, the rows must be linear independent. When we handle constraints for conditional independence, it is easy to see this because any row of the matrix has only one element equal to 1 (in the position of the parameter to constrain), and the other entries are zero. In general, in order to constrain CSIs, the rows of E must have more than one 1 entries. Wider constraints can include more than one constraint, such as explained in the proof of Theorem 4.4. In general, this is not a problem, and in the implementation of the matrix E, it can be useful to change the 1 entries into 0 entries when the row refers to the wider constraints. Furthermore, this simplification allows us to see that the constraints are not linear dependent easily. See also Nicolussi and Cazzaro [19] for further details. However, some constraints may imply more reliable conditions, when the addition of constraints for a CSI implies new conditional independence. Nevertheless, Example 4.6 discuss this topic.

A.2. Proofs

Proof of Lemma 1. If $\langle A, B|C\rangle \in \mathcal{J}(P)$ then $P(i_A|i_B, i_C) = P(i_A|i_C)$ for all $i_A \in \mathcal{I}_A$, $i_B \in \mathcal{I}_B$, and $i_C \in \mathcal{I}_C$ thus, information about the value of B being irrelevant to determine A. If $\langle A, C|B; \mathcal{K}_B\rangle \in \mathcal{J}(P)$ then, $P(i_A|i_B, i_C) = P(i_A|i_B)$ for any $i_A \in \mathcal{I}_A$, $i_B \in \mathcal{K}_B$, and $i_C \in \mathcal{I}_C$. If both the statements belong to $\mathcal{J}(P)$, the probability $P(i_A|i_B, i_C)$ must be equal to $P(i_A)$ when $i_B \in \mathcal{K}_B$ and $P(i_A|i_C)$ when $i_B \notin \mathcal{K}_B$. However, since $P(i_A|i_B, i_C)$ does not depend on i_B anymore, its values do not discriminate the value of the probability of $P(i_A|i_B, i_C)$. This implies that if $\langle A, B|C\rangle \in \mathcal{J}(P)$ and $\langle A, C|B; \mathcal{K}_B\rangle \in \mathcal{J}(P)$ then also $\langle A, C|B\rangle \in \mathcal{J}(P)$. Note that this last statement is stronger than the CSI $\langle A, C|B; \mathcal{K}_B\rangle \in \mathcal{J}(P)$. Thus, we prove that it is not possible to include in the same independence model $\langle A, B|C\rangle$ and $\langle A, C|B; \mathcal{K}_B\rangle \in \mathcal{J}(P)$ without including a stronger condition. □

Proof of Theorem 3.1. First, it is opportune to highlight a consideration. From Theorem 1 of Lauritzen and Sadeghi [14], the independence models $\mathcal{J}(\mathcal{G}^F)$ and all the $\mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ satisfy all the properties of Definition 2.1 since the full graph and the reduced graphs are chain graphs. Now we have to prove that this holds also for the independence models faithful to the SGC $\mathcal{J}(\mathcal{G})$. Furthermore, the following proof considers only the case of CSIs such as $\langle A, B|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ since it covers also the conditional independence statements when $\mathcal{K}_C = \mathcal{I}_C$ and the reference graph is $\mathcal{G}^F$.

(Symmetry). This easily comes from the fact that the $\mathcal{J}(\mathcal{G}^F)$ and $\mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ independence models satisfy this rule. Indeed, $\langle A, B|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ if and only if $\langle B, A|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and both belong to $\mathcal{J}(\mathcal{G})$.

(Decomposition). This property comes from the fact that independence models $\mathcal{J}(\mathcal{G}^F)$ and $\mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ satisfy this rule. Indeed, if $\langle A, B \cup D|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ then $\langle A, B|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and $\langle A, D|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and all these statements belong to $\mathcal{J}(\mathcal{G})$.

(*Weak union*). This property comes from the fact that independence models $\mathcal{J}(\mathcal{G}^F)$ and $\mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ satisfy this rule. Indeed, if $\langle A, B \cup D|C; \mathcal{K}_C\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ then $\langle A, B|C \cup D; \mathcal{K}_C \times \mathcal{I}_D\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and $\langle A, D|C \cup B; \mathcal{K}_C \times \mathcal{I}_B\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and all these statements belong to $\mathcal{J}(\mathcal{G})$.

(*Contraction*). About this property, we need to discriminate 2 cases.

(*case 1*): the restriction of the CSIs is limited to the set C, (that is $l \subseteq C$). Since $\mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ satisfies this property, $\langle A, B|C \cup D; \boldsymbol{i}_l \times \mathcal{I}_{C\setminus l} \times \mathcal{I}_D\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and $\langle A, D|C; \boldsymbol{i}_l \times \mathcal{I}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$, if and only if $\langle A, B \cup D|C; \boldsymbol{i}_l \times \mathcal{I}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and these statements hold also in $\mathcal{J}(\mathcal{G})$. In general, if $\langle A, B|C \cup D; \mathcal{K}_C^1 \times \mathcal{I}_D\rangle \in \mathcal{J}(\mathcal{G})$ and $\langle A, D|C; \mathcal{K}_C^2\rangle \in \mathcal{J}(\mathcal{G})$ then $\langle A, B \cup D|\mathcal{K}_C\rangle \in \mathcal{J}(\mathcal{G})$ where $\mathcal{K}_C = \mathcal{K}_C^1 \cap \mathcal{K}_C^2$. This comes from the property R1 in the Definition 3.6. On the other hand, if $\langle A, B \cup D|C; \mathcal{K}_C\rangle \in \mathcal{J}(\mathcal{G})$ then $\langle A, B|C \cup D; \mathcal{K}_C \times \mathcal{I}_D\rangle \in \mathcal{J}(\mathcal{G})$, from the weak union property, and $\langle A, D|C; \mathcal{K}_C\rangle \in \mathcal{J}(\mathcal{G})$, from the decomposition property. To this case it belongs also the special case where $\mathcal{K}_C^1$ and/or $\mathcal{K}_C^2$ are equal to $\mathcal{I}_C$.

(*case 2*): the restriction of the CSIs involves also nodes in D, (that is $m \cap D \neq \varnothing$). From Lemma 1, the statements $\langle A, B|C \cup D; \mathcal{I}_{C\setminus l} \times \mathcal{K}_l \times \mathcal{K}_m \times \mathcal{I}_{D\setminus m}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_{l\cup m}))$ and $\langle A, D|C; \mathbf{i}_l \times \mathcal{K}_{C\setminus l}\rangle \mathcal{J}(\mathcal{G}^R(\mathbf{i}_l))$ are not compatible.

(*Intersection*). (*case 1*): the restriction of the CSIs is limited to the set C, (that is $l \subseteq C$). If $\langle A, B|C \cup D; \boldsymbol{i}_l \times \mathcal{I}_{C\setminus l} \times \mathcal{I}_D\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and $\langle A, D|C \cup B; \boldsymbol{i}_l \times \mathcal{I}_{C\setminus l} \times \mathcal{I}_B\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ then $\langle A, B \cup D|C; \boldsymbol{i}_l \times \mathcal{I}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$. Thus, they belong also to $\mathcal{J}(\mathcal{G})$. In general, if $\langle A, B|C \cup D; \mathcal{K}_C^1 \times \mathcal{I}_D\rangle \in \mathcal{J}(\mathcal{G})$ and $\langle A, D|C \cup B; \mathcal{K}_C^2 \times \mathcal{I}_B\rangle \in \mathcal{J}(\mathcal{G})$ then $\langle A, B \cup D|\mathcal{K}_C\rangle \in \mathcal{J}(\mathcal{G})$ where $\mathcal{K}_C = \mathcal{K}_C^1 \cap \mathcal{K}_C^2$, in force of the property R2 in Definition 3.6. On the other hand, if $\langle A, B \cup D|C; \mathcal{K}_C\rangle \in \mathcal{J}(\mathcal{G})$ then $\langle A, B|C \cup D; \mathcal{K}_C \times \mathcal{I}_D\rangle \in \mathcal{J}(\mathcal{G})$ and $\langle A, D|C \cup B; \mathcal{K}_C \times \mathcal{I}_B\rangle \in \mathcal{J}(\mathcal{G})$ for the weak union. To this case, it belongs also the special case where $\mathcal{K}_C^1$ and/or $\mathcal{K}_C^2$ are equal to $\mathcal{I}_C$.

(*case 2*): the restriction of the CSIs involves also nodes in D, (that is $m \cap D \neq \varnothing$). The statements $\langle A, B|C \cup D; \mathcal{I}_{C\setminus l} \times \mathcal{K}_l \times \mathcal{K}_m \times \mathcal{I}_{D\setminus m}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_{l\cup m}))$ and $\langle A, D|C \cup B; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l} \times \mathcal{I}_B\rangle \mathcal{J}(\mathcal{G}^R(\mathbf{i}_l))$ are not compatible according to Lemma 1.

(*Composition*) if $\langle A, B|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and $\langle A, D|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ then $\langle A, B \cup D|C; \boldsymbol{i}_l \times \mathcal{K}_{C\setminus l}\rangle \in \mathcal{J}(\mathcal{G}^R(\boldsymbol{i}_l))$ and all these statements belong to $\mathcal{J}(\mathcal{G})$. In general, if $\langle A, B|C; \mathcal{K}_C^1\rangle \in \mathcal{J}(\mathcal{G})$ and $\langle A, D|C; \mathcal{K}_C^2\rangle \in \mathcal{J}(\mathcal{G})$ then $\langle A, B \cup D|\mathcal{K}_C\rangle \in \mathcal{J}(\mathcal{G})$ where $\mathcal{K}_C = \mathcal{K}_C^1 \cap \mathcal{K}_C^2$. □

Proof of Corollary 3.1. Sadeghi and Lauritzen [30] proved the equivalence between pairwise and global Markov properties for independence models satisfying all the properties in Definition 2.1. Theorem 3.1 proved that the independence model faithful to any stratified chain graph satisfies all the properties in Definition 2.1. □

Proof of Lemma 2. From condition *pS1* in Definition 3.2, the conditioning set of the CSI statement is $\mathrm{pre}(T_h)$, for any $\gamma, \delta \in T_h$. If there is at least a node $\xi \in \mathrm{pre}(T_h)\setminus(\mathrm{pa}_{\mathcal{G}}(\gamma) \cap \mathrm{pa}_{\mathcal{G}}(\delta))$, this implies that $(\xi \cup \gamma)$ and/or $(\xi \cup \delta)$ are(is) a non connected set. Thus, there is a set C such that the conditional independence statement(s) $\langle \xi, \gamma|C\rangle$ and/or $\langle \xi, \delta|C\rangle$ hold(s), and, according to Lemma 1, the CSI and these conditional independencies are incompatible. As a consequence, $l \in \mathrm{pa}_{\mathcal{G}}(\gamma) \cap \mathrm{pa}_{\mathcal{G}}(\delta)$.

From condition *pS2* in Definition 3.2, the conditioning set of the CSI statement is $\mathrm{pre}(T_h)\setminus\delta$, for $\gamma \in T_h$ and $\delta \in \mathrm{pre}(T_h)$. If there is at least a node $\xi \in \mathrm{pre}(T_h)\setminus(\mathrm{pa}_{\mathcal{G}}(\gamma))$ or such that ξ is not adjacent to δ, this implies that $(\xi \cup \gamma)$ and/or $(\xi \cup \delta)$ are(is) a non connected set(s). Thus, the previous reasoning still holds. □

Proof of Theorem 4.1. By applying formula (3) of Nicolussi and Cazzaro [20] (Corollary A.1) to the HMM parameters in formula (5) evaluated on the conditional distribution $\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}$, we obtain:

$$\eta_A^{A\cup\mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = \sum_{t\subseteq\mathrm{pa}_T(T_h)} \eta_{tA}^{A\cup\mathrm{pa}_T(T_h)}(\boldsymbol{i}_{tA}). \tag{17}$$

Notice that, according to formula (6), the generic addend on the right-hand side is $\beta_t^A(\boldsymbol{i}_t)$. □

Proof of Theorem 4.2. In this proof, we have to show that we can obtain all the HMM parameters from the parameters describing the regression models in formula (5). Note that proving the opposite is not necessary because, by definition, the regression parameters are a function of the HMM parameters.

We have to distinguish three cases.

Case 1: the parameters refer to the variables X_A, where $A \subseteq T_h$, and $\mathrm{pa}_T(T_h) = \varnothing$. All the HMM parameters η_A^A belong to the regression models according to formula (7).

Case 2: the parameters refer to the variables $X_{\mathcal{L}}$ such that $\mathcal{L} \subseteq (T_h \cup \mathrm{pa}_T(T_h))$ and $\mathrm{pa}_T(T_h)$ is not empty.

At first, we consider the regression parameters in formula (5) when $\boldsymbol{i}_{\mathrm{pa}_T(T_h)} = \mathbf{1}_{\mathrm{pa}_T(T_h)}$, that is, each variable in $X_{\mathrm{pa}_T(T_h)}$ assumes the first category. Since the parameter with at least one variable X_j in the first category $i_j = 1_j$ is equal to zero, formula (5) becomes

$$\eta_A^{A\cup\mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\mathbf{1}_{\mathrm{pa}_T(T_h)}) = \beta_{\varnothing}^A = \eta_A^{A\cup\mathrm{pa}_T(T_h)}(\boldsymbol{i}_A). \tag{18}$$

Then, by considering only one variable X_j, such that $j \in \mathrm{pa}_T(T_h)$, with $i_j \neq 1_j$ and the remaining $\mathrm{pa}_T(T_h)\setminus j$ setting equal to $\mathbf{1}_{\mathrm{pa}_T(T_h)\setminus j}$, we have

$$\eta_A^{A\cup\mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = \beta_{\varnothing}^A + \beta_j^A(i_j) = \eta_A^{A\cup\mathrm{pa}_T(T_h)}(\boldsymbol{i}_A) + \eta_{Aj}^{A\cup\mathrm{pa}_T(T_h)}(\boldsymbol{i}_{Aj}). \tag{19}$$

Then, from both (18) and (19), we can isolate the terms $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A)$ and $\eta_{Aj}^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_{Aj})$, whatever is the value i_j, $\forall \boldsymbol{i}_{Aj} \in \mathcal{I}_{Aj}$. By applying recursively this approach for all the variables X_J with $J \subseteq \mathrm{pa}_T(T_h)$, and for all the values $\boldsymbol{i}_J \in \mathcal{I}_J$, we obtain all the HMM parameters.

Case 3: the parameters refer to the variables $X_{\mathcal{L}}$ such that $\mathcal{L} \cap T_h \neq \varnothing$ and $\mathcal{L} \cap (\mathrm{pre}(T_h) \backslash \mathrm{pa}_T(T_h)) \neq \varnothing$. However, the nodes in $\mathcal{L}$ are m-separates given the nodes in $\mathrm{pa}_T(T_h)$, then these parameters $\boldsymbol{\eta}_{\mathcal{L}}^{\mathcal{M}}$ are null as each disjoint set corresponds to an independence statement, (see, for instance, Marchetti and Lupparelli [17]). □

Proof of Theorem 4.3. Before proceeding, it is worthwhile to remember the three remarks that we will use in this proof.

First, given an independence like $\langle A, B|C\rangle$, the probability distribution of X_{ABC} obeys the independence if, and only if, the HMM parameters $\eta_{abc}^{\mathcal{M}} = 0$, where $a \subseteq A$, $b \subseteq B$, $c \subseteq C$, $a, b \neq \varnothing$ and $\mathcal{M}$ is any subset of V, see Bergsma and Rudas [3].

Second, given a generic parameter $\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}})$, the choice of the unspecified category of the variable X_j with $j \in \mathcal{M}\backslash\mathcal{L}$ is arbitrary and we set equal to the first category without loss of generality, see Nicolussi and Cazzaro [19].

Finally, in force of the consideration in the proof of Theorem 4.2, all the parameters $\boldsymbol{\eta}_{vc}^{\mathcal{M}}$ are null if $v \subseteq T_h$ and $c \subseteq \mathrm{pre}(T_h)\backslash(\mathrm{pa}_T(T_h))$. As a consequence, when we need to constrain to zero parameters concerning subset of $T_h \cup \mathrm{pre}(T_h)$ with at least one non-empty element of T_h, we limit the discussion to the subset $T_h \cup \mathrm{pa}_T(T_h)$.

Considering these remarks, now we prove point by point the statements listed in the theorem.

Point i. Looking at *pM1* in Definition 3.2, when $A = \delta \cup \gamma$, with $\delta, \gamma \in T_h$, A is non connected set because there is no arc between the two nodes. Then, the parameters $\eta_{At}^{\mathcal{M}} = 0$ for $t \subseteq \mathrm{pa}_T(T_h)$, where $\mathcal{M} = A \cup \mathrm{pa}_T(T_h)$. Note that we are restricting the conditioning set $\mathrm{pre}(T_h)$ to the only $\mathrm{pa}_T(T_h)$ in force of the third consideration. The $\eta_{At}^{\mathcal{M}}(\boldsymbol{i}_{At})$ is exactly the parameters on the right-hand side of equation (6); thus, $\beta_t^A(\boldsymbol{i}_t) = 0$, $\forall t \subseteq \mathrm{pa}_T(T_h)$. By replacing this result in formula (5), we get that $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = 0$.

The same occur when $A = \{\gamma \cup \delta \cup \zeta\}$ and there are two missing edge between these three nodes. Of course one of these is an isolated node, says δ, thus A is a non connected set. In force of *pM1* we have $\langle X_\gamma, X_\delta|X_{\mathrm{pre}(T_h)}\rangle$ and $\langle X_\zeta, X_\delta|X_{\mathrm{pre}(T_h)}\rangle$ and in force of the composition property (S6) in Definition 2.1 we also have $\langle X_{\gamma\zeta}, X_\delta|X_{\mathrm{pre}(T_h)}\rangle$. In general, all the non connected subsets A of T_h represent an independence statement obtained via compositional property.

Point ii. Looking at *pM2* in Definition 3.2, when $\gamma \in T_h$ and $\delta \in \mathrm{pre}(T_h)\backslash\mathrm{pa}_{\mathcal{G}}(\gamma)$ are two non adjacent nodes, $\eta_{\gamma\delta t}^{\mathcal{M}} = 0$ for all $t \subseteq \mathrm{pa}_{\mathcal{G}}(\gamma)$, where $\mathcal{M} = \gamma \cup \mathrm{pa}_T(T_h)$.

Remember that a node $\gamma \in T_h$ is not adjacent to any node in $\mathrm{pre}(T_h)\backslash\mathrm{pa}_{\mathcal{G}(\gamma)}$. For the composition property in Definition 2.1, we get $\eta_{\gamma dt}^{\mathcal{M}} = 0$, for any $d \subseteq \mathrm{pa}_T(T_h)\backslash\mathrm{pa}_{\mathcal{G}}(\gamma)$. By using the equivalence in formula (6) the previous assertion becomes $\beta_d^\gamma = 0$, for any $d \subseteq \mathrm{pa}_T(T_h)\backslash\mathrm{pa}_{\mathcal{G}}(\gamma)$. Thus, the addends in the right-hand side of formula (5) reduce to $\sum_{t\subseteq \mathrm{pa}_{\mathcal{G}}(\gamma)} \beta_t^\gamma$. A further generalization can be made by considering all subsets A of T_h which regression models are not canceled by property (i). Even in this case, by applying the composition property to *pM2*, it can be derived that $\langle X_A, X_{\mathrm{pre}(T_h)\backslash\mathrm{pa}_{\mathcal{G}}(A)}|X_{\mathrm{pa}_{\mathcal{G}}(A)}\rangle$. Coherently with the previous passages we obtain $\eta_A^{\mathcal{M}}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = \sum_{t\subseteq \mathrm{pa}_{\mathcal{G}}(A)} \beta_t^A$.

The proofs of (iii) and (iv) follow from what has been said for point (i) and (ii) by considering two differences. Unlike the missing arcs, the stratum identifies conditional independence statements only for certain values, those for we can build for a reduced graph $\mathcal{G}^R(\boldsymbol{i}_l)$. Remember that, for any $\boldsymbol{i}_l \in \mathcal{K}_l$ and any stratum, we have a reduced graph $\mathcal{G}^R(\boldsymbol{i}_l)$. Secondly, the rule to constrain to zero the HMM parameters based on the *baseline* logits, is presented in formula (12).

Point iii. For any non-connected set A, $\sum_{t\subseteq \mathrm{pa}_T(T_h)} \eta_{At}^{\mathcal{M}}(\boldsymbol{i}_A, \boldsymbol{i}_t) = 0$, with $\mathcal{M} = A \cup \mathrm{pa}_T(T_h)$, for any $\boldsymbol{i}_A \in \mathcal{I}_A$ and $\boldsymbol{i}_t \in \mathcal{I}_t \cap (\mathcal{K}_l \times \mathcal{I}_{\mathrm{pa}_T(T_h)\setminus l})$. By using the regression parameters in formula (6), the previous sum is exactly the right-hand side of formula (5). Thus, we get $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = 0$ for any $\boldsymbol{i}_{\mathrm{pa}_T(T_h)} \in (\mathcal{K}_l \times \mathcal{I}_{\mathrm{pa}_T(T_h)\setminus l})$. Obviously, some non-connected A in $\mathcal{G}^R(\boldsymbol{i}_l)$ can be also non-connected in $\mathcal{G}^F$. However, in this case, simply the regression models in formula (5) are still set to zero according to (i).

Point iv. Let us consider all connected set $A \subseteq T_h$ in $\mathcal{G}^R(\boldsymbol{i}_l)$ for any $\boldsymbol{i}_l \in \mathcal{K}_l$ and for any stratum. We have that $\sum_{d\subseteq \mathrm{pa}_T(T_h)\setminus \mathrm{pa}_{\mathcal{G}}(A)} \eta_{Ad}^{\mathcal{M}}(\boldsymbol{i}_A, \boldsymbol{i}_d) = 0$ for all $\boldsymbol{i}_d \in \mathcal{I}_d \cap (\mathcal{K}_l \times \mathcal{I}_{\mathrm{pa}_T(T_h)\setminus l})$, where $\mathcal{M} = A \cup \mathrm{pa}_T(T_h)$. By using the equivalence in formula (6) the previous assertion becomes $\sum_{d\subseteq \mathrm{pa}_T(T_h)\setminus \mathrm{pa}_{\mathcal{G}}(A)} \beta_d^A = 0$. Thus, $\eta_A^{\mathcal{M}}(\boldsymbol{i}_A|\boldsymbol{i}_{\mathrm{pa}_T(T_h)}) = \sum_{t\subseteq \mathrm{pa}_{\mathcal{G}}(A)} \beta_t^A$ for all $\boldsymbol{i}_t \in \mathcal{I}_t \cap (\mathcal{K}_l \times \mathcal{I}_{\mathrm{pa}_T(T_h)\setminus l})$. □

Proof of Theorem 4.4. To provide the proof of this theorem, we closely follow the proof of Theorem 4.3, remembering that the difference between the two theorems lies in the different types of logits chosen and the alternative specification of the CSIs. In particular, we have to take into account the result in Lemma 1 of Nicolussi and Cazzaro [20], formula (2) and Corollary A.1, formula (4).

Point i. As discussed in the proof of Theorem 4.3, for any non-connected subset A of T_h, the HMM parameters

$$\eta_{At}^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_{At}) = 0 \quad \forall \boldsymbol{i}_{At} \in \mathcal{I}_{At} \text{ and } \forall t \subseteq \mathrm{pa}_T(T_h), \tag{20}$$

where A is any nonempty non-connected subset of T_h in the full chain graph $\mathcal{G}^F$. When $t = \varnothing$, we get that $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A) = 0$. Further, by considering the set t composed of only one vertex, say j, we get $\eta_{Aj}^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_{Aj}) = 0$. Let us apply the formula (2) to the previous result, obtaining

$$\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_j) - \eta_A^{A\cup \mathrm{pa}_T(T_h)}\big(\boldsymbol{i}_A|(\boldsymbol{i}_j - 1)\big) = 0. \tag{21}$$

When the variable X_j assumes the second category, the above difference becomes equal to $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|i_j = 2) - \eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A) = 0$. Since the second term on the left-hand side is null, we get that $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|i_j = 2) = 0$. Similarly, when the variable X_j assumes the third value $\boldsymbol{i}_j = 3$, the difference in formula (21) becomes equal to $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_j = 3) - \eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_j = 2) = 0$. But since we just proved that the term $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_j = 2)$ is equal to zero, $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_j = 3)$ is null. In the same way, we can prove that $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_j) = 0$ for all $\boldsymbol{i}_j \in \mathcal{I}_j$.

Now we generalize to whatever set $t \subseteq \mathrm{pa}_T(T_h)$. When each variable X_j with $j \in t$ assumes the second value $\boldsymbol{i}_j = 2$, in short, $\boldsymbol{i}_t = \mathbf{2}$, the formula (2) becomes $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_t = \mathbf{2}) - \sum_{\substack{J\subseteq C \\ J\neq\varnothing}} \eta_{AC\setminus J}^{\mathcal{M}}(\boldsymbol{i}_{AC\setminus J}) = 0$. All the terms in the sum (left-hand side of the equation) are null because of formula (20). This means that $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_t = \mathbf{2})$ is also equal to zero. When, there is one variable X_j assuming value $\boldsymbol{i}_j = 3$, leaving unchanged the categories of the variables $X_{t\setminus j}$ equal to $\boldsymbol{i}_{t\setminus j} = 2$, we can repeat the same process by obtaining $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{t\setminus j} = 2, \boldsymbol{i}_j = 3) = 0$. In this way, reducing one by one the categories of each variable, we can prove that the $\eta_A^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_t) = 0$ for all $c \subseteq \mathrm{pa}_T(T_h)$ and for all $\boldsymbol{i}_t \in \mathcal{I}_t$.

Point ii. As discussed in the proof of Theorem 4.3, we need the following constraints on the HMM parameters

$$\eta_{Adp}^{A\cup \mathrm{pa}_T(T_h)}(\boldsymbol{i}_{Adp}) = 0, \quad \forall \boldsymbol{i}_{Adp} \in \mathcal{I}_{Adp}, \forall d \subset \mathrm{pa}_T(T_h)\setminus \mathrm{pa}_{\mathcal{G}}(A) \text{ and } \forall p \subseteq \mathrm{pa}_{\mathcal{G}}(A), \tag{22}$$

where $A \subseteq T_h$ in the full chain graph $\mathcal{G}^F$ and where A and d must be necessarily nonempty. By applying the decomposition in formula (4) in Lemma 1 of Nicolussi and Cazzaro [20], where $L = A \cup d$ and $C = p$ to the constraints in formula (22), and by following the same consideration performed at *point i* of this proof, we get that $\eta_{Ad}^{A\text{pa}_T(T_h)}(\boldsymbol{i}_{Ad}|\boldsymbol{i}_p)$ is null. This means that, formula (3) in Corollary A.1 of Nicolussi and Cazzaro [20], in this case, becomes $\eta_A^{A\text{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\text{pa}_T(T_h)}) = \sum_{J\subseteq \text{pa}_{\mathcal{G}}(A)} \eta_{AJ}^{\text{pa}_T(T_h)}(\boldsymbol{i}_{AJ}|\boldsymbol{i}_{\text{pa}_{\mathcal{G}}(A)\setminus J})$. By using the regression parameters, we get $\eta_A^{A\text{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_{\text{pa}_T(T_h)}) = \sum_{J\subseteq \text{pa}_{\mathcal{G}}(A)} \beta_J^A(\boldsymbol{i}_J)$.

Point iii. In this case, by considering the new formulation of the CSIs as in formula (14), instead of the equality, we have that the CSI holds for all $\boldsymbol{i}_{\text{pa}_{\mathcal{G}}(A)} \preceq \boldsymbol{i}'_{\text{pa}_{\mathcal{G}}(A)}$. By taking into account the considerations performed at *point i*, it is easy to see that they are still valid $\forall \boldsymbol{i}_{\text{pa}_{\mathcal{G}}(A)} \preceq \boldsymbol{i}'_{\text{pa}_{\mathcal{G}}(A)}$. Then we get that $\eta_A^{A\cup\text{pa}_T(T_h)}(\boldsymbol{i}_A|\boldsymbol{i}_c) = 0$, for all $c \subseteq \text{pa}_T(T_h)$ and for all $\boldsymbol{i}_c \preceq \boldsymbol{i}'_c$.

Point iv. Exactly as for *point iii*, the considerations performed at *point ii* hold here when the categories of the conditioning variables are greater than or equal to $\boldsymbol{i}'_{\text{pa}_{\mathcal{G}}(A)}$. Thus, the constraints are $\sum_{t\subseteq \text{pa}_T(T_h)\setminus\text{pa}_{\mathcal{G}}(A)} \beta_t^A(\boldsymbol{i}_t = 0), \forall \boldsymbol{i}_t \preceq \boldsymbol{i}'_t$. □

Supplementary Material

Additional results for proofs and additional data from application (DOI: 10.3150/20-BEJ1302 SUPP; .pdf). We provide a lemma and a corollary slightly adapted from Nicolussi and Cazzaro [19]. These two results are used in the proofs of Theorem 4.1 and Theorem 4.4. We provide four tables of parameters concerning the application in Section 5.

References

[1] Aitchison, J. and Silvey, S.D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Stat.* **29** 813–828. MR0094873 https://doi.org/10.1214/aoms/1177706538

[2] Bartolucci, F., Colombi, R. and Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statist. Sinica* **17** 691–711. MR2398430

[3] Bergsma, W.P. and Rudas, T. (2002). Marginal models for categorical data. *Ann. Statist.* **30** 140–159. MR1892659 https://doi.org/10.1214/aos/1015362188

[4] Boutilier, C., Friedman, N., Goldszmidt, M. and Koller, D. (1996). Context-specific independence in Bayesian networks. In *Uncertainty in Artificial Intelligence* (*Portland, OR*, 1996) 115–123. San Francisco, CA: Morgan Kaufmann. MR1617129

[5] Cazzaro, M. and Colombi, R. (2008). Modelling two way contingency tables with recursive logits and odds ratios. *Stat. Methods Appl.* **17** 435–453. MR2447568 https://doi.org/10.1007/s10260-007-0068-2

[6] Cazzaro, M. and Colombi, R. (2014). Marginal nested interactions for contingency tables. *Comm. Statist. Theory Methods* **43** 2799–2814. MR3223712 https://doi.org/10.1080/03610926.2012.685550

[7] Colombi, R., Giordano, S. and Cazzaro, M. (2014). hmmm: An R package for hierarchical multinomial marginal models. *J. Stat. Softw.* **59** 1–25.

[8] Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statist. Sci.* **8** 204–218, 247–283. MR1243593

[9] Drton, M. (2009). Discrete chain graph models. *Bernoulli* **15** 736–753. MR2555197 https://doi.org/10.3150/08-BEJ172

[10] Højsgaard, S. (2003). Split models for contingency tables. *Comput. Statist. Data Anal.* **42** 621–645. MR1967060 https://doi.org/10.1016/S0167-9473(02)00119-6

[11] Højsgaard, S. (2004). Statistical inference in context specific interaction models for contingency tables. *Scand. J. Stat.* **31** 143–158. MR2042604 https://doi.org/10.1111/j.1467-9469.2004.00378.x

[12] ISTAT (2015). Italian innovation Survey 2002–2012. Available at http://www.istat.it/en/archive/87787.
[13] La Rocca, L. and Roverato, A. (2017). *Discrete Graphical Models*. *Handbook of Graphical Models*. *Handbooks of Modern Statistical Methods*. Boca Raton, FL: CRC Press/CRC.
[14] Lauritzen, S. and Sadeghi, K. (2018). Unifying Markov properties for graphical models. *Ann*. *Statist*. **46** 2251–2278. MR3845017 https://doi.org/10.1214/17-AOS1618
[15] Lauritzen, S.L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. New York: The Clarendon Press. MR1419991
[16] Marchetti, G.M. and Lupparelli, M. (2008). Parameterization and fitting of a class of discrete graphical models. *COMPSTAT* **2008** 117–128.
[17] Marchetti, G.M. and Lupparelli, M. (2011). Chain graph models of multivariate regression type for categorical data. *Bernoulli* **17** 827–844. MR2817607 https://doi.org/10.3150/10-BEJ300
[18] Nicolussi, F. (2013). Marginal parameterizations for conditional independence models and graphical models for categorical data. Ph.D. thesis, Univ. of Milan-Bicocca.
[19] Nicolussi, F. and Cazzaro, M. (2020). Context-specific independencies in hierarchical multinomial marginal models. *Stat*. *Methods Appl*. **29** 767–786. MR4174686 https://doi.org/10.1007/s10260-019-00503-8
[20] Nicolussi, F. and Cazzaro, M. (2021). Supplement to "Context-specific independencies in stratified chain regression graphical models". https://doi.org/10.3150/20-BEJ1302SUPP
[21] Nicolussi, F. and Colombi, R. (2017). Type II chain graph models for categorical data: A smooth subclass. *Bernoulli* **23** 863–883. MR3606753 https://doi.org/10.3150/15-BEJ762
[22] Nyman, H., Pensar, J. and Corander, J. (2016). Dependence Logic. 219–234. Springer.
[23] Nyman, H., Pensar, J., Koski, T. and Corander, J. (2016). Context-specific independence in graphical log-linear models. *Comput*. *Statist*. **31** 1493–1512. MR3573088 https://doi.org/10.1007/s00180-015-0606-6
[24] Pensar, J., Nyman, H., Koski, T. and Corander, J. (2015). Labeled directed acyclic graphs: A generalization of context-specific independence in directed graphical models. *Data Min*. *Knowl*. *Discov*. **29** 503–533. MR3312469 https://doi.org/10.1007/s10618-014-0355-0
[25] R Core Team (2014) R: a language and environment for statistical computing, Vienna, Austria.
[26] Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Ann*. *Statist*. **30** 962–1030. MR1926166 https://doi.org/10.1214/aos/1031689015
[27] Roverato, A. (2017). *Graphical Models for Categorical Data*. *SemStat Elements*. Cambridge: Cambridge Univ. Press. MR3751385 https://doi.org/10.1017/9781108277495
[28] Rudas, T., Bergsma, W.P. and Németh, R. (2010). Marginal log-linear parameterization of conditional independence models. *Biometrika* **97** 1006–1012. MR2746171 https://doi.org/10.1093/biomet/asq037
[29] Sadeghi, K. (2018). Markov properties of discrete determinantal point processes. Preprint. Available at arXiv:1810.02294.
[30] Sadeghi, K. and Lauritzen, S. (2014). Markov properties for mixed graphs. *Bernoulli* **20** 676–696. MR3178514 https://doi.org/10.3150/12-BEJ502
[31] Sadeghi, K. and Wermuth, N. (2016). Pairwise Markov properties for regression graphs. *Stat* **5** 286–294. MR3589267 https://doi.org/10.1002/sta4.122
[32] Wermuth, N. and Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **66** 687–717. MR2088296 https://doi.org/10.1111/j.1467-9868.2004.b5161.x
[33] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Chichester: Wiley. MR1112133

Received November 2019 and revised June 2020