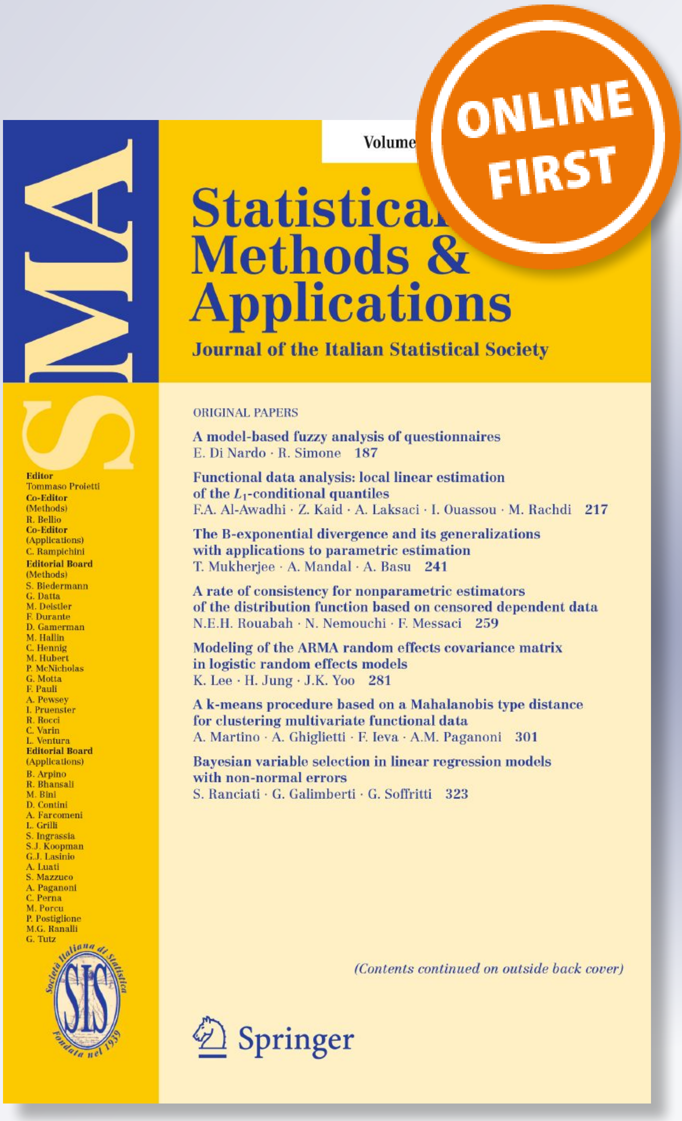
Context-specific independencies in hierarchical multinomial marginal models

Federica Nicolussi & Manuela Cazzaro

Statistical Methods & Applications
Journal of the Italian Statistical Society

ISSN 1618-2510

Stat Methods Appl
DOI 10.1007/s10260-019-00503-8



Springer

Statistical Methods & Applications
https://doi.org/10.1007/s10260-019-00503-8

ORIGINAL PAPER

Context-specific independencies in hierarchical multinomial marginal models

Federica Nicolussi[1] · **Manuela Cazzaro**[2]

Accepted: 23 November 2019

Abstract
This paper focuses on studying the relationships among a set of categorical (ordinal) variables collected in a contingency table. Besides the marginal and conditional (in)dependencies, thoroughly analyzed in the literature, we consider the context-specific independencies holding only in a subspace of the outcome space of the conditioning variables. To this purpose we consider the hierarchical multinomial marginal models and we provide several original results about the representation of context-specific independencies through these models. The theoretical results are supported by an application concerning the innovation degree of Italian enterprises.

Keywords Context-specific independence · Ordinal variable · Hierarchical multinomial marginal model

1 Introduction

In this work we deal with categorical (ordinal) variables collected in a contingency table and we propose a model able to capture different kinds of independence relationships involving variables. Several models have been proposed in the literature focusing on the independence or on the dependence structure. In particular, we refer to Marginal models, see e.g. (Bergsma and Rudas 2002), which impose constraints on marginal distributions of the tables in order to test different independence hypotheses. More specifically, we focus on hierarchical multinomial marginal (HMM) models, (Colombi et al. 2014) based on the work of Bartolucci et al. (2007) and extended by Cazzaro and Colombi (2014). The HMM models are specified by a set of marginal distributions of the contingency table together with a set of interactions: logits and

✉ Federica Nicolussi
federica.nicolussi@unimi.it

1 Department of Economics, Management and Quantitative Methods, University of Milano, Milano, Italy

2 Department of Statistics and Quantitative Methods, University of Milano Bicocca, Milano, Italy

Published online: 05 December 2019

higher order effects [i.e. contrasts of logarithms (of sums) of probabilities], defined within different marginal distributions. Well-known special cases of HMM models are the classical Log-Linear models, the Marginal models, (Bergsma and Rudas 2002), and the Multivariate Logistic models, (Glonek and McCullagh 1995). In particular, we take advantage of the different interactions that the HMM models are able to deal with, that are significant, especially, for ordinal variables, see Cazzaro and Colombi (2008, 2014). In this framework, we focus on the relationships among a set of categorical (ordinal) variables with the aim of testing marginal, conditional and context-specific independencies (CSIs), see, among others (Boutilier et al. 1996; Højsgaard 2004; Roverato 2017). The first two are well known relationships among variables collected in a contingency table; the CSI, instead, is a conditional independence which holds only in a subspace of the outcome space of the conditioning variables. For instance, given 3 variables X_1, X_2 and X_3, it may occur $X_1 \perp X_2|X_3 = 1$ while $X_1 \not\perp X_2|X_3 \neq 1$. It is interesting to study this kind of independence as it allows to focus on the category(ies) which discriminate(s) and really affect(s) the connection among the variables. Furthermore, adding also the CSIs to the independencies to test in a model admits to reduce the number of parameters needed for describing the joint probability distribution, in compliance with the principle of parsimony.

Nyman et al. (2016), provide the conditions and the constraints to investigate the CSIs in classical Log-Linear models where the independencies to test involve all the variables selected. In addition, La Rocca and Roverato (2017), show further results in order to test CSI hypotheses also on subsets of variables, by using the so-called *log-mean linear parametrization*. In this work, we focus on the HMM model and we provide the linear constraints to impose on its parameters in order to test, where possible at the same time, conditional, marginal, and context-specific hypotheses. Furthermore, we take into account parameters based on several types of logits, in order to consider the possible different nature of the variables. These different parameters, as we will see, involve different constraints.

The work follows this structure. In the next section, we give an overview of the HMM models. In Sect. 3, we introduce the representation of CSI via HMM models. We present the CSI under a double definition, one, as already presented in the literature, more suitable for categorical variables, and the new one that makes sense for ordinal variables. The reasons of this double definition of CSIs lie in the attempt to make the results more interpretable in compliance with the nature of the variables. As a matter of fact, with the formulation suitable for ordinal variables, the constraints on HMM parameters are more manageable and meaningful to interpret. In Sect. 3.1 we deal with the first formulation apt to categorical variables. At first, similarly to Nyman et al. (2016), but in a marginal framework, we provide the CSI constraints where the parameters are based on *baseline* logits, the parameters used in classical Log-Linear models. Furthermore, we provide, as new result, how it is possible to define a CSI by using other interactions more suitable for ordinal variables. In Sect. 3.2, we deal with the new formulation of CSIs that results to be more appropriate for ordinal variables. Finally, in Sect. 4 we provide an application to a real dataset on the innovation status of small and medium Italian firms, Istat (2012). Some conclusions are presented in Sect. 5. The proofs of all the theorems are located in the "Appendix" in order to improve the readability of the paper.

2 Hierarchical multinomial marginal models

Let us consider a vector of random variables $X_V = \left(X_j\right)_{j \in V}$ where the generic variable X_j takes value i_j in a set of finite categories $\mathcal{I}_j = (1, \ldots, i_j, \ldots, I_j)$. Let us denote with the symbol $|\cdot|$ the cardinality of a set. The contingency table of the $|V|$ variables is defined by $\mathcal{I}_V = \times_{j \in V} \mathcal{I}_j$ where the generic cell is defined as $\boldsymbol{i}_V = (i_j,\ j \in V)$. The strictly positive probability of a generic cell, $\boldsymbol{i}_V$ of $\mathcal{I}_V$, is denoted with $\pi(\boldsymbol{i}_V)$, thus the probability of the whole contingency table is represented by the vector $\boldsymbol{\pi}$, obtained by stacking each $\pi(\boldsymbol{i}_V)$ in the lexicographical order. Similarly, by considering a vector of variables in $X_{\mathcal{M}}$ with $\mathcal{M} \subseteq V$ which generate the marginal $\mathcal{M}$-contingency table $\mathcal{I}_{\mathcal{M}} = \times_{j \in \mathcal{M}} \mathcal{I}_j$, the marginal probability of the generic cell $\boldsymbol{i}_{\mathcal{M}}$ is $\pi(\boldsymbol{i}_{\mathcal{M}})$, obtained by summing with respect to the variables in $X_{V \setminus \mathcal{M}}$. Again, the probabilities of the $\mathcal{M}$-contingency table $\mathcal{I}_{\mathcal{M}}$, $\pi(\boldsymbol{i}_{\mathcal{M}})$, are stacked in $\boldsymbol{\pi}_{\mathcal{M}}$ in the lexicographical order.

A one-to-one function $\boldsymbol{\eta} = f(\boldsymbol{\pi})$ defines a parametrization of $\boldsymbol{\pi}$ in terms of $\boldsymbol{\eta}$. In the HMM models, the elements of $\boldsymbol{\eta}$ are the parameters based on different types of logits and defined on marginal distributions. Note that the classical Log-Linear models are built on the log-linear parameters based on the so-called *baseline* logits defined on the joint distribution of the involved variables. Indeed, the HMM parameters can be defined as contrasts among the logarithms of (sum of) probabilities of disjoint subsets of cells defined within different marginal distributions, associated to a non decreasing sequence of marginal sets of variables, $\mathcal{H} = \left\{\mathcal{M}_j\right\}_{j=1,\ldots,s}$, such that $\mathcal{M}_j \subseteq \mathcal{M}_i$ implies $j < i$, where $\mathcal{M}_s = V$. Given a set of variable X_V with the vector containing the cell probabilities of the joint distribution $\boldsymbol{\pi}$, a set of HMM parameters, listed in $\boldsymbol{\eta}$, is a parametrization of the probability distribution if the hierarchy and the completeness properties are satisfied, see Bergsma and Rudas (2002), Bartolucci et al. (2007). That is, in HMM models, every parameter is defined in one and only one marginal distribution (completeness property) and within the first marginal set in the non decreasing sequence $\mathcal{H}$ which contains it (hierarchy property).

The formula (1) shows the form of the HMM parameter referring to the variables in $X_{\mathcal{L}}$, fixed at the level $\boldsymbol{i}_{\mathcal{L}}$ and defined in the marginal table $\mathcal{I}_{\mathcal{M}}$. Hereafter we call $\mathcal{L}$ *interaction* set and $\mathcal{M}$ *marginal* set.

Thus, the HMM parameter has the following form:

$$\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}} | \boldsymbol{I}_{\mathcal{M} \setminus \mathcal{L}}) = \sum_{\mathcal{J} \subseteq \mathcal{L}} (-1)^{|\mathcal{L} \setminus \mathcal{J}|} \log \pi_{\mathcal{M}} \left(\boldsymbol{i}_{\mathcal{J}}, \boldsymbol{i}^*_{\mathcal{L} \setminus \mathcal{J}}, \boldsymbol{I}_{\mathcal{M} \setminus \mathcal{L}} \right), \tag{1}$$

where the symbol $\boldsymbol{I}_A$ denotes the set of the last level I_j of each variable X_j with $j \in A$ and $A \subseteq V$. The vector of indices marked with the asterisk, $\boldsymbol{i}^*$, changes in correspondence of the type of logits assigned to each variable. When, for all the variables X_j with $j \in \mathcal{L}$, we choose the *baseline* logits, hereafter *baseline* criterion, the cell $\boldsymbol{i}^*_{\mathcal{L}}$ is equal to a predetermined arbitrary cell, that, without loss of generality, we set equal to $\boldsymbol{I}_{\mathcal{L}}$. Further, in this paper, we consider also the parameters based on *local* and *continuation* logits, hereafter considered as *local* and *continuation* criterion for aggregating the probabilities within the parameters. In the case of *local* criterion, the cell $\boldsymbol{i}^*_{\mathcal{L}}$ is equal to the cell of coordinates $i_j + 1$ for all X_j with $j \in \mathcal{L}$, for brevity

$\boldsymbol{i}_{\mathcal{L}} + 1$. In the *continuation* criterion, the symbol $\boldsymbol{i}^*_{\mathcal{L}}$ denotes a list of certain cells of coordinates $\{i_j + 1, \ldots, I_j\}$, for all $j \in \mathcal{L}$. While the *baseline* criterion is useful when we deal with unordered categorical variables, the other two types of aggregation criteria are typically used when we deal with ordinal variables, see Bartolucci et al. (2007); Cazzaro and Colombi (2014). Note that, when the variables are dichotomous, the interactions have the same form whatever the aggregation criterion was chosen. Finally, the variables in $X_{\mathcal{M}\backslash\mathcal{L}}$ in formula (1), which are not involved in the parameter, are fixed equal to the levels $\boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}$, without loss of generality. Since there are no ambiguity, we omit the wording $\boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}$ from the definition of the parameter:

$$\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}}|\boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) = \eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}}). \tag{2}$$

Each parameter $\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}})$ is equal to zero if there is at least a variable X_j, with $j \in \mathcal{L}$, at level $i_j = I_j$. This result comes from formula (1).

As discussed before, the list of all parameters, $\forall\mathcal{L} \subseteq V$ and $\mathcal{M} \in \mathcal{H}$, constitutes the HMM parametrization of the vector $\boldsymbol{\pi}$. We stack all these parameters in the lexicographical order within the vector $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \left[\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}})\right]. \tag{3}$$

The following example shows the parametrization of two variables in correspondence of different aggregation criteria.

Hereafter, for brevity, the elements of a set, placed in the subscript or superscript, will be listed without parentheses and commas, so for instance $X_{(1,2)}$ becomes X_{12}.

Example 1 Let us consider the vector of two variables X_{12} collected in a 3×3 contingency table $\mathcal{I}_{\mathcal{M}}$, where $\mathcal{M} = (1, 2)$. The parameters in formula (2), according to the different aggregation criteria, are displayed in Table 1. Note that each row of Table 1 shows the different criteria used to code the variables in X_{12} in the 8 parameters of the HMM model: 2 logits for each univariate variable and 4 contrasts of logits. For simplicity, in every parametrization we use the same aggregation criterion for all variables, this, however, is not mandatory. In general it is possible to choose an aggregation criterion for each variable.

The vector of HMM parameters $\boldsymbol{\eta}$ defines the saturated model, where all the $|\mathcal{I}_V|-1$ parameters are free. Certain constraints to zero on HMM parameters, defined on suitable marginal distributions, highlight marginal and/or conditional independencies. For instance, given the vector of three variables X_{123}, in order to represent the conditional independence $X_1 \perp X_2|X_3$ we have to impose $\eta^{123}_{12}(\boldsymbol{i}_{12}) = \eta^{123}_{123}(\boldsymbol{i}_{123}) = 0$ for all $\boldsymbol{i}_{12} \in \mathcal{I}_{12}$ and $\boldsymbol{i}_{123} \in \mathcal{I}_{123}$.

It is worthwhile to note that not all the lists of marginal and/or conditional independencies are representable through HMM models. In the literature, this topic is thoroughly discussed, see among others (Drton 2009; Forcina 2012; Rudas et al. 2010). Bergsma and Rudas (2002) and Bartolucci et al. (2007) proved that, if a list of marginal and/or conditional independencies is representable through the parameters in formula (3), these parameters provide a parametrization of the vector $\boldsymbol{\pi}$.

Table 1 Logits and contrasts of logits in correspondence of different types of aggregation criteria, *baseline* (b), *local* (l) and *continuation* (c), respectively

Type	$\eta_1^{12}(i_1)$	$\eta_2^{12}(i_2)$	$\eta_{12}^{12}(\boldsymbol{i}_{12})$
b	$\log\left(\frac{\pi(i_1,3)}{\pi(3,3)}\right)$	$\log\left(\frac{\pi(3,i_2)}{\pi(3,3)}\right)$	$\log\left(\frac{\pi(\boldsymbol{i}_{12})\pi(3,3)}{\pi(3,i_2)\pi(i_1,3)}\right)$
l	$\log\left(\frac{\pi(i_1,3)}{\pi((i_1+1),3)}\right)$	$\log\left(\frac{\pi(3,i_2)}{\pi(3,(i_2+1))}\right)$	$\log\left(\frac{\pi(\boldsymbol{i}_{12})\pi(\boldsymbol{i}_{12}+1)}{\pi((i_1+1),i_2)\pi(i_1,(i_2+1))}\right)$
c	$\log\left(\frac{\pi(i_1,3)}{\pi(i_1+1,3)+\cdots+\pi(I_1,3)}\right)$	$\log\left(\frac{\pi(3,i_2)}{\pi(3,i_2+1)+\cdots+\pi(3,I_2)}\right)$	$\log\left(\frac{\pi(\boldsymbol{i}_{12})(\pi(i_1+1,i_2+1)+\cdots+\pi(I_1,I_2))}{(\pi(i_1+1,3)+\cdots+\pi(I_1,3))(\pi(3,i_2+1)+\cdots+\pi(3,I_2))}\right)$

3 Context-specific independence in HMM models

Let us suppose we want to investigate a CSI among the variables in the vector $X_{\mathcal{M}}$ with $\mathcal{M} \subseteq V$, saying

$$X_A \perp X_B | (X_C = \boldsymbol{i}'_C), \quad \boldsymbol{i}'_C \in \mathcal{K}_C, \tag{4}$$

where $A \cup B \cup C = \mathcal{M}$, and $\boldsymbol{i}'_C$ is the vector of certain level(s) of the variable(s) in X_C, such that $X_j = i'_j$ for all $j \in C$, and it takes value in the list of levels $\mathcal{K}_C \subseteq \mathcal{I}_C$ for which the independence in formula (4) holds. Henceforth, we refer to the CSI in formula (4), as a CSI with equality (in the conditioning set).

Remark 1 If $\mathcal{K}_C = \mathcal{I}_C$ in formula (4), the statement corresponds to the conditional independence $X_A \perp X_B | X_C$.

In general, a notation like $X_A \perp X_B | (X_C = \boldsymbol{i}'_C)$ for all $\boldsymbol{i}'_C \in (\mathcal{I}_{C_1} \times \mathcal{K}_{C_2})$, means that the independence holds for all levels $\boldsymbol{i}'_{C_1} \in \mathcal{I}_{C_1}$ combined with the levels $\boldsymbol{i}'_{C_2} \in \mathcal{K}_{C_2}$. This means that the independence holds between the variables in X_A and the variables in X_B given all the levels of the variables in X_{C_1} and the levels $\boldsymbol{i}'_{C_2} \in \mathcal{K}_{C_2}$ for the variables in X_{C_2}.

When we deal with ordinal variables, it is more interesting to investigate if the CSIs hold only for high or low values of the conditioning variables. We refer to these cases with the following notations:

$$X_A \perp X_B | (X_C \geq \boldsymbol{i}'_C) \tag{5}$$

or

$$X_A \perp X_B | (X_C \leq \boldsymbol{i}'_C). \tag{6}$$

These inequalities mean that the independence holds for all the values i_j of X_j in X_C that jointly are equal or greater(lower) than i'_j, in formula $\cap_{j \in C} X_j \geq i'_j$ or $\cap_{j \in C} X_j \leq i'_j$. Henceforth, we refer to the CSI in formula (5) and (6), as a CSI with inequality (in the conditioning set).

Remark 2 If $\boldsymbol{i}'_C$ is equal to $\mathbf{1}_C$ (the first cell of $\mathcal{I}_C$) in (5) or it is equal to $\boldsymbol{I}_C$ in (6), the corresponding statement is the conditional independence $X_A \perp X_B | X_C$.

The plausible CSIs according to formulas (5) and (6) are considered in the ones obtained from formula (4). However, these statements lead to advantages in the formulation of the constraints and in the interpretation of the parameters. First, (5) and (6) lead to more manageable constraints than (4). Indeed, when we express the CSIs with inequality such as in formulas (5) or (6), as we will see, the constraints set to zero certain parameters while, according to formula (4) with equality, the constraints set to zero *sums* of parameters that make them of difficult interpretation. In addition, to explain a CSI with equality is less intuitive than a CSI with inequality when the class $\mathcal{K}_C$ in formula (4) contains several levels. The choice of the definitions to use depends on the aim of the analysis. If we take advantage of the CSIs for reducing the number of free parameters these alternative definitions are worthless. On the other hand, if the aim lies on the study of the effect of certain levels, especially extreme values, of the conditioning variables that discriminate between independence and dependence, these alternative formulations advantageous the meanings.

In the following subsections we deal with the CSIs in HMM models when the parameters are based on different types of logits. A justification of the use of different parameters lies in the interpretation of unconstrained parameters. Let us suppose that we are interested in investigating the relationship between gender and work position conditioning to high levels of age. By testing this hypothesis, the aggregation criterion chosen for the age makes no difference. However, the interpretation of the unconstrained parameters referring also to the age is more informative if we use *local* or *continuation* criteria.

Section 3.1 deals with the CSI statement with equality such as in formula (4) when we use the parameters based on *baseline* or *local* logits. As it is shown in Table 1, when we use the *continuation* criterion, the parameters involve also sum of probabilities. This makes impossible to explicit constraints to define the CSIs with equality (4). However, when we deal with ordinal variables, handling with inequality, such as in formulas (5) or (6), is more appropriate. Hence, in Sect. 3.2 we provide the constraints needed to represent these CSIs when we use the *baseline*, *local* or *continuation* criteria.

3.1 Constraints on HMM parameters for the CSI with equality

As already stated, Nyman et al. (2016), provide the condition to define a CSI such as express in formula (4) in classical log-linear models, where all the variables are coded with the *baseline* approach and there is only one marginal set equal to the joint distribution $\mathcal{M} = V$. Furthermore, La Rocca and Roverato (2017), show the same result on the log-linear models and provide the constraints to define CSIs on log-mean linear parameters. In this work, we provide the conditions to define the CSIs on HMM models, that are, as we already said, a wider class of models. At first, we focus on the possibility to define a CSI also on marginal distributions, maintaining the baseline approach. In this case, the HMM models corresponds to the Marginal models, (Bergsma and Rudas 2002). As we will see, the obtained constraints are the same of the ones presented in the literature, but defined on suitable marginal distributions. Then, we take into account also different aggregation criteria, that characterize the HMM models, reaching out different results. Note that, for simplicity, we will consider all the variables coded with the same approach, in this case all *baseline*, but it is not mandatory, as we already said. Indeed, the constraints involved in the CSI statement are influenced only by the approach used to code just the variables in the conditioning set, X_C. In the following Theorem 1 we present the constraints that the HMM parameters must satisfied in order to impose the CSI with equality in formula (4) in a HMM model.

Theorem 1 *Let us consider a set of variables X_V, with probability distribution $\mathcal{P}$ parametrized through the parameters in formula* (3)*, where the baseline criterion is used. Then, the probability distribution $\mathcal{P}$ obeys to the CSI with equality in formula* (4) *if and only if the following constraints on the HMM parameters are satisfied:*

$$\sum_{c \subseteq C} \eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = 0 \quad \textit{where } \mathcal{L} = q \cup c, \quad \textit{and} \quad \boldsymbol{i}_{\mathcal{L}} \in \mathcal{I}_q \times \mathcal{K}_c, \tag{7}$$

and where q takes values in the class $\mathcal{Q} = \{q \subseteq (A \cup B) : A \cap q \neq \emptyset, B \cap q \neq \emptyset\}$.

Thus, the parameters involved in the constraints definition refer to the whole set of categories of the variables X_q and refer to the subset of categories in $\mathcal{K}_c$ for the variables X_c. Example 2 shows an application of the formula (7) in correspondence of two different CSI hypotheses.

Example 2 Let us consider the vector of four variables X_{1234} collected in the marginal contingency table $\mathcal{I}_{1234}$ of dimension $3 \times 3 \times 3 \times 3$ and let us suppose that the CSI $X_1 \perp X_2 | (X_{34} = (1, 1))$ holds. In this case, the three sets A, B, and C in formula (4) are $A = (1)$, $B = (2)$, and $C = (3, 4)$. Thus, the HMM parameters involved in formula (7) are $\eta_{12}^{1234}(\boldsymbol{i}_{12})$, $\eta_{123}^{1234}(\boldsymbol{i}_{123})$, $\eta_{124}^{1234}(\boldsymbol{i}_{124})$, and $\eta_{1234}^{1234}(\boldsymbol{i}_{1234})$, for all values $\boldsymbol{i}_{12} \in \mathcal{I}_{12}$, $\boldsymbol{i}_{123} \in \mathcal{I}_{12} \times \{(1)\}$, $\boldsymbol{i}_{124} \in \mathcal{I}_{12} \times \{(1)\}$, and $\boldsymbol{i}_{1234} \in \mathcal{I}_{12} \times \{(1, 1)\}$. Thus, for each $\boldsymbol{i}_{12} \in \mathcal{I}_{12}$, the parameters are

$$\begin{aligned}
\eta_{12}^{1234}(\boldsymbol{i}_{12}) &= \log\left(\frac{\pi(\boldsymbol{i}_{12}33)\pi(3333)}{\pi(\boldsymbol{i}_{1}333)\pi(3\boldsymbol{i}_{2}33)}\right) \\
\eta_{123}^{1234}(\boldsymbol{i}_{12}, 1) &= \log\left(\frac{\pi(\boldsymbol{i}_{12}13)\pi(\boldsymbol{i}_{1}333)\pi(3\boldsymbol{i}_{2}33)\pi(3313)}{\pi(\boldsymbol{i}_{12}33)\pi(\boldsymbol{i}_{1}313)\pi(3\boldsymbol{i}_{2}13)\pi(3333)}\right) \\
\eta_{124}^{1234}(\boldsymbol{i}_{12}, 1) &= \log\left(\frac{\pi(\boldsymbol{i}_{12}31)\pi(\boldsymbol{i}_{1}333)\pi(3\boldsymbol{i}_{2}33)\pi(3331)}{\pi(\boldsymbol{i}_{12}33)\pi(\boldsymbol{i}_{1}331)\pi(3\boldsymbol{i}_{2}31)\pi(3333)}\right) \\
\eta_{1234}^{1234}(\boldsymbol{i}_{12}, 1, 1) &= \log\left(\frac{\pi(3333)\pi(\boldsymbol{i}_{12}33)\pi(\boldsymbol{i}_{1}313)\pi(3\boldsymbol{i}_{2}13)\pi(\boldsymbol{i}_{1}331)\pi(3\boldsymbol{i}_{2}31)}{\pi(\boldsymbol{i}_{1}333)\pi(3\boldsymbol{i}_{2}33)\pi(3313)\pi(\boldsymbol{i}_{12}13)\pi(3331)\pi(\boldsymbol{i}_{12}31)}\right).
\end{aligned}$$

Trivial mathematical steps show that the sum of all these four parameters is equal to zero, for each $\boldsymbol{i}_{12} \in \mathcal{I}_{12}$.

In the case of CSI $X_1 \perp X_2 | (X_{34} = (1, 3))$, since $I_4 = 3$, the parameters involving the variable X_4 at the third category are zero by definition, thus formula (7) becomes

$$\eta_{123}^{1234}(\boldsymbol{i}_{123}) + \eta_{12}^{1234}(\boldsymbol{i}_{12}) = 0$$

where $\boldsymbol{i}_{123} \in \mathcal{I}_{12} \times \{(1)\}$ and $\boldsymbol{i}_{12} \in \mathcal{I}_{12}$.

Let us suppose that the conditioning set in (4) is composed only of ordinal variables so that we consider the *local* criterion for all variables. The CSI hypothesis can be described by the constraints listed in Theorem 2.

Theorem 2 *Let us consider a set of variables X_V, with probability distribution $\mathcal{P}$ parametrized through the parameters in formula* (3)*, where the local criterion is used. Then, the probability distribution $\mathcal{P}$ obeys to the CSI with equality in formula* (4) *if and only if the following constraints on the HMM parameters are satisfied:*

$$\sum_{c \subseteq C} \sum_{\boldsymbol{i}_c \geq \boldsymbol{i}'_c} \eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = 0 \quad \textit{where } \mathcal{L} = q \cup c, \quad \boldsymbol{i}_{\mathcal{L}} \in \mathcal{I}_q \times \mathcal{K}_c, \quad \textit{and} \quad \forall \boldsymbol{i}'_c \in \mathcal{K}_c \tag{8}$$

and where q takes values in the class $\mathcal{Q} = \{q \subseteq (A \cup B) : A \cap q \neq \emptyset, B \cap q \neq \emptyset\}$.

Note that, comparing with Theorem 1, we have here a double summation: the external one referring to the subsets of variables in C and the internal one referring to certain categories $\boldsymbol{i}_c$ which the parameter refers.

We apply formula (8) in the next example.

Example 3 Given a set of variables, let us suppose that we are interested in representing the CSI $X_1 \perp X_2 | X_3 = 2$. Then, we consider the marginal set $\mathcal{M} = (1, 2, 3)$, and the corresponding marginal contingency table $\mathcal{I}_{\mathcal{M}}$ of dimension $2 \times 2 \times 4$. Here, the sums in formula (8) are as follows:

$$\eta_{12}^{123}(1,1) + \left(\eta_{123}^{123}(1,1,2) + \eta_{123}^{123}(1,1,3)\right) =$$
$$= \log\left(\frac{\pi(224)\pi(114)}{\pi(124)\pi(214)}\right) + \log\left(\frac{\pi(123)\pi(213)\pi(222)\pi(112)}{\pi(223)\pi(113)\pi(122)\pi(212)}\right) +$$
$$+ \log\left(\frac{\pi(124)\pi(214)\pi(223)\pi(113)}{\pi(224)\pi(114)\pi(123)\pi(213)}\right).$$

By simplifying the probabilities in the different logarithms, we obtain

$$\log\left(\frac{\pi(222)\pi(112)}{\pi(122)\pi(212)}\right),$$

that is equal to zero in force of the CSI.

Remark 3 Given a CSI statement such as in formula (4), the number of zero constraints to impose on a HMM model is $\left[\prod_{j \in (A \cup B)} (I_j - 1)\right] \times |\mathcal{K}_C|$.

3.2 Constraints on HMM parameters for the CSI with inequality

When we deal with ordinal variables, it is helpful also to consider the subclass of CSI statements with the inequality such as displayed in formulas (5) and (6). Indeed, this approach presents some advantages. Firstly, the CSI statements are more immediate to understand with respect to the CSIs in formula (4). Further, as it will be described in Theorem 3, the HMM parameters involved in the constraints are all null (not only their sum). This, once again, makes the output of the parameters more interpretable.

Obviously, since the class of CSIs in formula (5) [or (6)] is a special case of the class of CSIs in formula (4), if the constraints in Theorems 1 and 2 are satisfied for each $\boldsymbol{i}_C \geq \boldsymbol{i}'_C$ ($\boldsymbol{i}_C \leq \boldsymbol{i}'_C$), then formula (5) [or (6)] holds too. However, for the *baseline*, *local* and *continuation* aggregation criteria, there is an easiest way to define the CSI in formula (5), as shown in the following Theorem 3.

Theorem 3 *Let us consider a set of variables* X_V*, with probability distribution* $\mathcal{P}$ *parametrized through the parameters in formula* (3)*, where the baseline, the local or the continuation criterion is used. Then, the probability distribution* $\mathcal{P}$ *obeys to the*

CSI with equality in formula (5) if and only if the following constraints on the HMM parameters are satisfied:

$$\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}}) = 0 \quad \textit{where } \mathcal{L} = q \cup c, \quad c \subseteq C, \quad \boldsymbol{i}_q \in \mathcal{I}_q \quad \textit{and } \boldsymbol{i}_c \geq \boldsymbol{i}'_c \tag{9}$$

and q takes values in the class $\mathcal{Q} = \{q \subseteq (A \cup B) : A \cap q \neq \emptyset, B \cap q \neq \emptyset\}$.

Example 4 By considering the same context of Example 3, let us suppose we are interested in defining the CSI $X_1 \perp X_2 | X_3 \geq 2$; then, according to Theorem 3, the following constraints must hold

$$\eta^{123}_{12}(1, 1) = 0, \quad \eta^{123}_{123}(1, 1, 2) = 0, \quad \eta^{123}_{123}(1, 1, 3) = 0.$$

The first parameter $\eta^{123}_{12}(1, 1)$, independently from the kind of aggregation criterion we choose, has the form

$$\eta^{123}_{12}(1, 1) = \log\left(\frac{\pi(114)\pi(224)}{\pi(124)\pi(214)}\right).$$

Note that, $X_1 \perp X_2 | X_3 \geq 2$ implies $X_1 \perp X_2 | X_3 = 4$. Then the previous parameter is equal to zero. This holds only when the variables not considered in the parameter $X_{\mathcal{M}\setminus\mathcal{L}}$ are set up to the last category. However, if we choose, in the model definition, to fix the variables $X_{\mathcal{M}\setminus\mathcal{L}}$ equal to the first category, it is enough to list the modalities of the variables in the reverse order, with attention to consider the CSI in (6) instead of the CSI in (5).

Concerning the second parameter, when we use the *baseline* criterion, we have

$$\eta^{123}_{123}(1, 1, 2) = \log\left(\frac{\pi(124)\pi(214)\pi(222)\pi(112)}{\pi(224)\pi(114)\pi(122)\pi(212)}\right);$$

when we use the *local* criterion, we have

$$\eta^{123}_{123}(1, 1, 2) = \log\left(\frac{\pi(123)\pi(213)\pi(222)\pi(112)}{\pi(223)\pi(113)\pi(122)\pi(212)}\right);$$

and finally, when we adopt the *continuation* criterion, we have

$$\eta^{123}_{123}(1, 1, 2) = \log\left(\frac{(\pi(123) + \pi(124))\,(\pi(213) + \pi(214))\,\pi(222)\pi(112)}{(\pi(223) + \pi(224))\,(\pi(113) + \pi(114))\,\pi(122)\pi(212)}\right). \tag{10}$$

Note that, whatever the aggregation criterion we choose, the variable X_3 appears only with categories 2, 3 and 4, according to which the CSI holds. In the case of *baseline* or *local* criterion it is easy to see that, by rewriting every joint probability $\pi(\boldsymbol{i}_{123})$ as the product $\pi(\boldsymbol{i}_{13})\pi(\boldsymbol{i}_{23})$, the numerator and the denominator simplified. Thus the nullity of the correspondent parameter is highlighted.

On the other hand, the parameters coded with the *continuation* criterion, as showed in Agresti (2013), can be conveyed as a second order contrast of baseline logits of the following conditional probabilities

$$\omega_{112} = P(X_{123}{=}(1,1,2)|X_1{=}1, X_2{=}1, X_3 \geq 2) = \frac{\pi(112)}{(\pi(112)+\pi(113)+\pi(114))}$$
$$\omega_{122} = P(X_{123}{=}(1,2,2)|X_1{=}1, X_2{=}2, X_3 \geq 2) = \frac{\pi(122)}{(\pi(122)+\pi(123)+\pi(124))}$$
$$\omega_{212} = P(X_{123}{=}(2,1,2)|X_1{=}2, X_2{=}1, X_3 \geq 2) = \frac{\pi(212)}{(\pi(212)+\pi(213)+\pi(214))}$$
$$\omega_{222} = P(X_{123}{=}(2,2,2)|X_1{=}2, X_2{=}2, X_3 \geq 2) = \frac{\pi(222)}{(\pi(222)+\pi(223)+\pi(224))}.$$

Few mathematical steps show that the parameter in (10) is equal to the contrast

$$\eta_{123}^{123}(1,1,2) = \log\left(\frac{\omega_{112}}{1-\omega_{112}}\frac{\omega_{222}}{1-\omega_{222}}\frac{1-\omega_{122}}{\omega_{122}}\frac{1-\omega_{212}}{\omega_{212}}\right). \qquad (11)$$

Since ωs probabilities concern distributions where the CSI holds, the parameter (11), based on baseline criterion, is null.

In the same way we proceed for the third parameter $\eta_{123}^{123}(1,1,3)$ that is equal to zero.

Remark 4 When we are interested in defining a CSI as expressed in formula (6), we can proceed in an analogous way, previously sorting in a descending order the categories of the variable of interest.

Example 5 Let us consider three variables taking values in $i_1 \in \{1,2\}$, $i_2 \in \{1,2\}$, and $i_3 \in \{1,2,3,4\}$, respectively. If we want to test the CSI $X_1 \perp X_2|X_3 \leq 2$ it is enough to invert the order of the categories. We replace the letter instead of the number for more clarity. Thus, we get $i_1 \in \{a=2, b=1\}$, $i_2 \in \{a=2, b=1\}$, and $i_3 \in \{a=4, b=3, c=2, d=1\}$ and the CSI becomes $X_1 \perp X_2|X_3 \geq c$. The parameters involved in the independence are the same of Example 4. For instance the first one is:

$$\eta_{12}^{123}(a,a) = \log\left(\frac{\pi(aad)\pi(bbd)}{\pi(abd)\pi(bad)}\right).$$

That actually is

$$\eta_{12}^{123}(2,2) = \log\left(\frac{\pi(221)\pi(111)}{\pi(211)\pi(121)}\right),$$

where the CSI holds.

Remark 5 Given a CSI statement such as in formula (5) or (6), the number of zero constraints to impose on a HMM model is $\sum_{c \subseteq C} \left[\prod_{j \in (A \cup B)} \left(I_j - 1\right)\right] \times (|\mathcal{K}_c| - 1)$ where $\boldsymbol{i}_C \in \mathcal{K}_C$ if and only if $\boldsymbol{i}_C \geq \boldsymbol{i}'_C$.

Example 6 In the Example 4 we considered the CSI $X_1 \perp X_2 | X_3 \geq 2$ where the variables have 2, 2, and 4 levels, respectively. Thus, the total number of constraints are $(1 \times 1) + (1 \times 1)(2) = 3$: $\eta_{12}^{123}(1, 1) = 0$, $\eta_{123}^{123}(1, 1, 2) = 0$, and $\eta_{123}^{123}(1, 1, 3) = 0$.

In general, we can decide to codify the variables heterogeneously, each of them with different types of aggregation criteria, in order to suit the nature of the single variable. Here, we present an example in order to show how to apply the different conditions on constraints when we deal with variables coded with diverse types of aggregation criteria.

Example 7 Let us consider a marginal set $\mathcal{M} = (1, 2, 3, 4)$ composed of 4 variables collected in a $2 \times 2 \times 4 \times 4$ contingency table $\mathcal{I}_{\mathcal{M}}$. We codify the variables with *baseline*, *baseline*, *local* and *continuation* aggregation criteria, respectively. We are interested in checking the CSI $X_1 \perp X_2 | X_{3,4} \geq (2, 2)$ that means that the CSI must hold when the variables X_3 and X_4 assume, respectively, the values $X_3 \geq 2$ and $X_4 \geq 2$, that is for the cells $\{(2, 2); (2, 3); (3, 2); (3, 3)\}$. In this case, noting that the variables in the conditioning set are coded with the *local* and the *continuation* aggregation criteria, the result dues to Theorem 3 implies that the parameters $\eta_{1234}^{1234}(\boldsymbol{i}_{1234})$ with $\boldsymbol{i}_{1234} \in \{(1, 1, 2, 2); (1, 1, 2, 3); (1, 1, 3, 2); (1, 1, 3, 3)\}$, $\eta_{123}^{1234}(\boldsymbol{i}_{123})$ with $\boldsymbol{i}_{123} \in \{(1, 1, 2); (1, 1, 3)\}$, $\eta_{124}^{1234}(\boldsymbol{i}_{124})$ with $\boldsymbol{i}_{124} \in \{(1, 1, 2); \ (1, 1, 3)\}$ and $\eta_{12}^{1234}(1, 1)$ involving the conditioning variables with values greater or equal to $(2, 2)$, have to be zero, how effectively is.

4 Application

In this section, we use the HMM model to study the small and medium Italian firms' innovation. In particular, our aim is to highlight the (in)dependence relationships among different aspects of the enterprises, by taking into account the possible order of the categories of the considered variables. For this reason, we take advantage of the CSIs as presented in Sect. 3.

We take into account the Italian Innovation Survey concerning the period starting from 2010 to 2012 (Istat 2012). We select 7 variables of interest, pertaining to different environments of the firms' life. The first type of variables is the firms' featuring: the *enterprise size*, **DIM** (1 = Small, 2 = Medium), the *percentage of graduate employers*, **DEG** (1 = 0% ⊢ 10%, 2 = 10% ⊢ 50%, 3 = 50% ⊢ 100%) and the *main market (in revenue terms)*, **MRKT** (1 = Regional, 2 = National, 3 = International), henceforth denoted as variables X_1, X_2 and X_3, respectively. Then, there are the variables concerning the innovation in some aspects of the enterprises: *innovation in marketing strategies*, **IMAR** (1 = Yes, 2 = No), *innovation in organization system*, **IOR** (1 = Yes, 2 = No) and *innovation in products or services or production line or investment in R&D*, **IPR** (1 = Yes, 2 = No), henceforth denoted as variables X_4, X_5 and X_6, respectively. Finally, we consider the *revenue growth* variable in 2012, **GROW** (1 = Yes, 2 = No) henceforth denoted as variable X_7.

The survey covers 18 697 firms, collected in a $2 \times 3 \times 3 \times 2 \times 2 \times 2 \times 2$ contingency table with only one empty cell on 288, thus the hypothesis of strictly positivity of the probability distribution can be credible.

Our aim is manifold, firstly we want to investigate if the presence/absence of innovation in some aspects of the enterprise life can be affected by the firms' features. Thus, we have to consider the independence $X_{123} \perp X_{456}$. Secondly, also the effect of the firms' features and the innovation variables on the revenue growth is another aspect that we want to study. Thus, the independence $X_{123456} \perp X_7$ must be considered.

A HMM parametrization, able to describe these two independencies, can be based on the hierarchical class of marginal sets $\mathcal{H} = \{(1, 2, 3); (1, 2, 3, 4, 5, 6), (1, 2, 3, 4, 5, 6, 7)\}$. The previous class $\mathcal{H}$ takes into account the nature of the variables and, at the same time, it satisfies the condition of Rudas et al. (2010), that assures the representativeness of a list of independencies. The validation of the HMM model with the previous marginals and the constraints induced by the two independencies is done with the likelihood ratio test G^2 which compares the model under investigation with the saturated (unconstrained) one. Under the null hypothesis the test statistic G^2 follows the χ^2 distribution with degree of freedom, df, equal to the difference between the free parameters in the two models (the saturated one and the model that we want to test). We reject the hypothesis that the selected model provides a good representation of the dataset when we get the *p value* lower than a chosen critical value.

The likelihood ratio test applied to the model, where both the conditional independencies hold, provides the following results: $G^2 = 3548.47$, $df = 262$, $p\,value = 0.000$. Since for these two variables the model is not a good representation, we proceed to test different nested models. The nested models can contain both conditional independencies or CSIs. The choice of the nested models to test was driven from the support of further statistical tools such as the mosaic plots.

As highlighted in Nyman et al. (2016), it can occur that the constraints to impose on the parameters in order to satisfy both conditional independencies and CSIs are too restrictive, underlying stronger independence relationships. In this case, we ascertained that each list of conditional independencies and/or CSIs was well represented by the constraints. That means we checked if there were no other independencies implied by these constraints.

The best fitting model, with $G^2 = 155.79$, $df = 132$, $p\,value = 0.08$ results the one where the following independencies hold simultaneously:

(a) $X_{24} \perp X_7 | X_{1356}$
(b) $X_2 \perp X_6 | X_{1345} \geq (2, 2, 2, 1)$
(c) $X_{13} \perp X_4 | X_{256} \geq (3, 1, 2)$.

Thus, in the best fitting model the revenue growth (X_7) results independent from the percentage of graduated employers (X_2) and the innovation in marketing strategies (X_4) given by all the other variables. Further, we have also the following CSIs. The percentage of graduated employers (X_2) is independent from the innovation in product or services or investments (X_6), when the dimension of the enterprise is big ($X_1 \geq 2$), when the main market is national or international ($X_3 \geq 2$), when there is no innovation

Table 2 List of the values of the second order parameters, $\eta_{\mathcal{L}}^{\mathcal{M}}(i_{\mathcal{L}})$, (logarithms of odds ratios) of the best fitting model

		DEG X_2		MRKT X_3		IMAR X_4	IOR X_5	IPR X_6	GROW X_7
		2	3	2	3	2	2	2	2
DIM X_1	2	0.10	0.92***	0.94***	0.86***	0.00	0.27	1.47***	− 0.04
DEG X_2	2			0.03	0.64***	0.32**	0.50*	0.00	0.00
	3			0.42***	0.46***	− 0.02	0.60***	0.00	0.00
MRKT X_3	2					0.00	− 0.15	0.85***	0.15*
	3					0.00	− 0.22	0.84***	− 0.54***
IMAR X_4	2						1.48***	1.90***	0.00
IOR X_5	2							1.20***	0.01
IPR X_6	2								0.00

The second column and row refer to the levels of the corresponding variables

The asterisk denotes the parameters that are significantly different from zero (*** with $\alpha \leq 0.01$; ** with $0.01 \leq \alpha \leq 0.05$; * with $0.05 \leq \alpha \leq 0.1$), according to the Wald test

in marketing strategies ($X_4 \geq 2$) and if there is or there is not innovation in the organization system ($X_5 \geq 1$). Finally, the dimension of the enterprise (X_1) and the main market (X_3) are jointly independent from the innovation in marketing strategies (X_4) when the firm has a high level of graduated employers ($X_2 \geq 3$), whatever is the innovation in the organization system ($X_5 \geq 1$), and when there is no innovation in product and services ($X_6 \geq 2$).

In Table 2 the second order parameters, logarithms of odds ratios, of the best fitting model are displayed. The asterisk denotes the parameters significantly different from zero (with a significance level α at most equal to 0.01) according to the Wald test. A brief consideration on the model is here discussed starting from these parameters. We use for all variables the *continuation* criterion. At first, note that, according to the conditional independence (a), the parameters $\eta_{27}^{V}(2,2)$, $\eta_{27}^{V}(3,2)$ and $\eta_{47}^{V}(2,2)$ are null. Obviously, also the parameters of greater order involving these pairs of variables are null. However, one of the most advantages in defining the CSIs as in formula (5) or (6) lies on the fact that all the parameters involved in the constraints are null [not only the sum such as for the case of CSIs in formula (4)]. Thus, also $\eta_{34}^{234567}(2,2)$, $\eta_{34}^{234567}(3,2)$, $\eta_{26}^{234567}(2,2)$, $\eta_{26}^{234567}(3,2)$ and $\eta_{14}^{V}(2,2)$ are null according to the CSIs (b) and (c). The remaining parameters inform us on how each pair of variables affect each other. Note that the parameter $\eta_{67}^{234567}(2,2)$ is null even if it not involved in any independencies. Worthy of note is the parameter $\eta_{46}^{123456}(2,2) = 1.90$ meaning that, the probability to have innovation in both marketing strategies and products and services and to not have innovation in both is greater than the probability of discordance of these two variables. The same trend, even if not so strong, is in almost all the remaining parameters. Thus, focusing on the relationships between the revenue growth and the other variables, the only one that seems to affect significantly the revenue growth is the main market where the firm works as we can see from $\eta_{37}^{V}(3,2) = -0.54$. In particular, by increasing the main market there is a negative trend between the innovation and bigger market.

All the analysis were carried out with the statistical software R, (R Core Team 2014), with the help the package hmmm, Colombi et al. (2014) for testing the HMM models.

5 Conclusion

In this work we provide several results concerning the framework of CSIs. Besides the classical definition of CSI we introduce a new point of view where the CSI holds for certain categories of few variables satisfying an inequality. While the classical definition is appropriate for unordered variables, the new one is more suitable for ordinal variables and it leads to benefits in the interpretation of the parameters. Concerning the classical definition of CSI we confirm the results for parameters based on *baseline* logits such as provided in Nyman et al. (2016), even in the marginal model context, and we also provide the results in the case of *local* criterion. On the other hand, when we deal with the definitions with inequality, we provide original results for different aggregation criteria: *baseline*, *local* and *continuation*. Although these new definitions consider a subset of CSIs with respect to (4), they make the parameters more meaningful since the involved ones are null, not only their sum.

The application shows a part of the potentiality of this work.

Appendix : Proofs and further results

We are going to prove Theorems 1, 2, and 3. Note that, in order to do that, we preliminary need to declare and demonstrate the following results: Lemma 1 and Corollary 1.

Lemma 1 *Given a HMM parameter $\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}})$, where the interaction set can be expressed as union of two incompatible sets, $\mathcal{L} = L \cup C$, belonging in $\mathcal{M}$, it can be decomposed as follows*

$$\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_{L}|\boldsymbol{i}_{C}) - \sum_{\substack{J \subseteq C \\ J \neq \emptyset}} \eta_{\mathcal{L}\backslash J}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash J}|\boldsymbol{i}^{*}_{J}). \tag{12}$$

Proof of Lemma 1 From the Proposition (1) of Bartolucci et al. (2007), each parameter $\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}})$, where $\mathcal{L} = L \cup C$ can be rewritten as

$$\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = \sum_{J \subseteq C} (-1)^{|C\backslash J|} \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_{L}|\boldsymbol{i}_{J}, \boldsymbol{i}^{*}_{C\backslash J}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}), \tag{13}$$

where $\eta_{L}^{\mathcal{M}}(\boldsymbol{i}_{L}|\boldsymbol{i}_{J}, \boldsymbol{i}^{*}_{C\backslash J}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}})$ is the HMM parameter $\boldsymbol{\eta}_{L}^{\mathcal{M}}$ evaluated in the conditional distribution where the variables in X_J assume values $\boldsymbol{i}_J$ and the variables in $X_{C\backslash J}$ are set to the categories $\boldsymbol{i}^{*}_{C\backslash J}$.

When the set C is composed of only one index, $C = \gamma_1$, the decomposition in formula (13) becomes

$$\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_L|\boldsymbol{i}_{\gamma_1}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) - \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_L|\boldsymbol{i}^*_{\gamma_1}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}), \tag{14}$$

that corresponds to formula (12).

When two variables belong to the set C, $C = \{\gamma_1, \gamma_2\}$, by applying formula (13) only to γ_1 we get

$$\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = \eta_{\mathcal{L}\backslash\gamma_1}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash\gamma_1}|\boldsymbol{i}_{\gamma_1}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) - \eta_{\mathcal{L}\backslash\gamma_1}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash\gamma_1}|\boldsymbol{i}^*_{\gamma_1}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}). \tag{15}$$

Note that, the first addend, on the right hand side, can be further decomposed by using the (13) as:

$$\eta_{\mathcal{L}\backslash\gamma_1}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash\gamma_1}|\boldsymbol{i}_{\gamma_1}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) = \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_L|\boldsymbol{i}_C, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) - \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_L|\boldsymbol{i}_{\gamma_1}, \boldsymbol{i}^*_{\gamma_2}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}). \tag{16}$$

Now, by considering the HMM parameter $\eta_{\mathcal{L}\backslash\gamma_2}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash\gamma_2}|\boldsymbol{i}^*_{\gamma_2}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}})$ and by applying the formula (13) to γ_1, we get

$$\eta_{\mathcal{L}\backslash\gamma_2}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash\gamma_2}|\boldsymbol{i}^*_{\gamma_2}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) = \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_L|\boldsymbol{i}_{\gamma_1}, \boldsymbol{i}^*_{\gamma_2}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) - \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_L|\boldsymbol{i}^*_C, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}). \tag{17}$$

It is easy to see that the last addend on the right hand side of the (16) is exactly the first addend on the right hand side of (17). Thus, by replacing the (16) and (17) in formula (15) we get:

$$\begin{aligned}\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = &- \eta_{\mathcal{L}\backslash\gamma_1}^{\mathcal{M}}(\boldsymbol{ii}_{\mathcal{L}\backslash\gamma_1}|\boldsymbol{ii}^*_{\gamma_1}, \boldsymbol{II}_{\mathcal{M}\backslash\mathcal{L}}) - \eta_{L}^{\mathcal{M}}(\boldsymbol{ii}_{\mathcal{L}}|\boldsymbol{ii}^*_C, \boldsymbol{II}_{\mathcal{M}\backslash\mathcal{L}}) \\ &- \eta_{\mathcal{L}\backslash\gamma_2}^{\mathcal{M}}(\boldsymbol{ii}_{\mathcal{L}\backslash\gamma_2}|\boldsymbol{ii}^*_{\gamma_2}, \boldsymbol{II}_{\mathcal{M}\backslash\mathcal{L}}) + \eta_{L}^{\mathcal{M}}(\boldsymbol{ii}_L|\boldsymbol{ii}_C, \boldsymbol{II}_{\mathcal{M}\backslash\mathcal{L}})\end{aligned} \tag{18}$$

that again corresponds to formula (12).

In general, when the set C is composed of k variables, $C = \{\gamma_1, \ldots, \gamma_k\}$, we apply formula (13), focusing on only one variable of C. Thus, at first step we get

$$\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = \eta_{\mathcal{L}\backslash\gamma_1}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash\gamma_1}|\boldsymbol{i}_{\gamma_1}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) - \eta_{\mathcal{L}\backslash\gamma_1}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash\gamma_1}|\boldsymbol{i}^*_{\gamma_1}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}). \tag{19}$$

Then, we apply formula (13) recursively, focusing on only one variable of C at a time, to any parameter in the formula without any index $\boldsymbol{i}^*$ in the conditioning set

$$\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = \eta_{L}^{\mathcal{M}}(\boldsymbol{i}_L|\boldsymbol{i}_C, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) - \sum_{j=1}^{k} \eta_{\mathcal{L}\backslash\gamma_{jp}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}\backslash\gamma_{jp}}|\boldsymbol{i}^*_{\gamma_j}, \boldsymbol{i}_{\gamma_{jp}\backslash\gamma_j}, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}). \tag{20}$$

where $\gamma_{jp} = \cup_{i=1}^{j}\gamma_i$.

Now, we take into account all the parameters having both $\boldsymbol{i}$ and $\boldsymbol{i}^*$ in the conditioning set. Let us denote them as $\eta_{L}^{\mathcal{M}}(\boldsymbol{i}_L|\boldsymbol{i}_A, \boldsymbol{i}^*_B, \boldsymbol{I}_{\mathcal{M}\backslash LAB})$. We can recognize it in the last term of the right hand side of the decomposition (21) obtained applying the (13) to $\eta^{\mathcal{M}}_{\mathcal{L}\setminus B}(\boldsymbol{i}_{\mathcal{L}\setminus B}|\boldsymbol{i}^*_B, \boldsymbol{I}_{\mathcal{M}\setminus LAB})$:

$$\eta^{\mathcal{M}}_{\mathcal{L}\setminus B}(\boldsymbol{i}_{\mathcal{L}\setminus B}|\boldsymbol{i}^*_B, \boldsymbol{I}_{\mathcal{M}\setminus LAB}) = \eta^{\mathcal{M}}_{L}(\boldsymbol{ii}_L|\boldsymbol{ii}_A, \boldsymbol{i}^*_B, \boldsymbol{II}_{\mathcal{M}\setminus LAB}) + \\ - \sum_{\substack{J \subseteq A \\ J \neq \emptyset}} \eta^{\mathcal{M}}_{L}(\boldsymbol{ii}_L|\boldsymbol{ii}_{A\setminus J}, \boldsymbol{ii}^*_{BJ}, \boldsymbol{II}_{\mathcal{M}\setminus LAB}). \tag{21}$$

Thus, we can isolate the term $\eta^{\mathcal{M}}_{L}(\boldsymbol{i}_L|\boldsymbol{i}_A, \boldsymbol{i}^*_B, \boldsymbol{I}_{\mathcal{M}\setminus LAB})$ as follows:

$$\eta^{\mathcal{M}}_{L}(\boldsymbol{i}_L|\boldsymbol{i}_A, \boldsymbol{i}^*_B, \boldsymbol{I}_{\mathcal{M}\setminus LAB}) = \sum_{J\subseteq A} \eta^{\mathcal{M}}_{L}(\boldsymbol{i}_L|\boldsymbol{i}_{A\setminus J}, \boldsymbol{i}^*_{BJ}, \boldsymbol{I}_{\mathcal{M}\setminus LAB}). \tag{22}$$

Now, in formula (20), we replace each addend like $\eta^{\mathcal{M}}_{L}(\boldsymbol{i}_L|\boldsymbol{i}_A, \boldsymbol{i}^*_B, \boldsymbol{I}_{\mathcal{M}\setminus LAB})$ with the expression in formula (22) and we apply this procedure recursively to each addend like $\eta^{\mathcal{M}}_{L}(\boldsymbol{i}_L|\boldsymbol{i}^*_{A\setminus J}, \boldsymbol{i}_{BJ}, \boldsymbol{I}_{\mathcal{M}\setminus LAB})$. In this way we finally obtain exactly formula (12). □

Corollary 1 *A parameter $\boldsymbol{\eta}^{\mathcal{M}}_{L}$ can be decomposed as the sum of greater order parameters as follows:*

$$\eta^{\mathcal{M}}_{L}(\boldsymbol{i}_L|\boldsymbol{i}_C) = \sum_{J\subseteq C} \eta^{\mathcal{M}}_{\mathcal{L}\setminus J}(\boldsymbol{i}_{\mathcal{L}\setminus J}|\boldsymbol{i}^*_J), \tag{23}$$

where $\mathcal{L} = L \cup C$ and $C \cap L = \emptyset$.

Proof of Corollary 1 The proof comes easily by isolating the first term in the right hand side of the formula (12) of Lemma 1. □

We are now ready to go into details of the proofs of the theorems.

Proof of Theorem 1 Let us consider the parameters $\boldsymbol{\eta}^{\mathcal{M}}_{\mathcal{L}}$ when $\mathcal{L} = (A \cup B \cup C) \subseteq \mathcal{M}$. From Lemma 1 we can decompose it as

$$\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}}) = \eta^{\mathcal{M}}_{\mathcal{L}\setminus C}(\boldsymbol{i}_{\mathcal{L}\setminus C}|\boldsymbol{i}_C, \boldsymbol{I}_{\mathcal{M}\setminus\mathcal{L}}) - \sum_{\substack{J\subseteq C \\ J\neq\emptyset}} \eta^{\mathcal{M}}_{\mathcal{L}\setminus J}(\boldsymbol{i}_{\mathcal{L}\setminus J}|\boldsymbol{i}^*_J, \boldsymbol{I}_{\mathcal{M}\setminus\mathcal{L}}) \tag{24}$$

where $\eta^{\mathcal{M}}_{\mathcal{L}\setminus C}(\boldsymbol{i}_{\mathcal{L}\setminus C}|\boldsymbol{i}_C, \boldsymbol{I}_{\mathcal{M}\setminus\mathcal{L}})$ is the marginal parameter $\boldsymbol{\eta}^{\mathcal{M}}_{\mathcal{L}\setminus C}$ evaluated in the conditional distribution $(A \cup B|C = \boldsymbol{i}_C)$. The term $\eta^{\mathcal{M}}_{\mathcal{L}\setminus C}(\boldsymbol{i}_{\mathcal{L}\setminus C}|\boldsymbol{i}_C, \boldsymbol{I}_{\mathcal{M}\setminus\mathcal{L}})$ is equal to zero if and only if the CSI in formula (4) holds. Thus,

$$\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}}) + \sum_{\substack{J \subseteq C \\ J \neq \emptyset}} \eta^{\mathcal{M}}_{\mathcal{L}\setminus J}(\boldsymbol{i}_{\mathcal{L}\setminus J}|\boldsymbol{i}^*_J, \boldsymbol{I}_{\mathcal{M}\setminus\mathcal{L}}) = 0 \\ \sum_{J\subseteq C} \eta^{\mathcal{M}}_{\mathcal{L}\setminus J}(\boldsymbol{i}_{\mathcal{L}\setminus J}|\boldsymbol{i}^*_J, \boldsymbol{I}_{\mathcal{M}\setminus\mathcal{L}}) = 0. \tag{25}$$

Note that in the case of *baseline* aggregation criterion, the cell $\boldsymbol{i}^*_J$ is equivalent to $\mathbf{I}_J$ thus, from formula (2), the parameter $\eta^{\mathcal{M}}_{\mathcal{L}\backslash J}(\boldsymbol{i}_{\mathcal{L}\backslash J}|\boldsymbol{i}^*_J, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}})$ is equal to $\eta^{\mathcal{M}}_{\mathcal{L}\backslash J}(\boldsymbol{i}_{\mathcal{L}\backslash J}, \boldsymbol{I}_{(\mathcal{M}\backslash\mathcal{L})J})$.

Finally, by considering that the previous decomposition holds for each set $q \in \mathcal{Q} = \{q \subseteq (A \cup B) : A \cap q \neq \emptyset, B \cap q \neq \emptyset\}$, the formula (7) comes. □

Proof of Theorem 2 By resuming the proof of Theorem 1, note that all the considerations until the decomposition in formula (25) still hold. However, by using the *local* aggregation criterion $\boldsymbol{i}^*_J \neq \mathbf{I}_J$ it is worthwhile to consider that the identity $\eta^{\mathcal{M}}_{\mathcal{L}\backslash J}(\boldsymbol{i}_{\mathcal{L}\backslash J}|\boldsymbol{i}^*_J, \boldsymbol{I}_{\mathcal{M}\backslash\mathcal{L}}) = \eta^{\mathcal{M}}_{\mathcal{L}\backslash J}(\boldsymbol{i}_{\mathcal{L}\backslash J}, \boldsymbol{I}_{(\mathcal{M}\backslash\mathcal{L})J})$ does not hold any more, such as in the *local* coding $\boldsymbol{i}^*_J$ is equal to $(i_j + 1)$ for all $j \in J$, in short-cut $\boldsymbol{i}_J + 1$. Further, the parameter $\eta^{\mathcal{M}}_{\mathcal{L}\backslash C}(\boldsymbol{i}_{\mathcal{L}\backslash C}, \boldsymbol{I}_{(\mathcal{M}\backslash\mathcal{L})C})$ is built in the conditional distribution where the variables in X_C assume the reference value $\mathbf{I}_C$. Note that $\eta^{\mathcal{M}}_{\mathcal{L}\backslash J}(\boldsymbol{i}_{\mathcal{L}\backslash J}|\boldsymbol{i}_J+1, \boldsymbol{I}_{(\mathcal{M}\backslash\mathcal{L})J})$ does not belong to the HMM parametrization. Now we remark that between the *baseline* parameters, $\boldsymbol{\eta}(\cdot)_b$, and the *local* parameters $\boldsymbol{\eta}(\cdot)_l$, the following relationship holds:

$$\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}'_{\mathcal{L}})_b = \sum_{\boldsymbol{i}_{\mathcal{L}} \geq \boldsymbol{i}'_{\mathcal{L}}} \eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}})_l. \tag{26}$$

When the variables in the conditioning set C are coded with the *local* approach, it is enough to apply the decomposition (26) only to the categories of the variables in the conditioning set C in order to return to the *baseline* approach. Thus we can rewrite (25) as:

$$\sum_{c \subseteq C} \sum_{\boldsymbol{i}_c \geq \boldsymbol{i}'_c} \eta^{\mathcal{M}}_{A\cup B\cup c}(\boldsymbol{i}_{A\cup B\cup c}|\mathbf{I}_{\mathcal{M}\backslash c}) = 0, \tag{27}$$

where $\eta^{\mathcal{M}}_{A\cup B\cup c}$ are the *local* parameters and they are exactly the same of formula (8). As in the proof of Theorem 1, the previous equivalence must hold for each subset q of $A \cup B$ with at least one element in A and one element in B. □

Proof of Theorem 3 Let us consider the inequality CSI statement as listed in formula (5).

When the set of categories $\boldsymbol{i}'_C$ in formula (5) is equal to $\mathbf{I}_C$, i.e. when the CSI holds only when all variables in C assume the last level, the parameters $\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}})$ are null. Indeed, when $\mathcal{L} = q \cup c$, with $c \subseteq C$ and $c \neq \emptyset$ the parameter is equal to $\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_q, \mathbf{I}_c)$ that is null by definition, whatever we code the variables (*baseline*, *local* or *continuation*). However, when $c = \emptyset$, the parameter becomes $\eta^{\mathcal{M}}_q(\boldsymbol{i}_q)$ that is a contrast of logits or an higher order parameter evaluated in the (conditional) contingency table of $X_q|X_C = I_C$. Hence, the parameter is null if and only if the independence holds, it is shown in the Example 4. Thus, since the $\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_q, \mathbf{I}_c)$ $\forall q \in \mathcal{Q}$ and $\forall c \subseteq C$ are evaluated in the conditional distribution $X_C = \mathbf{I}_C$ where the CSI holds, these parameters are equal to zero.

When the $\boldsymbol{i}'_C$ is equal to $(\mathbf{I}_{C\backslash j}, \mathbf{I}_j - 1)$, that is the level of each variable is equal to the last level but the variable j assumes the level $\mathbf{I}_j - 1$, as before, we have that all the parameters $\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}})$ with $\mathcal{L} = q \cup c$ and $c \subseteq C\backslash j$ are equal to zero. However, note that in the parameter $\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{qj})$, whatever the aggregation criteria is chosen (*baseline*, *local* or *continuation*), the variable X_j takes value $I_j - 1$ or I_j. Since in each of these distributions the CSI (5) holds, also this parameter is equal to zero and vice versa.

In general, when the CSI in (5) holds for a generic $\boldsymbol{i}'_C$, the parameters $\eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}})$, with $\mathcal{L} = q \cup c$ for any $\boldsymbol{i}_c$ greater or equal to $\boldsymbol{i}'_c$, involve the categories of each variable X_j in X_C, i_j or I_j (*baseline* approach), or $i_j + 1$ (*local* approach) or $((i_j + 1) + \cdots + I_j)$ (*continuation* approach). Since in all these cells the CSI holds, the parameters are equal to zero and vice versa. □

References

Albert JH (1996) Bayesian selection of log-linear models. Can J Stat 24(3):327–347

Agresti A (2013) Categorical data analysis. Wiley, Hoboken

Bartolucci F, Colombi R, Forcina A (2007) An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. Stat Sin 17:691–711

Bergsma WP, Rudas T (2002) Marginal models for categorical data. Ann Stat 30(1):140–159

Boutilier C, Friedman N, Goldszmidt M, Koller D (1996) Context-specific independence in Bayesian networks. In: Proceedings of the twelfth international conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp 115–123

Cazzaro M, Colombi R (2008) Modelling two way contingency tables with recursive logits and odds ratios. Stat Methods Appl 17(4):435–453

Cazzaro M, Colombi R (2014) Marginal nested interactions for contingency tables. Commun Stat Theory Methods 43(13):2799–2814

Colombi R, Forcina A (2014) A class of smooth models satisfying marginal and context specific conditional independencies. J Multivar Anal 126:75–85

Colombi R, Giordano S, Cazzaro M (2014) hmmm: an R package for hierarchical multinomial marginal models. J Stat Softw 59(11):1–25

Dellaportas P, Forster JJ (1999) Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika 86(3):615–633

Drton M (2009) Likelihood ratio tests and singularities. Ann Stat 37(2):979–1012

Forcina A (2012) Smoothness of conditional independence models for discrete data. J Multivar Anal 106:49–56

Glonek GFV, McCullagh P (1995) Multivariate logistic models. J R Stat Soc Ser B (Methodological) 57:533–546

Højsgaard S (2004) Statistical inference in context specific interaction models for contingency tables. Scand J Stat 31(1):143–158

Istat. Italian innovation survey on SM enterprises (2012)

La Rocca L, Roverato A (2017) Discrete graphical models. Handbook of graphical models. Handbooks of modern statistical methods. Chapman and Hall/CRC, Boca Raton

Lauritzen SL (1996) Graphical models, vol 17. Clarendon Press, Oxford

Meyer D, Zeileis A, Hornik K (2017) vcd: Visualizing Categorical Data. R package version 1.4-4

Nicolussi F, Cazzaro M (2017) Context-specific independencies for ordinal variables in chain regression models. arXiv:1712.05229

Ntzoufras I, Tarantola C (2013) Conjugate and conditional conjugate Bayesian analysis of discrete graphical models of marginal independence. Comput Stat Data Anal 66:161–177

Ntzoufras I, Tarantola C, Lupparelli M (2018) Probability based independence sampler for Bayesian quantitative learning in graphical log-linear marginal models. Working paper No. 149. University of Pavia, Department of Economics and Management

Nyman H, Pensar J, Koski T, Corander J (2014) Stratified graphical models—context-specific independence in graphical models. Bayesian Anal 9(4):883–908

Nyman H, Pensar J, Koski T, Corander J (2016) Context specific independence in graphical log-linear models. Comput Stat 31(4):1493–1512

R Core Team (2014) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria

Roverato A (2017) Graphical models for categorical data. Cambridge University Press, Cambridge

Rudas T, Bergsma WP, Németh R (2010) Marginal log-linear parameterization of conditional independence models. Biometrika 97(4):1006–1012

Sadeghi K, Rinaldo A (2018) Markov properties of discrete determinantal point processes. arxiv:1810.02294v2