# Log-mean linear models for binary data

Alberto Roverato*, Monia Lupparelli† and Luca La Rocca‡

February 9, 2012

**Abstract**

This paper is devoted to the theory and application of a novel class of models for binary data, which we call log-mean linear (LML) models. The characterizing feature of these models is that they are specified by linear constraints on the LML parameter, defined as a log-linear expansion of the mean parameter of the multivariate Bernoulli distribution. We show that marginal independence relationships between variables can be specified by setting certain LML interactions to zero and, more specifically, that graphical models of marginal independence are LML models. LML models are code dependent, in the sense that they are not invariant with respect to relabelling of variable values. As a consequence, they allow us to specify sub-models defined by code-specific independencies, which are independencies in sub-populations of interest. This special feature of LML models has useful applications. Firstly, it provides a flexible way to specify parsimonious sub-models of marginal independence models. The main advantage of this approach concerns the interpretation of the sub-model, which is fully characterized by independence relationships, either marginal or code-specific. Secondly, the code-specific nature of these models can be exploited to focus on a fixed, arbitrary, cell of the probability table and on the corresponding sub-population. This leads to an innovative family of models, which we call pivotal code-specific LML models, that is especially useful when the interest of researchers is focused on a small sub-population obtained by stratifying individuals according to some features. The application of LML models is illustrated on two datasets, one of which concerns the use of pivotal code-specific LML models in the field of personalized medicine.

*Keywords*: Contingency table; Graphical Markov model; Marginal independence; Mean parameterization; Parsimonious model.

*University of Bologna (alberto.roverato@unibo.it)

†University of Bologna (monia.lupparelli@unibo.it)

‡University of Modena and Reggio Emilia (luca.larocca@unimore.it).

# 1 Introduction

A straightforward way to parameterize the probability distribution of a set of categorical variables is by means of their probability table. Probabilities are easy to interpret but have the drawback that sub-models of interest typically involve non-linear constraints on these parameters. For instance, conditional independence relationships can be specified by requiring certain factorizations of the cell probabilities; see Lauritzen (1996) and Cox and Wermuth (1996). For this reason, it is useful to develop alternative parameterizations such that sub-models of interest correspond to linear subspaces of the parameter space of the saturated model. Log-linear parameters are computed as a log-linear expansion of the cell probabilities (see Agresti, 2002) and can be seen as a special case of the wider family of marginal log-linear parameters of Bergsma and Rudas (2002). These parameters are defined by first specifying a suitable sequence of marginal distributions and then by selecting a subset of the log-linear parameters computed from each margin of this sequence. The traditional log-linear parameters are obtained when the specified sequence only includes the "marginal" distribution of all the variables. Otherwise, when all possible marginal distributions are included in this sequence, one obtains the multivariate logistic parameters first introduced by Glonek and McCullagh (1995).

It is well-established that pairwise conditional independence relationships correspond to certain zero entries in the vector of log-linear parameters and, more generally, undirected graphical models for discrete data are sub-families of log-linear models; see Lauritzen (1996, Chap. 4). Several recent papers investigate the use of marginal-log linear models to specify other classes of graphical models as well as models defined by arbitrary collections of conditional independencies; see for instance Evans and Richardson (2011), Forcina et al. (2010) and Rudas et al. (2010).

A special case of interest is the family of graphical models of marginal independence introduced by Cox and Wermuth (1993, 1996) with the name of *covariance graph models*, but later addressed in the literature also as *bidirected graph models* following Richardson (2003). These models have appeared in several applied contexts as described in Drton and Richardson (2008) and references therein. Defining a parameterization for discrete marginal independence models represents a challenging research area both because pairwise independencies do not imply higher order independencies as, for instance, in the Gaussian case, and because a parameterization for these models should also allow the flexible implementation of additional substantive constraints to facilitate the selection of parsimonious models; see Evans and Richardson (2011). In the case of binary data, Drton and Richardson (2008) showed that bidirected graph models can be parameterized by imposing multiplicative constraints

on the mean parameter, which they called the *Möbius parameter*, of the multivariate Bernoulli distribution. Successively, Lupparelli et al. (2009) showed that bidirected graph models for discrete data are a subclass of multivariate logistic models.

In this paper we consider binary data and introduce a novel parameterization based on a log-linear expansion of the mean parameter of the multivariate Bernoulli distribution that we call the *log-mean linear* (LML) parameterization. Similarly to the multivariate logistic parameterization, the computation of LML parameters is based on all marginal distributions, but we remark that they are not marginal log-linear parameters, with the advantage that their computation from cell probabilities is more efficient. The LML parameterization defines a parameter space where the multiplicative constraints in the Möbius parameters of Drton and Richardson (2008) correspond to linear subspaces and, from this perspective, they resemble the connection between log-linear parameters and cell probabilities. It is therefore natural to investigate the family of LML models obtained by imposing linear constraints on the parameter space of the saturated model. We show that marginal independence between variables can be specified by setting certain LML parameters to zero and, more specifically, that graphical models of marginal independence are LML models. Our approach to the problem seems especially appealing because it avoids some disadvantages of both the Möbius parameterization and the multivariate logistic one.

The LML models are code dependent in the sense that they are not invariant with respect to relabelling of the variables. We show that this makes it possible to specify sub-models defined by code-specific independencies, which are independencies in sub-populations of interest. This specific feature of LML models has useful applications. Firstly, it provides an alternative way to specify parsimonious bidirected graph sub-models with the advantage that all the constrains can be easily interpreted as independencies. More specifically, the model is defined by the collection of marginal independencies encoded by a bidirected graph under a given Markov property together with an additional collection of code-specific independencies. A simulation study suggests that the coding of variables can be specified in such a way that statistical procedures for testing code-specific independencies have appealing asymptotic properties. Secondly, the code-specific nature of these models can be exploited to focus on a specific cell of the probability table and the corresponding sub-population. This leads to an innovative family of models, which we call *pivotal code-specific* LML models, that is especially useful when the interest of researchers is focused on a small sub-population specified by stratifying individuals according to some features. The application of LML models is illustrated on two datasets, one of which concerns the use of pivotal code-specific LML models in the field of personalized medicine.

Although this work is devoted to binary variables, we also show that our results provide the building blocks for the extension of LML models to categorical variables with an arbitrary number of levels.

The remainder of the paper is organized as follows. In Section 2 we review the most common parameterizations of the multivariate Bernoulli distribution and their use in graphical modelling with special attention to marginal independence models. In Section 3 we introduce the LML parameterization and the associated class of models. Section 4 shows that LML models can be used to define marginal independence models such as bidirected graph models. Section 5 is devoted to the theory of LML models defined with code-specific independencies. Section 6 gives two different applications of LML models. Section 7 deals with the non-binary case and, finally, Section 8 contains a discussion. Basic lemmas and long proofs are deferred to Appendices A and B, respectively, whereas Appendix C gives an algorithm for maximum likelihood estimation in LML models.

# 2 Preliminaries

In this section we introduce the multivariate Bernoulli distribution and describe its most commonly used parameterizations. Furthermore, we review the theory related to graphical models of marginal independence at the level requested for this paper; we refer to Richardson (2003) and Drton and Richardson (2008) for further details.

## 2.1 Parameterizations for binary data

Given the finite set $V = \{1, \ldots, p\}$, with $|V| = p$, let $X_V = (X_v)_{v \in V}$ be a random vector of binary variables taking values in the set $\mathcal{I}_V = \{0,1\}^p$. We call $\mathcal{I}_V$ a $2^p$-table and its elements $i_V \in \mathcal{I}_V$ the cells of the table. In this way, $X_V$ follows a multivariate Bernoulli distribution with probability table $\pi(i_V)$, $i_V \in \mathcal{I}_V$, which we assume to be strictly positive. From the fact that $\mathcal{I}_V = \{0,1\}^p = \{(1_D, 0_{V \setminus D}) | D \subseteq V\}$, it follows that we can write the probability table as a vector $\pi^V = (\pi_D)_{D \subseteq V}$ with entries $\pi^V_D = P(X_D = 1_D, X_{V \setminus D} = 0_{V \setminus D})$. We refer to $\pi^V$ as to the probability parameter of $X_V$ and recall that it belongs to the $(2^p - 1)$-dimensional simplex, which we write as $\pi^V \in \Pi^V$. For every $U \subseteq V$, with $U \neq \emptyset$, $X_U = (X_v)_{v \in U}$ denotes the marginal Bernoulli random vector taking values in the set $\mathcal{I}_U = \{0,1\}^{|U|}$, and the quantities $\pi^U$ and $\Pi^U$ are defined accordingly. Hereafter, whenever $U = V$ we simplify the notation and omit the superscript, so that $\pi \equiv \pi^V$ and $\Pi \equiv \Pi^V$.

We call $\theta \equiv \theta^V$ a *parameter* of $X_V$ if it is a vector in $R^{2^p}$ that characterizes the

joint probability distribution of $X_V$, and use the convention that the entries of $\theta$ (called interactions) are indexed by the subsets of $V$, i.e., $\theta = (\theta_D)_{D \subseteq V}$. If $\omega \equiv \omega^V$ is an alternative parameter of $X_V$, then a result known as *Möbius inversion* states that

$$\omega_D = \sum_{E \subseteq D} \theta_E, \quad \forall D \subseteq V \quad \Leftrightarrow \quad \theta_D = \sum_{E \subseteq D} (-1)^{|D \setminus E|} \omega_E, \quad \forall D \subseteq V; \tag{1}$$

see, among others, Lauritzen (1996, Appendix A). Let $\mathbb{Z} \equiv \mathbb{Z}^V$ and $\mathbb{M} \equiv \mathbb{M}^V$ be two $(2^p \times 2^p)$ matrices with entries indexed by the subsets of $V \times V$ and given by

$$\mathbb{Z}_{D,H} = 1(D \subseteq H) \quad \text{and} \quad \mathbb{M}_{D,H} = (-1)^{|H \setminus D|} 1(D \subseteq H),$$

respectively, where $1(\cdot)$ denotes the indicator function. Then, the equivalence (1) can be written in matrix form as

$$\omega = \mathbb{Z}^\top \theta \quad \text{iff} \quad \theta = \mathbb{M}^\top \omega, \tag{2}$$

and Möbius inversion follows by noticing that $\mathbb{M} = \mathbb{Z}^{-1}$. In the literature $\mathbb{M}$ is usually called the *Möbius matrix* whereas $\mathbb{Z}$ is the *zeta-matrix*.

In the remaining part of this subsection, we review some well-known alternative parameterizations for the distribution of $X_V$, each defined by a smooth invertible mapping from $\Pi$ onto a smooth $(2^p - 1)$-dimensional manifold of $R^{2^p}$. For simplicity, we denote both the mapping and the alternative parameter it defines by the same (greek) letter. We remark that, if the inverse mapping can be analytically computed, then the likelihood function under multinomial sampling can be written in closed form as a function of the alternative parameter; Poisson sampling can be dealt with by extending the mapping domain to $R_+^{2^p}$.

Multivariate Bernoulli distributions form a regular exponential family with *canonical log-linear parameter* $\lambda \equiv \lambda^V$ computed as

$$\lambda = \mathbb{M}^\top \log \pi. \tag{3}$$

The mapping $\lambda$ defined by (3) can be easily inverted by exploiting Möbius inversion to obtain $\pi = \exp \mathbb{Z}^\top \lambda$, for all $\lambda \in \lambda(\Pi)$. Notice that $\lambda(\Pi)$ has a simple structure, since $(\lambda_D)_{D \neq \emptyset}$ is free to vary in $R^{2^p - 1}$ and $\lambda_\emptyset$ is a smooth function of its elements. The parameterization $\lambda$ captures conditional features of the distribution of $X_V$ and is used to define the class of log-linear models, which is widely used for the analysis of discrete data (see Agresti, 2002) and includes, as a special case, the class of undirected graphical models (see Lauritzen, 1996, Chap. 4).

The *mean parameter* of the multivariate Bernoulli distribution is $\mu = (\mu_D)_{D \subseteq V}$, where $\mu_\emptyset = 1$ (on grounds of convention) and $\mu_D = P(X_D = 1_D)$ otherwise. This was

called the *Möbius parameter* by Drton and Richardson (2008), because one finds

$$\mu = \mathbb{Z}\pi. \tag{4}$$

The linear mapping $\mu$ defined by (4) is trivially Möbius-inverted to obtain $\pi = \mathbb{M}\mu$, for all $\mu \in \mu(\Pi)$. However, it should be said that the structure of $\mu(\Pi)$ is rather involved, and actually well-understood only for small $p$. The parameterization $\mu$ satisfies the *upward compatibility property*, i.e., it is invariant with respect to marginalization. Drton and Richardson (2008) used the mean parameterization in the context of graphical models of marginal independence.

Bergsma and Rudas (2002) developed a wide class of mixed parameterizations, that is parameterizations capturing both marginal and conditional distributional features, named *marginal log-linear parameterizations*. Broadly speaking, any marginal log-linear parameter is obtained by stacking subvectors of log-linear parameters computed in suitable marginal distributions. Thus, this class of parameterizations includes as special, extreme, cases the log-linear parameterization $\lambda$, where a single margin is used, and the *multivariate logistic parameterization* of Glonek and McCullagh (1995), $\eta = (\eta_D)_{D \subseteq V}$, where each $\eta_D$ is computed in the margin $X_D$. The parameterization $\eta$ clearly satisfies the upward compatibility property, and it is typically used for modelling marginal distributions; see McCullagh and Nelder (1989, Chap. 6). More generally, marginal log-linear parameterizations can be useful in several contexts (see Bergsma et al., 2009) and, in particular, they have been recently used for different types of graphical models of marginal independence (Lupparelli et al., 2009; Rudas et al., 2010; Marchetti and Lupparelli, 2011; Evans and Richardson, 2011). A disadvantage of marginal log-linear parameterizations is that their inverse mappings cannot be analytically computed (but for the special case of $\lambda$).

## 2.2 Marginal independence and bidirected graph models

Marginal independence models aim to capture marginal independence relationships between variables. Dealing with marginal independence is especially challenging in the discrete case because pairwise independencies do not imply higher order independencies as, for instance, in the Gaussian case. In this subsection we focus on graphical models of marginal independence and, following Richardson (2003), we use the convention that the independence structure of variables is represented by a bidirected graph. It is also worth recalling that graphical models of marginal independence have been previously discussed by Cox and Wermuth (1993), who adopt a different graphical representation with undirected dashed edges.

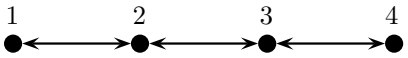



Figure 1: Example of bidirected graph with disconnected sets $\{1,3\}$, $\{1,4\}$, $\{2,4\}$, $\{1,2,4\}$ and $\{1,3,4\}$, encoding the marginal independencies $X_{\{1,2\}} \perp\!\!\!\perp X_4$ and $X_1 \perp\!\!\!\perp X_{\{3,4\}}$ under the connected set Markov property.

A bidirected graph $\mathcal{G} = (V, E)$ is defined by a set $V = \{1, \ldots, p\}$ of nodes and a set $E$ of edges drawn as bidirected; see Richardson (2003). Two nodes $j, k$ are *adjacent* if $jk$ is an edge of $\mathcal{G}$, whereas two edges are adjacent if they have an end-node in common. A *path* from a node $j$ to a node $k$ is a sequence of adjacent edges connecting $j$ and $k$ for which the corresponding sequence of nodes contains no repetitions. For any subset $D \subseteq V$ of nodes, $D$ is said to be *connected* in $\mathcal{G}$ if there is a path between every couple $j, k \in D$, otherwise it is said to be *disconnected* in $\mathcal{G}$. Any disconnected set $D \subseteq V$ can be uniquely partitioned into its (maximal) *connected components* $C_1, \ldots, C_r$ such that $D = C_1 \cup \cdots \cup C_r$, $C_i \cap C_j = \emptyset$ whenever $i \neq j$, $C_i$ is connected for all $i$, and $C_i \cup \{v\}$ is disconnected for all $v \in D \backslash C_i$.

A bidirected graph model is the family of probability distributions for $X_V$ that satisfy a given Markov property with respect to a bidirected graph $\mathcal{G}$. Given three disjoint subsets $A, B, C \subseteq V$, we use the standard notation $X_A \perp\!\!\!\perp X_B \mid X_C$ (Dawid, 1979) to represent the statement "$X_A$ is independent of $X_B$ given $X_C$". The distribution of $X_V$ satisfies the *pairwise Markov property* if, for every missing edge $jk$, it holds that $X_j \perp\!\!\!\perp X_k$. The distribution of $X_V$ satisfies the *connected set Markov property* (Richardson, 2003) if, for every disconnected set $D$, the subvectors corresponding to its connected components $X_{C_1}, \ldots, X_{C_r}$ are mutually independent. For instance, in the bidirected graph of Figure 1, the set $\{1,2,4\}$ is disconnected with connected components $\{1,2\}$ and $\{4\}$. Then, under the connected set Markov property $X_{\{1,2\}} \perp\!\!\!\perp X_4$; implying $X_1 \perp\!\!\!\perp X_4$ and $X_2 \perp\!\!\!\perp X_4$. Notice that the connected set Markov property implies the pairwise Markov property, whereas the converse is not true in general, even for strictly positive distributions. We denote by $\mathbf{B}(\mathcal{G})$ the bidirected graph model for $X_V$ defined by $\mathcal{G}$ under the connected set Markov property.

Parameterizations for the class $\mathbf{B}(\mathcal{G})$ have been studied by Drton and Richardson (2008), where $\mathbf{B}(\mathcal{G})$ is parameterized by imposing non-linear constraints on the Möbius parameterization $\mu$, and by Lupparelli et al. (2009), where $\mathbf{B}(\mathcal{G})$ is parameterized by imposing linear constraints on the multivariate logistic parameterization $\eta$; see also Evans and Richardson (2011). In the next section we introduce a novel parameterization for binary data, which we find useful to introduce a new family of

models, and we compare its features to the parameterizations reviewed above.

# 3 Log-mean linear models

We introduce a new class of models for the multivariate Bernoulli distribution based on the notion of Log-Mean Linear (LML) parameter, denoted by $\gamma = (\gamma_D)_{D \subseteq V}$. Each element $\gamma_D$ of $\gamma$ is a log-linear expansion of a subvector, namely $(\mu_E)_{E \subseteq D}$, of the mean parameter:

$$\gamma_D = \sum_{E \subseteq D} (-1)^{|D \setminus E|} \log(\mu_E), \tag{5}$$

so that in vector form we have

$$\gamma = \mathbb{M}^\top \log \mu. \tag{6}$$

Notice that by replacing $\mu$ with $\pi$ in (6) one obtains the canonical log-linear parameter $\lambda$ in (3).

It is worth describing in detail the elements of $\gamma$ corresponding to sets with low cardinality (its low-order interactions). Firstly, and trivially, $\gamma_\emptyset = \log \mu_\emptyset$ is always equal to 0. Secondly, for every $j \in V$, the main LML effect $\gamma_{\{j\}} = \log \mu_{\{j\}}$ is always negative, because $\mu_{\{j\}}$ is a probability. Then, for every $j, k \in V$, the two-way LML interaction

$$\gamma_{\{j,k\}} = \log \frac{\mu_{\{j,k\}}}{\mu_{\{j\}}\mu_{\{k\}}}$$

coincides with the logarithm of the second order dependence ratio introduced by Ekholm et al. (1995); see also Ekholm et al. (2000) and Darroch and Speed (1983), where dependence ratios were used in models named Lancaster additive. Finally, the three-way LML interaction, for every triple $j, k, z \in V$, is equal to

$$\gamma_{\{j,k,z\}} = \log \frac{\mu_{\{j,k,z\}}\mu_{\{j\}}\mu_{\{k\}}\mu_{\{z\}}}{\mu_{\{j,k\}}\mu_{\{j,z\}}\mu_{\{k,z\}}}$$

and thus differs from the third order dependence ratio; the same is true for each $\gamma_D$ with $|D| \geq 3$. Note that, already from two-way LML interactions, it is apparent that $\gamma$ is not a marginal log-linear parameter of Bergsma and Rudas (2002).

We now formally define the LML parameterization as a mapping from $\Pi$.

**Definition 1** *For a vector $X_V$ of binary variables, the* log-mean linear parameterization $\gamma$ *is defined by the mapping*

$$\gamma = \mathbb{M}^\top \log \mathbb{Z}\pi, \quad \pi \in \Pi. \tag{7}$$

Notice that (7) follows from (6) and (4). Although (7) is of the form of the marginal log-linear parameterizazion of Bergsma and Rudas (2002), in the latter the contrast and marginalization matrices corresponding to $\mathbb{M}$ and $\mathbb{Z}$ in (7) are rectangular matrices of size $t \times 2^p$ with $t \gg 2^p$, so that the inverse transformation is not available in closed form. On the other hand, in our case the inverse transformation can be analytically computed by applying Möbius inversion twice to obtain

$$\pi = \mathbb{M} \exp \mathbb{Z}^\top \gamma, \quad \gamma \in \gamma(\Pi). \tag{8}$$

Clearly, the bijection specified by (7) and (8) is smooth, so that it constitutes a valid reparameterization. Finally, we remark that, as well as the mean and multivariate logistic parameterizations, the LML parameterization satisfies the upward compatibility property.

Given a full rank matrix $\mathbb{H}$ of size $(2^p \times k)$, with $k \leq 2^p$, whose rows are indexed by the subsets of $V$, we define the LML model constrained by $\mathbb{H}$ as follows.

**Definition 2** *For a vector $X_V$ of binary variables and a full rank $(2^p \times k)$ matrix $\mathbb{H}$, the* log-mean linear model $\mathbf{\Gamma}(\mathbb{H})$ *is the family of probability distributions for $X_V$ such that $\mathbb{H}^\top \gamma = 0$.*

It is not difficult to construct a matrix $\mathbb{H}$ such that $\mathbf{\Gamma}(\mathbb{H})$ is empty. However, the family $\mathbf{\Gamma}(\mathbb{H})$ is non-empty if the linear constraints neither involve $\gamma_\emptyset$ nor the main effect $\gamma_{\{j\}}$, for every $j \in V$. More formally, a sufficient condition for $\mathbf{\Gamma}(\mathbb{H})$ to be non-empty is that the rows of $\mathbb{H}$ indexed by $D \subseteq V$ with $|D| \leq 1$ be all equal to 0; see Section 4.1.

**Proposition 1** *Any non-empty log-mean linear model $\mathbf{\Gamma}(\mathbb{H})$ is a curved exponential family of dimension $(2^p - k - 1)$.*

*Proof.* This follows from the mapping defining the parameterization $\gamma$ being smooth, and the matrix $\mathbb{H}$ imposing a $k$-dimensional linear constraint on the parameter $\gamma$. □

Maximum likelihood estimation for LML models under a multinomial or Poisson sampling scheme is a constrained optimization problem, which can be carried out by using standard algorithms. In particular, we adopt an iterative method also used for fitting marginal log-linear models; see Appendix C for details. It should be noted that in our case the algorithm is computationlly more efficient than for marginal log-linear models, because, as remarked above, rectangular matrices of size $t \times 2^p$ with $t \gg 2^p$ are replaced by square matrices of size $2^p \times 2^p$.

The LML parameterization, like the mean parameterization $\mu$, is not symmetric under relabelling of the two states taken by the random variables (Drton and

Richardson, 2008). As a consequence, a key issue concerning the specification of $\mathbb{H}$ is the possible dependence of $\mathbf{\Gamma}(\mathbb{H})$ on variable coding and the forthcoming Sections 4 and 5 show how the matrix $\mathbb{H}$ can be constructed so as to obtain useful models.

# 4 Marginal independence models

In this section we show that the LML parameterization $\gamma$ can be used to encode marginal independencies and, furthermore, that bidirected graph models are LML models. Hence, the LML parameterization can be used in alternative to the approaches developed by Drton and Richardson (2008) and Lupparelli et al. (2009). Our approach is appealing because it combines the advantages of the Möbius parameterization $\mu$ and of the multivariate logistic parameterization $\eta$: the inverse map $\gamma \mapsto \pi$ can be analytically computed, as for $\mu$, and the model is defined by means of linear constraints, as for $\eta$.

## 4.1 LML models and marginal independence

The following theorem shows how suitable linear constraints on the LML parameter correspond to marginal independencies.

**Theorem 2** *For a vector $X_V$ of binary variables with probability parameter $\pi \in \Pi$, let $\mu = \mu(\pi)$ in (4) and $\gamma = \gamma(\pi)$ in (7). Then, for a pair of disjoint, nonempty, proper subsets $A$ and $B$ of $V$, the following conditions are equivalent:*

*(i)* $X_A \perp\!\!\!\perp X_B$;

*(ii)* $\mu_{A' \cup B'} = \mu_{A'} \times \mu_{B'}$ *for every* $A' \subseteq A$ *and* $B' \subseteq B$;

*(iii)* $\gamma_{A' \cup B'} = 0$ *for every* $A' \subseteq A$ *and* $B' \subseteq B$ *such that* $A' \neq \emptyset$ *and* $B' \neq \emptyset$.

*Proof.* See Appendix B.1. □

We remark that equivalence $(i) \Leftrightarrow (ii)$ of Theorem 2 follows immediately from Theorem 1 of Drton and Richardson (2008).

The next result generalizes Theorem 2 to the case of three or more subvectors.

**Corollary 3** *For a sequence $A_1, \ldots, A_r$ of $r \geq 2$ pairwise disjoint, nonempty, subsets of $V$, let $\mathcal{D} = \{D | D \subseteq A_1 \cup \cdots \cup A_r$ with $D \not\subseteq A_i$ for $i = 1, \ldots, r\}$. Then $X_{A_1}, \ldots, X_{A_r}$ are mutually independent if and only if $(\gamma_D)_{D \in \mathcal{D}} = 0$.*

*Proof.* See Appendix B.2. □

A matrix $\mathbb{H}$, such that $\mathbf{\Gamma}(\mathbb{H})$ is the LML model defined by the constraints of Corollary 3, has size $2^p \times |\mathcal{D}|$; the rows of $\mathbb{H}$ are indexed by the subsets of $V$ while we index its columns by the elements of $\mathcal{D}$. The model of mutual independence is then specified by setting $\mathbb{H}_{D,D} = 1$ for every $D \in \mathcal{D}$, and all the remaining entries to zero. In this case, clearly, variable coding is uninfluential. Lupparelli et al. (2009) showed that $X_{A_1}, \ldots, X_{A_r}$ in Corollary 3 are mutually independent if and only if $(\eta_D)_{D\in\mathcal{D}} = 0$. On the other hand, Drton and Richardson (2008) showed that the same independence relationship holds if and only if $\mu_D = \prod_{i=1}^r \mu_{A_i \cap D}$ for all $D \in \mathcal{D}$. Consider for instance two disjoint subsets $A = \{j, k\}$ and $B = \{z\}$. In this case $\mathcal{D} = \{\{j, z\}, \{k, z\}, \{j, k, z\}\}$, so that $X_A \perp\!\!\!\perp X_B$ if and only if $\gamma_{\{j,z\}} = \gamma_{\{k,z\}} = \gamma_{\{j,k,z\}} = 0$. This corresponds to the LML model $\mathbf{\Gamma}(\mathbb{H})$ defined by the $2^p \times 3$ full rank matrix $\mathbb{H}$ whose columns are indexed by $\mathcal{D}$ and with all zero entries but for $\mathbb{H}_{\{j,z\},\{j,z\}} = \mathbb{H}_{\{k,z\},\{k,z\}} = \mathbb{H}_{\{j,k,z\},\{j,k,z\}} = 1$. The same independence model can be defined by either the linear constraints $\eta_{\{j,z\}} = \eta_{\{k,z\}} = \eta_{\{j,k,z\}} = 0$ or the non-linear constraints

$$\mu_{\{j,z\}} = \mu_{\{j\}} \times \mu_{\{z\}}, \quad \mu_{\{k,z\}} = \mu_{\{k\}} \times \mu_{\{z\}} \quad \text{and} \quad \mu_{\{j,k,z\}} = \mu_{\{j,k\}} \times \mu_{\{z\}}.$$

An interesting special case of Corollary 3 is given below.

**Corollary 4** *For a subset $A \subseteq V$ with $|A| > 1$, the variables in $X_A$ are mutually independent if and only if $\gamma_D = 0$ for every $D \subseteq A$ such that $|D| > 1$.*

*Proof.* It is enough to apply Corollary 3 by taking $A = A_1 \cup \cdots \cup A_r$ with $|A_i| = 1$ for every $i = 1, \ldots, r$. □

We stated in Section 3 that $\mathbf{\Gamma}(\mathbb{H})$ is non-empty whenever the rows indexed by $D \subseteq V$ with $|D| \leq 1$ are equal to zero. This fact follows from Corollary 4, because the latter implies that for mutually independent variables $X_1, \ldots, X_p$ the constraint $\mathbb{H}^\top \gamma = 0$ is satisfied.

## 4.2 Bidirected graph models are LML models

It follows from Theorem 2 that the probability distribution of $X_V$ satisfies the pairwise Markov property with respect to a bidirected graph $\mathcal{G} = (V, E)$ if and only if $\gamma_{\{j,k\}} = 0$ whenever the edge $jk$ is missing in $\mathcal{G}$. The following theorem shows that bidirected graph models for binary data are LML models also under the connected set Markov property.

**Theorem 5** *The distribution of a vector of binary variables $X_V$ belongs to the bidirected graph model $\boldsymbol{B}(\mathcal{G})$ if and only if its log-mean linear parameter $\gamma$ is such that $\gamma_D = 0$ for every set $D$ disconnected in $\mathcal{G}$.*

*Proof.* See Appendix B.3. □

For instance, let $\mathcal{G}$ be the graph in Figure 1. Its family of disconnected sets is

$$\mathcal{D} = \{\{1,3\},\{1,4\},\{2,4\},\{1,2,4\},\{1,3,4\}\}.$$

and the bidirected graph model $\mathbf{B}(\mathcal{G})$ is defined by the linear constraints

$$\gamma_{\{1,3\}} = \gamma_{\{1,4\}} = \gamma_{\{2,4\}} = \gamma_{\{1,2,4\}} = \gamma_{\{1,3,4\}} = 0,$$

corresponding to the marginal independencies $X_{\{1,2\}} \perp\!\!\!\perp X_4$ and $X_1 \perp\!\!\!\perp X_{\{3,4\}}$.

# 5 Code-specific independencies and LML models

We have shown in the previous section that useful LML models can be defined by specifying a collection of marginal independencies (like those encoded by a bidirected graph under the connected set Markov property). LML models are code dependent, in general, but variable coding is uninfluential, as far as only marginal independence constraints are specified. In this section we show that interesting LML models can also be defined by specifying a collection of independence relationships on special conditional distributions which depend on variable coding; we call these relationships *code-specific independencies* and the models they define code-specific LML models. We then describe two possible applications of such models: the first application concerns the specification of parsimonious sub-models of bidirected graph models which are easily interpretable because they are fully defined by a collection of independencies, either marginal or code-specific; in the second application the choice of a suitable coding leads to a class of LML models focused on a specific sub-population of interest.

## 5.1 Pivot cell and partial tables

In order to define a coding of $X_V$ it is necessary and sufficient to specify, for every $v \in V$, the level of $X_v$ that takes value 1. This is equivalent to fixing, among the $2^p$ cells of the probability table, the cell to be coded as $1_V$; we call this cell the *pivot cell* and its probability $\pi_V$ the *pivot probability* of the coding.

Let $U \subset V$ and denote by $W$ the complement of $U$ with respect to $V$, that is, $W = V \backslash U$. The probability distribution of $X_U | \{X_W = i_W\}$ where $i_W \in \mathcal{I}_W$ is determined by the *conditional probability table*

$$P(X_U = i_U | X_W = i_W) = \frac{P(X_U = i_U, X_W = i_W)}{P(X_W = i_W)}, \quad i_U \in \mathcal{I}_U. \tag{9}$$

Hence, the probability parameter characterizing the distribution of $X_U | \{X_W = i_W\}$ in (9) is obtained by extracting a subvector of $\pi$ that is then normalized. We note that such a subvector of $\pi$ contains the pivot probability of the coding if and only if $i_W = 1_W$ and in the following we only consider conditional distributions of this kind, that is, of the form $X_U | \{X_W = 1_W\}$. We denote by $\pi^{U|\{X_W=1\}} \in \Pi^U$ the probability parameter of the binary variables $X_U | \{X_W = 1_W\}$, whose entries are

$$\pi_D^{U|\{X_W=1\}} = P(X_D = 1_D, X_{U \backslash D} = 0_{U \backslash D} | X_W = 1_W) = \frac{\pi_{D \cup W}}{\mu_W}, \quad D \subseteq U.$$

Similarly, the mean and LML parameters are $\mu^{\{X_W=1\}} = \mu(\pi^{U|\{X_W=1\}})$ and $\gamma^{\{X_W=1\}} = \gamma(\pi^{U|\{X_W=1\}})$, respectively, where in this case the mappings in (4) and (6) involve the matrices $\mathbb{M}^U$ and $\mathbb{Z}^U$, because $\pi^{U|\{X_W=1\}} \in \Pi^U$. It follows that the theory developed in the previous section can be directly applied to specify independence models for the distribution of $X_U | \{X_W = 1_W\}$. More specifically, an LML model for $X_U | \{X_W = 1_W\}$ is identified by a $2^{|U|} \times k$ full-rank matrix $\mathbb{H}$ via the linear constraint $\mathbb{H}^\top \gamma^{\{X_W=1\}} = 0$.

The main result of this section states that $\gamma^{\{X_W=1\}}$ is a linear transformation of $\gamma$, so that every linear constraint in $\gamma^{\{X_W=1\}}$ is equivalent to a linear constraint in $\gamma$.

**Theorem 6** *Let $\pi \in \Pi$ be the probability parameter of a vector $X_V$ of binary variables. For a nonempty subset $U \subset V$, with $W = V \backslash U$, let $\pi^{U|\{X_W=1\}} \in \Pi^U$ be the probability parameter of $X_U | \{X_W = 1_W\}$. Then, the following linear relationship between $\gamma = \gamma(\pi)$ and $\gamma^{\{X_W=1\}} = \gamma(\pi^{U|\{X_W=1\}})$ holds:*

$$\gamma_D^{\{X_W=1\}} = \sum_{W' \subseteq W} \gamma_{D \cup W'} \tag{10}$$

*for every $D \subseteq U$ with $D \neq \emptyset$.*

*Proof.* See Appendix B.4 □

Theorem 6 says that there exists a $2^p \times 2^{|U|}$ matrix $\mathbb{K} \equiv \mathbb{K}^W$ such that $\gamma^{\{X_W=1\}} = \mathbb{K}^\top \gamma$ and consequently $\mathbb{H}^\top \gamma^{\{X_W=1\}} = 0$ if and only if $(\mathbb{K}\mathbb{H})^\top \gamma = 0$. Denoting by $\mathbb{K}_{\bullet, D}$, for $D \subseteq U$, the column of $\mathbb{K}$ indexed by $D$ (containing $2^p$ entries indexed by the subsets of $V$) the matrix $\mathbb{K}$ can be constructed as follows. First, we set $\mathbb{K}_{\emptyset,\emptyset} = 1$ and all

other entries of $\mathbb{K}_{\bullet,\emptyset}$ equal to zero, so that $\gamma_{\emptyset}^{\{X_W=1\}} = \mathbb{K}_{\bullet,\emptyset}^{\top}\gamma = 0$ as required. Then, for every $D \subseteq U$ with $D \neq \emptyset$, we set $\mathbb{K}_{E,D} = 1$ if $E = D \cup W'$ with $W' \subseteq W$, and $\mathbb{K}_{E,D} = 0$ otherwise. In this way we can write

$$\mathbb{K}_{\bullet,D}^{\top}\gamma = \sum_{W' \subseteq W} \gamma_{D \cup W'} \tag{11}$$

and (10) can be written in matrix form as $\gamma_{D}^{\{X_W=1\}} = \mathbb{K}_{\bullet,D}^{\top}\gamma$.

As a consequence of Theorem 6, it is possible to specify a bidirected graph model for $X_U|\{X_W = 1_W\}$ by means of a linear constraint on the LML parameter of $X_V$. Specifically, we have the following result under the connected set Markov property.

**Corollary 7** *Under the assumptions of Theorem 6, for a bidirected graph $\mathcal{G} = (U, E)$, the probability distribution of $X_U|\{X_W = 1_W\}$ belongs to the bidirected graph model $\boldsymbol{B}(\mathcal{G})$ if and only if any of the following, equivalent, conditions are satisfied:*

*(i)* $\gamma_{D}^{\{X_W=1\}} = 0$ *for every set $D$ disconnected in $\mathcal{G}$;*

*(ii)* $\sum_{W' \subseteq W} \gamma_{D \cup W'} = 0$ *for every set $D$ disconnected in $\mathcal{G}$.*

*Proof.* This follows immediately from Theorems 5 and 6. □

Condition (ii) of Corollary 7 can be written in matrix form as $(\mathbb{K}_{\bullet,D}^{\top}\gamma)_{D \in \mathcal{D}_U} = 0$, where $\mathcal{D}_U$ is the family of disconnected subsets of $U$ (in $\mathcal{G}$).

More generally, we can consider the *collection of code-specific independencies*, defined as the set of all independence relationships of the form $X_A \perp\!\!\!\perp X_B|\{X_C = 1_C\}$, where $A, B, C$ is an arbitrary tern of mutually disjoint, nonempty, subsets of $V$. Then, an immediate consequence of Theorem 6 is that code-specific independencies correspond to linear constraints on $\gamma$ and therefore to LML models for $X_V$.

**Corollary 8** *If $X_V$ is a vector of binary variables, then for every tern $A, B, C$ of mutually disjoint, nonempty, subsets of $V$ the following are equivalent:*

*(i)* $X_A \perp\!\!\!\perp X_B|\{X_C = 1_C\}$;

*(ii)* $\gamma_{A' \cup B'}^{\{X_C=1\}} = 0$ *for every $A' \subseteq A$ and $B' \subseteq B$ such that $A' \neq \emptyset, B' \neq \emptyset$;*

*(iii)* $\sum_{C' \subseteq C} \gamma_{A' \cup B' \cup C'} = 0$ *for every $A' \subseteq A$ and $B' \subseteq B$ such that $A' \neq \emptyset, B' \neq \emptyset$.*

*Proof.* The equivalence (i)⇔(ii) follows from Theorem 2, whereas the equivalence (ii)⇔(iii) holds true by Theorem 6. □

Condition (iii) in Corollary 8 can be written in matrix form as $(\mathbb{K}_{\bullet,A' \cup B'}^{C})^{\top}\gamma = 0$ for every non-empty $A' \subseteq A$ and $B' \subseteq B$. Corollary 8 shows that an LML model can be specified by any set of code-specific independencies and the rest of this section is devoted to two applications of this result.

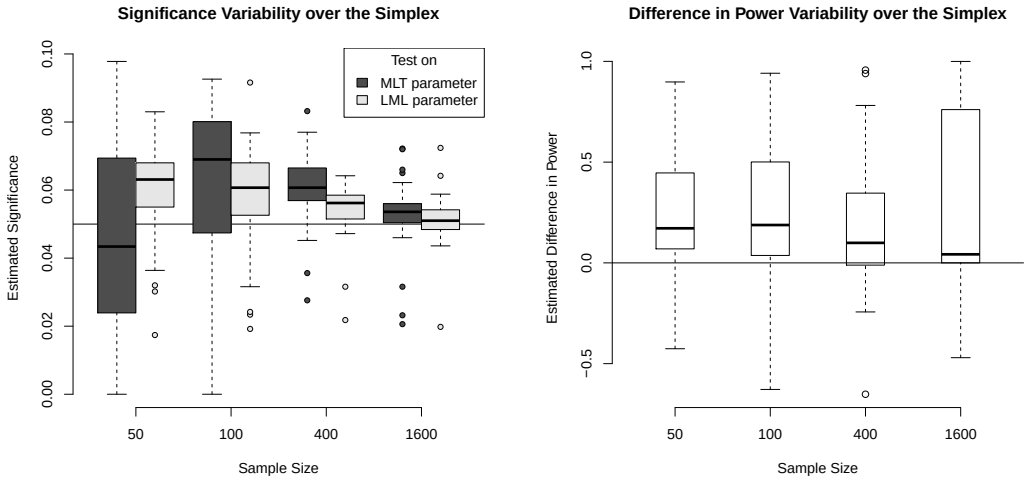
## 5.2 Parsimonious code-specific bidirected graph sub-models

In graphical models the number of parameters depends on sparseness of the graph. However, unlike other families of graphical models such as models for either undirected graphs or directed acyclic graphs, the number of parameters in a bidirected graph model can be relatively large even for sparse graphs; see Richardson (2009) and Evans and Richardson (2011). As a consequence, even though bidirected graph models are recommended when the observed variables are jointly affected by unobserved variables (Richardson, 2003), an analysis restricted to the family of bidirected graphs may result in an overparameterized model. Hence, a convenient parameterization of bidirected graph models should allow the flexible implementation of marginal independence constraints together with additional substantive constraints leading to parsimonious sub-models.

Parameterizations based on marginal log-linear parameters, such as the multivariate logistic parameterization, can be used to specify parsimonious sub-models by setting higher order interactions to zero; see Lupparelli et al. (2009) and Evans and Richardson (2011). In this way, parsimonious modelling can be achieved, but the interpretation of the constraints not directly associated with marginal independence is difficult. On the other hand, not being able to specify parsimonious sub-models is perhaps the main drawback of the Möbius parameterization of Drton and Richardson (2008), which is therefore not flexible enough for this task.

The LML parameterization has the immediate advantage that there are two different ways in which it can be used to specify bidirected graph sub-models. On the one hand, it is still possible to set higher order interactions to zero, but the interpretation of such constraints remains difficult. On the other hand, it is possible to include additional constraints in the form of code-specific independencies. This is appealing because code-specific independencies have a straightforward interpretation, and furthermore the resulting model is an independence model in the sense that it is fully defined by independence relationships: (i) the collection of marginal independencies encoded by the bidirected graph under a given Markov property; (ii) a collection of code-specific independencies.

One should observe that the coding is arbitrary, whenever all cells are on equal footing, and distinct codings will result in distinct models (based on distinct sets of code-specific independencies). Since this flexibility originates an indeterminacy in the analysis, we find it appropriate to define a criterion for choosing the coding of the variables when all cells are on equal footing. Specifically, we propose the *maximal count coding*, which consists in setting as the pivot cell the cell of the table with the largest count. Compared to alternative codings, we expect to test code-specific




(a) (b)

Figure 2: Simulation results: (a) distribution of the estimated significance level; (b) distribution of the estimated difference in power.

independencies in partial tables with more observations, which results in increased efficiency. This aspect represents a further advantage of using the LML parameterization for parsimonious bidirected graph modelling, compared to any marginal log-linear parameterization and, in particular, to the multivariate logistic one.

This feature is illustrated through a series of simulations aimed to compare the performance of the multivariate logistic and LML parameterizations to achieve parsimonious models, the former by setting zero higher order interactions and the latter by specifying code-specific independencies. In particular, given a set of four binary variables indexed by $V = \{A, B, C, D\}$, we compared the performance in testing the hypothesis $\eta_V = 0$ in the multivariate logistic parameterization with the performance in testing the hypothesis $X_C \perp\!\!\!\perp X_D | \{X_A = 1, X_B = 1\}$ in the LML parameterization under the maximal count coding. Both hypotheses correspond to a single linear constraint and both are implied by the conditional independence $X_C \perp\!\!\!\perp X_D | \{X_A, X_B\}$. We remark that the choice of the coding is uninfluential for testing zero multivariate logistic interactions.

We generated a sequence of probability vectors $\pi_i$, $i = 1, \ldots, 40$, uniformly over the simplex, and satisfying the constraint $X_C \perp\!\!\!\perp X_D | \{X_A, X_B\}$. Then, for each probability parameter $\pi_i$, we sampled 5.000 multinomial random vectors $n_j$, $j = 1, \ldots, 5.000$, of size $N$. Finally, for each random sample $n_j$, we tested the two above hypotheses at the $\alpha = 0.05$ nominal significance level, using the deviance of the corresponding model and the chi-squared distribution with one degree of freedom.

For each probability parameter $\pi_i$, we estimated the actual (finite sample) significance level $\hat{\alpha}_i^{\eta}$ and $\hat{\alpha}_i^{\gamma}$ of the two tests through the percentage of rejected models in the 5.000 random samples, thus obtaining two distributions of estimates over the conditional independence model. The whole procedure was repeated for $N = 50, 100, 400, 1600$ and, for every sample size, Figure 2(a) compares the two box plots of the estimated significance levels. The plot clearly shows a lower variability in the estimated significance level and a faster convergence to the nominal value (0.05) for the test on the LML parameter.

We also compared the two tests in terms of power. To this aim, we replicated the same series of simulations using a sequence of unconstrained probability vectors $\pi_i$, $i = 1, \ldots, 40$, generated uniformly over the simplex. Then, for every $i = 1, \ldots, 40$ we estimated the type II error of the two tests, $\hat{\beta}_i^{\eta}$ and $\hat{\beta}_i^{\gamma}$, through the percentage of accepted models in the 5.000 random samples and, from these, the difference in power $\hat{\delta}_i = \hat{\beta}_i^{\eta} - \hat{\beta}_i^{\gamma}$ . Figure 2(b) gives, for every sample size $N = 50, 100, 400, 1600$, the box plot of the quantities $\hat{\delta}_i$, $i = 1, \ldots, 40$. The plot shows a clear gain in power for the test based on the LML parameter with respect to the test based on the multivariate logistic parameter.

Our simulation results suggest that the property of LML models of being code-dependent can be exploited to improve the efficiency of model selection. Therefore, LML models can be used to search parsimonious bidirected graph models in a lower dimensional space defined by constraints which are both interpretable and can be tested with powerful inferential procedures. In Section 6.1 we present an effective application of models defined by both marginal and code-specific independencies, under maximal count coding.

## 5.3 Pivotal code-specific LML models

In this subsection we describe an application of LML models that fully exploits the advantages deriving from their code dependent nature. In statistical applications it is often of interest to investigate the behaviour of small sub-populations. For instance, one of the challenges of modern medicine is *personalized medicine*, in which therapies are tailored to the exact biological state of an individual; see Nicholson (2006). From a statistical viewpoint, this involves the stratification of individuals according to some features $X_V$ so as to identify a sub-population of interest, which in our perspective corresponds to a single cell of the cross-classified table. Researchers are then interested in statistical models that may highlight relevant features of such a sub-population, and in this context LML models provide a powerful tool of analysis. More specifically, if the variables are suitably coded, LML models allow one to focus on the probability

that a unit of the population belongs to the sub-population of interest and to specify a rich class of factorizations for such probability.

The first step of the analysis requires that the levels of $X_V$ be coded so that the sub-population of interest corresponds to the pivot cell of the table and, accordingly, the probability of belonging to such sub-population is the pivot probability of the coding. Indeed, the key feature of LML models is that the cells of the table are not on an equal footing. The pivot cell has a central role and every code-specific independence implies a factorization of either the pivot probability or of a probability in a marginal table corresponding to a cell that contains the sub-population of interest. For instance, the code-specific independence $X_A \perp\!\!\!\perp X_B | \{X_C = 1_C\}$, with possibly $C = \emptyset$, implies

$$P(X_{A\cup B\cup C} = 1_{A\cup B\cup C}) = \frac{P(X_{A\cup C} = 1_{A\cup C})P(X_{B\cup C} = 1_{B\cup C})}{P(X_C = 1_C)}, \tag{12}$$

where we use the convention that $P(X_C = 1_C) = 1$ for $C = \emptyset$. If $A \cup B \cup C = V$, then $P(X_{A\cup B\cup C} = 1_{A\cup B\cup C}) = \pi_V = \mu_V$ is the pivot probability, but also the case $A \cup B \cup C \subset V$ is of interest, as clarified in the application of Section 6.2.

We remark that information on the above factorization of the pivot probability could also be obtained by fitting a graphical model to the data, because the conditional independence relation $X_A \perp\!\!\!\perp X_B | X_C$ implies $X_A \perp\!\!\!\perp X_B | \{X_C = 1_C\}$ and, in turn, the factorization in (12). However, the converse implication does not hold and therefore LML models can encode factorizations of the pivot probability that cannot be encoded by any graphical model. Furthermore, consider the case where different factorizations of the pivot probability are suggested by different classes of graphical models. In this case such factorizations can be encoded into a single LML model, which can then be used to investigate to what extent they may simultaneously hold.

# 6 Applications

In this section we describe two applications of LML models. The first one refers to Section 5.2 and shows how LML models can be used to identify interpretable parsimonious bidirected graph sub-models. The second concerns the selection of a pivotal code-specific LML model as described in Section 5.3.

## 6.1 Coppen data

Table 1 shows data from Coppen (1966) for a set of four binary variables concerning symptoms of 362 psychiatric patients. The symptoms are: $X_1 \equiv$ *stability* (0 = extroverted, 1 = introverted); $X_2 \equiv$ *validity* (0 = energetic, 1 = psychasthenic);

Table 1: Data from Coppen (1966) on symptoms of psychiatric patients.

| | | $X_3$ | 0 | | 1 | |
|---|---|---|---|---|---|---|
| $X_2$ | $X_4$ | $X_1$ | 0 | 1 | 0 | 1 |
| 1 | 0 | | 30 | 22 | 22 | 8 |
| | 1 | | 32 | 16 | 27 | 12 |
| 0 | 0 | | 15 | 23 | 25 | 14 |
| | 1 | | 9 | 14 | 46 | 47 |

$X_3 \equiv$ *acute depression* (0 = yes, 1 = no); $X_4 \equiv$ *solidity* (0 = hysteric, 1 = rigid). The maximal cell coding has been applied.

We analyse these data because they have been analysed twice, in the graphical modelling literature, with two different classes of models: Wermuth (1976) analysed them by means of conditional independence models, and found the model in Figure 3(a). More recently, Lupparelli et al. (2008) analysed the same data using marginal independence models, and obtained the model in Figure 3(b). Both the undirected 4-chain in Figure 3(a) and the bidirected 4-chain in Figure 3(b) achieve a good fit, but they encode different independencies.

Since there is no probability distribution for binary variables that simultaneously satisfies all and only the (conditional) independencies encoded by the two graphs in Figure 3, it is not possible to reconcile the different results of the two analyses without adding further independence relationships. However, such an addition would make little sense here, because it would lead to a model with poor fit. LML models allow us to follow a different approach: by including in a single model all the marginal independencies encoded by the bidirected graph and the code-specific independencies


Figure 3: Independence models for Coppen data: (a) the 4-chain undirected graph model $X_1 \perp\!\!\!\perp X_{\{3,4\}}|X_2$, $X_4 \perp\!\!\!\perp X_{\{1,2\}}|X_3$; $\chi^2_{(8)} = 13.9$ ($p$-value = 0.09), $BIC = -33.26$; (b) the 4-chain bidirected graph model $X_1 \perp\!\!\!\perp X_{\{3,4\}}$, $X_4 \perp\!\!\!\perp X_{\{1,2\}}$; $\chi^2_{(5)} = 8.6$ ($p$-value = 0.13), $BIC = -20.85$. BIC values with saturated model as baseline ($BIC_{SAT} = 0$).

suggested by the undirected graph and the chosen variable coding, we partially recover the conditional independencies encoded by the undirected graph in a bidirected graph sub-model.

By Theorem 5 the bidirected 4-chain represents an LML model, namely, the model given by the constraints $\gamma_{13} = \gamma_{14} = \gamma_{24} = \gamma_{134} = \gamma_{124} = 0$. We can simplify this model by exploiting the conditional independencies encoded in the undirected 4-chain to include further independence constraints in some partial tables:

$$X_1 \perp\!\!\!\perp X_{\{3,4\}} | \{X_2 = 1\} \quad \text{and} \quad X_{\{1,2\}} \perp\!\!\!\perp X_4 | \{X_3 = 1\}.$$

As a consequence of Theorem 6, these additional independencies correspond to six sum-to-zero constraints on the LML parameter. However, because of redundancies with the marginal independence model, these independencies can be included by adding only three linear constraints:

$$\gamma_{13} + \gamma_{123} = 0, \quad \gamma_{134} + \gamma_{1234} = 0 \quad \text{and} \quad \gamma_{24} + \gamma_{234} = 0.$$

Together with the original constraints, these amount to assuming $\gamma_{123} = \gamma_{1234} = \gamma_{234} = 0$. The resulting LML model fits the data with $\chi^2_{(8)} = 11.45$ ($p$-value $= 0.18$) and $BIC = -35.68$, thus improving on both models in Figure 3.

## 6.2 HIV data

Table 2 contains data from Guaraldi et al. (2011) for four binary variables observed on 2 860 HIV-positive patients: $H \equiv$ *hypertension* (0 = yes, 1 = no); $E \equiv$ *cardiovascular event* (0 = no, 1 = yes); $A \equiv$ *age* (0 = greater than or equal to 45, 1 = less than 45); $G \equiv$ *gender* (0 = female, 1 = male). In this application, researchers are interested in the sub-population of young males without hypertension having had some kind of cardiovascular event (such as a stroke or a heart attack). For this reason we have coded the variables in such a way that this sub-population corresponds to the pivot cell $(1, 1, 1, 1)$ of Table 2, which only contains 6 observations.

Cardiovascular events among HIV-positive young men with no hypertension are much less rare than cardiovascular events in the corresponding HIV-negative sub-population. Consequently, researchers are interested in the effect of specific HIV-related risk factors, such as CD4 counts, drugs and length of HIV infection, on the response probability $\pi_V$. This can be assessed by testing the hypothesis that the logistic regression coefficient corresponding to a given risk factor is equal to zero, possibly in the presence of other covariates. However, as routinely occurs in personalized medicine, the sub-population of interest is very small and it is well-known

Table 2: Data from Guaraldi et al. (2011) concerning HIV-positive patients. The pivot cell is in bold.

| | | $A$ | 0 | | 1 | |
|---|---|---|---|---|---|---|
| $H$ | $E$ | $G$ | 0 | 1 | 0 | 1 |
| 0 | 0 | | 93 | 320 | 33 | 120 |
| | 1 | | 4 | 51 | 3 | 10 |
| 1 | 0 | | 296 | 614 | 626 | 665 |
| | 1 | | 3 | 13 | 3 | **6** |

that the sample size required to achieve an adequate power in such a test increases as $\pi_V$ gets closer to zero; see Agresti (2002, Section 6.5.2) and Whittemore (1981). More concretely, Peduzzi et al. (1996) provided guidelines on the minimum sample size required in terms of *events per variable*, and suggested that in a logistic regression analysis there should be at least 10 events for every covariate. According to this rule, the 6 events observed in the pivot cell of Table 2 are not sufficient to reliably identify the effect of any risk factor, and it becomes of interest to assess whether it makes sense to focus on cells of marginal tables with higher number of observations. For this reason, we use LML models to investigate the independence structure of the sub-population corresponding to the pivot cell and find simplifying factorizations of the pivot probability.

We considered the distribution of $X_U|\{X_W = 1\}$ for every $U$ indexing a subvector of $X_V = (H, E, A, G)$ such that $W = V \backslash U$, with $|W| = 1, 2$, and for each such distribution we checked for marginal independencies, that is, for code-specific independencies of $X_V$. We obtained in this way the code-specific independencies listed in the third column of Table 3. Then, we fitted LML models starting from the saturated model and introducing each code-specific independence of Table 3, from top to bottom, in a stepwise manner, rejecting the code-specific independencies leading to a small $p$-value/increase in BIC. The last three columns of Table 3 report the deviance, the degrees of freedom and the BIC index of each model $\mathbf{\Gamma}(\mathbb{H})$ taken into consideration. An adequate fit with $\chi^2_{(4)} = 7.71$ ($p$-value $= 0.10$) is achieved by the LML model that jointly satisfies the code-specific independencies

$$H \perp\!\!\!\perp (A, G)|\{E = 1\} \quad \text{and} \quad E \perp\!\!\!\perp G|\{A = 1\}.$$

Hence, according to the selected LML model, the joint probability of the pivot cell can be factorized as

$$\pi_{HEAG} \;=\; \frac{\pi_{HE++} \times \pi_{+EAG}}{\pi_{+E++}} \tag{13}$$

Table 3: Marginal independence models for the partial distributions associated to $X_U|\{X_W = 1\}$ and stepwise search of an overall LMLmodel for the full vector $X_V$. The BIC values are computed with the saturated model as baseline ($BIC_{SAT} = 0$). The term *present* means that the code-specific constraint is already contained in the selected model.

| $X_U$ | $X_W$ | Code-specific indep. | Deviance | DF | $p$-value | $BIC$ |
|---|---|---|---|---|---|---|
| $(H, E, A)$ | $G$ | – | | | | |
| $(H, E, G)$ | $A$ | $E \perp\!\!\!\perp G \vert \{A = 1\}$ | 3.09 | 1 | 0.08 | -4.87 |
| $(H, A, G)$ | $E$ | $H \perp\!\!\!\perp (A, G) \vert \{E = 1\}$ | 7.71 | 4 | 0.10 | -24.13 |
| $(E, A, G)$ | $H$ | $E \perp\!\!\!\perp (A, G) \vert \{H = 1\}$ | 114.82 | 7 | 0.00 | 59.11 |
| $(H, E)$ | $(A, G)$ | – | | | | |
| $(H, A)$ | $(E, G)$ | $H \perp\!\!\!\perp A \vert \{E = 1, G = 1\}$ | *present* | | | |
| $(H, G)$ | $(E, A)$ | $H \perp\!\!\!\perp G \vert \{E = 1, A = 1\}$ | *present* | | | |
| $(E, A)$ | $(H, G)$ | $E \perp\!\!\!\perp A \vert \{H = 1, G = 1\}$ | 65.09 | 5 | 0.00 | 25.29 |
| $(E, G)$ | $(H, A)$ | $E \perp\!\!\!\perp G \vert \{H = 1, A = 1\}$ | 50.43 | 5 | 0.00 | 10.64 |
| $(A, G)$ | $(H, E)$ | $A \perp\!\!\!\perp G \vert \{H = 1, E = 1\}$ | 48.75 | 5 | 0.00 | 8.96 |

and, moreover, the pivot probability induced on the marginal table of $(E, A, G)$ can be factorized as

$$\pi_{+EAG} = \frac{\pi_{+EA+} \times \pi_{++AG}}{\pi_{++A+}}. \tag{14}$$

Notice that, for the sake of concreteness, here we slightly changed our indexing notation, so that for instance $\pi_{HEAG} = P(H = 1, E = 1, A = 1, G = 1)$ and $\pi_{HE++} = P(H = 1, E = 1)$.

Since the factorizations (13) and (14) hold simultaneously, we can replace $\pi_{+EAG}$ in (13) with (14) and show that the selected LML model is defined by the factorization

$$\pi_{HEAG} \quad = \quad \frac{\pi_{HE++} \times \pi_{+EA+} \times \pi_{++AG}}{\pi_{+E++} \times \pi_{++A+}} \tag{15}$$

corresponding to the log linear expansion of $\mu_V = \pi_V$

$$\log \pi_{HEAG} = \gamma_\emptyset + \gamma_H + \gamma_E + \gamma_A + \gamma_G + \gamma_{HE} + \gamma_{EA} + \gamma_{AG},$$

where the missing interactions are due to the fact that the factorization (13) is encoded by the linear constraints $\gamma_{HAG} + \gamma_{HEAG} = 0$, $\gamma_{HA} + \gamma_{HEA} = 0$ and $\gamma_{HG} + \gamma_{HEG} = 0$, whereas the factorization (14) is encoded by the linear constraint $\gamma_{EG} + \gamma_{EAG} = 0$.

It follows from our analysis that the pivot probability can be written as a function of probabilities corresponding to sub-populations that are larger than the sub-population of interest. More specifically, the cells corresponding to $\pi_{HE++}$, $\pi_{+EA+}$

and $\pi_{++AG}$ contain 25, 22 and 801 events, respectively, and it can thus be helpful to investigate the role of risk factors in these sub-populations.

We remark that the amount of simplification of the pivot probability achieved by (15) cannot be achieved by applying ordinary graphical models. Indeed, we analysed the data with bidirected graph models, but only the saturated model provided a good fit. We also analysed the data with undirected graph models, and the only model with a good fit was the model $E \perp\!\!\!\perp A|(H, G)$, with $\chi^2_{(4)} = 8.45$ ($p$-value $= 0.08$). The latter model identifies the code-specific independence $E \perp\!\!\!\perp A|\{H = 1, G = 1\}$, which is also identified by our procedure, although it is not included in the selected model.

# 7 Extension to non-binary categorical data

This paper is devoted to the theory of LML models for binary variables, nevertheless it is useful to discuss briefly the extension of these models to the case where the variables have an arbitrary number of levels. The main point is that LML models are intrinsically binary in the sense that their generalization can be conceptually obtained by means of a sequence of dichotomizations of the non-binary variables and then by iteratively applying the procedures developed for the binary case.

Assume that $X_V = (X_v)_{v\in V}$ is a discrete random vector with $X_v$ taking values in $\{0, 1, \ldots, d_v\}$ so that the state space of $X_V$ is $\mathcal{I}_V = \times_{v\in V} \{0, 1, \ldots, d_v\}$. For every $v \in V$ and $i \equiv i_V \in \mathcal{I}_V$ we introduce the Bernoulli random variable $X_v^{(i)}$ defined as

$$X_v^{(i)} = \begin{cases} 1 & \text{if } X_v = i_v \\ 0 & \text{otherwise} \end{cases}$$

where, for $D \subseteq V$, $i_D$ is the subset of the levels of $i$ taken by the variables in $X_D$ and, accordingly, $i_v$ is the level of $X_v$. In this way, for every $i \in \mathcal{I}_V$, the random vector $X_V^{(i)}$ follows a multivariate Bernoulli distribution with probability parameter $\pi^{(i)} = (\pi_D^{(i)})_{D\subseteq V}$ and mean parameter $\mu^{(i)} = \mathbb{Z}\pi^{(i)}$. We remark that here $\pi_D^{(i)} = P(X_D^{(i)} = 1_D, X_{V\setminus D}^{(i)} = 0_{V\setminus D})$ and, since $X_D^{(i)} = 1_D$ if and only if $X_D = i_D$, then the probability table of $X_V$ can be written as $\{\pi_V^{(i)}\}_{i\in\mathcal{I}_V}$.

Let $\{1, \ldots, d_v\}$ be the state space of $X_v$ with the level "0" removed. Drton (2009, eqn. 3.3) considered the restricted state space $\mathcal{J}_D = \times_{v\in D} \{1, \ldots, d_v\}$ for $D \subseteq V$ and defined the *saturated Möbius parameters* as the collection of marginal probabilities given by

$$P(X_D = j_D) \quad \text{for every } j_D \in \mathcal{J}_D \text{ and } D \subseteq V \text{ with } D \neq \emptyset.$$

Drton (2009, Lemma 7) showed that there exists a bijective linear map from the cell probabilities to saturated Möbius parameters and provided a closed form expression

for the inverse map. Furthermore, Drton (2009, Theorem 8) generalized to the non-binary case the connection between bidirected graph models and factorization of saturated Möbius parameters. It is straightforward to see that $\mu_D^{(j)} = P(X_D^{(j)} = 1_D) = P(X_D = j_D)$ and therefore, in our notation, the saturated Möbius parameters can be written as the collection of vectors

$$\mu^{(j)} \quad \text{for every } j \equiv j_V \in \mathcal{J}_V. \tag{16}$$

The representation in (16) is redundant because for every $j, j' \in \mathcal{J}_V$ such that $j_D = j'_D$ it holds that $\mu_D^{(j)} = \mu_D^{(j')}$. However, it is very useful because it writes the saturated Möbius parameters as a collection of Möbius parameters for binary variables. We can thus parameterize the distribution of $X_V$ by means of the collection of LML parameters defined as

$$\gamma^{(j)} = \mathbb{M}^\top \log \mu^{(j)} \quad \text{for every } j \in \mathcal{J}_V \tag{17}$$

and, as a consequence, several properties of the LML parameters for the non-binary case follow immediately form the iterative application, for every $j \in \mathcal{J}_V$ of the corresponding properties for the binary case.

For instance, let $A$ and $B$ be two disjoint, proper, subsets of $V$. Then, $X_A \perp\!\!\!\perp X_B$ if and only if $X_A^{(j)} \perp\!\!\!\perp X_B^{(j)}$ for every $j \in \mathcal{J}_V$; see also Drton (2009, Theorem 8). Hence, by iteratively applying Theorem 2 for all $j \in \mathcal{J}_V$ one obtains that $X_A \perp\!\!\!\perp X_B$ if and only if $\gamma^{(j)}_{A' \cup B'} = 0$ for every $A' \subseteq A$ and $B' \subseteq B$ such that $A' \neq \emptyset$ and $B' \neq \emptyset$ and for every $j \in \mathcal{J}_V$.

In a similar way, it is possible to generalize the definition of code-specific independencies by iteratively applying Theorem 6. However, a formal generalization of this idea is less straightforward and it is deferred to a future paper. Here, it is worth remarking that a vector $\gamma^{(j)}$ is directly associated with the cell $j \in \mathcal{J}_V$ thereby making it possible to focus on the factorization of the probability $\pi_V^{(j)}$. In other words, every cell $j \in \mathcal{J}_V$ is a pivot cell of the table. The LML parameters satisfy the upward compatibility property also in the non-binary case but, interestingly, in this case they also satisfy the additional property that they remain unchanged if some levels of the variables are merged to the level "0".

# 8 Discussion

We have introduced the LML parameterization for binary data and described two areas of application of the family of models defined by imposing linear constraints on such parameters. We deem that the full potential of these models is shown in

applications where the main interest is for a sub-population corresponding to a given cell of the table as illustrated in the analysis of the HIV data in Section 6.2. The generalization of this kind of applications to non-binary categorical variables is promising because it allows one to focus simultaneously on more than one pivot cell.

The second area of application concerns bidirected graph models where LML models are shown to provide a worthwhile alternative to marginal log-linear models. The advantages concern the efficiency of the algorithm for the computation of MLEs but, more importantly, the possibility to identify interpretable parsimonious submodels by searching for code-specific independencies. In this kind of application, variable coding is arbitrary and distinct codings will typically result in models defined by different sets of code-specific independencies. However, it is not uncommon for a model selection procedure to produce results that, to some extent, depend on the way it is implemented, and we deem that the interpretability of the selected model and efficiency of testing procedures are of great importance and worth some sacrifice. Furthermore, when the sample size is small (compared to the dimension of the table) and therefore the distribution of the test statistics is not well-approximated by its asymptotic version, then the coding can be driven by cell counts in such a way that code-specific independencies are searched in partial tables with larger cell counts thereby improving the asymptotic properties of statistical procedures.

In the recent statistical literature, a certain degree of attention has been posed on models encoding conditional independencies in partial tables. Højsgaard (2004) introduced the family of *context specific interaction models* which generalize the theory of log-linear models for contingency table so as to make it possible to search for pairwise conditional independencies in partial tables. Context specific interaction models are not code-specific but they are not suited to deal with constraints defined on marginal distributions. Huang and Frey (2011) developed the theory of *Cumulative Distribution Networks* (CDNs) which, for certain graph structures, encode the same marginal independences as the bidirected graph with the same connectivity, together with some additional conditional independence constraints. In the case of CDNs defined over ordered discrete variables, such additional conditional independencies include also the so called *min-independence* that is a special case of code-specific independence. Unlike CDNs, in LML models the collection of code-specific independencies specified by the model is arbitrary and not constrained by the graph structure. Nevertheless, it would be of interest to investigate whether in the binary case CDNs can be defined as a subclass of LML models.

The applications of Section 6 implement naïve model selection procedures, but the general issue of developing model search strategies for the exploration of independence

structures in partial tables is still open, and would be crucial to deal with large tables.

## Acknowledgments

We gratefully acknowledge useful discussions with David R. Cox, Mathias Drton, Antonio Forcina, Feifang Hu, Giovanni M. Marchetti, and Nanny Wermuth.

## References

Agresti, A. (2002). *Categorical data analysis* (Second ed.). Hoboken, NJ: Wiley.

Aitchison, J. and S. D. Silvey (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics 29*(3), 813–828.

Bergsma, W., M. Croon, and J. Hagenaars (2009). *Marginal models for dependent, clustered, and longitudinal categorical data*. London, UK: Springer.

Bergsma, W. P. (1997). *Marginal models for categorical data*. Ph.d thesis, Tilburg University, Tilburg, NL.

Bergsma, W. P. and T. Rudas (2002). Marginal log-linear models for categorical data. *Annals of Statistics 30*(1), 140–159.

Bertsekas, D. P. (1982). *Constrained optimization and Lagrange multiplier methods*. New York: Academic Press.

Coppen, A. (1966). The Mark-Nyman temperament scale: an English translation. *Brit. J. Med. Psychol. 39*(1), 55–59.

Cox, D. R. and N. Wermuth (1993). Linear dependencies represented by chain graphs. *Statistical Science 8*(3), 204–218.

Cox, D. R. and N. Wermuth (1996). *Multivariate dependencies. Models, analysis and interpretation*. London: Chapman and Hall.

Darroch, J. N. and T. P. Speed (1983). Additive and multiplicative models and interactions. *Annals of Statistics 11*(3), 724–738.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B (Methodological) 41*(1), 1–31.

Drton, M. (2009). Discrete chain graph models. *Bernoulli 15*(3), 736–753.

Drton, M. and T. Richardson (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(2), 287–309.

Ekholm, A., J. W. McDonald, and P. W. F. Smith (2000). Association models for a multivariate binary response. *Biometrics 56*(3), 712–718.

Ekholm, A., P. W. F. Smith, and J. W. McDonald (1995). Marginal regression analysis of a multivariate binary response. *Biometrika 82*(4), 847–854.

Evans, R. J. and A. Forcina (2011). Two algorithms for fitting constrained marginal models. Technical report, arXiv:1110.2894v1[stat.CO].

Evans, R. J. and T. S. Richardson (2011). Marginal log-linear parameters for graphical Markov models. Technical report, arXiv:1105.6075v1[stat.ME].

Forcina, A., M. Lupparelli, and G. M. Marchetti (2010). Marginal parameterizations of discrete models defined by a set of conditional independencies. *Journal of Multivariate Analysis 101*(10), 2519–2527.

Glonek, G. J. N. and P. McCullagh (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B (Methodological) 57*(3), 533–546.

Guaraldi, G., G. Orlando, S. Zona, M. Menozzi, F. Carli, E. Garlassi, A. Berti, E. Rossi, A. Roverato, and F. Palella (2011). Premature age-related comorbidities among hiv-infected persons compared with the general population. *Clinical Infectious Diseases 53*(11), 1120–1126.

Højsgaard, S. (2004). Statistical inference in context specific interaction models for contingency tables. *Scandinavian journal of statistics 31*(1), 143–158.

Huang, J. and B. Frey (2011). Cumulative distribution networks and the derivative-sum-product algorithm: Models and inference for cumulative distribution functions on graphs. *The Journal of Machine Learning Research 12*, 301–348.

Lang, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics 24*(2), 726–752.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, UK: Clarendon Press.

Lupparelli, M., G. M. Marchetti, and W. P. Bergsma (2008). Parameterizations and fitting of bi-directed graph models to categorical data. Technical report, arXiv:0801.1440v1.

Lupparelli, M., G. M. Marchetti, and W. P. Bergsma (2009). Parameterizations and fitting of bi-directed graph models to categorical data. *Scandinavian Journal of Statistics 36*(3), 559–576.

Marchetti, G. M. and M. Lupparelli (2011). Chain graph models of multivariate regression type for categorical data. *Bernoulli 17*(3), 827–844.

McCullagh, P. and A. Nelder, J. (1989). *Generalized linear models* (Second ed.). London, UK: Chapman and Hall.

Nicholson, J. (2006). Global systems biology, personalized medicine and molecular epidemiology. *Molecular Systems Biology 2*(52), 1–6.

Peduzzi, P., J. Concato, E. Kemper, T. Holford, and A. Feinstein (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology 49*(12), 1373–1379.

Richardson, T. (2009). A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 462–470. AUAI Press, Arlington (VA).

Richardson, T. S. (2003). Markov property for acyclic directed mixed graphs. *Scandinavian Journal of Statistics 30*(1), 145–157.

Rudas, T., W. Bergsma, and R. Nemeth (2010). Marginal log-linear parameterization of conditional independence models. *Biometrika 97*(4), 1006–1012.

Wermuth, N. (1976). Model search among multiplicative models. *Biometrics 32*(2), 253–263.

Whittemore, A. (1981). Sample size for logistic regression with small response probability. *Journal of the American Statistical Association 76*(373), 27–32.

# A Basic lemmas

**Lemma 1** *For any set $D \neq \emptyset$ it holds that $\sum_{E \subseteq D} (-1)^{|E|} = \sum_{E \subseteq D} (-1)^{|D \setminus E|} = 0$.*

*Proof.* It is well-known, and can be proven by induction, that any non-empty set $D$ has the same number of even and odd subsets. □

**Lemma 2** *Let $g(\cdot)$ be a real-valued function defined on the subsets of a set $D$. If two nonempty, disjoint, proper subsets $A$ and $B$ of $D$ exist, such that $A \cup B = D$ and $g(E) = g(E \cap A) + g(E \cap B)$ for every $E \subseteq D$, then $\sum_{E \subseteq D} (-1)^{|D \setminus E|}\, g(E) = 0$.*

*Proof.* If we set $h = \sum_{E \subseteq D} (-1)^{|D \setminus E|}\, g(E)$, then we have to show that $h = 0$. Since $A$ and $B$ form a partition of $D$, we can write

$$h \;=\; \sum_{A' \subseteq A} \sum_{B' \subseteq B} (-1)^{|(A \cup B) \setminus (A' \cup B')|}\, g(A' \cup B'),$$

where $A' = E \cap A$ and $B' = E \cap B$. Then, from the fact that $A \cap B = A' \cap B' = A' \cap B = B' \cap A = \emptyset$ it follows both that $(-1)^{|(A \cup B) \setminus (A' \cup B')|} = (-1)^{|A \setminus A'|} \times (-1)^{|B \setminus B'|}$ and that $g(A' \cup B') = g(A') + g(B')$. Hence, we obtain

$$\begin{aligned} h &= \sum_{A' \subseteq A} \sum_{B' \subseteq B} (-1)^{|A \setminus A'|} (-1)^{|B \setminus B'|}\; \{g(A') + g(B')\} \\ &= \sum_{A' \subseteq A} (-1)^{|A \setminus A'|} \sum_{B' \subseteq B} (-1)^{|B \setminus B'|}\; \{g(A') + g(B')\} \\ &= \sum_{A' \subseteq A} (-1)^{|A \setminus A'|} \left\{ g(A') \sum_{B' \subseteq B} (-1)^{|B \setminus B'|} + \sum_{B' \subseteq B} (-1)^{|B \setminus B'|} g(B') \right\}. \end{aligned}$$

By assumption $B \neq \emptyset$, so that Lemma 1 implies $\sum_{B' \subseteq B} (-1)^{|B \setminus B'|} = 0$ and thus

$$h \;=\; \sum_{A' \subseteq A} (-1)^{|A \setminus A'|} \left\{ \sum_{B' \subseteq B} (-1)^{|B \setminus B'|} g(B') \right\}.$$

Since we also have $A \neq \emptyset$, Lemma 1 also implies that $\sum_{A' \subseteq A} (-1)^{|A \setminus A'|} = 0$ and therefore that $h = 0$, as required. □

# B Long proofs

## B.1 Proof of Theorem 2

We first show (i)⇔(ii). The implication (i)⇒(ii) is straightforward. To prove that (i)⇐(ii) we use the same argument as in the proof of Theorem 1 of Drton and Richardson (2008), which for completeness we now give in detail.

We want to show that for every $i_{A \cup B} \in \mathcal{I}_{A \cup B}$ it holds that

$$P(X_{A \cup B} = i_{A \cup B}) = P(X_A = i_A) P(X_B = i_B) \tag{18}$$

and we do this by induction on the number of 0s in $i_{A \cup B}$, which we denote by $k$, with $0 \leq k \leq |A \cup B|$. More precisely, point (ii) implies that the factorization (18) is

satisfied for $k = 0$, also when $A$ and $B$ are replaced with proper subsets, and we show that if such factorization is satisfied for every $k < j \leq |A \cup B|$ then it is also true for $k = j$. Since $j > 0$, there exists $v \in A \cup B$ such that $i_v = 0$ and, in the following, we assume without loss of generality that $v \in A$, and set $A' = A \backslash \{v\}$. Hence,

$$\begin{aligned} P(X_{A\cup B} = i_{A\cup B}) &= P(X_{A'\cup B} = i_{A'\cup B}) - P(X_{A'\cup B} = i_{A'\cup B}, X_v = 1) \\ &= P(X_{A'} = i_{A'})P(X_B = i_B) - P(X_{A'} = i_{A'}, X_v = 1)P(X_B = i_B) \\ &= \{P(X_{A'} = i_{A'}) - P(X_{A'} = i_{A'}, X_v = 1)\} P(X_B = i_B) \\ &= P(X_A = i_A)P(X_B = i_B) \end{aligned}$$

as required; note that the factorizations in the second equality follow from (ii) and the inductive assumption, because the number of 0s in $i_{A'\cup B}$ is $j-1$, and furthermore that for the case $A' = \emptyset$ we use the convention $P(X_{A'} = i_{A'}) = 1$ and $P(X_{A'} = i_{A'}, X_v = 1) = P(X_v = 1)$.

We now show (ii)$\Leftrightarrow$(iii). The implication (ii)$\Rightarrow$(iii) follows by noticing that $\gamma_D = \sum_{E\subseteq D} (-1)^{|D\backslash E|}\, g(E)$, where $g(E) = \log \mu_E$. Hence, if we set $D = A' \cup B'$, with $A'$ and $B'$ as in (iii), the statement in (ii) implies that for every $E \subseteq D$

$$g(E) = \log \mu_E = \log \mu_{A'\cap E} + \log \mu_{B'\cap E} = g(A' \cap E) + g(B' \cap E)$$

so that the equality $\gamma_D = 0$ follows immediately from Lemma 2. We next show that (ii)$\Leftarrow$(iii) by induction on the cardinality of $A \cup B$, which we again denote by $k$.

We first notice that the identity $\mu_{A\cup B} = \mu_A \times \mu_B$ is trivially true whenever either $A = \emptyset$ or $B = \emptyset$ because $\mu_\emptyset = 1$. Then, if $|A \cup B| = 2$, so that $|A| = |B| = 1$, $\gamma_{A\cup B} = 0$ implies $\mu_{A\cup B} = \mu_A \times \mu_B$ as an immediate consequence of the identity $\gamma_{A\cup B} = \log \frac{\mu_{A\cup B}}{\mu_A \mu_B}$. Finally, we show that if the result is true for $|A \cup B| < k$ then it also holds for $|A \cup B| = k$. To this aim, it is useful to introduce the vector $\mu^*$ indexed by $E \subseteq A \cup B$ defined as follows:

$$\mu^* = \begin{cases} \mu_E & \text{for } E \subset A \cup B; \\ \mu_A \times \mu_B & \text{for } E = A \cup B. \end{cases}$$

Condition (iii) is recursive and, therefore, if it is satisfied for $A$ and $B$ then it is also satisfied for every $A' \subseteq A$ and $B' \subseteq B$ such that $|A' \cup B'| < k$, that is, such that $A' \cup B' \subset A \cup B$. As a consequence, the inductive assumption implies that $\mu_{A'\cup B'} = \mu_{A'} \times \mu_{B'}$ for every $A' \subseteq A$ and $B' \subseteq B$ such that $A' \cup B' \neq A \cup B$, and this in turn has two implications: firstly, we only have to prove that (iii) implies $\mu_{A\cup B} = \mu_A \times \mu_B$; secondly, we have $\sum_{E\subseteq A\cup B} (-1)^{|(A\cup B)\backslash E|}\, \log \mu^*_E = 0$ by Lemma 2.

Hence, we can write

$$\begin{aligned}
\gamma_{A\cup B} &= \sum_{E\subseteq A\cup B} (-1)^{|(A\cup B)\setminus E|}\ \log \mu_E \\
&= \log \mu_{A\cup B} + \sum_{E\subset A\cup B} (-1)^{|(A\cup B)\setminus E|}\ \log \mu^*_E \\
&= \log \mu_{A\cup B} - \log \mu_A - \log \mu_B + \sum_{E\subseteq A\cup B} (-1)^{|(A\cup B)\setminus E|}\ \log \mu^*_E \\
&= \log \mu_{A\cup B} - \log \mu_A - \log \mu_B \qquad (19)
\end{aligned}$$

and since (iii) implies that $\gamma_{A\cup B} = 0$ then (19) leads to $\mu_{A\cup B} = \mu_A \times \mu_B$, and the proof is complete.

## B.2 Proof of Corollary 3

For $i = 1, \ldots, r$, we introduce the sets $A_{-i} = \bigcup_{j\neq i} A_j$ and $\mathcal{D}_i = \{D | D \subseteq A_i \cup A_{-i},$ with both $D \cap A_i \neq \emptyset$ and $D \cap A_{-i} \neq \emptyset\}$ and note that, by Theorem 2, $X_{A_i} \perp\!\!\!\perp X_{A_{-i}}$ if and only if $\gamma_D = 0$ for every $D \in \mathcal{D}_i$. The mutual independence $X_{A_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{A_r}$ is equivalent to $X_{A_i} \perp\!\!\!\perp X_{A_{-i}}$ for every $i = 1, \ldots, r$ and, by Theorem 2, the latter holds true if and only if $\gamma_D = 0$ for every $D \in \bigcup_{i=1}^r \mathcal{D}_i$. Hence, to prove the desired result we have to show that $\mathcal{D} = \bigcup_{i=1}^r \mathcal{D}_i$.

It is straightforward to see that $\mathcal{D}_i \subseteq \mathcal{D}$ for every $i = 1, \ldots, r$, so that $\mathcal{D} \supseteq \bigcup_{i=1}^r \mathcal{D}_i$. The reverse inclusion $\mathcal{D} \subseteq \bigcup_{i=1}^r \mathcal{D}_i$ can be shown by noticing that for any $D \in \mathcal{D}$ one can always find at least one set $A_i$ such that $D \cap A_i \neq \emptyset$; since $D \not\subseteq A_i$ by construction, it holds that $D \cap A_{-i} \neq \emptyset$ and therefore that $D \in \mathcal{D}_i$. Hence, we have $D \in \bigcup_{i=1}^r \mathcal{D}_i$ for every $D \in \mathcal{D}$, and this completes the proof.

## B.3 Proof of Theorem 5

Every set $D \subseteq V$ that is disconnected in $\mathcal{G}$ can be partitioned uniquely into inclusion maximal connected sets $\tilde{D}_1, \ldots, \tilde{D}_r$ with $r \geq 2$. It is shown in Lemma 1 of Drton and Richardson (2008) that $\pi \in \mathbf{B}(\mathcal{G})$ if and only if $X_{\tilde{D}_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{\tilde{D}_r}$ for every disconnected set $D \subseteq V$. Hence, it is sufficient to prove that the mutual independence $X_{\tilde{D}_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{\tilde{D}_r}$ holds for every disconnected set $D$ in $\mathcal{G}$ if and only if $\gamma_D = 0$ for every disconnected set $D$ in $\mathcal{G}$.

We assume that $D = \tilde{D}_1 \cup \cdots \cup \tilde{D}_r$ is an arbitrary subset of $V$ that is disconnected in $\mathcal{G}$ and note that, in this case, also every set $E \subseteq \tilde{D}_1 \cup \cdots \cup \tilde{D}_r$ such that $E \not\subseteq \tilde{D}_i$ for every $i = 1, \ldots, r$ is disconnected in $\mathcal{G}$. Then, if $X_{\tilde{D}_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{\tilde{D}_r}$ it follows from Corollary 3 that also $\gamma_D = 0$. On the other hand, if every element of $\gamma$ corresponding

to a disconnected set is equal to zero, then $\gamma_E = 0$ for every $E \subseteq \tilde{D}_1 \cup \cdots \cup \tilde{D}_r$ such that $E \not\subseteq \tilde{D}_i$ for every $i = 1, \ldots, r$ and, by Corollary 3, this implies that $X_{\tilde{D}_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{\tilde{D}_r}$.

## B.4 Proof of theorem 6

For every pair of subsets $D \subseteq U$ and $W' \subseteq W = V \backslash U$ it follows from the definition of LML parameter in (6) that

$$\gamma_{D \cup W'} = \sum_{E \subseteq D \cup W'} (-1)^{|(D \cup W') \backslash E|} \log \mu_E,$$

which is equivalent to

$$\gamma_{D \cup W'} = \sum_{E \subseteq D \cup W} \left\{ (-1)^{|(D \cup W') \backslash E|} \log \mu_E \right\} \mathbf{1}(E \subseteq D \cup W') \tag{20}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. Hence, if we define $\omega(E,\, D \cup W') = (-1)^{|(D \cup W') \backslash E|} \times \mathbf{1}(E \subseteq D \cup W')$, then we can rewrite (20) as

$$\gamma_{D \cup W'} = \sum_{E \subseteq D \cup W} \omega(E,\, D \cup W') \log \mu_E.$$

Accordingly, the quantity $\sum_{W' \subseteq W} \gamma_{D \cup W'}$ in (10) can be written as

$$\begin{aligned} \sum_{W' \subseteq W} \gamma_{D \cup W'} &= \sum_{W' \subseteq W} \sum_{E \subseteq D \cup W} \omega(E,\, D \cup W') \log \mu_E \\ &= \sum_{E \subseteq D \cup W} \log \mu_E \sum_{W' \subseteq W} \omega(E,\, D \cup W'). \end{aligned} \tag{21}$$

We now focus on the internal sum of (21),

$$\sum_{W' \subseteq W} \omega(E,\, D \cup W') = \sum_{W' \subseteq W} (-1)^{|(D \cup W') \backslash E|} \times \mathbf{1}(E \subseteq D \cup W'). \tag{22}$$

Since $E$, $D$ and $W$ in (22) are fixed and such that $D \cap W = \emptyset$ and $E \subseteq D \cup W$, we can depict the structure of these sets as in Figure 4. The sum on the right hand side of (22) is taken over all subsets $W' \subseteq W$ but its terms are different from zero if and only if $W'$ is such that $E \subseteq D \cup W'$, that is, if and only if there exists a subset $H \subseteq W \backslash E$ such that $W' = (E \backslash D) \cup H$. Formally, the following equality between sets holds,

$$\{W' \subseteq W |\; E \subseteq D \cup W'\} = \{W' = (E \backslash D) \cup H |\; H \subseteq W \backslash E\},$$

and, consequently, we can write (22) in the form

$$\sum_{W' \subseteq W} \omega(E,\, D \cup W') = \sum_{H \subseteq W \backslash E} (-1)^{|\{D \cup (E \backslash D) \cup H\} \backslash E|}.$$

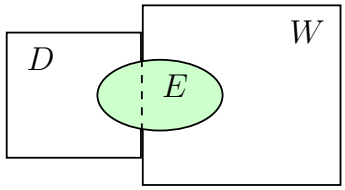



Figure 4: Subset structure in (22).

It is easy to see that $\{D \cup (E \backslash D) \cup H\} \backslash E = (D \cup H) \backslash E$ and from the fact that both $E \cap H = \emptyset$ and $D \cap H = \emptyset$ it follows that $(D \cup H) \backslash E = (D \backslash E) \cup H$ and, furthermore, that $(-1)^{|(D\backslash E)\cup H|} = (-1)^{|D\backslash E|} \times (-1)^{|H|}$. We can thus write

$$\begin{aligned} \sum_{W' \subseteq W} \omega(E,\, D \cup W') &= \sum_{H \subseteq W \backslash E} (-1)^{|(D\backslash E)|} \times (-1)^{|H|} \\ &= (-1)^{|(D\backslash E)|} \sum_{H \subseteq W \backslash E} (-1)^{|H|}. \end{aligned}$$

Now the sum $\sum_{H\subseteq W\backslash E} (-1)^{|H|}$ is equal to 1 for $W \backslash E = \emptyset$, that is, for $W \subseteq E$, and it is otherwise equal to 0 by Lemma 1. Hence, the final form of (22) is

$$\sum_{W' \subseteq W} \omega(E,\, D \cup W') = (-1)^{|(D\backslash E)|}\, \mathbf{1}(W \subseteq E). \tag{23}$$

Substituting (23) in (21) leads to

$$\begin{aligned} \sum_{W' \subseteq W} \gamma_{D \cup W'} &= \sum_{E \subseteq D \cup W} \log \mu_E\, (-1)^{|(D\backslash E)|}\, \mathbf{1}(W \subseteq E) \\ &= \sum_{F \subseteq D} \log \mu_{F \cup W}\, (-1)^{|D \backslash (F \cup W)|}, \end{aligned}$$

where $F = E \cap D$, and since $D \cap W = \emptyset$ it holds that $D \backslash (F \cup W) = D \backslash F$ so that

$$\sum_{W' \subseteq W} \gamma_{D \cup W'} = \sum_{F \subseteq D} (-1)^{|D\backslash F|}\, \log \mu_{F \cup W}. \tag{24}$$

We have assumed $D \neq \emptyset$ and, in this case, by Lemma 1

$$\sum_{F \subseteq D} (-1)^{|D\backslash F|} \log \mu_W = \log \mu_W \left\{ \sum_{F \subseteq D} (-1)^{|D\backslash F|} \right\} = 0 \tag{25}$$

so that we can subtract (25) from (24) and complete the proof as follows:

$$\begin{aligned} \sum_{W' \subseteq W} \gamma_{D \cup W'} &= \sum_{F \subseteq D} (-1)^{|D\backslash F|}\, \log \mu_{F \cup W} - \sum_{F \subseteq D} (-1)^{|D\backslash F|} \log \mu_W \\ &= \sum_{F \subseteq D} (-1)^{|D\backslash F|}\, \log \frac{\mu_{F \cup W}}{\mu_W} \\ &= \gamma_D^{\{X_W = 1\}}. \end{aligned}$$

# C Algorithm for ML estimation in LML models

Let $n^V = (n_D)_{D \subseteq V}$ be a vector of cell counts observed under multinomial sampling from a binary random vector $X_V$ with probability parameter $\pi^V > 0$. If we denote by $\psi^V = N\pi^V$ the expected value of $n^V$, where $N = \mathbf{1}^\top n^V$ is the total observed count (sample size) and $\mathbf{1}$ is the unit vector of size $R^{2^{|V|}}$. We can deal with maximum likelihood estimation of $\pi^V$ by considering $n^V$ as coming from Poisson sampling with parameter $\psi^V > 0$ and, in this case, we will find $\hat{\psi}^V_V = N$ and $N^{-1}\hat{\psi}^V = \hat{\pi}^V$. Thus, using the reparameterization $\omega^V = \log \psi^V$ to remove the positivity constraint on $\psi^V$, we can write the log-likelihood function (up to a constant term) as

$$\ell(\omega; n) = n^\top \omega - \mathbf{1}^\top \exp(\omega), \quad \omega \in R^{2^{|V|}},$$

where $n = n^V$ and $\omega = \omega^V$.

The LML parameter $\gamma$ is obtained from $\omega$ through the reparameterization $\gamma = \mathbb{M}^\top \log\{\mathbb{Z} \exp(\omega)\}$, $\omega \in R^{2^{|V|}}$, so that the linear constraint on $\gamma$ defined by $\mathbb{H}^\top \gamma = 0$ can be transformed into the following non-linear constraint on $\omega$:

$$g(\omega) = \mathbb{H}^\top \mathbb{M}^\top \log\{\mathbb{Z} \exp(\omega)\} = 0.$$

Maximum likelihood estimation in the LML model defined by $\mathbb{H}$ can thus be formulated as the problem of maximizing the objective function $\ell(\omega; n)$, with respect to $\omega$, subject to the constraint $g(\omega) = 0$.

A well-known method for the above constrained optimization problem looks for a saddle point of the Lagrangian function $\ell(\omega; n) + \tau g(\omega)$, where $\tau$ is a $k$-dimensional vector of unknown Lagrange multipliers, by solving for $\omega$ and $\tau$ the gradient equation

$$\frac{\partial \ell(\omega; n)}{\partial \omega} + \frac{\partial g(\omega)}{\partial \omega} \tau = 0$$

together with the constraint equation $g(\omega) = 0$. If $\hat{\omega}$ is a local maximum of $\ell(\omega; n)$ subject to $g(\omega) = 0$, and $\partial g(\omega)/\partial \omega$ is a full rank matrix, then a classical result (Bertsekas, 1982) guarantees that there exists a unique $\hat{\tau}$ such that the gradient equation is satisfied by $(\hat{\omega}, \hat{\tau})$. In the following we assume that the maximum likelihood estimate of interest is a local (constrained) maximum.

The gradient equation requires that the gradient of $\ell$, that is, the score vector

$$s(\omega; n) = \frac{\partial \ell(\omega; n)}{\partial \omega} = n - \exp(\omega),$$

be orthogonal to the constraining manifold defined by $g(\omega) = 0$, that is, belong to the vector space spanned by the columns of

$$\begin{aligned} \mathbb{G}(\omega) = \frac{\partial g(\omega)}{\partial \omega} &= \frac{\partial\{\mathbb{Z} \exp(\omega)\}}{\partial \omega} \frac{\partial \log\{\mathbb{Z} \exp(\omega)\}}{\partial\{\mathbb{Z} \exp(\omega)\}} \mathbb{M}\mathbb{H} \\ &= \operatorname{diag} \exp(\omega)\, \mathbb{Z}^\top [\operatorname{diag}\{\mathbb{Z} \exp(\omega)\}]^{-1} \mathbb{M}\mathbb{H}, \end{aligned}$$

where $\operatorname{diag} v$ is the diagonal matrix with diagonal entries taken from the vector $v$. We remark that $\mathbb{G}(\omega)$ has full rank, for all $\omega \in R^{2^{|V|}}$, because $\mathbb{H}$ has full rank by construction.

Since no closed-form solution of the system formed by the gradient and constraint equations is available (in our case) we resort to an iterative procedure inspired by Aitchison and Silvey (1958) and Lang (1996). Specifically, we use the Fisher-score-like updating equation

$$\begin{bmatrix} \omega^{t+1} \\ \tau^{t+1} \end{bmatrix} = \begin{bmatrix} \omega^{t} \\ 0 \end{bmatrix} + \begin{bmatrix} \mathbb{F}(\omega^t) & -\mathbb{G}(\omega^t) \\ -\mathbb{G}(\omega^t)^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} s(\omega^t; n) \\ g(\omega^t) \end{bmatrix}$$

to take step $t+1$ of the procedure, where $\omega^t$ and $\tau^t$ (unused) are the estimates of $\omega$ and $\tau$ (respectively) at step $t$, and $\mathbb{F}(\omega)$ is the Fisher information matrix

$$\mathbb{F}(\omega) = -E\left\{\frac{\partial s(\omega; n)}{\partial \omega}\right\} = -E\{-\operatorname{diag}\exp(\omega)\} = \operatorname{diag}\exp(\omega)$$

at $\omega \in R^{2^{|V|}}$. The above updating equation is obtained using a first order expansion of $s(\omega; n)$ and $g(\omega)$ about $\omega^t$; see Evans and Forcina (2011) for details.

The matrix inversion in the updating equation can be solved block-wise as follows (Aitchison and Silvey, 1958):

$$\begin{bmatrix} \mathbb{F}(\omega^t) & -\mathbb{G}(\omega^t) \\ -\mathbb{G}(\omega^t)^\top & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbb{R} & \mathbb{Q} \\ \mathbb{Q}^\top & -\mathbb{P}^{-1} \end{bmatrix},$$

where

$$\begin{aligned} \mathbb{P} &= \mathbb{G}(\omega^t)^\top \mathbb{F}(\omega^t)^{-1} \mathbb{G}(\omega^t), \\ \mathbb{Q} &= -\mathbb{F}(\omega^t)^{-1} \mathbb{G}(\omega^t) \mathbb{P}^{-1}, \\ \mathbb{R} &= \mathbb{F}(\omega^t)^{-1} + \mathbb{F}(\omega^t)^{-1} \mathbb{G}(\omega^t) \mathbb{Q}^\top. \end{aligned}$$

Then, introducing the relative score vector

$$e(\omega^t; n) = \mathbb{F}(\omega^t)^{-1} s(\omega^t; n) = \{\operatorname{diag}\exp(\omega^t)\}^{-1}\{n - \exp(\omega^t)\},$$

the updating equation can be split and simplified as

$$\begin{aligned} \tau^{t+1} &= -\mathbb{P}^{-1}\{\mathbb{G}(\omega^t)^\top e(\omega^t; n) + g(\omega^t)\}, \\ \omega^{t+1} &= \omega^t + e(\omega^t; n) + \mathbb{F}(\omega^t)^{-1} \mathbb{G}(\omega^t) \tau^{t+1}, \end{aligned}$$

so that the instrumental role of Lagrange multipliers becomes apparent, and it is clear that the algorithm actually runs in the space of $\omega$. Notice that the updates take place in the rectangular space $\mathbf{R}^{2^{|V|}}$, so that there is no risk of out of range estimation.

Since the algorithm does not always converge when the starting estimate $\omega^0$ is not close enough to $\hat{\omega}$, it is necessary to introduce a step size into the updating equation. The standard approach to choosing a step size in unconstrained optimization problems is to use a value for which the objective function to be maximized increases. However, since in our case we are looking for a saddle point of the Lagrangian function, we need to adjust the standard strategy. Specifically, Bergsma (1997) suggests to introduce a step size in the updating equation for $\omega$, which becomes

$$\omega^{t+1} \quad = \quad \omega^t + \text{step}^t\{e(\omega^t; n) + F(\omega^t)^{-1}\mathbb{G}(\omega^t)\tau^{t+1}\},$$

with $0 < \text{step}^t \leq 1$, while the updating equation for $\tau$ is unchanged, in light of the fact that $\tau^{t+1}$ is computed from scratch at each iteration. Our choice of $\text{step}^t$ is based on a simple step halving criterion, which has proven satisfactory for our needs, but more sophisticated criteria are available. At convergence we obtain $\hat{\gamma} = \mathbb{M}^\top \log\{\mathbb{Z}\exp(\hat{\omega})\}$ with asymptotic covariance matrix

$$\text{cov}(\hat{\gamma}) = \mathbb{J}^\top\mathbb{R}\mathbb{J},$$

where $\mathbb{J} = \text{diag}\exp(\hat{\omega})\,\mathbb{Z}^\top[\text{diag}\{\mathbb{Z}\exp(\hat{\omega})\}]^{-1}\mathbb{M}$ is the Jacobian of the map $\omega \mapsto \gamma$.

Finally, concerning the choice of the initial estimate $\omega^0$, we start from the maximum likelihood estimate under the saturated model: this choice is believed to result in quick convergence, because it makes the algorithm start close to the data, and our experience confirms this belief. If zero cell counts are present, we smooth the data by means of a convex combination with the mutual independence table having the same sample size and univariate counts as the data, using weights 0.5 for this "prior" table and $N$ for the actually observed table.