Log-mean linear models for binary data

Alberto Roverato[*]
Monia Lupparelli[†]
Luca La Rocca[‡]

June 2, 2012

Abstract

This paper introduces a novel class of models for binary data, which we call log-mean linear models. The characterizing feature of these models is that they are specified by linear constraints on the log-mean linear 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 log-mean linear interactions to zero and, more specifically, that graphical models of marginal independence are log-mean linear models. Our approach overcomes some drawbacks of the existing parameterizations of graphical models of marginal independence.

Keywords: Contingency table; Graphical Markov model; Marginal independence; Mean parameter

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 sub-spaces of the parameter space of the saturated model. In particular, we are interested in graphical models of marginal independence, which were introduced by Cox and Wermuth (1993, 1996) with the name of covariance graph models, but later addressed

[*]University of Bologna (alberto.roverato@unibo.it)
[†]University of Bologna (monia.lupparelli@unibo.it)
[‡]University of Modena and Reggio Emilia (luca.larocca@unimore.it)

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.

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, which we call the log-mean linear parameterization. We then define the family of log-mean linear 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 log-mean linear interactions to zero and, more specifically, that graphical models of marginal independence are log-mean linear models.

In the discrete case, two alternative parameterizations of bidirected graph models are available: the Möbius parameterization (Drton and Richardson, 2008) and the multivariate logistic parameterization (Glonek and McCullagh, 1995; Lupparelli et al., 2009). Our approach avoids some disadvantages of both these parameterizations: log-mean linear interactions can be interpreted as measures of association, which allows one to specify interesting sub-models not readily available using the Möbius parameterization, and the likelihood function can be written in closed form, which is not possible with the multivariate logistic parameterization. Furthermore, the log-mean linear approach to bidirected graph modelling is computationally more efficient than the multivariate logistic one.

2 Preliminaries

2.1 Parameterizations for binary data

Given the finite set $V = \{1, \dots, p\}$, with $\mid V \mid \; = p$, 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. Since $\mathcal{I}_V = \{0, 1\}^p = \{(1_D, 0_{V \setminus D}) \mid D \subseteq V\}$, we can write the probability table as a vector $\pi = (\pi_D)_{D \subseteq V}$ with entries $\pi = \text{pr}(X_D = 1_D, X_{V \setminus D} = 0_{V \setminus D})$. We refer to π as to the probability parameter of X_V and recall that it belongs to the $(2^p - 1)$-dimensional simplex, which we write as $\pi \in \Pi$.

In general, we call θ 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 θ (called interactions) are indexed by the subsets of V, i.e., $\theta = (\theta_D)_{D \subseteq V}$. If ω is an alternative parameter of X_V, then a result known as Möbius inversion states that

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

see, among others, Lauritzen (1996, Appendix A). Let Z and M be two $(2^p \times 2^p)$ matrices with entries indexed by the subsets of $V \times V$ and given by $Z_{D,H} = 1(D \subseteq H)$ and

$M_{D,H} = (-1)^{|H \setminus D|} 1(D \subseteq H)$, respectively, where $1(\cdot)$ denotes the indicator function. Then, the equivalence (1) can be written in matrix form as $\omega = Z^{\mathrm{T}} \theta$ if and only if $\theta = M^{\mathrm{T}} \omega$, and Möbius inversion follows by noticing that $M = Z^{-1}$.

We now review some well-known alternative parameterizations for the distribution of X_V, each defined by a smooth invertible mapping from Π 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.

Multivariate Bernoulli distributions form a regular exponential family with canonical log-linear parameter λ computed as $\lambda = M^T \log \pi$. The parameterization λ captures conditional features of the distribution of X_V and is used to define the class of log-linear models, which 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 $\mu = Z\pi$. The linear mapping $\pi \mapsto \mu$ is trivially Möbius-inverted to obtain $\pi = M\mu$, for all $\mu \in \mu(\Pi)$. However, the structure of $\mu(\Pi)$ is rather involved, and actually well-understood only for small p. The parameterization μ captures marginal distributional features of X_V and thus satisfies the upward compatibility property, i.e., it is invariant with respect to marginalization.

Ekholm et al. (1995), in a context of regression analysis, proposed to modify the mean parameter by replacing each entry μ_D of μ such that $| D |> 1$ with the corresponding dependence ratio defined as $\tau_D = \mu_D / (\prod_{v \in D} \mu_{\{v\}})$; see also Ekholm et al. (2000) and Darroch and Speed (1983), where these ratios were used in models named Lancaster additive. We define $\tau_D = \mu_D$ for $| D |\leq 1$ and call $\tau = (\tau_D)_{D \subseteq V}$ the dependence ratio parameter.

Bergsma and Rudas (2002) developed a wide class of parameterizations capturing both marginal and conditional distributional features, named marginal log-linear parameterizations, which have been applied in several contexts; see Bergsma et al. (2009). Broadly speaking, any marginal log-linear parameter is obtained by stacking subvectors of log-linear parameters computed in suitable marginal distributions. This class of parameterizations includes as special, extreme, cases the log-linear parameterization λ, where a single margin is used, and the multivariate logistic parameterization of Glonek and McCullagh (1995), denoted by $\eta = (\eta_D)_{D \subseteq V}$, where each η_D is computed in the margin X_D. The parameterization η clearly satisfies the upward compatibility property, while the structure of $\eta(\Pi)$ is rather involved. A disadvantage of these parameterizations is that their inverse mappings cannot be analytically computed (but for the special case of λ).

2.2 Bidirected graph models

Graphical models of marginal independence aim to capture marginal independence relationships between variables. Following Richardson (2003), we use the convention that the independence

Figure 1: Bidirected graph with disconnected sets $\{1,3\}$, $\{1,4\}$, $\{2,4\}$, $\{1,2,4\}$ and $\{1,3,4\}$, encoding the independencies $X_{\{1,2\}} \perp\!\!\!\perp X_4$ and $X_1 \perp\!\!\!\perp X_{\{3,4\}}$.

structure of variables is represented by a bidirected graph. Nevertheless, we recall that these same models have been previously discussed by Cox and Wermuth (1993) adopting a different graphical representation with undirected dashed edges.

A bidirected graph $\mathcal{G} = (V, E)$ is defined by a set $V = \{1, \dots, p\}$ of nodes and a set E of edges drawn as bidirected. A set $D \subseteq V$ is said to be connected in $\mathcal{G}$ if it induces a connected subgraph and it is said to be disconnected otherwise. Any disconnected set $D \subseteq V$ can be uniquely partitioned into its connected components $C_1, \dots, C_r$ such that $D = C_1 \cup \cdots \cup C_r$; see Richardson (2003) for technical details.

A bidirected graph model is the family of probability distributions for X_V satisfying a given Markov property with respect to a bidirected graph $\mathcal{G}$. 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 $X_{C_1}, \dots, X_{C_r}$ are mutually independent; in symbols $X_{C_1} \perp\!\!\!\perp X_{C_2} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{C_r}$. We denote by $B(\mathcal{G})$ the bidirected graph model for X_V defined by $\mathcal{G}$ under the connected set Markov property. See Figure 1 for an example.

Parameterizations for the class $B(\mathcal{G})$ have been studied by Drton and Richardson (2008) and Lupparelli et al. (2009), where $B(\mathcal{G})$ is defined by imposing multiplicative constraints on μ and linear constraints on η, respectively; see also Forcina et al. (2010), Rudas et al. (2010), Evans and Richardson (2012) and Marchetti and Lupparelli (2011).

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 parameter, denoted by $\gamma = (\gamma_D)_{D \subseteq V}$. Each element γ_D of γ is a log-linear expansion of a subvector, namely $(\mu_E)_{E \subseteq D}$, of the mean parameter:

$$\gamma_D = \sum_{E \subseteq D} (-1)^{|D \setminus E|} \log(\mu_E), \tag{2}$$

so that in vector form we have $\gamma = M^{\mathrm{T}} \log \mu$. Notice that by replacing μ with π in (2) one obtains the canonical log-linear parameter λ. Indeed, we will show in the next section that the log-mean linear parameterization defines a parameter space where the multiplicative constraints on the Möbius parameter of Drton and Richardson (2008) correspond to linear subspaces and, from this perspective, it resembles the connection between log-linear interactions and cell probabilities. It should be stressed, however, that the parameter space where γ lives has, like $\mu(\Pi)$ and $\eta(\Pi)$, a rather involved structure.

It is worth describing in detail the elements of γ corresponding to sets with low cardinality (its low-order interactions). Firstly, and trivially, $\gamma_\emptyset = \log \mu_\emptyset$ is always zero. Secondly, for every $j \in V$, the main log-mean linear 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 log-mean linear interaction $\gamma_{\{j,k\}} = \log\{\mu_{\{j,k\}}/(\mu_{\{j\}}\mu_{\{k\}})\}$ coincides with the logarithm of the second-order dependence ratio. Finally, for every triple $j, k, z \in V$, the three-way interaction is

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

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

We now formally define the log-mean linear parameterization as a mapping from Π.

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

$$\gamma = M^{\mathrm{T}} \log Z\pi, \quad \pi \in \Pi. \tag{3}$$

The multivariate logistic parameter η can also be computed as $\eta = C \log(L\pi)$ for a suitable choice of matrices C and L, so that the mapping $\pi \mapsto \eta$ resembles (3), but with the major difference that C and L 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 $\pi = M \exp Z^{\mathrm{T}}\gamma$. Clearly, the bijection specified by $\pi \mapsto \gamma$ is smooth, so that it constitutes a valid reparameterization. Finally, like μ and η, the parameterization γ satisfies the upward compatibility property.

We next define log-mean linear models as follows.

Definition 2 *For a vector X_V of binary variables and a full rank $(2^p \times k)$ matrix H, where $k < 2^p$ and the rows of H are indexed by the subsets of V, the log-mean linear model $\Gamma(H)$ is the family of probability distributions for X_V such that $H^T\gamma = 0$.*

It is not difficult to construct a matrix H such that $\Gamma(H)$ is empty. However, the family $\Gamma(H)$ 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 $\Gamma(H)$ to be non-empty is that the rows of H indexed by $D \subseteq V$ with $\mid D \mid\leq 1$ be all equal to zero; see § 4.

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

Proof. This follows from the mapping defining the parameterization γ being smooth, and the matrix H imposing a k-dimensional linear constraint on the parameter γ. □

Maximum likelihood estimation for log-mean linear models under a Multinomial or Poisson sampling scheme is a constrained optimization problem, which can be solved by means of standard algorithms. Specifically, we adopt an iterative method typically used for fitting marginal log-linear models which also gives the asymptotic standard errors; see Appendix B for details. In our case, the algorithm is computationally more efficient than for marginal log-linear models, especially when these are obtained by constraining the multivariate logistic parameter, 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 elements of γ, as well as those of μ and of τ, are not symmetric under relabelling of the two states taken by the random variables, because they measure event specific association. Ekholm et al. (1995, § 4) show that in some contexts this feature may amount to an advantage; see also the application in § 5. Furthermore, this is not an issue in the definition of bidirected graph models, which is illustrated in the next section.

4 Log-mean linear models and marginal independence

We show that the log-mean linear parameterization γ can be used to encode marginal independencies and, also, that bidirected graph models are log-mean linear models. Hence, the log-mean linear 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 μ and of the multivariate logistic parameterization η: the inverse map $\gamma \mapsto \pi$ can be analytically computed, as for μ, and the model is defined by means of linear constraints, as for η.

The following theorem shows how suitable linear constraints on the log-mean linear parameter correspond to marginal independencies; see Appendix A for a proof.

Theorem 1 *For a vector X_V of binary variables with probability parameter $\pi \in \Pi$, let $\mu = \mu(\pi)$ and $\gamma = \gamma(\pi)$. 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* $B' \neq \emptyset$.

We remark that the equivalence (i)⇔(ii) of Theorem 1 follows immediately from Theorem 1 of Drton and Richardson (2008). Furthermore, it is straightforward to see that (ii) could be restated by replacing the μ-interactions with the corresponding τ-interactions.

The next result generalizes Theorem 1 to the case of three or more subvectors; see Appendix A for a proof.

Corollary 1 *For a sequence $A_1, \ldots, A_r$ of $r \geq 2$ pairwise disjoint, nonempty, subsets of V, let $\mathcal{D} = \{D \mid D \subseteq A_1 \cup \cdots \cup A_r$ 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 $(\gamma_D)_{D \in \mathcal{D}} = 0$.*

An interesting special case of Corollary 1 is given below; see Appendix A for a proof.

Corollary 2 *For a subset $A \subseteq V$ with $\mid A \mid > 1$, the variables in X_A are mutually independent if and only if $\gamma_D = 0$ for every $D \subseteq A$ such that $\mid D \mid > 1$.*

We stated in § 3 that $\Gamma(H)$ is non-empty whenever the rows indexed by $D \subseteq V$ with $\mid D \mid \leq 1$ are equal to zero. This fact derives from Corollary 2, because the distribution of mutually independent variables satisfies the constraint $H^T \gamma = 0$.

It follows from Theorem 1 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 j and k are disjoint nodes in $\mathcal{G}$. The following theorem shows that bidirected graph models for binary data are log-mean linear models also under the connected set Markov property; see Appendix A for a proof.

Theorem 2 *The distribution of a vector of binary variables X_V belongs to the bidirected graph model $B(\mathcal{G})$ if and only if its log-mean linear parameter γ is such that $\gamma_D = 0$ for every set D disconnected in $\mathcal{G}$.*

For instance, if $\mathcal{G}$ is the graph in Figure 1 the bidirected graph model $B(\mathcal{G})$ is defined by the linear constraints $\gamma_{\{1,3\}} = \gamma_{\{1,4\}} = \gamma_{\{2,4\}} = \gamma_{\{1,2,4\}} = \gamma_{\{1,3,4\}} = 0$.

5 Application

Table 1 shows data from Coppen (1966) for a set of four binary variables concerning symptoms of 362 psychiatric patients. Wermuth (1976) analysed these data within the family of decomposable undirected graphical models, but a visual inspection of Table 6 of Wermuth (1976) suggests that also investigating the marginal independence structure may be useful. For this reason, we performed an exhaustive model search within the family of bidirected graph models and selected the model with optimal value of the Bayesian information criterion among those whose p-value, computed on the basis of the asymptotic chi-squared distribution of the deviance, is not smaller than 0.05. The selected model has deviance $\chi^2_{(5)} = 8.6$ ($p = 0.13$, $BIC = -20.85$) and corresponds to the graph of Figure 1, where $X_1 =$ Stability, $X_2 =$ Validity, $X_3 =$ Depression and $X_4 =$ Solidity.

The application of bidirected graph models is typically motivated by the fact that the observed Markov structure can be represented by a data generating processes with latent variables.

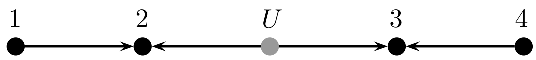

Figure 2: A generating model for Coppen's data.

In particular, the independence structure of the selected model is compatible, among others, with the generating process represented in Figure 2, where U is a latent factor. Under the generating model in Figure 2, one may be interested in investigating substantive research hypotheses on the role of the latent. For instance, U might be a binary variable representing a necessary factor for Depression: $\{U = \text{on}\}$ might be a necessary condition for patients to have acute depression, that is, for $\{X_3 = \text{yes}\}$. Formally, we might have $\text{pr}(X_3 = \text{yes} \mid U = \text{off}) = 0$, whereas $0 < \text{pr}(X_3 = \text{yes} \mid U = \text{on}) < 1$; see Ekholm et al. (2000, § 3.1).

If, in the above generating process, U represents a necessary factor, then the context-specific independence $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid \{X_3 = \text{yes}\}$ holds but, typically, $X_{\{1,2\}} \not\perp\!\!\!\perp X_4 \mid \{X_3 = \text{no}\}$. Furthermore, if the levels of X_3 = Depression are coded so that yes = 1, the above context-specific independence is satisfied in the selected marginal independence model if and only if some additional log-mean linear interactions are equal to zero, namely, $\gamma_{\{2,3,4\}} = \gamma_{\{1,2,3,4\}} = 0$; see Appendix C for details. Thus, by properly coding the levels of X_3, we can specify a log-mean linear model that encodes the independence structure of the graph in Figure 2 together with the context-specific independence implied by the assumption that U is a necessary factor for $\{X_3 = \text{yes}\}$. This model has deviance $\chi^2_{(7)} = 17.08$ ($p = 0.02$, $BIC = -24.16$) and, therefore, the necessary factor hypothesis is only weakly supported by the data. We remark that this log-mean linear model is invariant with respect to the coding of $X_{\{1,2,4\}}$, because it is fully specified by the constraints $X_{\{1,2\}} \perp\!\!\!\perp X_4$, $X_1 \perp\!\!\!\perp X_{\{3,4\}}$ and $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid \{X_3 = 1\}$. On the other hand, the log-mean linear model specified by the same zero constraints, but coding the levels of X_3 = Depression so that no = 1, allows one to verify the hypothesis that U is a necessary factor for the absence of depression, that is, for $\{X_3 = \text{no}\}$. The latter log-mean linear model provides an adequate fit with deviance $\chi^2_{(7)} = 9.3$ ($p = 0.23$, $BIC = -31.94$) so that the hypothesis is not contradicted by the data.

Table 1: Data from Coppen (1966) on four symptoms of 362 psychiatric patients.

		Solidity	hysteric		rigid	
Stability	Depression	Validity	psychasthenic	energetic	psychasthenic	energetic
extroverted	no		12	47	8	14
	yes		16	14	22	23
introverted	no		27	46	22	25
	yes		32	9	30	15

6 Discussion

Our log-linear expansion of μ provides the first instance of a parameterization for binary data, not belonging to the class of marginal log-linear parameterizations, which allows one to specify bidirected graph models through linear constraints.

We deem that the log-mean linear parameterization represents an appealing candidate for the implementation of Bayesian procedures for this class of models because the likelihood function under Multinomial or Poission sampling is explicitly available and marginal independencies correspond to zero-interactions. However, there are still difficulties related to the involved structure of the parameter space, which is a common trait of marginal parameterizations.

The specification of log-mean linear models encoding substantive research hypotheses, possibly by exploiting the asymmetry of our parameterization with respect to variable coding which we briefly touched upon in § 5, represents an open research area. Clearly, log-mean linear models can incorporate any linear constraint on $\log(\tau)$, because the latter is a linear transformation of γ. Some instances of substantive research assumptions that can be expressed in this way, such as, for instance, horizontal and vertical homogeneity of dependence ratios, can be found in Ekholm et al. (1995) and Ekholm et al. (2000).

Acknowledgments

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

Appendices

A Proofs of technical results

The following Lemma is instrumental in proving Theorem 1.

Lemma 1 *Let $g(\cdot)$ be a real-valued function defined on the sub-sets of a set D. If two non-empty, disjoint, proper sub-sets 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. We start this proof by recalling a well-known fact. It can be proven by induction that any non-empty set D has the same number of even and odd sub-sets. Consequently, it holds that

$$\sum_{E \subseteq D} (-1)^{|E|} = \sum_{E \subseteq D} (-1)^{|D \setminus E|} = 0 \quad \text{for all set } D \neq \emptyset. \tag{4}$$

We will use this fact twice in the remainder of this 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 equation (4) 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 $A \neq \emptyset$, equation (4) also implies that $\sum_{A'\subseteq A} (-1)^{|A\setminus A'|} = 0$ and therefore that $h = 0$, as required. □

Proof of Theorem 1

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

$$P(X_{A\cup B} = i_{A\cup B}) = P(X_A = i_A)P(X_B = i_B) \tag{5}$$

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 (5) 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 $A' = A\setminus\{v\}$. 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)⇔(iii). The implication (ii)⇒(iii) follows by noticing that

$$\gamma_D = \sum_{E\subseteq D} (-1)^{|D\setminus E|}\, g(E),$$

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 1. We next show that (ii)⇐(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 $\gamma_{A\cup B}=\log\{\mu_{A\cup B}/\mu_A\mu_B\}$. 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 μ^* 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 $\sum_{E\subseteq A\cup B}(-1)^{|(A\cup B)\setminus E|}\ \log\mu^*_E=0$ by Lemma 1. Hence, we can write

$$\begin{aligned}
\gamma_{A\cup B} &= \sum_{E\subseteq A\cup B}(-1)^{|(A\cup B)\setminus E|}\ \log\mu_E\\
&= \log\mu_{A\cup B}+\sum_{E\subset A\cup B}(-1)^{|(A\cup B)\setminus E|}\ \log\mu^*_E\\
&= \log\mu_{A\cup B}-\log\mu_A-\log\mu_B+\sum_{E\subseteq A\cup B}(-1)^{|(A\cup B)\setminus E|}\ \log\mu^*_E\\
&= \log\mu_{A\cup B}-\log\mu_A-\log\mu_B \qquad (6)
\end{aligned}$$

and since (iii) implies that $\gamma_{A\cup B}=0$ then (6) leads to $\mu_{A\cup B}=\mu_A\times\mu_B$, and the proof is complete.

Proof of Corollary 1

For $i=1,\dots,r$, 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 1, $X_{A_i}\perp\!\!\!\perp X_{A_{-i}}$ 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 1, 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.

Proof of Corollary 2

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

Proof of Theorem 2

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 $\pi \in B(\mathcal{G})$ 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 1 that also $\gamma_D = 0$. On the other hand, if every element of γ 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 1, this implies that $X_{\tilde{D}_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{\tilde{D}_r}$.

B Algorithm for maximum likelihood estimation

Let $n = (n_D)_{D \subseteq V}$ be a vector of cell counts observed under Multinomial sampling from a binary random vector X_V with probability parameter $\pi > 0$. If we denote by $\psi = N\pi$ the expected value of n, where $N = \mathbf{1}^{\mathrm{T}} n$ 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 π by considering n as coming from Poisson sampling with parameter $\psi > 0$ and, in this case, we will find $\mathbf{1}^T \hat{\psi} = N$ and $N^{-1}\hat{\psi} = \hat{\pi}$. Thus, using the reparameterization $\omega = \log \psi$ to remove the positivity constraint on ψ, we can write the log-likelihood function (up to a constant term) as

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

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

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

Maximum likelihood estimation in the log-mean linear model defined by $\mathbb{H}$ can thus be formulated as the problem of maximizing the objective function $\ell(\omega; n)$, with respect to ω, 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 τ is a k-dimensional vector of unknown Lagrange multipliers, by solving for ω and τ 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 ℓ, 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

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

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

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

to take step $t+1$ of the procedure, where ω^t and τ^t (unused) are the estimates of ω and τ (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 ω^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):

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

where

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

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

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

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

Since the algorithm does not always converge when the starting estimate ω^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 ω, which becomes

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

with $0 < \text{step}^t \leq 1$, while the updating equation for τ is unchanged, in light of the fact that τ^{t+1} is computed from scratch at each iteration. Our choice of 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 $\hat{\gamma} = \mathbb{M}^{\mathrm{T}}\log\{\mathbb{Z}\exp(\hat{\omega})\}$ with asymptotic covariance matrix

$$\operatorname{asy}\operatorname{cov}(\hat{\gamma}) = \mathbb{J}^{\mathrm{T}}\mathbb{R}\mathbb{J},$$

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

Finally, concerning the choice of the initial estimate ω^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.

C Details on the application

In this section we provide a formal description of some technical details of our application of log-mean linear models to the data by Coppen (1966).

Under the connected set Markov property, the bidirected graph in Figure 1 encodes the marginal independencies $X_{\{1,2\}} \perp\!\!\!\perp X_4$ and $X_1 \perp\!\!\!\perp X_{\{3,4\}}$, which are satisfied if and only if

$$\gamma_{\{1,3\}} = \gamma_{\{1,4\}} = \gamma_{\{2,4\}} = \gamma_{\{1,2,4\}} = \gamma_{\{1,3,4\}} = 0; \tag{7}$$

note that variable coding in uninfluential here. The directed acyclic graph in Figure 2 is a possible data generating process for the above bidirected graph model, because the directed Markov property (Lauritzen, 1996, § 3.2.2) implies, among others, the same marginal independencies and, moreover, it is associated with the recursive factorization

$$\mathrm{pr}(X_V = x_V, U = u) = \mathrm{pr}(x_2 \mid x_1, u)\mathrm{pr}(x_3 \mid x_4, u)\mathrm{pr}(x_1)\mathrm{pr}(x_4)\mathrm{pr}(u), \tag{8}$$

where $u \in \{\text{on}, \text{off}\}$ and $x_3 \in \{\text{yes}, \text{no}\}$; see Lauritzen (1996) and Drton and Richardson (2008) for details.

We claimed that, if the latent U is a necessary factor for Depression, that is, $\mathrm{pr}(X_3 = \text{yes} \mid U = \text{off}) = 0$, then $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid \{X_3 = \text{yes}\}$. This follows by noticing that

$$\mathrm{pr}(X_{\{1,2,4\}} = x_{\{1,2,4\}}, U = u \mid X_3 = \text{yes}) \propto \mathrm{pr}(X_{\{1,2,4\}} = x_{\{1,2,4\}}, X_3 = \text{yes}, U = u)$$

so that marginalizing over U one obtains

$$\mathrm{pr}(X_{\{1,2,4\}} = x_{\{1,2,4\}} \mid X_3 = \text{yes}) \propto \mathrm{pr}(X_{\{1,2,4\}} = x_{\{1,2,4\}}, X_3 = \text{yes}, U = \text{on}), \tag{9}$$

because $\mathrm{pr}(X_{\{1,2,4\}} = x_{\{1,2,4\}}, X_3 = \text{yes}, U = \text{off}) = 0$ by the definition of necessary factor. Hence, one can factorize the the right hand side of (9) as in (8) and the required context-specific independence follows immediately from the application of the factorization criterion; see Lauritzen (1996, eqn. (3.6)).

We now show that, if the levels of the variable $X3$=Depression are coded so that yes $= 1$, then the bidirected graph model in Figure 1 satisfies the additional context-specific independence $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid \{X_3 = 1\}$ if and only if, in addition to (7), it holds that $\gamma_{\{2,3,4\}} = \gamma_{\{1,2,3,4\}} = 0$. To this aim, we first notice that for the conditional distribution of $X_{\{1,2,4\}} \mid \{X_3 = 1\}$ the mean parameter, denoted by $\mu^{(3)}$, has entries

$$\mu^{(3)}_D = \mathrm{pr}(X_D = 1_D \mid X_3 = 1) = \frac{\mathrm{pr}(X_D = 1_D, X_3 = 1)}{\mathrm{pr}(X_3 = 1)} = \frac{\mu_{D\cup\{3\}}}{\mu_{\{3\}}} \tag{10}$$

for every $D \subseteq \{1, 2, 4\}$. From (10) it is possible to compute the corresponding log-mean linear parameter, denoted by $\gamma^{(3)}$, as a function of μ. In particular, if one computes $\gamma^{(3)}_{\{1,4\}}$, $\gamma^{(3)}_{\{2,4\}}$, $\gamma^{(3)}_{\{1,2,4\}}$ and then $\gamma_{\{1,3,4\}}$, $\gamma_{\{2,3,4\}}$, $\gamma_{\{1,2,3,4\}}$ by exploiting the factorizations of μ implied by (7) and Theorem 1, then it is straightforward to see that

$$\gamma^{(3)}_{\{1,4\}} = \gamma_{\{1,3,4\}} = 0, \qquad \gamma^{(3)}_{\{2,4\}} = \gamma_{\{2,3,4\}} \quad \text{and} \quad \gamma^{(3)}_{\{1,2,4\}} = \gamma_{\{1,2,3,4\}}. \tag{11}$$

The context-specific independence $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid \{X_3 = 1\}$ is a marginal independence in the distribution of $X_{\{1,2,4\}} \mid \{X_3 = 1\}$ and thus, by Theorem 1, it holds if and only if $\gamma^{(3)}_{\{1,4\}} = \gamma^{(3)}_{\{2,4\}} = \gamma^{(3)}_{\{1,2,4\}} = 0$. Therefore, if (7) holds true, it follows from (11) that $\gamma_{\{2,3,4\}} = \gamma_{\{1,2,3,4\}} = 0$ is a necessary and sufficient condition for $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid \{X_3 = 1\}$ to hold.

References

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.

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 (2012). Marginal log-linear parameters for graphical Markov models. Technical report, arXiv:1105.6075v2[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.

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 (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.

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.