Log-Mean Linear Parameterizations for Smooth Independence Models

Monia Lupparelli, Luca La Rocca and Alberto Roverato

Key words: marginal independence, smooth model, sparse table

1 Introduction

In categorical data analysis the choice of suitable parameterizations is a relevant aspect for several reasons: (i) the parameter space is often involved, (ii) its dimension rapidly increases with the number of variables, (iii) tables are sparse for high dimensional data, (iv) models specified by non-linear constraints on joint probabilities can result in non-smooth models. There is, in particular, an interest in parameterizations defining smooth and interpretable models by means of linear constraints on the parameter space. These considerations motivate the increasing attention for novel parameterizations; see [5], [1] and [9].

We focus on the log-mean linear (LML) parameterization recently introduced by [9] for the binary case and then generalized by [8], which is suitable for models of marginal independence, also known as bi-directed graph or covariance graph models; see [3] and [4]. These models investigate the marginal independence structures of the variables and are very useful in high dimensional data analysis, where working in low-dimension sub-spaces is highly desirable.

Monia Lupparelli
University of Bologna, Department of Statistical Sciences, Via Belle Arti 41, 40126 Bologna, Italy, e-mail: monia.lupparelli@unibo.it

Luca La Rocca
University of Modena and Reggio Emilia, Department of Computer, Mathematical and Physical Sciences, Via Campi 213/b, 41125 Modena, Italy, e-mail: luca.larocca@unimore.it

Alberto Roverato
University of Bologna, Department of Statistical Sciences, Via Belle Arti 41 40126 Bologna, Italy, e-mail: alberto.roverato@unibo.it

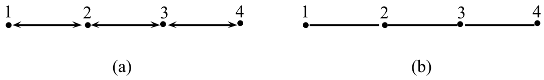

Fig. 1 Independence models for Coppen data: (a) a bi-directed graph giving $Y_1 \perp\!\!\!\perp \{Y_3,Y_4\}$ and $\{Y_1,Y_2\} \perp\!\!\!\perp Y_4$, under the connected set Markov property, with $\chi^2_{(5)} = 8.6$ (p-value $= 0.13$, BIC $= -20.85$); (b) an undirected graph giving $Y_1 \perp\!\!\!\perp \{Y_3,Y_4\}|Y_2$ and $\{Y_1,Y_2\} \perp\!\!\!\perp Y_4|Y_3$, under the global Markov property, with $\chi^2_{(8)} = 13.9$ (p-value $= 0.09$, BIC $= -33.26$).

We show, through an example, how linear constraints on the LML parameterization allow us to specify, at the same time, marginal independencies and partial conditional independencies, thus obtaining a class of smooth parsimonious bi-directed models defined in a lower dimensional space where the constraints have a clear interpretation; see [2] on a similar topic. In addition, through simulations, we show that a convenient choice of variable coding focusses LML models on partial tables with relatively large counts, which results in increased efficiency.

2 Smooth Parsimonious LML Models for Bi-Directed Graphs

We consider a vector $Y_V = (Y_v)_{v\in V}$ of discrete random variables taking values $i_V \in \mathscr{I}_V$, where $\mathscr{I}_V = \times_{v\in V}\, \{0,1,\ldots,d_v\}$, according to a Multinomial distribution with strictly positive probability parameter $\pi_V = (\pi^{i_V})_{i_V\in\mathscr{I}_V}$, where $\pi^{i_V} = P(Y_V = i_V)$ and $\sum_{i_V\in\mathscr{I}_V} \pi^{i_V} = 1$; π_V belongs to the $|\mathscr{I}_V| - 1$ dimensional simplex Π_V. For every $D \subseteq V$, Y_D takes values $i_D \in \mathscr{I}_D$ with $\mathscr{I}_D$ defined accordingly. The mean parameter is the vector $\mu_V = (\mu^{j_D}, j_D \in \mathscr{J}_D)_{D\subseteq V}, \mu_V \in \mu(\Pi_V)$, where $\mu^{j_D} = P(Y_D = j_D)$, $\mu^{j_\emptyset} = 1$, and $\mathscr{J}_D = \times_{v\in D}\, \{1,\ldots,d_v\}$. The LML parameter proposed by [9] and [8] is the vector $\gamma_V = (\gamma^{j_D}, j_D \in \mathscr{J}_D)_{D\subseteq V}$ defined by the smooth mapping $\Pi_V \to \gamma(\Pi_V)$

$$\gamma^{j_D} = \sum_{E\subseteq D} (-1)^{|D\setminus E|} \log(\mu^{j_E}); \tag{1}$$

for every $D \subseteq V$ we define $\gamma_D = (\gamma^{j_D})_{j_D\in\mathscr{J}_D}$, which is a subvector of γ_V.

Let $\mathscr{B} = (V,E)$ be a bi-directed graph defined by a finite set V of nodes and a symmetric set of edges $E \subseteq V \times V$ drawn as bi-directed. Under the *pairwise Markov property*, for the vector Y_V, a missing edge between a pair of nodes $(u,v) \notin E$ corresponds to the marginal independence $Y_u \perp\!\!\!\perp Y_v$. The set of all independencies encoded by $\mathscr{B}$ can be derived using the *connected set Markov property*: given any disconnected set $D \subseteq V$ of nodes in $\mathscr{B}$, the vectors associated to its connected components $Y_{C_1},\ldots,Y_{C_r}$ are mutually independent; see Fig. 1(a) for an illustration and [4] for technical details. Given a graph $\mathscr{B}$, the probability distribution of Y_V satisfies the connected set Markov property iff the vector $\gamma_D = 0$ for every disconnected set D of $\mathscr{B}$; see [8, Thr. 4.1]. Parameterizations for these models have also been studied by [4] and [6] using respectively the mean and multivariate logistic (MLT) parameter.

We consider the Coppen data set including four binary variables concerning symptoms of 362 psychiatric patients: $Y_1 \equiv$ stability (0 = extroverted, 1 = introverted); $Y_2 \equiv$ validity (0 = energetic, 1 = psychasthenic); $Y_3 \equiv$ acute depression (0 = yes, 1 = no); $Y_4 \equiv$ solidity (0 = hysteric, 1 = rigid). These data were analysed by [10], finding the conditional independence model in Fig. 1(b). More recently, [6] and [9] obtained the model in Fig. 1(a) using marginal independence models. Both models achieve a good fit, but they encode different independencies that can be combined only adding further independence relationships; however this operation requires some care because it may lead to non-smooth models; see [1, Ex. 7].

LML models represent a tool which allows us to partially combine the independencies under the two graph models into a single smooth model. In details, we can define an LML model under two sets of linear constraints:

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

$$\gamma_{\{1,3\}} + \gamma_{\{1,2,3\}} = 0, \quad \gamma_{\{2,4\}} + \gamma_{\{2,3,4\}} = 0, \quad \gamma_{\{1,3,4\}} + \gamma_{\{1,2,3,4\}} = 0. \tag{3}$$

Constraints in (2) define the bi-directed graph model, while constraints in (3) define the independencies $Y_1 \perp\!\!\!\perp \{Y_3, Y_4\}|\{Y_2 = 1\}$ and $Y_4 \perp\!\!\!\perp \{Y_1, Y_2\}|\{Y_3 = 1\}$ both implied by the undirected graph; for proofs and details about partial conditional independencies see [7, Thr. 6, Cor. 8]. In this way, we achieve a smooth parsimonious bi-directed graph model with $\chi^2_{(8)} = 11.45$ (p-value $= 0.18$, BIC $= -35.68$) where all constraints have a clear interpretation in term of independencies.

Partial conditional independencies which can be tested using LML models are of the form $\{Y_C = 1_C\}$ and thus depend on the coding of the variables. We propose to adopt the criterion of *maximal count coding*, so that hypotheses are tested in partial tables with many observations: given a set of binary variables, we will code them so that the cell with the largest count corresponds to all variables taking level 1. We deem this approach should improve the efficiency of inference, especially for large and sparse tables. This feature is illustrated by the following simulation study, which compares the performance of the LML and MLT parameterizations in achieving parsimonious models; the latter parameterization is denoted by η and attains parsimoniousness by setting to zero higher order interactions.

3 A Simulation Study

Consider four binary variables indexed by $V = \{A, B, C, D\}$. We compare the performance in testing the hypothesis $\eta_V = 0$ using the MLT parameter with the performance in testing the hypothesis $Y_C \perp\!\!\!\perp Y_D \{Y_A = 1, Y_B = 1\}$ using the LML parameter with maximal count coding. Both hypotheses are implied by the independence $Y_C \perp\!\!\!\perp Y_D | \{Y_A, Y_B\}$. We generated a sequence of probability vectors π_k, $k = 1, \ldots, 40$, satisfying the constraint $Y_C \perp\!\!\!\perp Y_D | \{Y_A, Y_B\}$. For each π_k, we sampled 5000 multinomial vectors n_w, $w = 1, \ldots, 5000$, of size N. For each random sample n_w, we tested the two above hypotheses at $\alpha = 0.05$ nominal significance level, using the $\chi^2_{(1)}$ dis-

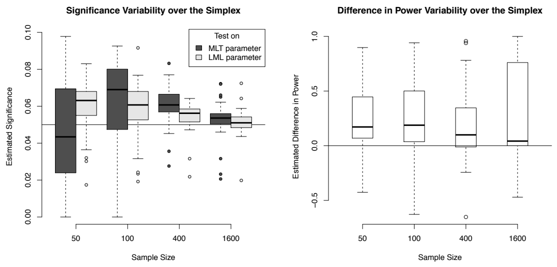

Fig. 2 (a) Box-plots of the estimated significance levels. (b) Box-plot of the difference in power.

tribution. For each π_k, we estimated the finite sample significance level $\hat{\alpha}_k^{\eta}$ and $\hat{\alpha}_k^{\gamma}$ of the two tests through the proportion of rejected models in the 5000 random samples, thus obtaining two distributions of estimates. The procedure was repeated for $N = 50$, 100, 400, 1600. Fig. 2(a) compares for every N the two box-plots of the estimated significance levels. The plot shows a lower variability in the estimates 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. We replicated our simulations using a sequence of 40 unconstrained probability vectors π_k. For every k, we estimated the type II error of the two tests, $\hat{\beta}_k^{\eta}$ and $\hat{\beta}_k^{\gamma}$, through the proportion of accepted models in the 5000 random samples. Fig. 2(b) reports, for every N, the box-plot of the differences in power $\hat{\delta}_k = \hat{\beta}_k^{\eta} - \hat{\beta}_k^{\gamma}$, $k = 1, \ldots, 40$. The plot shows a clear gain in power for the test based on the LML parameter.

References

1. Bergsma, W. P. and Rudas, T.: Marginal log-linear models for categorical data. Annals of Statistics **1**, 140–159 (2002)
2. Colombi, R. and Forcina, A.: A class of smooth models satisfying marginal and context specific conditional independencies. arXiv:1210.8050v1 (2012)
3. Cox, D. R. and Wermuth, N.: Linear dependencies represented by chain graphs. Statistical Science **8**, 204–218 (1993)
4. Drton, M. and Richardson, T.S.: Binary models for marginal independence. Journal of the Royal Statistical Society: Series B **70**, 287–309 (2008)
5. Ekholm A., McDonald J. W. and Smith P. W. F.: Association models for a multivariate binary response. Biometrics **56**, 712–718 (2000)
6. Lupparelli M., Marchetti G. M. and Bergsma W. P.: Parameterizations and fitting of bi-directed graph models to categorical data. arXiv:0801.1440v1 (2008)
7. Roverato, A., Lupparelli, M. and La Rocca, L.: Log-mean linear models for binary data. arXiv:1109.6239v2 (2012)
8. Roverato, A.: Dichotomization invariant log-mean linear parameterization for discrete graphical models of marginal independence. arXiv:1302.4641 (2013)
9. Roverato, A., Lupparelli, M. and La Rocca, L.: Log-mean linear models for binary data. Biometrika **100**, 485–494 (2013)
10. Wermuth, N.:Model search among multiplicative models. Biometrics **32**, 253–263 (1976)