Multivariate analysis of binary ordinal data using graphical models

Analisi multivariata di dati binari ordinali attraverso l'impiego di modelli grafici

Camilla Caroni, Fabio Alberto Comazzi, Andrea Deretti, Federico Castelletti

Abstract In this contribution we address the issue of inferring dependence relations between ordered binary variables. We propose a Bayesian model specification which assumes that ordinal data are generated by discretization of latent Gaussian data and that the joint density of the latent variables satisfies the conditional independencies imposed by a graphical model. We then propose an MCMC strategy for joint structural (graph) learning and parameter estimation.

Abstract *In questo contributo si considera il problema di stima di relazioni di dipendenza tra variabili binarie ordinali. Si propone un modello bayesiano, il quale assume che dati ordinali siano generati attraverso discretizzazione di osservazioni latenti gaussiane e che la distribuzione congiunta delle variabili latenti soddisfi relazioni di dipendenza imposte da un modello grafico. Si propone quindi un algoritmo di tipo MCMC per l'apprendimento della struttura grafica e l'inferenza sui parametri del modello.*

Key words: Structural learning, Ordinal data, Graphical model

Camilla Caroni
Politecnico di Milano, camilla.caroni@mail.polimi.it

Fabio Alberto Comazzi
Politecnico di Milano, fabioalberto.comazzi@mail.polimi.it

Andrea Deretti
Politecnico di Milano, andrea.deretti@mail.polimi.it

Federico Castelletti
Università Cattolica del Sacro Cuore, federico.castelletti@unicatt.it

1 Introduction

Modelling dependence relations between variables represents an important issue in many applied domains, and a variety of statistical methods have been proposed for this purpose. Typically the structure of dependencies is unknown and accordingly it must be inferred from the available data. To this end, probabilistic graphical models [6] adopt a graph-based representation of conditional independence relations among variables which are embedded in the joint density through a suitable graph-factorization. Specifically, a graph $\mathscr{G} = (V, E)$ consists of a set of nodes V associated to variables in the system, and a set of edges E, representing dependence relations. The goal is therefore to *learn* the graphical structure $\mathscr{G}$ given the data.

Most of the available methodologies however, both frequentist and Bayesian, deal under the assumption that Gaussian or categorical data are collected; see for instance [4], [5] and [7] for a Bayesian approach. More recently, a few methods for *mixed* (Gaussian and categorical) data have been proposed; see for instance [2] for a frequentist methodology. Still in the categorical framework, the majority of the literature has considered the case of *unordered* data which are represented as contingency tables of counts. In the Bayesian framework, methodologies based on suitable extensions of the Multinomial-Dirichlet model have been proposed for the analysis of unordered categorical data; see for instance [1] and references therein.

This contribution summarizes our first attempt to build a Bayesian framework for the multivariate analysis of categorical *ordinal* variables. For simplicity we assume all variables being *binary*, while in the last section of the work we provide some remarks for possible extensions to a general setting with ordered *polytomous* variables. Ordinal variables, namely variables whose levels are arranged according to a given ordering are quite common in many contexts, specifically in social sciences and psychology. In this framework, the adoption of standard methods for the analysis of categorical unordered data, although possible, is clearly not satisfactory, since the whole information encoded in the ordering of the data is lost.

The key assumption of our method is that the observed categorical variables $X_1, \dots X_q$ are generated by *discretization* of latent Gaussian random variables $Z_1, \dots,$ Z_q, whose joint density satisfies the conditional independence relations imposed by the graph $\mathscr{G}$. The main advantage of the proposed methodology is that we are able to infer both the underlying network structure as well as the *sign* of the dependence relation between pairs of variables. Finally, being fully Bayesian, a coherent quantification of the uncertainty around parameter estimates is provided.

2 Model formulation

In this section we introduce our model specification, which is based on the assumption that ordinal binary data are generated by thresholding of latent Gaussian observations.

Let $(X_1,\dots,X_q)$ be a collection of ordinal (binary) variables such that $X_j \in \{0,1\}$ for each $j=1,\dots,q$. Let also $(Z_1,\dots,Z_q)$ be q continuous variables, each associated with one of the q binary variables. We assume that each binary variable is obtained by *discretization* of its latent counterpart as

$$X_j = \begin{cases} 0 & \text{if } Z_j < \theta_0^{(j)} \\ 1 & \text{if } Z_j \geq \theta_0^{(j)} \end{cases} \tag{1}$$

where Z_j is the (latent) Gaussian random variable associated with X_j and $\theta_0^{(j)} \in (-\infty,\infty)$ represents an unknown cut-off parameter. Observed data then consist of $\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}$ where $\mathbf{x}^{(i)} = \left(x_1^{(i)},\dots,x_q^{(i)}\right)^\top$ denotes the i-th realization of the random vector $(X_1,\dots X_q)$, row of the (n,q) data matrix $\mathbb{X}$.

In what follows we assume that the joint distribution of the latent variables is multivariate Gaussian and that the model-parameter (covariance matrix Σ) satisfies the constraints imposed by a graphical model $\mathscr{G}$. Specifically, we rely on decomposable undirected graphical models (UGs) [3] and write

$$\begin{aligned} (Z_1,\dots,Z_q) \mid \Sigma,\mathscr{G} &\sim \mathscr{N}_q(\mathbf{0},\Sigma) \\ \Sigma \mid \mathscr{G} &\sim \text{HIW}(b,D), \end{aligned} \tag{2}$$

where HIW denotes the *Hyper Inverse Wishart* prior distribution. In addition, under the decomposable UG $\mathscr{G}$, the joint density of $(Z_1,\dots,Z_q)$ factorizes as

$$p(\mathbf{z} \mid \Sigma,\mathscr{G}) \propto \frac{\prod_{C\in\mathscr{C}} |\Sigma_C|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{z}_C^T\Sigma_C^{-1}\mathbf{z}_C\right\}}{\prod_{S\in\mathscr{S}} |\Sigma_S|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{z}_S^T\Sigma_S^{-1}\mathbf{z}_S\right\}}, \tag{3}$$

where $\mathscr{C}$ and $\mathscr{S}$ denotes the sets of cliques and separators respectively of $\mathscr{G}$. We refer the reader to [3] for full details.

Because of (1), the *augmented* density of $(X_1,\dots,X_q)$ and $(Z_1,\dots,Z_q)$ can be written as

$$\begin{aligned} p(\mathbf{x},\mathbf{z} \mid \Sigma,\Theta,\mathscr{G}) &= p(\mathbf{z} \mid \Sigma,\mathscr{G}) p(\mathbf{x} \mid \mathbf{z},\Theta) \\ &= p(\mathbf{z} \mid \Sigma,\mathscr{G}) \mathbf{1}\left\{\mathbf{z} \in C(\mathbf{x},\Theta)\right\}, \end{aligned} \tag{4}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $C(\mathbf{x},\Theta)$ is defined as

$$C(\mathbf{x},\Theta) = \left[\theta_{x_1-1}^{(1)},\theta_{x_1}^{(1)}\right) \times \left[\theta_{x_2-1}^{(2)},\theta_{x_2}^{(2)}\right) \times \cdots \times \left[\theta_{x_q-1}^{(q)},\theta_{x_q}^{(q)}\right),$$

and for each $j=1,\dots,q$ we adopt the notation $\theta_{-1}^{(j)} = -\infty$, $\theta_1^{(j)} = +\infty$. Given (3) we can write explicitly

$$p(\mathbf{x},\mathbf{z} \mid \Sigma,\Theta,\mathscr{G}) \propto \frac{\prod_{C\in\mathscr{C}} |\Sigma_C|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{z}_C^T\Sigma_C^{-1}\mathbf{z}_C\right\}}{\prod_{S\in\mathscr{S}} |\Sigma_S|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{z}_S^T\Sigma_S^{-1}\mathbf{z}_S\right\}} \mathbf{1}\left\{\mathbf{z} \in C(\mathbf{x},\Theta)\right\}. \tag{5}$$

Consider now the (n,q) data matrix $\mathbb{X}$ collecting n i.i.d. (binary) observations $\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}$ from (4) and the corresponding (n,q) latent data matrix $\mathbb{Z}$. Then, we can write the *augmented likelihood* as

$$p(\mathbb{X},\mathbb{Z} \mid \Sigma,\Theta,\mathscr{G}) = p(\mathbb{Z} \mid \Sigma,\mathscr{G}) \prod_{i=1}^{n} \mathbf{1}\left\{\mathbf{z}^{(i)} \in C\big(\mathbf{x}^{(i)},\Theta\big)\right\}$$
$$\propto \frac{\prod_{C\in\mathscr{C}} |\Sigma_C|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma_C^{-1}S_C)\right\}}{\prod_{S\in\mathscr{S}} |\Sigma_S|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma_S^{-1}S_S)\right\}} \prod_{i=1}^{n} \mathbf{1}\left\{\mathbf{z}^{(i)} \in C(\mathbf{x}^{(i)},\Theta)\right\},$$

where $S = \mathbb{Z}^T\mathbb{Z}$ and S_C denotes the sub-matrix of S with columns and rows indexed by $C \subseteq \{1,\dots,q\}$.

Now remember that $\theta_0^{(j)}$, $j = 1,\dots,q$, are (unknown) random thresholds linking the latent data to the binary observations. As a prior distribution we then assume

$$\theta_0^{(j)} \overset{\text{iid}}{\sim} \mathscr{N}(0,\tau^2), \quad j = 1,\dots,q. \tag{6}$$

We complete our model specification by assigning a prior to graph $\mathscr{G}$, for any $\mathscr{G} \in \mathbb{S}_q$, the space of decomposable UGs on q nodes. Specifically, let $A_{u,v}$ be the 0-1 (u,v)-element of the lower triangular adjacency matrix of $\mathscr{G}$. We assign hierarchically

$$A_{u,v} \mid \pi \overset{iid}{\sim} \mathrm{Ber}(\pi),$$
$$\pi \sim \mathrm{Beta}(a,b),$$

which corresponds to a multiplicity-correction prior on graph $\mathscr{G}$; see also [8].

3 Posterior inference

Starting from the model formulation introduced in the previous section we can consider the posterior distribution

$$p(\Sigma,\Theta,\mathscr{G},\mathbb{Z} \mid \mathbb{X}) = p(\mathbb{Z} \mid \Sigma,\mathscr{G}) \prod_{i=1}^{n} \mathbf{1}\left\{\mathbf{z}^{(i)} \in C(\mathbf{x}^{(i)},\Theta)\right\} p(\Sigma \mid \mathscr{G}) p(\mathscr{G}) \prod_{j=1}^{q} p(\theta_0^{(j)}),$$

where the latent data $\mathbb{Z}$, since unobserved, are also included among the "parameters" of the model. Because direct sampling from the latter distribution is not possible, we propose to implement a block Gibbs sampler scheme coupled with a Metropolis Hasting step to sample from the full conditional distributions of the model parameters. For convenience, we consider the two sets of parameters $(\Sigma,\mathscr{G})$ and $(\mathbb{Z},\Theta)$.

To start with, the full conditional of $(\Sigma,\mathscr{G})$ can be written as

$$\begin{aligned} p(\Sigma, \mathscr{G} \mid \Theta, \mathbb{Z}, \mathbb{X}) &= p(\Sigma, \mathscr{G} \mid \mathbb{Z}) \\ &\propto p(\mathbb{Z} \mid \Sigma, \mathscr{G}) p(\Sigma \mid \mathscr{G}) p(\mathscr{G}), \end{aligned} \tag{7}$$

where $p(\Sigma, \mathscr{G} \mid \Theta, \mathbb{Z}, \mathbb{X})$ reduces to $p(\Sigma, \mathscr{G} \mid \mathbb{Z})$ because Σ and $\mathscr{G}$ are independent of $\mathbb{X}$ and Θ given $\mathbb{Z}$, that is once the latent data are "observed". The latter expression corresponds to a joint posterior of $(\Sigma, \mathscr{G})$ in a Gaussian graphical model with a HIW (conjugate) prior; accordingly, direct sampling from $p(\Sigma, \mathscr{G} \mid \mathbb{Z})$ is possible following the MCMC scheme presented in [9].

The full conditional of $(\mathbb{Z}, \Theta)$ can be instead written as

$$p(\mathbb{Z}, \Theta \mid \Sigma, \mathscr{G}, \mathbb{X}) = p(\mathbb{Z} \mid \Theta, \Sigma, \mathscr{G}, \mathbb{X}) p(\Theta \mid \Sigma, \mathscr{G}, \mathbb{X}). \tag{8}$$

Specifically, we can write the first term as

$$p(\mathbb{Z} \mid \Theta, \Sigma, \mathscr{G}, \mathbb{X}) \propto \prod_{i=1}^{n} d\mathscr{N}_q(\mathbf{z}^{(i)} \mid \mathbf{0}, \Sigma) \mathbf{1}\left\{ \mathbf{z}^{(i)} \in C(\mathbf{x}^{(i)}, \Theta) \right\},$$

where $d\mathscr{N}_q(\cdot)$ denotes the density function of the multivariate Normal distribution $\mathscr{N}_q(\mathbf{0}, \Sigma)$. Accordingly, we can sample each latent observation $\mathbf{z}^{(i)}$, $i = 1, \dots n$, independently from a suitable multivariate Normal distribution truncated at the region $C(\mathbf{x}^{(i)}, \Theta)$. The second term can be written as

$$p(\Theta \mid \Sigma, \mathscr{G}, \mathbb{X}) \propto \prod_{i=1}^{n} \left\{ \Phi_q(\theta_{x_{ij}}^{(1)}, \dots, \theta_{x_{ij}}^{(q)} \mid \Sigma) - \Phi_q(\theta_{x_{ij}-1}^{(1)}, \dots, \theta_{x_{ij}-1}^{(q)} \mid \Sigma) \right\} \cdot \prod_{j=1}^{q} p(\theta_0^{(j)}),$$

where $\Phi_q(\cdot)$ is the c.d.f. of $\mathscr{N}_q(\mathbf{0}, \Sigma)$. For $j = 1, \dots, q$, we can update $\theta_0^{(j)}$ sequentially using a Metropolis-Hastings scheme based on the following steps:

- draw $\left(\theta_0^{(j)}\right)^*$ from a suitable proposal distribution $q\left(\left(\theta_0^{(j)}\right)^* \mid \theta_0^{(j)} \right)$, for instance $\mathscr{N}\left(\theta_0^{(j)}, \sigma_0^2\right)$;
- given the current value of $\theta_0^{(j)}$ accept $(\theta_0^{(j)})^*$ with probability

$$\alpha_j = \min \left\{ 1 \, ; \, \frac{p\left(\left(\theta_0^{(j)}\right)^* , \, \theta_0^{-j} \mid \Sigma, \mathscr{G}, \mathbb{X} \right)}{p\left(\theta_0^{(j)} \, , \, \theta_0^{-j} \mid \Sigma, \mathscr{G}, \mathbb{X} \right)} \cdot \frac{q\left(\theta_0^{(j)} \mid \left(\theta_0^{-j}\right)^* \right)}{q\left(\left(\theta_0^{-j}\right)^* \mid \theta_0^{(j)} \right)} \right\}$$

where $\theta_0^{-j} = \left\{ \theta_0^{(k)}, k \neq j \right\}$ denotes the collection of all cut-offs excluding the j-th.

4 Conclusions

We have presented a Bayesian methodology for learning dependence relations among multivariate ordinal binary data. Our method assumes that the ordinal cate-

gorical data are generated by thresholding of latent Gaussian data whose sampling distribution is Markov w.r.t. to a graph. Main advantage of the proposed method is that, differently from alternative methods based on Multinomial-Dirichlet models, we can provide a correlation-type measure between each pair of ordinal variables, which is encoded in the covariance matrix of the latent variables, Σ.

An extension of the method to ordinal variables $X_1, \ldots, X_q$, each with an arbitrary number of levels is possible and requires a suitable adaptation of the proposed framework. In particular, assuming that $X_j \in \{0, 1, \ldots, K_j\}$, one could introduce for each variable X_j a collection of cut-offs $\theta_0^{(j)}, \ldots, \theta_{K_j-1}^{(j)}$ such that

$$X_j = \begin{cases} 0 & \text{if } -\infty < Z_j \leq \theta_0^{(j)} \\ 1 & \text{if } \theta_0^{(j)} < Z_j \leq \theta_1^{(j)} \\ \ldots & \\ K_j & \text{if } \theta_{K_j-1}^{(j)} < Z_j < +\infty \end{cases} \tag{9}$$

where Z_j is the latent variable associated with X_j, and priors to the new cut-off parameters can be assigned independently following Equation (6).

We remark that this extension may augment the computation cost of the proposed algorithm, especially in high-dimensional (large q) settings, and specifically in the update of the latent data which requires sampling from truncated multivariate Normal distributions. Efficient computational implementations are currently under investigation.

References

1. Castelletti, F., Peluso, S.: Equivalence class selection of categorical graphical models. Computational Statistics & Data Analysis **164**, 107304 (2021)
2. Cheng, J., Li, T., Levina, E., Zhu, J.: High-Dimensional Mixed Graphical Models. Journal of Computational and Graphical Statistics **26**, 367–378 (2017)
3. Dawid, A. P., Lauritzen, S. L.: Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. The Annals of Statistics **21**, 272–1317 (1993)
4. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**, 432–441 (2008)
5. Kalisch, M., Bühlmann, P.: Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. Journal of Machine Learning Research **8**, 613–636 (2007)
6. Lauritzen, S. L.: Graphical Models, Oxford University Press (1996)
7. Mohammadi, A., Wit, E. C.: Bayesian Structure Learning in Sparse Gaussian Graphical Models. Bayesian Analysis **10**, 109–138 (2015)
8. Scott. J. G., Berger, J. O.: Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics **38**, 2587–2619 (2010)
9. Wang, H., Li, S. Z.: Efficient Gaussian graphical model determination under G-Wishart prior distributions. Electronic Journal of Statistics **6**, 168–198 (2012)