Clustering multivariate categorical data: a graphical model-based approach

Clustering di dati categoriali multivariati: un approccio basato su modelli grafici

Francesco Rettore, Michele Russo, Luca Zerman, Federico Castelletti

Abstract Clustering multivariate data using mixture models is a well-studied topic in statistics. In this contribution we propose a Bayesian framework for clustering categorical data which makes use of graphical models to account for possible group-specific dependence relations between variables. Our mixture model formulation allows to simultaneously infer a clustering structure of the units and the network of dependencies between variables.

Abstract *Il clustering di dati multivariati attraverso modelli mistura è un problema ampiamente studiato nella letteratura statistica. In questo contributo proponiamo un modello bayesiano per il clustering di dati categoriali che attraverso l'adozione di modelli grafici consente di considerare relazioni di dipendenza tra variabili specifiche di ciascun gruppo. Il modello mistura proposto permette di individuare gruppi latenti di osservazioni e al tempo stesso di stimare la struttura di dipendenza tra variabili del sistema.*

Key words: Model-based clustering, Categorical data, Graphical model

Francesco Rettore
Politecnico di Milano, e-mail: francesco.rettore@mail.polimi.it

Michele Russo
Politecnico di Milano, e-mail: michele6.russo@mail.polimi.it

Luca Zerman
Politecnico di Milano, e-mail: luca.zerman@mail.polimi.it

Federico Castelletti
Università Cattolica del Sacro Cuore, e-mail: federico.castelletti@unicatt.it

1 Introduction

Clustering individuals is a pervasive issue in statistics, with applications in several domains, and primarily social sciences. Traditionally, the two main clustering frameworks are represented by *distance-based* techniques that are opposed to *model-based* methods. The former implement a suitable measure of distance between observations and assign units to the same cluster whenever these are "close" in terms of the adopted metric; in this setting the most popular heuristic method is represented by the original k-means algorithm of [5] together with several of its extensions. Differently, model-based clustering techniques are probabilistic models based on *mixtures*, where each mixture component corresponds to a given cluster; see for instance [4].

In this work we assume that *multivariate* categorical observations have been collected and propose a model-based framework for clustering the available data. In doing so, we fully account for dependence relations between variables that are encoded in the data. Specifically, we represent conditional independence relations by using a graphical model-based approach [6]. Graphical models are probabilistic models for a collection of random variables which provide a powerful tool to impose dependence relations between variables in the joint density through a graph structure $\mathscr{G}$. The latter is made up of a set of *nodes* V, each corresponding to a variable in the system, and a set of *edges* E, representing dependence relations between nodes.

We propose a Bayesian non-parametric mixture model based on a Dirichlet Process prior [3], where each component of the mixture corresponds to a suitable categorical graphical model. Our model formulation allows to simultaneously learn dependence relations between variables and infer a clustering structure among individuals represented by latent groups of units sharing the same set of dependence statements. The rest of this contribution is organized as follows. In *Section 2* we introduce graphical models for categorical data under a Bayesian framework. In *Section 3* we describe our mixture of categorical graphical models, while in *Section 4* we provide computational details relative to a Markov chain Monte Carlo (MCMC) scheme for posterior inference on clustering and graphs.

2 Categorical graphical models

Let $(X_1,\dots,X_q)$ be a categorical random vector such that for each $j = 1\dots,q$, $X_j \in \mathscr{X}_j$, the set of levels of the categorical variable X_j. Let also $\mathbb{X}$ be an (n,q) dataset whose rows are i.i.d. realizations of the random vector $(X_1,\dots,X_q)$, namely $\mathbf{x}^{(i)} = \left(x_1^{(i)},\dots,x_q^{(i)}\right)^\top$, for $i = 1,\dots n$. It follows that each $\mathbf{x}^{(i)} \in \mathscr{X} := \bigotimes_{j=1}^q \mathscr{X}_j$, the product space generated by the levels of the q categorical variables. We then let $\pi(\mathbf{x}) \in (0,1)$ be the probability to observe a given configuration $\mathbf{x} \in \mathscr{X}$. Starting from the data matrix $\mathbb{X}$, we can compute for each $\mathbf{x} \in \mathscr{X}$ the corresponding *configuration count* $n(\mathbf{x})$ defined as

$$n(\mathbf{x}) = \sum_{i=1}^{n} \mathbf{1}\{\mathbf{x}^{(i)} = \mathbf{x}\}. \tag{1}$$

In addition, for a given subset $S \subseteq \{1,\dots,q\}$ and each $\mathbf{x}_S \in \mathscr{X}_S$, the corresponding *marginal configuration count* can be computed as

$$n(\mathbf{x}_S) = \sum_{i=1}^{n} \mathbf{1}\big\{\mathbf{x}^{(i)}(S) = \mathbf{x}_S\big\} = \sum_{\mathbf{x}\in\mathscr{X}} n(\mathbf{x})\mathbf{1}\big\{\mathbf{x}(S) = \mathbf{x}_S\big\}, \tag{2}$$

where $\mathbf{x}(S)$ is the sub-vector of $\mathbf{x}$ with components indexed by S. Similarly, we can define the *marginal probability*

$$\pi(\mathbf{x}_S) = \sum_{\mathbf{x}\in\mathscr{X}} \pi(\mathbf{x})\mathbf{1}\big\{\mathbf{x}(S) = \mathbf{x}_S\big\}. \tag{3}$$

By assuming independence among the n observations, we can write the likelihood function as

$$\begin{aligned} p(\mathbf{N} \mid \theta) = \prod_{i=1}^{n}\prod_{\mathbf{x}\in\mathscr{X}} \pi(\mathbf{x})^{\mathbf{1}\{\mathbf{x}^{(i)}=\mathbf{x}\}} &= \prod_{\mathbf{x}\in\mathscr{X}} \pi(\mathbf{x})^{\sum_{i=1}^{n}\mathbf{1}\{\mathbf{x}^{(i)}=\mathbf{x}\}} \\ &= \prod_{\mathbf{x}\in\mathscr{X}} \pi(\mathbf{x})^{n(\mathbf{x})}. \end{aligned} \tag{4}$$

where θ is the model parameter collecting the probabilities $\{\pi(\mathbf{x}), \mathbf{x} \in \mathscr{X}\}$.

Let now $\mathscr{G} = (V,E)$ be a decomposable undirected graph (UG) on the set of nodes $V = \{1,\dots,q\}$ and $\mathbb{G}_q$ the space of all decomposable graphs with q nodes. A decomposable UG is uniquely characterized by its set of *cliques* and *separators*; see for instance [6]. More importantly, as we associate each categorical variable X_j to a node in $\mathscr{G}$, the conditional independencies encoded in $\mathscr{G}$ are imposed to the joint density of $(X_1,\dots,X_q)$ and the likelihood function factorizes as

$$p(\mathbf{N} \mid \theta, \mathscr{G}) = \frac{\prod_{C\in\mathscr{C}} p(\mathbf{N}_C \mid \theta_C)}{\prod_{S\in\mathscr{S}} p(\mathbf{N}_S \mid \theta_S)}, \tag{5}$$

where $\mathscr{C}$ and $\mathscr{S}$ denote, respectively, the sets of cliques and separators of $\mathscr{G}$. Also, $\mathbf{N}_C$ and θ_C represent the contingency tables of marginal configuration counts and probabilities, $\{n(\mathbf{x}_C), \mathbf{x}_C \in \mathscr{X}_C\}$ and $\{\pi(\mathbf{x}_C), \mathbf{x}_C \in \mathscr{X}_C\}$ respectively; see also [2].

We complete the model specification by assigning a prior to θ. Specifically, we assume that conditionally on $\mathscr{G}$, θ follows a *Hyper Dirichlet* (HD) distribution with hyperparameter $\mathbf{A} = \{a(\mathbf{x}), \mathbf{x} \in \mathscr{X}\}$, namely

$$p(\theta \mid \mathscr{G}) = \frac{\prod_{C\in\mathscr{C}} p(\theta_C \mid \mathscr{G})}{\prod_{S\in\mathscr{S}} p(\theta_S \mid \mathscr{G})}, \tag{6}$$

where for each $S \in \mathscr{S}$, $\theta_S \mid \mathbf{A}_S \sim \mathrm{Dir}(\mathbf{A}_S)$, with $\mathbf{A}_S = \{a(\mathbf{x}_S), \mathbf{x}_S \in \mathscr{X}_S\}$ and

$$a(\mathbf{x}_S) = \sum_{\mathbf{x} \in \mathscr{X}} a(\mathbf{x}) \mathbf{1}\big\{\mathbf{x}(S) = \mathbf{x}_S\big\}; \tag{7}$$

similarly for each $C \in \mathscr{C}$; see [2] for full details. Finally, a prior $p(\mathscr{G})$, for any $\mathscr{G} \in \mathbb{S}_q$, the set of all decomposable UGs on q nodes, can be assigned through a Beta-Binomial distribution on the number of edges in the graph; see for instance [1].

The HD distribution provides a conjugate prior for the model parameter θ; accordingly, the posterior distribution of θ is still HD and the *marginal likelihood* of $\mathscr{G}$, $m(\mathbf{N} \mid \mathscr{G}) = \int p(\mathbf{N}, \theta \mid \mathscr{G}) p(\theta \mid \mathscr{G})\, d\theta$, is available in closed form. The latter feature is essential for the implementation of the MCMC scheme for posterior inference on graph structures and clustering introduced at the end of the following section.

3 Mixture model formulation

We extend the previous framework to a *mixture* of categorical graphical models thus allowing for possible heterogeneous dependence relations among subjects that are linked to a latent clustering structure of the data. We base our model formulation on a *Dirichlet Process* (DP) prior, by assuming

$$\begin{aligned} \mathbf{x}^{(i)} \mid (\theta_i, \mathscr{G}_i) &\sim p\big(\mathbf{x}^{(i)} \mid \theta_i, \mathscr{G}_i\big), \\ (\theta_i, \mathscr{G}_i) \mid M &\sim M, \\ M &\sim DP(M_0, \alpha), \end{aligned} \tag{8}$$

where in particular $DP(M_0, \alpha)$ represents the Dirichlet Process with base distribution M_0 and concentration parameter α. We then take $p(\theta, \mathscr{G}) = p(\theta \mid \mathscr{G}) p(\mathscr{G})$ as the baseline measure M_0, with $p(\theta \mid \mathscr{G})$, $p(\mathscr{G})$ defined as in *Section 2*.

A well known equivalent representation of the previous model can be obtained by taking the limit as K goes to infinity of a finite mixture model with K components (clusters) of the form

$$\begin{aligned} \mathbf{x}^{(i)} \mid c_i, \{\theta_i, \mathscr{G}_i\}_{k=1}^{K} &\sim p\big(\mathbf{x}^{(i)} \mid \theta_{c_i}, \mathscr{G}_{c_i}\big), \\ (\theta_k, \mathscr{G}_k) &\sim M_0, \\ c_i \mid \mathbf{p} &\sim \text{Discrete}(p_1, \ldots, p_K), \\ \mathbf{p} &\sim \text{Dir}(\alpha/K, \ldots, \alpha/K). \end{aligned} \tag{9}$$

In particular, each $c_i \in \{1, ..., K\}$ is a random variable indexing the cluster associated with the i-th observation; for each $k = 1, \ldots, K$, $(\theta_k, \mathscr{G}_k)$ are instead cluster-specific parameters.

The latter representation allows for the implementation of a *collapsed* MCMC scheme where parameters θ_k's are integrated out and the algorithm approximates a marginal posterior distribution over the space of graphs and cluster indicators $c_1, \dots, c_n$.

4 Computational details

Our sampling scheme for posterior inference relies on Algorithm 2 of [7], which applies to conjugate models whereas model-dependent parameters can be integrated out. Accordingly, the target is represented by the joint posterior

$$p\left(K, \{c_i\}_{i=1}^{n}, \{\mathscr{G}_k\}_{k=1}^{K}, \mid \mathbb{X}\right), \tag{10}$$

where K is the (random) number of clusters.

Given an initial state where the data $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ are divided into K clusters (equivalently, the indicator variables $c_1, \dots, c_n$ have been fixed) and graphs $\mathscr{G}_1, \dots, \mathscr{G}_K$ are associated with each of the K components, an MCMC algorithm can be implemented by reiterating the following two steps.

As a first step we update the indicator variables $c_1, \dots, c_n$ and, implicitly, the number of clusters K through a Gibbs sampling scheme which sequentially samples each c_i $(i = 1, \dots, n)$ from its full conditional distribution. Specifically,

if $c_i = c_j$ for some $j \neq i$,

$$P(c_i = k \mid \mathbf{c}_{-i}, \mathbb{X}, \mathscr{G}_1, \dots, \mathscr{G}_K) \propto \frac{n_{-i,k}}{n-1+\alpha} \int p(\mathbf{x}^{(i)} \mid \theta_k, \mathscr{G}_k)\, dH_{-i,k}(\theta_k \mid \mathscr{G}_k);$$

if $c_i \neq c_j \; \forall j \neq i$,

$$P(c_i \neq c_j \forall j \neq i \mid \mathbf{c}_{-i}, \mathbb{X}, \mathscr{G}_1, \dots, \mathscr{G}_K) \propto \frac{\alpha}{n-1+\alpha} \int p(\mathbf{x}^{(i)} \mid \theta_k, \mathscr{G}^*)\, dM_0(\theta_k \mid \mathscr{G}^*),$$

where in particular:

- $n_{-i,k} = \sum_{j \neq i} \mathbf{1}\{c_j = k\}$, i.e. the number of indicator variables (excluding c_i) that are equal to k;
- $H_{-i,k}$ denotes the posterior distribution of θ_k based on the prior M_0 and given the data $\{\mathbf{x}^{(j)}, j \neq i, c_j = k\}$ and graph $\mathscr{G}_k$;
- $\mathscr{G}^*$ corresponds to an empty cluster (graph) which is randomly sampled from the baseline measure on $\mathbb{S}_q$.

Since the two integrals correspond, respectively, to posterior and prior predictive distributions and the categorical HD model is conjugate, we can provide closed-form expressions for both the two terms (we omit details).

Conditionally to the cluster-indicator variables $c_1, \dots, c_n$, each graph $\mathscr{G}_k$, $k = 1, ..., K$, can be updated as follows. Consider first a partition of the n observations into K clusters, $\{C_1, \dots, C_K\}$, where

$$C_k = \{\mathbf{x}^{(i)} : c_i = k\}, \quad k = 1, \dots, K.$$

Each graph $\mathscr{G}_k$ can be updated through a Metropolis-Hastings step. Specifically, we first propose a new graph $\mathscr{G}_k'$ from a suitable proposal distribution $q(\mathscr{G}_k' \mid \mathscr{G}_k)$ and then we accept $\mathscr{G}_k'$ with probability:

$$\alpha_k = \min\left\{1, \frac{m(C_k \mid \mathscr{G}_k')}{m(C_k \mid \mathscr{G}_k)} \cdot \frac{p(\mathscr{G}_k')}{p(\mathscr{G}_k)} \cdot \frac{q(\mathscr{G}_k \mid \mathscr{G}_k')}{q(\mathscr{G}_k' \mid \mathscr{G}_k)}\right\},$$

where $m(C_k \mid \mathscr{G}_k)$ denotes the marginal likelihood of graph G_k given the data C_k.

References

1. Carlos M. Carvalho and James G. Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96:497–512, 2009.
2. A. Philip Dawid and Steffen L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21:1272–1317, 1993.
3. Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.
4. Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.
5. J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.
6. Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
7. Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249 – 265, 2000.