Graphical Models for Categorical Data

Camilla Caroni, Fabio Comazzi, Andrea Deretti, Francesco Rettore, Michele Russo, Luca Zerman

Politecnico di Milano

Tutors: Prof.ssa Lucia Paci, Prof. Federico Castelletti

January 10, 2022

Overview

Introduction

Main Goal: Perform clustering of multivariate categorical data using a mixture of graphical models.

Definition

A graphical model is a probabilistic model for a collection of random variables based on a graph structure.

In particular, we will work with undirected decomposable graphs.

Outline

1. **Inference on the graph:** Find the most suitable graph $\mathcal{G}$ representing the conditional independence relationships among the variables in the dataset. In particular, we consider two models:
   - *Multinomial-Dirichlet model*;
   - *Latent Normal Inverse-Wishart model*;

2. **Clustering:** Cluster the observations of the dataset via a Dirichlet Process mixture of graphical models.

Inference on the Graph: General Framework

Ingredients for inference on the graph:

- **Data Model:** $\mathsf{X}_1, \ldots \mathsf{X}_\mathsf{q} \mid \underline{\theta}, \mathcal{G} \sim p(\underline{\mathsf{x}} \mid \underline{\theta}, \mathcal{G})$
- **Prior on graph-dependent parameter $\underline{\theta}$ (given the graph):** $p(\underline{\theta} \mid \mathcal{G})$
- **Prior on graph $\mathcal{G}$:** $p(\mathcal{G})$

To proceed, we need to compute the posterior probability of $\mathcal{G}$ given the data $\mathbf{X}$:

$$p(\mathcal{G} \mid \mathbf{X}) \propto p(\mathbf{X} \mid \mathcal{G}) p(\mathcal{G})$$

where $p(\mathbf{X} \mid \mathcal{G}) = \int p(\mathbf{X} \mid \underline{\theta}, \mathcal{G}) p(\underline{\theta} \mid \mathcal{G}) d\underline{\theta}$ is the marginal likelihood.

Prior on the Graph

We considered three different choices for $p(\mathcal{G})$:

- **Uniform prior:** Assigns equal probabilities to all the graphs.
- **Binomial prior:** Assumes $A_{u,v} \mid \pi \overset{iid}{\sim} Be(\pi), \pi \in (0,1)$ where $A_{u,v}$ is the (u,v)-element of the upper-triangular adjacency matrix of $\mathcal{G}$.
- **Beta-Binomial prior:** Assumes $A_{u,v} \mid \pi \overset{iid}{\sim} Be(\pi), \pi \sim Beta(a,b)$.

Remark: The last two choices are particularly convenient to include prior information about the sparsity of the graph (whenever available).

Multinomial-Dirichlet Model

The *Multinomial-Dirichlet model* is defined as:

$$\mathbb{X} \mid \underline{\theta}, \mathcal{G} \sim p(\underline{x}^{(1)}, \ldots \underline{x}^{(n)} \mid \underline{\theta}, \mathcal{G}) = \frac{\prod_{C \in \mathcal{C}} \prod_{\underline{x}_C \in \mathcal{X}_C} \pi(\underline{x}_C)^{n(\underline{x}_C)}}{\prod_{S \in \mathcal{S}} \prod_{\underline{x}_S \in \mathcal{X}_S} \pi(\underline{x}_S)^{n(\underline{x}_S)}}$$

$$\underline{\theta} \mid \mathcal{G} \sim Hyper - Dirichlet(A) \text{ s.t. } \forall C \in \mathcal{C}, \forall S \in \mathcal{S} :$$

$$p(\theta_C \mid \mathcal{G}) = \frac{\Gamma(\sum_{\underline{x}_C \in \mathcal{X}_C} a(\underline{x}_C))}{\prod_{\underline{x}_C \in \mathcal{X}_C} \Gamma(a(\underline{x}_C))} \prod_{\underline{x}_C \in \mathcal{X}_C} \pi(\underline{x}_C)^{a(\underline{x}_C)-1}$$

$$p(\theta_S \mid \mathcal{G}) = \frac{\Gamma(\sum_{\underline{x}_S \in \mathcal{X}_S} a(\underline{x}_S))}{\prod_{\underline{x}_S \in \mathcal{X}_S} \Gamma(a(\underline{x}_S))} \prod_{\underline{x}_S \in \mathcal{X}_S} \pi(\underline{x}_S)^{a(\underline{x}_S)-1}$$

$$\mathcal{G} \sim p(\mathcal{G})$$

where $\mathcal{C}$ and $\mathcal{S}$ are the set of the cliques and the set of the separators of the graph respectively.

Multinomial-Dirichlet Model: Marginal Likelihood

In this case the marginal likelihood is available in closed form[1]:

$$m(N \mid \mathcal{G}) = \frac{\prod_{C \in \mathcal{C}} m(N_C \mid \mathcal{G})}{\prod_{S \in \mathcal{S}} m(N_S \mid \mathcal{G})}$$

$$m(N_C \mid \mathcal{G}) = \frac{\Gamma(\sum_{\underline{x}_C \in \mathcal{X}_C} a(\underline{x}_C))}{\Gamma(\sum_{\underline{x}_C \in \mathcal{X}_C} a(\underline{x}_C) + n(\underline{x}_C))} \prod_{\underline{x}_C \in \mathcal{X}_C} \frac{\Gamma(a(\underline{x}_C) + n(\underline{x}_C))}{\Gamma(a(\underline{x}_C))}$$

$$m(N_S \mid \mathcal{G}) = \frac{\Gamma(\sum_{\underline{x}_S \in \mathcal{X}_S} a(\underline{x}_S))}{\Gamma(\sum_{\underline{x}_S \in \mathcal{X}_S} a(\underline{x}_S) + n(\underline{x}_S))} \prod_{\underline{x}_S \in \mathcal{X}_S} \frac{\Gamma(a(\underline{x}_S) + n(\underline{x}_S))}{\Gamma(a(\underline{x}_S))}$$

where $n(\underline{x}) = \sum_{i=1}^{n} \mathbb{I}\{\underline{x}^{(i)} = \underline{x}\}$.

---

[1] S. L. Lauritzen et al.: Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models, 1993.

Multinomial-Dirichlet Model: MH Algorithm

Algorithm MH algorithm for the Multinomial-Dirichlet Model

Input: $\mathcal{G}^{(0)}$ (the initial candidate graph), M (the number of MCMC iterations)

Output: An MCMC sample $\{\mathcal{G}^{(t)}\}_{t=1}^{M}$ from $p(\mathcal{G} \mid N)$

for $t \leftarrow 1$ **to** M **do**

set $\mathcal{G} \leftarrow \mathcal{G}^{(t-1)}$;

draw a new candidate $\mathcal{G}'$ from $q(\mathcal{G}' \mid \mathcal{G})$;

compute $\alpha(\mathcal{G}' \mid \mathcal{G}) = \min\left\{1, \frac{m(N|\mathcal{G}')}{m(N|\mathcal{G})} \cdot \frac{p(\mathcal{G}')}{p(\mathcal{G})} \cdot \frac{q(\mathcal{G}|\mathcal{G}')}{q(\mathcal{G}'|\mathcal{G})}\right\}$;

update $\mathcal{G}^{(t)} = \begin{cases} \mathcal{G}', & \text{with probability } \alpha \\ \mathcal{G}^{(t-1)}, & \text{with probability } 1-\alpha \end{cases}$

end

Proposal Distribution

Problem: We need to define a suitable proposal distribution $q(\mathcal{G}' \mid \mathcal{G})$ from which we sample a new proposal graph ($\mathcal{G}'$) starting from the current state of the chain.

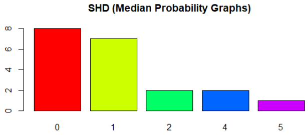
In particular, we build $\mathcal{G}'$ by either adding (a) or removing (b) an edge from $\mathcal{G}$ according to the following scheme:

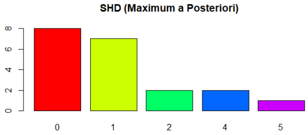
1. Construct the space $\mathcal{O}_{\mathcal{G}}$ of all possible *undirected* graphs obtained by (a) or (b) (starting from $\mathcal{G}$).
2. Uniformly draw a graph $\mathcal{G}'$ from $\mathcal{O}_{\mathcal{G}}$.
3. If $\mathcal{G}'$ is decomposable propose it, otherwise go back to step (2).

Remark: Such approach is computationally efficient as it guarantees that $\frac{q(\mathcal{G}|\mathcal{G}')}{q(\mathcal{G}'|\mathcal{G})} = 1$ so that we do not have to evaluate $q(\cdot)$ for the computation of the acceptance probability α in the MH algorithm.

Multinomial-Dirichlet Model: Assessment of the Results

Methodology: We create 20 categorical datasets starting from 20 randomly generated decomposable graphs and we run the MH algorithm on such datasets. After that, we compute the *Structural Hamming Distance* between the original graph and the one estimated from the chain.





Results: In 75% of the cases the reconstructed graph is very similar to the original one ($SHD \leq 1$).

Latent Normal Inverse-Wishart Model

In this case we introduce latent Gaussian random variables $(Z_1, ..., Z_q)$ in order to define a conjugate model and to be able to sample from the graph posterior distribution in an easy way.

Therefore, we approach the problem as follows:

1. **Sampling of the latent Gaussian data**;
2. **Inference on the graph given the Gaussian data**.

Latent Normal Inv-Wish Model: Inference on the Graph

The *Normal-Inverse-Wishart model* is defined as:

$$\mathbb{Z} \mid \Sigma, \mathcal{G} \sim p(\underline{z}^{(1)}, \ldots \underline{z}^{(n)} \mid \Sigma, \mathcal{G}) \propto \frac{\prod_{C \in \mathcal{C}} |\Sigma_C|^{-\frac{n}{2}} \cdot e^{-\frac{1}{2}\sum_{i=1}^{n} \underline{z}_C^{(i)^T} \Sigma_C^{-1} \underline{z}_C^{(i)}}}{\prod_{S \in \mathcal{S}} |\Sigma_S|^{-\frac{n}{2}} \cdot e^{-\frac{1}{2}\sum_{i=1}^{n} \underline{z}_S^{(i)^T} \Sigma_S^{-1} \underline{z}_S^{(i)}}}$$

$$\Sigma \mid \mathcal{G} \sim HIW(b, D) \text{ s.t. } \forall C \in \mathcal{C}, \forall S \in \mathcal{S}:$$

$$p(\Sigma_C \mid \mathcal{G}) \propto |\Sigma_C|^{-(\frac{b}{2}+|C|)} \cdot e^{-\frac{1}{2}tr(\Sigma_C^{-1} D_C)}$$

$$p(\Sigma_S \mid \mathcal{G}) \propto |\Sigma_S|^{-(\frac{b}{2}+|S|)} \cdot e^{-\frac{1}{2}tr(\Sigma_S^{-1} D_S)}$$

$$\mathcal{G} \sim p(\mathcal{G})$$

where $\mathcal{C}$ and $\mathcal{S}$ are the set of the cliques and the set of the separators of the graph respectively.

Latent Normal Inv-Wish Model: Marginal Likelihood

In this case the marginal likelihood is available in closed form[1]:

$$m(\underline{z}^{(1)}, \dots \underline{z}^{(n)} \mid \mathcal{G}) = (2\pi)^{-nq/2} \frac{h(\mathcal{G}, b, D)}{h(\mathcal{G}, b^*, D^*)}$$

$$b^* = b + n$$

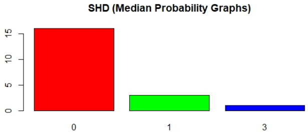
$$D^* = D + \sum_{i=1}^{n} \underline{z}^{(i)} \underline{z}^{(i)^T}$$

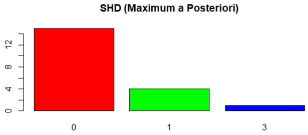
where $h(\mathcal{G}, b, D) = \frac{\prod_{C \in \mathcal{C}} |\frac{1}{2} D_C|^{\frac{b+|C|-1}{2}} \Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right)^{-1}}{\prod_{S \in \mathcal{S}} |\frac{1}{2} D_S|^{\frac{b+|S|-1}{2}} \Gamma_{|S|}\left(\frac{b+|S|-1}{2}\right)^{-1}}$ and $\Gamma_p(\cdot)$ denotes the *multivariate gamma function*.

[1] C. M. Carvalho, J. G. Scott: Objective Bayesian Model Selection in Gaussian Graphical Models, 2009

Latent Normal Inv-Wish Model: Assessment of the Results

Methodology: We create 20 gaussian datasets starting from 20 randomly generated decomposable graphs and we run the MH algorithm on such datasets. After that, we compute the *Structural Hamming Distance* between the original graph and the one estimated from the chain.





Results: The reconstructed graph is equal to the original one ($SHD = 0$) in the majority of the cases.

Next Steps

The final steps of the project are the following:

- **Complete the Latent Normal-Inverse-Wishart model for categorical data** (assuming that categorical data are generated by discretization of their latent counterparts);
- **Develop a mixture model whose components are Multinomial Dirichlet models or Latent Normal-Inverse-Wishart models**.

The implemented algorithms and the tests we performed are available at the following link.

GitHub