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

November 12, 2021

Overview

1. Introduction
2. Models
   - Categorical with Hyper Dirichlet prior
   - Gaussian with Hyper Inverse Wishart prior
   - Graphical Model Selection
3. Future Steps

Introduction

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

Definition

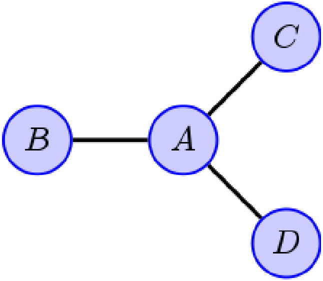
A graphical model is a probabilistic model for a collection of random variables based on a graph structure.

A graph is made up of a set of nodes (representing variables) and edges (representing ”dependence” relations between nodes).

Therefore, it can be used to model conditional dependence structures between random variables.

In particular, we will consider undirected decomposable graphs.

Example



Here we can read from the graph that B, C and D are conditionally independent given A.

We will write $B \perp\!\!\!\perp C \perp\!\!\!\perp D \mid A \; [\mathcal{G}]$, meaning *$B$, C and D are conditionally independent given A in $\mathcal{G}$.*

N.B.: *Conditional independencies can be read-off from the graph using graphical criteria such as d-separation (see References).*

A Bayesian approach for Model Selection

Ingredients for Model Selection:

- Data Model: $\mathrm{Y}_1, \ldots \mathrm{Y}_\mathrm{q} \mid \underline{\theta}, \mathcal{G} \overset{iid}{\sim} p(\underline{\mathrm{y}} \mid \underline{\theta}, \mathcal{G})$
- Prior on graph-dependent parameter $\underline{\theta}$ (given the graph): $p(\underline{\theta} \mid \mathcal{G})$
- Prior on graph $\mathcal{G}$: $p(\mathcal{G})$

In order to perform model selection we need to compute the posterior probability of $\mathcal{G}$ given the data $\mathbf{Y}$:

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

where $p(\mathbf{Y} \mid \mathcal{G}) = \int p(\mathbf{Y} \mid \underline{\theta}, \mathcal{G}) p(\underline{\theta} \mid \mathcal{G}) d\underline{\theta}$ is the marginal likelihood of graph $\mathcal{G}$ for $\mathcal{G} \in \mathcal{S}_q$, the set of all decomposable graphs on q nodes.

Factorization Property of Decomposable Graphs

Definition

A complete subset that is maximal with respect to inclusion is called a *clique*. We denote by $\mathcal{C}$ the set of all cliques of a graph.

Definition

Let $\mathcal{C} = \{C_1, ..., C_K\}$. For $k = 2, ..., K$ we define *separators* of the graph the sets:

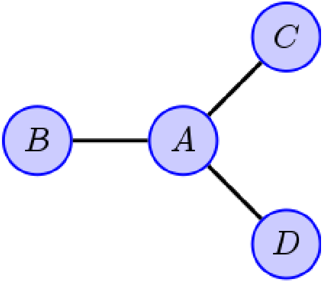
$$S_k = C_k \cap H_{k-1}$$

where $H_{k-1} = C_1 \cup ... \cup C_{k-1}$. We denote by $\mathcal{S}$ the set of all separators of a graph.

The factorization property of decomposable graphs allows to write the joint density as:

$$p(\mathbf{Y} \mid \underline{\theta}, \mathcal{G}) = \frac{\prod_{C \in \mathcal{C}} p(\mathbf{Y}_C \mid \underline{\theta}_C)}{\prod_{S \in \mathcal{S}} p(\mathbf{Y}_S \mid \underline{\theta}_S)}$$

Example



The conditional dependencies relations for this graph are the following:

$$p(B, C, D \mid A) = p(B \mid A)p(C \mid A)p(D \mid A) \iff$$

$$\iff p(A, B, C, D) = \frac{p(A, B)p(A, C)p(A, D)}{p(A)p(A)}$$

so that $\mathcal{C} = \{\{A, B\}, \{A, C\}, \{A, D\}\}$ and $\mathcal{S} = \{\{A\}, \{A\}\}$.

Prior Specification

To model categorical data, there are two approaches we are planning to follow. They are based on:

- **Multivariate categorical (multinomial) distributions** with (Hyper) Dirichlet prior;
- **Latent multivariate Gaussian distributions** with (Hyper) Inverse-Wishart prior (inference via data augmentation).

Categorical variables with (Hyper) Dirichlet prior

Let $\underline{\mathrm{Y}} = (Y_1, ..., Y_q)$ be a collection of categorical random variables. Each Y_i has set of levels $\mathcal{X}_i$.
Then:

$$\underline{\mathrm{Y}} \in \bigotimes_{i=1}^{q} \mathcal{X}_i, \text{ the set of all possible configurations of } \underline{\mathrm{Y}}.$$

Let $V := dim\left(\bigotimes_{i=1}^{q} \mathcal{X}_i\right)$, so that $\underline{\theta} \in \Sigma_V$ (the V-dimensional simplex), *i.e.* a vector of joint probabilities (one for each configuration of $\underline{\mathrm{Y}}$).

It can be shown that if $\underline{\theta} \mid \mathcal{G} \sim \textit{Hyp-Dir}(\underline{\alpha})$ then the marginal likelihood $p(\underline{\mathrm{Y}} \mid \mathcal{G})$ can be derived in closed form.

Remark

Hyper Dirichlet distributions provide conjugate priors for $\underline{\theta}$, satisfying the conditional independence constraints imposed by the graph.

Gaussian variables with (Hyper) Inverse Wishart prior

Assume the categorical (binary for simplicity) variables have been generated from a collection of latent Gaussian random variables:

$$Z_1, ..., Z_q \mid \Sigma_{\mathcal{G}}, \mathcal{G} \overset{iid}{\sim} \mathcal{N}_q(\underline{0}, \Sigma_{\mathcal{G}})$$

It can be shown that if $\Sigma_{\mathcal{G}} \sim \textit{Hyp-Inv-Wish}(b, D)$ then the marginal likelihood $p(\underline{\mathrm{Y}} \mid \mathcal{G})$ can be derived in closed form.

The (binary) categorical variables $Y_1, ..., Y_q$ are then obtained as follows:

$$Y_i = \begin{cases} 0, & \text{if } \mathrm{Z_i} \geqslant \theta_i \\ 1, & \text{if } \mathrm{Z_i} \leqslant \theta_i \end{cases}$$

where $\theta_i, i = 1, \ldots q$ are unknown cut-offs. We assign $\theta_i \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$.

Remark

Hyper Inverse Wishart distributions provide conjugate priors for $\underline{\theta}$, satisfying the conditional independence constraints imposed by the graph.

Bayesian Graphical Model Selection

In both cases the posterior is known up to a normalizing constant:

$$p(\mathcal{G} \mid \underline{\mathrm{Y}}) \propto p(\underline{\mathrm{Y}} \mid \mathcal{G}) p(\mathcal{G})$$

Problem: The dimension of the graph space $\mathcal{S}_q$ grows super-exponentially with the number of random variables $\Longrightarrow$ it is unfeasible to explicitly compute the posterior for each possible graph.

Therefore we will resort to a Metropolis-Hastings algorithm in order to make inference on $\mathcal{G}$ given the data.

How to specify a convenient proposal for the MH algorithm?

General idea: A new $\mathcal{G}'$ (decomposable) can be obtained by adding or removing an edge $u - v$ from $\mathcal{G}$.

Next Steps

Our immediate future steps will be:

- **Better specify a convenient proposal density** $q(\mathcal{G}' \mid \mathcal{G})$ **and implement the MH algorithm**;

- **Consider the case of a mixture of graphical models to approach the problem of cluster analysis**.

References

Main references:

- *d*-**Separation**: J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, 1988.
- **Hyper Dirichlet Hyper Inverse-Wishart**: S. L. Lauritzen et al. Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models, 1993.
- **Bayesian Graphical Model Selection**: F. Castelletti. Bayesian Model Selection of Gaussian Directed Acyclic Graph Structures, 2020.
- **Mixture of Graphical Models**: Rodriguez et al. Sparse covariance estimation in heterogeneous samples, 2011.