Investigating False Positives in Covariate-Dependent Graphical Model

Isaac R, Renat S, Gözde S

Section 1

Background and Introduction

Graphical Modeling

Gozde:

Undirected graphical models enables to model multivariate distributions. Suppose we observe a p-dimensional sample $x = (x_1, \dots, x_p)$ from a multivariate Gaussian distribution with a non-singular covariance matrix. Then the conditional independence structure of the distribution can be represented with a graph G. The graph $G = (V, E)$ is characterized by a node set $V = (1, \dots, p)$ corresponding to the p variables, and an edge set E such that $(i, j) \in E$ if and only if x_i and x_j are conditionally dependent given all other variable. The goal is to estimate the underlying graph G from given n idd observations $x_1, \dots, x_p$.

Adding Covariates

Gozde
Several methods developed under this assumption however, in practice, the observations may not be identically distributed. In this paper they suppose the variability in the graph structure across observations depending on additional covariate information. Let $X \in \mathbb{R}^{n \times p}$ stand for the data matrix corresponding to n individuals on p variables. We denote the rows $X_i \in \mathbb{R}^p$ corresponding the observation for individual i and the columns $x_j \in \mathbb{R}^n$.
The main goal of this paper is to learn the graph structure G from a collection of p-variate independent samples X_i, *as a function of some extraneous covariates* z_i corresponding to the samples. The only assumption on the dependence structure is that the graph parameters vary smoothly with respect to the covariates, that is, if z_i and z_j are similar, then the graph structure corresponding to X_i and X_j will be similar.

Existing Literature

Gozde

There are several approaches to model heterogeneous graphs

- Without using covariate information: These methods depend on the criteria of first splitting the data into homogeneous groups and sharing information withing groups
- Adding the covariates into the mean structure of Gaussian graphical models as multiple linear regressions such that the mean is a continuous function of covariates. This approaches studied from a Bayesian perspective and a frequentist perspective. For this approach still uses the homogeneous graph structure for all observation which we do not want.
- Modeling the underlying covariance matrix as a function of the covariates. The main difficulty of this approach is to enforce sparsity in the precision matrix while being positive definite, as the sparsity in the covariance matrix does not normally carry to the precision matrix through matrix inversion.

The W-PL Approach

Gozde

In this method, a weighted pseudo-likelihood (W-PL) function to obtain a posterior distribution for the graph structure for a fixed individual, with the weights defined as a function of the covariates. Suppose there are n individuals, indexed $i = 1, .., n$. Let the i-th observation in the data set X be denoted as $X_i = (x_{i,1}, ..., x_{i,p})$, which corresponds to the i-th individual. Let $x_{i,-j} \in \mathbb{R}^{p-1}$ denote the vector of the i-th observation including all variables except $x_{i,j}$. This approach tries to model the conditional distribution of each of the x_j's given all other variables, denoted by $X_{-j} \in \mathbb{R}^{n\times(p-1)}$. Let the $p-1$-dimensional vector β_j indicate the regression effect on X_{-j} on x_j. Then the conditional likelihood of x_j denoted by $L(j)$ can be written as

$$L(j) = p(x_j | X_{-j}, \beta_j) \sim \prod_{i=1}^{n} \exp\left\{-(x_{i,j} - x_{i,-j}^T \beta_j)^2 / 2\sigma^2\right\}, \qquad (1)$$

with a possibly sparse coefficient vector β_j.

The W-PL Approach

Then for a fixed graph G the pseudo-likelihood can be calculated as

$$L(G) = \prod_{j=1}^{n} L(j) = \prod_{j=1}^{n} p(x_j | X_{-j}, \beta_j). \tag{2}$$

In this paper different from the previous methods, they define a weighted version of this conditional likelihood for each individual. They assume that the underlying graph structure is a function of extraneous covariates z. Thus, we allow the coefficient vector β_j's to be different for different individuals,depending on the extraneous covariates. $\beta_j^l \in \mathbb{R}^{p-1}$ denotes the coefficient vector corresponding to the regression of the variable x_j on the remaining variables for individual l. Let z_i denote the covariate vector associated with the i-th individual and define $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_n)$. Next, relative to the covariate z, we assign weights $w(z, \mathbf{z}_i) = \phi_\tau(\|z - \mathbf{z}_l\|)$ to every individual where ϕ_τ is the Gaussian density with mean 0 and variance τ^2. When $z = z_l$ corresponds to the

The W-PL Approach

```
## Spike and Slab for Bayesian Approach

\note{
 Gozde

 Next, we put a prior distribution for the coefficient paramet

$$
p_0(\beta_{j,k}^l, \gamma_{j,k}^l) = \prod_{k =1, k \neq j}^n
$$

```{=tex}
}
```
```