Investigating False Positives in Covariate-Dependent Graphical Model

Isaac R, Renat S, Gözde S

Section 1

Background and Introduction

Graphical Modeling

- Undirected graphical models are used 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.
- The conditional independence structure of the distribution can be represented with a graph G.
- Node set $V = (1, \dots, p)$ corresponding to the p variables,
- Edge set E such that $(i, j) \in E$ if and only if x_i and x_j are conditionally dependent given all other variables.
- **Goal:** estimate the underlying graph G from given n i.d.d. observations $x_1, \dots, x_n$.

Graphical Modeling

- Undirected graphical models are used 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.
- The conditional independence structure of the distribution can be represented with a graph G.
- Node set $V = (1, .., p)$ corresponding to the p variables,
- Edge set E such that $(i, j) \in E$ if and only if x_i and x_j are conditionally dependent given all other variables.
- **Goal:** estimate the underlying graph G from given n i.d.d. observations $x_1, \dots, x_n$.

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

- The observations may not be identically distributed
- In this paper(Sutanoy Dasgupta 2022) they suppose the variability in the graph structure across observations depending on additional covariate information.
- $X \in \mathbb{R}^{n \times p}$ stands for the data matrix corresponding to n individuals on p variables
- Rows $X_i \in \mathbb{R}^p$ corresponding the observation for individual i
- Columns $x_j \in \mathbb{R}^n$ corresponding the observation for variable j

- The observations may not be identically distributed
- In this paper(Sutanoy Dasgupta 2022) they suppose the variability in the graph structure across observations depending on additional covariate information.
- $X \in \mathbb{R}^{n \times p}$ stands for the data matrix corresponding to n individuals on p variables
- Rows $X_i \in \mathbb{R}^p$ corresponding the observation for individual i
- Columns $x_j \in \mathbb{R}^n$ corresponding the observation for variable j

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.

Adding Covariates

- 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.

Investigating False Positives in Covariate-Dependent Graphical Model

Background and Introduction

Adding Covariates

Gozde

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.

Adding Covariates

- 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

- 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.
  - Still uses the homogeneous graph structure for all observation
- 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

Existing Literature
- 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.
  - Still uses the homogeneous graph structure for all observation
- 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

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 (Brief Introduction to Pseudo-likelihood approach)

- Suppose there are n individuals, indexed $i = 1, \ldots, n$.
- Let $X_i = (x_{i,1}, \ldots, x_{i,p})$, 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 $x_{i,j}$.
- Model the conditional distribution of each of the x_j's given all other variables, denoted by $X_{-j} \in \mathbb{R}^{n \times (p-1)}$.

The W-PL Approach (Brief Introduction to Pseudo-likelihood approach)

- Suppose there are n individuals, indexed $i = 1, \dots, n$.
- Let $X_i = (x_{i,1}, \dots, x_{i,p})$, 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 $x_{i,j}$.
- Model the conditional distribution of each of the x_j's given all other variables, denoted by $X_{-j} \in \mathbb{R}^{n\times(p-1)}$.

Gozde

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

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

with a possibly sparse coefficient vector β_j. Then for a fixed graph G the pseudo-likelihood can be calculated as

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

The W-PL Approach (Brief Introduction to Pseudo-likelihood approach)

- Let the $(p-1)$-dimensional vector β_j indicate the regression effect on X_{-j} on x_j.
- The conditional likelihood of x_j

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

- for a fixed graph G the pseudo-likelihood $L(G)$

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

The W-PL Approach (Brief Introduction to Pseudo-likelihood approach)

- Let the $(p-1)$-dimensional vector β_j indicate the regression effect on X_{-j} on x_j.
- The conditional likelihood of x_j

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

- for a fixed graph G the pseudo-likelihood $L(G)$

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

Gozde

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

$$L(j) = p(x_j|X_{-j}, \beta_j) \propto \prod_{i=1}^{n} \exp\left\{-(x_{i,j} - x_{i,-j}^T \beta_j)^2/2\sigma^2\right\}, \qquad (3)$$

with a possibly sparse coefficient vector β_j. Then for a fixed graph G the pseudo-likelihood can be calculated as

$$L(G) = \prod_{j=1}^{n} L(j) = \prod_{j=1}^{n} p(x_j|X_{-j}, \beta_j). \qquad (4)$$

The W-PL Approach (Brief Introduction to Pseudo-likelihood approach)

- We now define a weighted version of this conditional likelihood for each individual:
- 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 z.
- $\beta_j^l \in \mathbb{R}^{p-1}$: the coefficient vector corresponding to the regression of the variable x_j on the remaining variables for individual l.

- We now define a weighted version of this conditional likelihood for each individual:
- 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 z.
- $\beta_j^l \in \mathbb{R}^{p-1}$: the coefficient vector corresponding to the regression of the variable x_j on the remaining variables for individual l.

Gozde

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 l-th individual in the study, we use the notation $w_l(\mathbf{z}_i) = w(\mathbf{z}_l, \mathbf{z}_i)$ to denote the weight associated with the i-th individual.

Proposed conditional working model: for $i = 1, \dots, n$,

$$x_{ij}|x_{i,-j}, \mathbf{z}, z \sim N(x_{i,-j}^T \beta_j(z), \sigma^2/w(z, \mathbf{z}_i))$$

The W-PL Approach

- Let z_i denote the covariate vector associated with the i-th individual and $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$
- Define weights $w(z, \mathbf{z}_i) = \phi_\tau(\|z - z_i\|)$ to every individual
- When $z = z_l$ corresponds to the l-th individual, we use the notation $w_l(\mathbf{z}_i) = w(\mathbf{z}_l, \mathbf{z}_i)$ to denote the weight associated with the i-th individual.
- **Proposed conditional working model:** for $i = 1, \ldots, n$,

$$x_{ij} | x_{i,-j}, \mathbf{z}, z \sim N(x_{i,-j}^T \beta_j(z), \sigma^2 / w(z, \mathbf{z}_i))$$

The W-PL Approach

- Let z_i denote the covariate vector associated with the i-th individual and $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$
- Define weights $w(z, \mathbf{z}_i) = \phi_\tau(\|z - z_i\|)$ to every individual
- When $z = z_l$ corresponds to the l-th individual, we use the notation $w_l(\mathbf{z}_i) = w(\mathbf{z}_l, \mathbf{z}_i)$ to denote the weight associated with the i-th individual.
- **Proposed conditional working model:** for $i = 1, \ldots, n$,

$$x_{ij}|x_{i,-j}, \mathbf{z}, z \sim N(x_{i,-j}^T \beta_j(z), \sigma^2/w(z, \mathbf{z}_i))$$

Gozde

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 l-th individual in the study, we use the notation $w_l(\mathbf{z}_i) = w(\mathbf{z}_l, \mathbf{z}_i)$ to denote the weight associated with the i-th individual.

Proposed conditional working model: for $i = 1, \ldots, n$,

$$x_{ij}|x_{i,-j}, \mathbf{z}, z \sim N(x_{i,-j}^T \beta_j(z), \sigma^2/w(z, \mathbf{z}_i))$$

Next let $W(z, \mathbf{z})$ denote the diagonal matrix $\text{Diag}(w(z, \mathbf{z}_1), \ldots, w(z, \mathbf{z}_n))$.

The W-PL Approach

- Let $W(z, \mathbf{z})$ be the diagonal matrix $\text{Diag}(w(z, \mathbf{z}_1), \ldots, w(z, \mathbf{z}_n))$.
- The weighted conditional distribution function can be calculated as

$$p^w(x_j|X_{-j},\beta_j(\mathbf{z}), \mathbf{z}, z) = \left(\prod_{i=1}^{n} \sqrt{\frac{w(z, \mathbf{z}_i)}{2\pi\sigma_*^2}} \right) \exp\left\{ -\frac{(x_j - X_{-j}\beta_j(z))^T W(z, \mathbf{z})(x_j - X_{-j}\beta_j(z))}{2\sigma_*^2} \right\}$$

- The weighted pseudo-likelihood for the graph $G(z)$ corresponding to a covariate value z:

$$L^w(G(z)) = \prod_{j=1}^{n} p^w(x_j|\beta_j(\mathbf{z}), X_{-j}, \mathbf{z}, z)$$

The W-PL Approach

- Let $W(z, \mathbf{z})$ be the diagonal matrix $\text{Diag}(w(z, \mathbf{z}_1), \dots, w(z, \mathbf{z}_n))$.
- The weighted conditional distribution function can be calculated as

$$p^w(x_j|X_{-j}, \beta_j(\mathbf{z}), \mathbf{z}, z) = \left(\prod_{i=1}^{n} \sqrt{\frac{w(z, \mathbf{z}_i)}{2\pi\sigma_*^2}}\right) \exp\left\{-\frac{(x_j - X_{-j}\beta_j(z))^T W(z, \mathbf{z})(x_j - X_{-j}\beta_j(z))}{2\sigma_*^2}\right\}$$

- The weighted pseudo-likelihood for the graph $G(z)$ corresponding to a covariate value z:

$$L^w(G(z)) = \prod_{j=1}^{n} p^w(x_j|\beta_j(\mathbf{z}), X_{-j}, \mathbf{z}, z)$$

Gozde

Next let $W(z, \mathbf{z})$ denote the diagonal matrix $\text{Diag}(w(z, \mathbf{z}_1), \dots, w(z, \mathbf{z}_n))$. The weighted conditional distribution function can be calculated as

$$p^w(x_j|X_{-j}, \beta_j(\mathbf{z}), \mathbf{z}, z) = \left(\prod_{i=1}^{n} \sqrt{\frac{w(z, \mathbf{z}_i)}{2\pi\sigma_*^2}}\right) \exp\left\{-\frac{(x_j - X_{-j}\beta_j(z))^T W(z, \mathbf{z})(z} \tag{5}$$

Then using the previous pseudo-likelihood for the graph G, we now give the weighted pseudo-likelihood for the graph $G(z)$ corresponding to a covariate value z,

$$L^w(G(z)) = \prod_{j=1}^{n} p^w(x_j|\beta_j(\mathbf{z}), X_{-j}, \mathbf{z}, z)$$

Next, we put a prior distribution for the coefficient parameters corresponding to the regression problem described before. Fix an observa-

Spike and Slab for Bayesian Approach

- Next we want to put a prior distribution for the coefficient parameters corresponding to the regression problem described before:
- Fix an observation $l \in \{1, \dots, n\}$ and a variable $j \in \{1, \dots, p\}$: We take a spike-and-slab prior on the parameter β_j^l for $k \in \{1, \dots, p\}$,
- $\beta_{j,k}^l$ is assumed to come from a zero-mean Gaussian density with variance component $\sigma^2 \sigma_\beta^2$ with probability π
- And $\beta_{j,k}^l$ equals to zero with probability $1 - \pi$.

Spike and Slab for Bayesian Approach

- Next we want to put a prior distribution for the coefficient parameters corresponding to the regression problem described before:
- Fix an observation $l \in \{1, \dots, n\}$ and a variable $j \in \{1, \dots, p\}$: We take a spike-and-slab prior on the parameter β_j^l for $k \in \{1, \dots, p\}$,
- $\beta_{j,k}^l$ is assumed to come from a zero-mean Gaussian density with variance component $\sigma^2\sigma_\beta^2$ with probability π
- And $\beta_{j,k}^l$ equals to zero with probability $1 - \pi$.

Gozde

Next,we put a prior distribution for the coefficient parameters corresponding to the regression problem described before. Fix an observation $l \in \{1, \dots, n\}$ and a variable $j \in \{1, \dots, p\}$. Then a spike-and-slab prior on the parameter β_j^l. So for $k \in \{1, \dots, p\}$, $\beta_{j,k}^l$ is assumed to come from a zero-mean Gaussian density with variance component $\sigma^2\sigma_\beta^2$ with probability π and equals to zero with probability $1 - \pi$. Let $\gamma_{j,k}^l$ be the indicator of nonzero $\beta_{j,k}^l$ and we denote it as $\gamma_{j,k}^l = 1\{\beta_{j,k}^l \neq 0\}$ which can be treated as Bernoulli random variables with a common probability of success π. Then we define $\gamma_j^l = (\gamma_{j,1}^l, \dots, \gamma_{j,p}^l)$ and $\Gamma^l = \{\gamma_{j,k}^l, j = 1, \dots, p\}$. Then prior distribution for $(\beta_{j,k}^l, \gamma_{j,k}^l)$ can be written as

$$p_0(\beta_{j,k}^l, \gamma_{j,k}^l) = \prod_{k=1, k\neq j}^{n} \delta_0(\beta_{j,k}^l)^{1-\gamma_{j,k}^l} N(0, \beta_{j,k}^l; 0, \sigma^2\sigma_\beta^2) \prod_{k=1, k\neq j}^{n} \pi^{\gamma_{j,k}^l} (1-\pi)^{\gamma_{j,k}^l}.$$

Spike and Slab for Bayesian Approach

- Let $\gamma_{j,k}^l$ be the indicator of nonzero $\beta_{j,k}^l$ and we define $\gamma_j^l = (\gamma_{j,1}^l, \dots, \gamma_{j,p}^l)$ and $\Gamma^l = \{\gamma_{j,k}^l, j = 1, \dots, p\}$.
- Prior distribution $p_0(\beta_{j,k}^l, \gamma_{j,k}^l)$ for $(\beta_{j,k}^l, \gamma_{j,k}^l)$:

$$\prod_{k=1,k\neq j}^{p} \delta_0(\beta_{j,k}^l)^{1-\gamma_{j,k}^l} N(\beta_{j,k}^l; 0, \sigma^2\sigma_\beta^2) \prod_{k=1,k\neq j}^{p} \pi^{\gamma_{j,k}^l}(1-\pi)^{\gamma_{j,k}^l}.$$

Investigating False Positives in Covariate-Dependent Graphical Model

Spike and Slab for Bayesian Approach

- Let $\gamma_{j,k}^l$ be the indicator of nonzero $\beta_{j,k}^l$ and we define $\gamma_j^l = (\gamma_{j,1}^l, \dots, \gamma_{j,p}^l)$ and $\Gamma^l = \{\gamma_{j,k}^l, j = 1, \dots, p\}$.
- Prior distribution $p_0(\beta_{j,k}^l, \gamma_{j,k}^l)$ for $(\beta_{j,k}^l, \gamma_{j,k}^l)$:

$$\prod_{k=1,k\neq j}^{p} \delta_0(\beta_{j,k}^l)^{1-\gamma_{j,k}^l} N(\beta_{j,k}^l; 0, \sigma^2\sigma_\beta^2) \prod_{k=1,k\neq j}^{p} \pi^{\gamma_{j,k}^l}(1-\pi)^{\gamma_{j,k}^l}.$$

Gozde Let $\gamma_{j,k}^l$ be the indicator of nonzero $\beta_{j,k}^l$ and we denote it as $\gamma_{j,k}^l = 1\{\beta_{j,k}^l \neq 0\}$ which can be treated as Bernoulli random variables with a common probability of success π. Then we define $\gamma_j^l = (\gamma_{j,1}^l, \dots, \gamma_{j,p}^l)$ and $\Gamma^l = \{\gamma_{j,k}^l, j = 1, \dots, p\}$. Then prior distribution for $(\beta_{j,k}^l, \gamma_{j,k}^l)$ can be written as

$$p_0(\beta_{j,k}^l, \gamma_{j,k}^l) = \prod_{k=1,k\neq j}^{n} \delta_0(\beta_{j,k}^l)^{1-\gamma_{j,k}^l} N(0, \beta_{j,k}^l; 0, \sigma^2\sigma_\beta^2) \prod_{k=1,k\neq j}^{n} \pi^{\gamma_{j,k}^l}(1-\pi)^{\gamma_{j,k}^l}.$$

Calculate the posterior distribution

- The full posterior is then given by:

$$p(\beta_j^l, \gamma_j^l | X, Z) \propto \exp\left\{ \sum_{i=1}^{n} \left(x_{ij} - \sum_{k \neq j} x_{ik} \beta_{j,k}^l \right)^2 w_l(\mathbf{z}_i) \right\} p_0(\beta_j^l, \gamma_j^l)$$

- Unfortunately, the posterior distribution is intractable.
- We are going to use variational inference to approximate the posterior distribution.

2022-11-30

Investigating False Positives in Covariate-Dependent Graphical Model

Background and Introduction

Calculate the posterior distribution

Renat

Calculate the posterior distribution

- The full posterior is then given by:

$$p(\beta_j^l, \gamma_j^l | X, Z) \propto \exp\left\{\sum_{i=1}^{n}\left(x_{ij} - \sum_{k \neq j} x_{ik}\beta_{j,k}^l\right)^2 w_l(\mathbf{z}_i)\right\} p_0(\beta_j^l, \gamma_j^l)$$

- Unfortunately, the posterior distribution is intractable.
- We are going to use variational inference to approximate the posterior distribution.

CAVI Updates

- Recall that the variational inference can be formulated as:

$$(\phi_j^l)^* = \arg\min_{\phi_j^l} D_{KL}\left[q(\beta_j^l, \gamma_j^l|\phi_j^l)||p(\beta_j^l, \gamma_j^l|X, Z)\right]$$

- We use the blocked mean-field approach which assumes that the variational distribution is factorized as follows:

$$q(\beta_j^l, \gamma_j^l|\phi_j^l) = \prod_{j \neq k} q_k(\beta_{j,k}^l, \phi_{j,k}^l; \phi_{j,k}^l)$$

- Further for each factor q_k of the variational distribution we assume individual spike-and-slab density:

$$q_k(\beta_{j,k}^l, \gamma_{j,k}^l; \phi_{j,k}^l) = N(\beta_{j,k}^l; \mu_{j,k}^l, (s_{j,k}^l)^2)^{\gamma_{j,k}^l} \delta_0(\beta_{j,k}^l)^{1-\gamma_{j,k}^l} (\alpha_{j,k}^l)^{\gamma_{j,k}^l} (1-\alpha_{j,k}^l)^{1-\gamma_{j,k}^l}$$

Investigating False Positives in Covariate-Dependent Graphical Model

Background and Introduction

CAVI Updates

Renat

CAVI Updates

- Recall that the variational inference can be formulated as:

$$(\phi_j^l)^* = \arg\min_{\phi_j^l} D_{KL}\left[q(\beta_j^l, \gamma_j^l|\phi_j^l)||p(\beta_j^l, \gamma_j^l|X, Z)\right]$$

- We use the blocked mean-field approach which assumes that the variational distribution is factorized as follows:

$$q(\beta_j^l, \gamma_j^l|\phi_j^l) = \prod_{j \neq k} q_k(\beta_{j,k}^l, \phi_{j,k}^l; \phi_{j,k}^l)$$

- Further for each factor q_k of the variational distribution we assume individual spike-and-slab density:

$$q_k(\beta_{j,k}^l, \gamma_{j,k}^l; \phi_{j,k}^l) = N(\beta_{j,k}^l; \mu_{j,k}^l, (s_{j,k}^l)^2)^{\gamma_{j,k}^l} \delta_0(\beta_{j,k}^l)^{1-\gamma_{j,k}^l} \\ (\alpha_{j,k}^l)^{\gamma_{j,k}^l}(1-\alpha_{j,k}^l)^{1-\gamma_{j,k}^l}$$

CAVI Updates

- Deriving the ELBO based on the assumed variational distribution and taking partial derivatives with respect to the variational parameters we obtain the following CAVI updates:

$$(s^l_{j,k})^2 = \frac{\sigma^2}{1/\sigma^2_\beta + \sum_{i=1}^n x^2_{ik} w_l(\mathbf{z}_i)}$$

$$\mu^l_{j,k} = \frac{(s^l_{j,k})^2}{\sigma^2} \sum_{i=1}^n \left\{ w_l(\mathbf{z}_i) x_{ik} \left(x_{ij} - \sum_{m \neq j,k} x_{im} \mu^l_{j,m} \alpha^l_{j,m} \right) \right\}$$

$$\text{logit}(\alpha^l_{j,k}) = \text{logit}(\pi) + \left(\frac{\mu^l_{j,k}}{\sqrt{2} s^l_{j,k}} \right)^2 + \log \frac{s^l_{j,k}}{\sigma \sigma_\beta}$$

- The resulting inclusion probabilities $\alpha^l_{j,k}$ are symmetrized to obtain valid graphical model as:

$$\alpha^l_{j,k} = \frac{1}{2} \left(\alpha^l_{j,k} + \alpha^l_{k,j} \right)$$

2022-11-30

Investigating False Positives in Covariate-Dependent Graphical Model

Background and Introduction

CAVI Updates

Renat

CAVI Updates

- Deriving the ELBO based on the assumed variational distribution and taking partial derivatives with respect to the variational parameters we obtain the following CAVI updates:

$$(s^l_{j,k})^2 = \frac{\sigma^2}{1/\sigma^2_\beta + \sum_{i=1}^n x^2_{ik} w_l(\mathbf{z}_i)}$$

$$\mu^l_{j,k} = \frac{(s^l_{j,k})^2}{\sigma^2} \sum_{i=1}^n \left\{ w_l(\mathbf{z}_i) x_{ik} \left(x_{ij} - \sum_{m \neq j,k} x_{im} \mu^l_{j,m} \alpha^l_{j,m} \right) \right\}$$

$$\text{logit}(\alpha^l_{j,k}) = \text{logit}(\pi) + \left(\frac{\mu^l_{j,k}}{\sqrt{2} s^l_{j,k}} \right)^2 + \log \frac{s^l_{j,k}}{\sigma \sigma_\beta}$$

- The resulting inclusion probabilities $\alpha^l_{j,k}$ are symmetrized to obtain valid graphical model as:

$$\alpha^l_{j,k} = \frac{1}{2} \left(\alpha^l_{j,k} + \alpha^l_{k,j} \right)$$

The False Positive Problem

- This model demonstrates superior sensitivity than competing methods(Helwig et al. 2022)
- It suffers from lower specificity
  - Compared with competing methods from the `mgm` (Haslbeck and Waldorp 2020) and `JGL` (Danaher 2018) packages, specificity gets substantially worse with p.
- For gene expression data, this could lead to worse outcomes than a less sensitive and more specific model
  - The cost of carrying out experiments which show a lack of a predicted relation may be very expensive
- We want to **increase the specificity and preserve the model's sensitivity and speed**.

The False Positive Problem

- This model demonstrates superior sensitivity than competing methods(Helwig et al. 2022)
- It suffers from lower specificity
  - Compared with competing methods from the `mgm` (Haslbeck and Waldorp 2020) and `JGL` (Danaher 2018) packages, specificity gets substantially worse with p.
- For gene expression data, this could lead to worse outcomes than a less sensitive and more specific model
  - The cost of carrying out experiments which show a lack of a predicted relation may be very expensive
- We want to **increase the specificity and preserve the model's sensitivity and speed**.

Isaac

Despite this model demonstrating superior sensitivity to true dependence relations than competing methods, it suffers from lower specificity. Compared with competing methods from the ‘mgm‘ and ‘JGL‘packages, the specificity gets substantially worse as the number of features increases. In the case of analyzing gene expression data, this could lead to worse outcomes than a less sensitive and more specific model since the cost of carrying out experiments which show a lack of a predicted relation may be very expensive. Further, validating a true but relatively weak relationship may not be desirable considering the cost. Ideally we want to increase the specificity of the model without substantially hurting the model's sensitivity and speed.

The False Positive Problem

p	n	Package	Sensitivity (sd)	FP/graph (sd)
25	150	‘covdepGE‘	0.80 (0.08)	1.29 (0.98)
25	150	‘JGL‘	0.72 (0.19)	1.96 (1.92)
25	150	‘mgm‘	0.61 (0.11)	0.09 (0.22)
50	150	‘covdepGE‘	0.74 (0.10)	4.36 (2.22)
50	150	‘JGL‘	0.66 (0.22)	1.69 (1.90)
50	150	‘mgm‘	0.52 (0.10)	0.06 (0.15)
100	300	‘covdepGE‘	0.85 (0.06)	18.21 (3.89)
100	300	‘JGL‘	0.74 (0.11)	1.52 (1.20)
100	300	‘mgm‘	0.68 (0.09)	0.02 (0.08)

Table 1: Comparison of existing methodologies

Investigating False Positives in Covariate-Dependent Graphical Model

p	n	Package	Sensitivity (sd)	FP/graph (sd)
25	150	‘covdepGE‘	0.80 (0.08)	1.29 (0.98)
25	150	‘JGL‘	0.72 (0.19)	1.96 (1.92)
25	150	‘mgm‘	0.61 (0.11)	0.09 (0.22)
50	150	‘covdepGE‘	0.74 (0.10)	4.36 (2.22)
50	150	‘JGL‘	0.66 (0.22)	1.69 (1.90)
50	150	‘mgm‘	0.52 (0.10)	0.06 (0.15)
100	300	‘covdepGE‘	0.85 (0.06)	18.21 (3.89)
100	300	‘JGL‘	0.74 (0.11)	1.52 (1.20)
100	300	‘mgm‘	0.68 (0.09)	0.02 (0.08)

Table 1: Comparison of existing methodologies

Isaac

We’ll look at a variety of different modifications to the algorithm in order of least to most extensive changes necessary, and use the ‘covdepGE‘ package in order to generate data for simulation studies. For some of these solutions, the package’s functions will be used without modification and changes will occur outside of the inference algorithm. Otherwise, any changes to the functions will be explicitly noted.

Section 2

Attempted Solutions

Feature & Covariate Scaling

- We believe the underlying cause of the false positive problem is having the same prior inclusion probability π across all spike-and-slab regressions.
- The first modification we'll make is to scale the features and covariates. The idea is that by scaling the features and covariates, we can make a common prior π a more reasonable choice.
  - Min-max scaling ($\sim U(0,1)$),
  - Standardization ($\sim N(0,1)$).
- Pros:
  - Easy to implement.
  - No changes to the inference algorithm.
- Cons:
  - Ad-hoc solution.

2022-11-30

Investigating False Positives in Covariate-Dependent Graphical Model

Attempted Solutions

Feature & Covariate Scaling

Renat

Feature & Covariate Scaling

- We believe the underlying cause of the false positive problem is having the same prior inclusion probability π across all spike-and-slab regressions.
- The first modification we'll make is to scale the features and covariates. The idea is that by scaling the features and covariates, we can make a common prior π a more reasonable choice.
  - Min-max scaling ($\sim U(0,1)$),
  - Standardization ($\sim N(0,1)$).
- Pros:
  - Easy to implement.
  - No changes to the inference algorithm.
- Cons:
  - Ad-hoc solution.

Multiple Prior Inclusion Probabilities

- The second approach is to allow for multiple prior inclusion probabilities π:
  - Separate π for each individual;
  - Cluster and assign separate π for each cluster.
- Pros:
  - Theoretically and intuitively plausible.
- Cons:
  - Requires additional steps (clustering, assigning π to each cluster).
  - May be computationally prohibitive (due to hyper-parameter search for different π's).

Investigating False Positives in Covariate-Dependent Graphical Model

Attempted Solutions

Multiple Prior Inclusion Probabilities

Renat

Multiple Prior Inclusion Probabilities

- The second approach is to allow for multiple prior inclusion probabilities π:
  - Separate π for each individual;
  - Cluster and assign separate π for each cluster.
- Pros:
  - Theoretically and intuitively plausible.
- Cons:
  - Requires additional steps (clustering, assigning π to each cluster).
  - May be computationally prohibitive (due to hyper-parameter search for different π's).

Section 3

Simulations

Normalization (Baseline)

- For each covariate, apply a Z-transform
  - $z(Z) = \frac{Z - \bar{Z}}{\sigma_Z}$
- For each variable, center at 0
  - $c(X) = X - \bar{X}$
- Simulations are replicated 100 times each

p	n	Baseline FPs (sd)	Baseline FNs (sd)
5	90	0.34 (0.54)	0.99 (0.53)
15	90	0.82 (0.94)	1.41 (0.69)
25	150	1.29 (0.98)	0.91 (0.38)
50	150	4.36 (2.22)	1.19 (0.48)

Table 2: False Positives and False Negatives per sample - Normalized Z, Centered X

Normalization (Baseline)

- For each covariate, apply a Z-transform
  - $z(Z) = \frac{Z - \bar{Z}}{\sigma_Z}$
- For each variable, center at 0
  - $c(X) = X - \bar{X}$
- Simulations are replicated 100 times each

p	n	Baseline FPs (sd)	Baseline FNs (sd)
5	90	0.34 (0.54)	0.99 (0.53)
15	90	0.82 (0.94)	1.41 (0.69)
25	150	1.29 (0.98)	0.91 (0.38)
50	150	4.36 (2.22)	1.19 (0.48)

Table 2: False Positives and False Negatives per sample - Normalized Z, Centered X

Isaac The first approach will be to use a different or additional approach to feature scaling on X and/or Z in order to try and make a singular prior inclusion probability more appropriate. Currently, the default behavior in the ‘covdepGE‘ function is to perform a columnwise Z-score Normalization on Z and a columnwise 0 centering on X. For brevity we’ll call this procedure ”normalization”. The baseline performance under this scheme is given below. All experiments were run under 4 different setups each having different values for p and n, and data simulated using the ‘generateData‘ function. To assess sensitivity and specificity, we’ll examine the number of false positives per sample and number of false negatives per sample across 100 replications of each simulation setup. So, in all cases lower numbers are desirable. First, we’ll look at the baseline performance of the existing function with no changes to the default behavior.

Min/Max Scaling

- What if we used an alternative method of feature scaling than the default normalization?
- Define Min-Max scaling of a vector by $f(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$
- Rescaling versus Normalization

Min/Max Scaling

- What if we used an alternative method of feature scaling than the default normalization?
- Define Min-Max scaling of a vector by $f(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$
- Rescaling versus Normalization

Isaac

Next, we wanted to consider a different scaling method. We tried 3 situations; applying min-max scaling columnwise to both X and Z, only Z, and only X. Notably, a major difference between Z-score normalization and 'range scaling' is that we aren't changing the shape of the distribution. However, we do still condense the range of the variables to be within [0,1]. It's also worth noting that this approach is not robust to outliers, but for our simulation examples this isn't considered.

Unfortunately, applying only a Max-Min scaling to any of these setups is ineffective. When we apply this scaling to X, we do get a perfect false positive rate; but only because it fails to include any edges at all. Using the raw X and a scaled Z leads to uniformly worse sensitivity and specificity.

p	n	Baseline FPs (sd)	MM on Z	MM on X	MM on Z, X
5	90	0.34 (0.54)	0.81 (0.63)	0 (0)	0 (0)
15	90	0.82 (0.94)	4.08 (1.44)	0 (0)	0 (0)
25	150	1.29 (0.98)	11.89 (2.5)	0 (0)	0 (0)
50	150	4.36 (2.22)	19.86 (3.59)	0 (0)	0 (0)

Table 3: False **positives** per sample - Max Min Scaling only; (sd)

p	n	Baseline FNs (sd)	MM on Z	MM on X	MM on Z, X
5	90	0.99 (0.53)	1.14 (0.49)	4.67 (0)	4.67 (0)
15	90	1.41 (0.69)	2.35 (0.78)	4.67 (0)	4.67 (0)
25	150	0.91 (0.38)	1.71 (0.7)	4.67 (0)	4.67 (0)
50	150	1.19 (0.48)	2.42 (0.71)	4.67 (0)	4.67 (0)

Table 4: False **negatives** per sample - Max Min Scaling only; (sd)

Combination of Scaling & Normalization

- Can we get the best of both worlds?
- Max-min transform followed by Z-transform
  - $\tilde{Z} = (z \circ f)(Z)$
- Max-min transform followed by 0-centering
  - $\tilde{X} = (c \circ f)(X)$

Combination of Scaling & Normalization

- Can we get the best of both worlds?
- Max-min transform followed by Z-transform
  - $\tilde{Z} = (z \circ f)(Z)$
- Max-min transform followed by 0-centering
  - $\tilde{X} = (c \circ f)(X)$

Isaac

So, what if we could get the best of both worlds? First do a max-min transform to scale, then do the z-transform (or mean 0 center transform in the case of X). This way we are still dealing with nice Gaussian distributions after the transformations, but gain some control over the range of X.

As it turns out, this works and ended up being our best result. In particular, doing this procedure on both X and Z was best overall and resulted in slightly better specificity and little change in sensitivity

p	n	Baseline FPs (sd)	MM/N Z	MM/N X	MM/N Z, X
5	90	0.34 (0.54)	0.24 (0.47)	0.26 (0.47)	0.24 (0.44)
15	90	0.82 (0.94)	0.63 (0.83)	0.61 (0.77)	0.61 (0.77)
25	150	1.29 (0.98)	1.34 (0.99)	1.34 (0.97)	1.35 (0.98)
50	150	4.36 (2.22)	4.08 (1.89)	4.07 (1.84)	4.07 (1.82)

Table 5: False **positives** per sample - Max Min Scaling + Normalization; (sd)

p	n	Baseline FNs (sd)	MM/N Z	MM/N X	MM/N Z, X
5	90	0.99 (0.53)	1.03 (0.45)	1 (0.47)	1.02 (0.46)
15	90	1.41 (0.69)	1.53 (0.65)	1.48 (0.63)	1.47 (0.65)
25	150	0.91 (0.38)	0.95 (0.46)	0.92 (0.44)	0.92 (0.44)
50	150	1.19 (0.48)	1.15 (0.44)	1.12 (0.43)	1.11 (0.43)

Table 6: False **negatives** per sample - Max Min Scaling + Normalization; (sd)

Oracle Clustering for Multiple π

Suppose we know the true number of clusters K and the true cluster assignments C. This is true in the simulation setting.

p	n	Baseline FPs (sd)	Mean (sd)
5	90	0.34 (0.54)	0.45 (0.62)
15	90	0.82 (0.94)	1.25 (0.97)
25	150	1.29 (0.98)	3.24 (1.39)
50	150	4.36 (2.22)	9.16 (2.42)

Table 7: False **positives** per sample - Multiple PIP with Oracle Clustering

p	n	Baseline FNs (sd)	Mean (sd)
5	90	0.99 (0.53)	1.09 (0.58)
15	90	1.41 (0.69)	1.53 (0.66)
25	150	0.91 (0.38)	0.99 (0.42)
50	150	1.19 (0.48)	1.01 (0.36)

Table 8: False **negatives** per sample - Multiple PIP with Oracle Clustering

2022-11-30

Investigating False Positives in Covariate-Dependent Graphical Model

Simulations

Oracle Clustering for Multiple π

Renat

Oracle Clustering for Multiple π

Suppose we know the true number of clusters K and the true cluster assignments C. This is true in the simulation setting.

p	n	Baseline FPs (sd)	Mean (sd)
5	90	0.34 (0.54)	0.45 (0.62)
15	90	0.82 (0.94)	1.25 (0.97)
25	150	1.29 (0.98)	3.24 (1.39)
50	150	4.36 (2.22)	9.16 (2.42)

Table 7: False **positives** per sample - Multiple PIP with Oracle Clustering

p	n	Baseline FNs (sd)	Mean (sd)
5	90	0.99 (0.53)	1.09 (0.58)
15	90	1.41 (0.69)	1.53 (0.66)
25	150	0.91 (0.38)	0.99 (0.42)
50	150	1.19 (0.48)	1.01 (0.36)

Table 8: False **negatives** per sample - Multiple PIP with Oracle Clustering

Hierarchical Clustering for Multiple π

We can also try empirical clustering approaches.

p	n	clusts	Baseline Mean (sd)	Mean (sd)
5	90	2	0.34 (0.54)	0.47 (0.66)
5	90	3	0.34 (0.54)	0.47 (0.68)
5	90	6	0.34 (0.54)	0.48 (0.64)
15	90	2	0.82 (0.94)	1.07 (0.98)
15	90	3	0.82 (0.94)	1.21 (0.99)
15	90	6	0.82 (0.94)	1.41 (1)
25	150	2	1.29 (0.98)	2.71 (1.37)
25	150	3	1.29 (0.98)	3.2 (1.32)
25	150	6	1.29 (0.98)	3.54 (1.36)
50	150	2	4.36 (2.22)	7.76 (2.3)
50	150	3	4.36 (2.22)	8.78 (2.43)
50	150	6	4.36 (2.22)	10.14 (2.46)

Table 9: False **positives** per sample - Multiple PIP with Hierarchical Clustering

2022-11-30

Investigating False Positives in Covariate-Dependent Graphical Model

Simulations

Hierarchical Clustering for Multiple π

Renat

Hierarchical Clustering for Multiple π

We can also try empirical clustering approaches.

p	n	clusts	Baseline Mean (sd)	Mean (sd)
5	90	2	0.34 (0.54)	0.47 (0.66)
5	90	3	0.34 (0.54)	0.47 (0.68)
5	90	6	0.34 (0.54)	0.48 (0.64)
15	90	2	0.82 (0.94)	1.07 (0.98)
15	90	3	0.82 (0.94)	1.21 (0.99)
15	90	6	0.82 (0.94)	1.41 (1)
25	150	2	1.29 (0.98)	2.71 (1.37)
25	150	3	1.29 (0.98)	3.2 (1.32)
25	150	6	1.29 (0.98)	3.54 (1.36)
50	150	2	4.36 (2.22)	7.76 (2.3)
50	150	3	4.36 (2.22)	8.78 (2.43)
50	150	6	4.36 (2.22)	10.14 (2.46)

Table 9: False **positives** per sample - Multiple PIP with Hierarchical Clustering

Hierarchical Clustering for Multiple $\pi(cont.)$

p	n	clusts	Baseline Mean (sd)	Mean (sd)
5	90	2	0.99 (0.53)	1.08 (0.61)
5	90	3	0.99 (0.53)	1.09 (0.6)
5	90	6	0.99 (0.53)	1.09 (0.61)
15	90	2	1.41 (0.69)	1.53 (0.69)
15	90	3	1.41 (0.69)	1.51 (0.67)
15	90	6	1.41 (0.69)	1.52 (0.67)
25	150	2	0.91 (0.38)	1 (0.42)
25	150	3	0.91 (0.38)	0.99 (0.43)
25	150	6	0.91 (0.38)	0.99 (0.42)
50	150	2	1.19 (0.48)	1.06 (0.37)
50	150	3	1.19 (0.48)	1.04 (0.37)
50	150	6	1.19 (0.48)	1 (0.37)

Table 10: False **negatives** per sample - Multiple PIP with Hierarchical Clustering

Investigating False Positives in Covariate-Dependent Graphical Model

Simulations

Hierarchical Clustering for Multiple $\pi(cont.)$

Renat

Hierarchical Clustering for Multiple $\pi(cont.)$

p	n	clusts	Baseline Mean (sd)	Mean (sd)
5	90	2	0.99 (0.53)	1.08 (0.61)
5	90	3	0.99 (0.53)	1.09 (0.6)
5	90	6	0.99 (0.53)	1.09 (0.61)
15	90	2	1.41 (0.69)	1.53 (0.69)
15	90	3	1.41 (0.69)	1.51 (0.67)
15	90	6	1.41 (0.69)	1.52 (0.67)
25	150	2	0.91 (0.38)	1 (0.42)
25	150	3	0.91 (0.38)	0.99 (0.43)
25	150	6	0.91 (0.38)	0.99 (0.42)
50	150	2	1.19 (0.48)	1.06 (0.37)
50	150	3	1.19 (0.48)	1.04 (0.37)
50	150	6	1.19 (0.48)	1 (0.37)

Table 10: False **negatives** per sample - Multiple PIP with Hierarchical Clustering

Individual π for Each Observation

- We experimented with giving every observation its own prior inclusion probability.
- Simulations became computationally unfeasible due to the dimension of the parameter space.
- After multiple days of running the simulation crashed due to excessive memory consumption.
- Based on the previous results with the oracle clustering and hierarchical clustering, likely worse.

Individual π for Each Observation

- We experimented with giving every observation its own prior inclusion probability.
- Simulations became computationally unfeasible due to the dimension of the parameter space.
- After multiple days of running the simulation crashed due to excessive memory consumption.
- Based on the previous results with the oracle clustering and hierarchical clustering, likely worse.

Isaac What if every observation had its own π? We experimented with giving every observation its own prior inclusion probability by specifying the cluster mapping as $\{1, 2, ..., n\}$ however the simulations became computationally unfeasible due to the dimension of the parameter space being grid searched over scaling linearly with n. After multiple days of running the simulation crashed due to excessive memory consumption (>32GB). Based on the previous results with the oracle clustering and hierarchical clustering, and a much smaller example with $p = 5, n = 10$, we believe it highly unlikely that having a different prior inclusion probability for every observation would improve the false positive rate.

Section 4

Discussions

Why does Max/Min Alone Fail?

- Our assumptions are being violated:
  - Realizations of X may not quite be zero-mean Gaussian and we are not correcting for this
  - Max-Min scaling on X exacerbates this issue by constantly centering at a positive value
  - Max-Min scaling is not robust to outliers, so our scaled Z is inconsistent and the weighting function performs poorly

Why does Max/Min Alone Fail?

- Our assumptions are being violated:
  - Realizations of X may not quite be zero-mean Gaussian and we are not correcting for this
  - Max-Min scaling on X exacerbates this issue by constantly centering at a positive value
  - Max-Min scaling is not robust to outliers, so our scaled Z is inconsistent and the weighting function performs poorly

Isaac We believe that the max min scaling alone performs very badly because the assumptions of our algorithm are violated. In particular, our X is not zero-mean Gaussian and we aren't correcting for this fact (and are even making it worse when we move it to always be positively centered). And since max min is not robust to outliers, and our true Z is generated from uniform distributions, rescaling is likely only going to cause us to capture less information about how the conditional dependency is scaling with Z.

Why does Max/Min + Normalization Succeed?

- Doesn't require distributional changes
  - It isn't a major change over the baseline, assumptions are unchanged
- Gives control over the range of variables in X without affecting its shape
  - Tightening the range of X may make the assumption of a common prior inclusion probability more reasonable
- Still allows for controlling the shape of Z for good weight performance, and normalization helps with outlier robustness

Why does Max/Min + Normalization Succeed?

- Doesn't require distributional changes
  - It isn't a major change over the baseline, assumptions are unchanged
- Gives control over the range of variables in X without affecting its shape
  - Tightening the range of X may make the assumption of a common prior inclusion probability more reasonable
- Still allows for controlling the shape of Z for good weight performance, and normalization helps with outlier robustness

Isaac We believe that max min scaling combined with the Z-tranform on covariates and centering on variables works because we are able to control the range of X without having to affect the shapes of the distributions involved. And, we are still controlling the shape of Z so our weight function should perform well.

Why does Multiple π Values Fail?

- Overfitting to the data.
  - We see that the false negatives go down as the number of clusters goes up.
  - On the other hand, the false positives go up as the number of clusters goes up.
  - This is because the number of clusters is increasing the number of parameters in the model.

Investigating False Positives in Covariate-Dependent Graphical Model

Discussions

Why does Multiple π Values Fail?

Renat

Why does Multiple π Values Fail?

- Overfitting to the data.
  - We see that the false negatives go down as the number of clusters goes up.
  - On the other hand, the false positives go up as the number of clusters goes up.
  - This is because the number of clusters is increasing the number of parameters in the model.

Additional Experiments to Try

- Model Mispecification
  - Injecting non-Gaussianity into X - does the scaling improve model robustness?
  - Contaminating X with an unrelated, independent Gaussian distribution - can max/min scaling help?
  - Sharp Discontinuities in Z - can max/min scaling help?

Additional Experiments to Try

- Model Mispecification
  - Injecting non-Gaussianity into X - does the scaling improve model robustness?
  - Contaminating X with an unrelated, independent Gaussian distribution - can max/min scaling help?
  - Sharp Discontinuities in Z - can max/min scaling help?

Isaac In order to verify our results, there are a few more experiments we want to try running. In particular, we want to see whether we can break the Gaussian assumption of our true data generating function. The hope is that the scaling on X will improve the model's robustness to a distribution with fatter tails such as a t distribution with low degrees of freedom. Similarly, we want to try adding a small percentage of 'contaminated' observations who are drawn from an unrelated, independent Gaussian distribution to the one we are trying to work with. We again hope that the additional scaling we do can help combat the effects of the bad data. Finally, we want to see if we can use the additional scaling of Z to account for potential non-smoothness; for example if Z was double exponentially distributed.

References

Danaher, Patrick. 2018. "JGL: Performs the Joint Graphical Lasso for Sparse Inverse Covariance Estimation on Multiple Classes." https://CRAN.R-project.org/package=JGL.

Haslbeck, Jonas M. B., and Lourens J. Waldorp. 2020. "{Mgm}: Estimating Time-Varying Mixed Graphical Models in High-Dimensional Data" 93. https://doi.org/10.18637/jss.v093.i08.

Helwig, Jacob, Sutanoy Dasgupta, Peng Zhao, Bani Mallick, and Debdeep Pati. 2022. "covdepGE: Covariate Dependent Graph Estimation." https://CRAN.R-project.org/package=covdepGE.

Sutanoy Dasgupta, Prasenjit Ghost, Peng Zhao. 2022. "An Approximate Bayesian Approach to Covariate-Dependent Graphical Modeling."