An Approximate Bayesian Approach to Covariate-dependent Graphical Modeling

February 8, 2023

Abstract

Gaussian graphical models typically assume a homogeneous structure across all subjects, which is often restrictive in applications. In this article, we propose a weighted pseudo-likelihood approach for graphical modeling which allows different subjects to have different graphical structures depending on extraneous covariates. The pseudo-likelihood approach replaces the joint distribution by a product of the conditional distributions of each variable. We cast the conditional distribution as a heteroscedastic regression problem, with covariate-dependent variance terms, to enable information borrowing directly from the data instead of a hierarchical framework. This allows independent graphical modeling for each subject, while retaining the benefits of a hierarchical Bayes model and being computationally tractable. An efficient embarrassingly parallel variational algorithm is developed to approximate the posterior and obtain estimates of the graphs. Using a fractional variational framework, we derive asymptotic risk bounds for the estimate in terms of a novel variant of the α-Rényi divergence. We theoretically demonstrate the advantages of information borrowing across covariates over independent modeling. We show the practical advantages of the approach through simulation studies and illustrate the dependence structure in protein expression levels on breast cancer patients using CNV information as covariates.

Keywords: Bayesian Gaussian graphical model, heterogeneous graphs, mean-field, pseudo-likelihood, variational inference.

1 Introduction

Undirected graphical models provide a widely used framework for modeling multivariate distributions, with applications ranging across diverse disciplines such as statistical physics, bioinformatics, computational biology and sociology. Here, one exploits the structure in the distribution in the form of assumptions of conditional independence among the involved variables. Suppose we observe a p-dimensional sample $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ 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 $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ is characterized by a node set $\mathrm{V} = (1, 2, \ldots, p)$ corresponding to the p variables, and an edge set E such that $(i, j) \in \mathrm{E}$ if, and only if, x_i and x_j are conditionally dependent given all other variables.

Several methods have been developed with the goal of estimating this underlying graph G given n independent and identically distributed observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, such as Friedman et al. (2008); Yuan & Lin (2007); Giudici & Green (1999). However, in practice, the n observations might not be identically distributed, that is, there is no *homogeneous* underlying graph describing the conditional dependence structure among the variables for all the observations. It is imperative, therefore, to develop efficient graph modeling schemes that can take into account the variability in the graph structure across observations depending on additional covariate information.

1.1 Current Literature

Perhaps surprisingly, the literature on handling this *heterogeneity* in the underlying graph structure is relatively sparse. Some approaches attempt to model heterogeneous graphs without using covariate information, as in Guo et al. (2011); Danaher et al. (2014); Peterson et al. (2015); Ha et al. (2015); Ren et al. (2022). These methods depend on the criteria of first splitting the data into homogeneous groups and then sharing information within and across groups as appropriate. However, a clear criterion for the choice of homogeneous groups is difficult to obtain without

extraneous data, and the performance can suffer when the identified groups have small samples. A second approach focuses on adding the covariates into the mean structure of Gaussian graphical models as multiple linear regressions such that the mean is a continuous function of the covariates. Bhadra & Mallick (2013) proposed a Bayesian joint model for estimating the mean structure and the graph together. Yin & Li (2011); Cai et al. (2013); Lee & Liu (2012) studied similar models from a frequentist perspective. Such an approach estimates the graph after eliminating the effects of the covariates from its mean structure. However, the graph structure is still assumed to be homogeneous for all observations. Another approach for estimating heterogeneous graphs is to model the underlying covariance matrix as a function of the covariates, as considered in Hoff & Niu (2012); Fox & Dunson (2015); Pourahmadi (1999, 2000, 2013); Zhang & Leng (2012). The main challenge here 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. Recently, Liu et al. (2010) developed a graph-valued regression model that partitions the covariate space into different groups by classification and regression trees (CART). This method assumes that there exists a true partition of the covariate space such that the graph structure is homogeneous inside each of the partitions. Second, tree structures may not be flexible enough to capture the true partition, even if such a partition exists. Ni et al. (2019) proposed a graphical regression method that estimates covariate-dependent continuously varying directed acyclic graphs (DAGs). But, the conditional dependence structure cannot be extended to undirected graphs. In related literature, Kolar et al. (2010a); Wang & Kolar (2014) developed a penalized kernel smoothing method for conditional precision matrices under an additional simplifying assumption that the precision matrix is a function of a low-dimensional index variable. Kolar et al. (2010b); Zhou et al. (2010); Qiu et al. (2016) proposed methods for inferring time-varying graphs, which are however, difficult to extend to non-time indexed covariates.

1.2 Proposed formulation

In what follows, $\mathbf{X} \in \mathbb{R}^{n \times p}$ refers to the data matrix corresponding to n individuals on p variables, with the rows $\mathbf{X}_i \in \mathbb{R}^p$ corresponding to the observation on individual i. The columns $x_j \in \mathbb{R}^n, j = \{1, \ldots, p\}$ correspond to the p variables. The main goal of this paper is to learn the graph structure G from a collection of p-variate independent samples $\mathbf{X}_i$, as a function of some extraneous covariates $\mathbf{z}_i$ 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 $\mathbf{z}_i$ and $\mathbf{z}_j$ are similar, then the graph structure corresponding to $\mathbf{X}_i$ and $\mathbf{X}_j$ will be similar. To the best of our knowledge, there is no method available in the literature that can model the graph itself as a continuous function of covariates without putting additional restrictive or simplifying assumptions on the dependence structure of the graphs on the covariates. A natural way to achieve the sharing of information through covariates is to consider a hierarchical model in a Bayesian setting. Embedding a complex graphical modeling framework in a hierarchical setting involves manifold challenges. In the following, we develop a novel weighted pseudo-likelihood based approach that obviates these challenges, which is computationally efficient and yet retains all the benefits of a hierarchical model.

Our modeling scheme can be organized into two main steps. First, we use a novel weighted pseudo-likelihood (W-PL) function (described in Section 2) to obtain a posterior distribution for the graph structure for a fixed individual, with the weights defined as a function of the covariates. The idea of a pseudo-likelihood approach is to tackle each of the variables $x_j, j = \{1, 2, \ldots, p\}$ separately instead of trying to jointly model them. See, for instance, Meinshausen et al. (2006) and Atchadé (2019) for a more detailed discussion in this context. It is important to note that the pseudo-likelihood model is not a valid probability model, as the conditional distributions for a multivariate Gaussian distribution do not determine the joint distribution. However, consistency and other benefits of the pseudo-likelihood models have been extensively explored in Besag (1975), demonstrating the efficacy of such an approach. The standard pseudo-likelihood

approach replaces the original joint likelihood function by the product of the conditional likelihoods of the random variables x_js. Thus, this approach casts the conditional distribution of each of the variables x_j given the remaining variables as a standard homoscedastic regression problem. Instead, we cast the conditional distribution as a *weighted* regression problem, by introducing covariate-dependent weights in the error variance term, which leads to a weighted pseudo-likelihood function.

Second, we use a variational algorithm to efficiently approximate the posterior distribution and obtain an estimate of the graph for a fixed individual. Repeating this process for every individual, we obtain an empirical distribution of the graph structure over the support of the covariates associated with the individuals. The advantages of this two-step approach are manifold.

Embarrassingly parallel: The approach allows independent estimation of the graph structure parameters for different individuals, by sharing information across individuals directly from the data rather than through the parameters themselves.

Borrowing of information: Observe that the standard approach to sharing information across the parameters would be to consider a full-blown hierarchical Bayesian model. To illustrate this idea, consider the following simple setup as an example. Assume we have k groups $\{y_{ij}, i = 1, \ldots, n_j, j = 1, \ldots, k\}$ and $\bar{y}_j$ denotes the mean of the j-th group. Instead of considering a hierarchical model to estimate the true group mean θ_j, consider for every fixed $j \in \{1, \ldots, k\}$,

$$\bar{y}_l \mid \theta_j \sim \mathcal{N}(\theta_j, (\sigma^2/n_l) w(\bar{y}_j, \bar{y}_l)^{-1}), l = 1, \ldots, k,$$

for some similarity function $w(x, y)$ which takes higher values for $x \approx y$, and lower values as x moves further away from y. Then, assuming a balanced sample, the weighted MLE of θ_j is simply

$$\hat{\theta}_j = \frac{\sum_{l=1}^{k} \bar{y}_l w(\bar{y}_j, \bar{y}_l)}{\sum_{l=1}^{k} w(\bar{y}_j, \bar{y}_l)}. \tag{1}$$

In (1), the information is still shared among the different groups, but the sharing of information comes directly through the data instead of a common prior. This is because of the nature of the weighted MLE which borrows more information from subjects with similar group means. As a result, the weighted likelihood of the j-th and k-th true group means θ_j and θ_k would be similar if $\bar{y}_j$ and $\bar{y}_k$ are similar. This approach avoids the computational overhead of a hierarchical Bayesian model and forms the basis of our covariate-dependent graphical model where we share information across the model parameters directly through the data via a weighted (pseudo)-likelihood function, rather than the standard hierarchical modeling framework.

Finally, we derive risk bounds for the variational estimator by casting it in a fractional variational framework adopting Yang et al. (2020), using a novel variant of the α-Rényi divergence. In particular, assuming a careful interplay between the sparsity and smoothness in the conditional regression coefficients, we showed that the W-PL framework achieves optimal variational risk bounds irrespective of whether the underlying distribution is homogeneous or heterogeneous across different covariates. Our theory also shows how W-PL improves on the independent modeling framework in the case of imbalanced samples corresponding to different covariate levels, leveraging its ability to borrow information from the entire data while estimating the graph for a specific covariate level. In the fractional variational framework, the term α controls the relative trade-off between the model fit and the prior regularization term. In the current study, the variational estimator corresponds to $\alpha = 1$. However, for technical simplicity, we restrict the theoretical analysis to estimators with $\alpha < 1$. The results for $\alpha = 1$ can be derived under stronger assumptions on prior tails, as discussed in Yang et al. (2020). However, that extension does not alter the main message of the theoretical results, and has been omitted.

The rest of the paper is organized as follows. Section 2 describes the proposed weighted pseudo-likelihood approach. Section 3 describes the variational algorithm used to approximate the posterior distribution obtained in Section 2. Section 4 provides variational risk bounds for the parameter estimates for both discrete and continuous covariates and demonstrates the advantage

of our W-PL model over a standard approach that assumes the observations for each covariate level to be independent. A thorough simulation study is conducted in Section 5. Finally, Section 6 illustrates the performance of the approach to estimate the dependence structure in protein expression levels in cancer patients using copy number variation values as covariates.

2 A weighted pseudo-likelihood (W-PL) approach

Likelihood based approaches provide a sound basis for comparing the plausibility of different graphs given the observations. Unfortunately, likelihood based approaches to modeling graph structures are intractable in general for non-chordal graphs because of an intractable normalizing term. This has sparked a lot of interest in tractable learning of non-chordal graphs in high dimensions. A pseudo-likelihood approach discussed in Besag (1975, 1977) became a convenient alternative to likelihood approaches for modeling the underlying graph dependence structure. Over the last few years, the pseudo-likelihood approach has been used widely for learning Markov random fields and neighborhood detection in Markov fields, as in Ji et al. (1996); Csiszár & Talata (2006) and others. Heckerman et al. (1995); Freno et al. (2009) discussed a pseudo-likelihood based model class to learn the dependence structure in Bayesian networks. Recently, Pensar et al. (2017) used the idea of marginal pseudo-likelihood and proved the consistency of the pseudo-marginal likelihood estimator in learning the dependence structure (neighborhood detection) of the Markov network. The set of *neighbors* of a variable x_j are variables such that given these neighbor variables, the conditional distribution of x_j is independent of all other variables. Besag (1975) argued the consistency of the maximum pseudo-likelihood estimator as the dimension of the random variable $\mathbf{X}$ increases, under the assumption that the number of neighbors is small and finitely bounded. Besag (1977) further studies the efficiency of the pseudo-likelihood estimators under Gaussian schemes.

The pseudo-likelihood approach can be described as follows: Suppose there are n individuals, indexed $i = 1, 2, \ldots, n$, in a study. Let the i-th observation in the dataset $\mathbf{X}$ be denoted as

$\mathbf{X}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$, 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 $x_{i,j}$. This approach tries to model the conditional distribution of each of the x_j's given all other variables, denoted by $\mathbf{X}_{-j} \in \mathbb{R}^{n\times(p-1)}$. Let the $p-1$ dimensional vector β_j indicate the regression effect of $\mathbf{X}_{-j}$ on x_j. The assumption commonly used here is that the conditional distribution of a variable x_j depends on only a few of the remaining variables referred to as the neighbors of x_j, and can be completely specified in terms of a regression function comprising of the neighbors as predictors. Here, we further assume a Gaussian likelihood. Then, the conditional likelihood of x_j, denoted by $\mathrm{L}(j)$, can be written as

$$\mathrm{L}(j) = p(x_j \mid \mathbf{X}_{-j}, \beta_j) \propto \prod_{i=1}^{n} \exp\left\{-(x_{i,j} - x_{i,-j}^{\mathrm{T}}\beta_j)^2/2\sigma^2\right\} \tag{2}$$

with a possibly sparse coefficient vector β_j. Consequently, the pseudo-likelihood $\mathrm{L}(\mathrm{G})$ for a fixed graph G can be written as

$$\mathrm{L}(\mathrm{G}) = \prod_{j=1}^{p} \mathrm{L}(j) = \prod_{j=1}^{p} p(x_j \mid \mathbf{X}_{-j}, \beta_j). \tag{3}$$

If the true data generating distribution $p(\mathbf{X})$ is a zero-mean multivariate Gaussian with a precision matrix Ω^*, i.e.,

$$\mathbf{X} \sim \mathcal{N}(0, \Omega^{*-1}), \tag{4}$$

then it is well-known that the conditional distribution of $x_j \mid \mathbf{X}_{-j}$ is given by (2) with $\beta_{jk} = -\Omega^*_{jk}/\Omega^*_{jj}$. However, (3) is the product of conditional distributions and is not a valid probability density in general. However, it serves as an effective computational tool for estimating the precision matrix.

In this paper, we propose a novel adaptation of this approach, and define a weighted version of this conditional likelihood for each individual in the study. We assume that the underlying graph structure is a function of extraneous covariates z. That is, given a covariate $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_n)'$, our population model assumes the true data generating distribution $p(\mathbf{x}_i)$ is a zero-mean multivariate

Gaussian with a precision matrix $\Omega^*(\mathbf{z}_i)$ for $i = 1, ..., n$.

$$\mathbf{x}_i \mid \mathbf{z}_i \sim \mathcal{N}(0, \Omega^{*-1}(\mathbf{z}_i)), \quad i = 1, ..., n, \quad \mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]'. \tag{5}$$

Thus, we allow the coefficient vector β_j's to be different for different individuals, depending on the extraneous covariates. We use the notation $\beta_j^l \in \mathbb{R}^{p-1}$ to denote the coefficient vector corresponding to the regression of the variable x_j on the remaining variables for individual l. More generally, we use the notation $\beta_j(\mathrm{z})$ to denote the coefficient vector given an arbitrary covariate z. Let $\mathbf{z}_i$ denote the covariate vector associated with the i-th individual in the study, and define $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$. Next, relative to the covariate z, we assign weights $\mathrm{w}(\mathrm{z}, \mathbf{z}_i) = \phi_\tau(\|\mathrm{z} - \mathbf{z}_l\|)$ to every individual in the study, where ϕ_τ is the Gaussian density with mean 0 and variance τ^2. When $\mathrm{z} = \mathbf{z}_l$ corresponds to the l-th individual in the study, we use the notation $\mathrm{w}_l(\mathbf{z}_i) = \mathrm{w}(\mathbf{z}_l, \mathbf{z}_i)$ to denote the weight associated with the i-th individual in the study. Next, we provide a brief overview of our approach, and then move on to more details about the steps involved.

1. For a fixed individual l, we attach weights $\mathrm{w}_l(\mathbf{z}_i), i \in \{1, 2, \ldots, n\}$, to every individual in the study relative to the l-th individual, and perform a weighted regression of the j-th variable on the remaining variables.

2. The likelihood function of regression parameters $\beta_j^l \in \mathbb{R}^{(p-1)}$ corresponding to the j-th variable for the l-th individual is given by $\prod_{i=1}^n \exp\left\{-(x_{i,j} - x_{i,-j}^{\mathrm{T}}\beta_j^l)^2 \mathrm{w}_l(\mathbf{z}_i)/2\sigma^2\right\}$. We place a suitable spike-and-slab prior on the parameters to enforce sparsity on the conditional dependency structure of x_j given x_{-j}.

3. The collection of vectors $\beta_j^l, j \in \{1, 2, \ldots, p\}$ together form the parameter of interest B^l for the l-th individual in the study, which is used to learn the undirected conditional dependency structure among $x_1, \ldots, x_p$, for the l-th individual.

4. Rather than a fully Bayesian approach, a variational approximation is made to the posterior

to obtain estimates of the coefficients.

Note that the likelihood of the parameter β_j, associated with individual l, is independent of the regression parameters associated with other individuals in the study, allowing parallel estimation of the parameters associated with the n different individuals in the study. However, information is being borrowed from every individual for each regression parameter estimation through the associated weights which attach more importance to individuals with covariates similar to the individual l and lower weights to individuals with different covariate values.

In what follows, we describe the steps involved in the pseudo-likelihood approach for graph estimation given an arbitrary covariate z in more detail. First, we introduce some more notations. Throughout, i, l, l_1 and l_2 have been used as indices for the n individuals in the study (or their corresponding observations), while j and k have been used to index the p variables. We propose the following conditional working model

$$x_{i,j} \mid x_{i,-j}, \mathbf{z}, \mathrm{z} \sim \mathcal{N}(x_{i,-j}^{\mathrm{T}} \beta_j(\mathrm{z}), \sigma^2 / \mathrm{w}(\mathrm{z}, \mathbf{z}_i)), \quad i = 1, \ldots, n. \tag{6}$$

This results in a weighted version of the conditional likelihood form shown in (2). Let x_j denote the n observations on the j-th variable. Then, the weighted conditional likelihood of x_j given a covariate value z, denoted by $\mathrm{L}_j^w(\mathrm{z})$, is given by

$$\mathrm{L}_j^w(\mathrm{z}) = p^w(x_j \mid \mathbf{X}_{-j}, \beta_j(\mathrm{z}), \mathbf{z}, \mathrm{z}) \quad \propto \quad \prod_{i=1}^{n} \exp\left\{ -\frac{(x_{i,j} - x_{i,-j}^{\mathrm{T}} \beta_j(\mathrm{z}))^2 \mathrm{w}(\mathrm{z}, \mathbf{z}_i)}{2\sigma^2} \right\}.$$

Here, the superscript w is used to indicate that the conditional distribution p involves a weighted likelihood function. Thus the weighted pseudo-likelihood for the graph $\mathrm{G}(\mathrm{z})$ corresponding to a covariate value z, denoted by $\mathrm{L}^w(\mathrm{G}(\mathrm{z}))$ can be written as

$$\mathrm{L}^w(\mathrm{G}(\mathrm{z})) = \prod_{j=1}^{p} \mathrm{L}_j^w(\mathrm{z}) = \prod_{j=1}^{p} p^w(x_j \mid \mathbf{X}_{-j}, \beta_j(\mathrm{z}), \mathbf{z}, \mathrm{z}).$$

Next, suppose that we are interested in the underlying graph structure of an individual l in the study, that is, $\mathrm{z} = \mathbf{z}_l$. Then, the weighted conditional likelihood of x_j for the l-th individual in the study, denoted by $\mathrm{L}_l^w(j)$, is given by

$$\mathrm{L}_l^w(j) = p^w(x_j \mid \mathbf{X}_{-j}, \beta_j^l, \mathbf{z}) \;\; \propto \;\; \prod_{i=1}^{n} \exp\left\{ -\frac{(x_{i,j} - x_{i,-j}^{\mathrm{T}} \beta_j^l)^2 \mathrm{w}_l(\mathbf{z}_i)}{2\sigma^2} \right\}.$$

Also, the weighted pseudo-likelihood for the graph G^l corresponding to the l-th individual, denoted by $\mathrm{L}^w(\mathrm{G}^l)$ can be written as

$$\mathrm{L}^w(\mathrm{G}^l) = \prod_{j=1}^{p} \mathrm{L}_l^w(j) = \prod_{j=1}^{p} p^w(x_j \mid \mathbf{X}_{-j}, \beta_j^l, \mathbf{z}). \tag{7}$$

For the graph G^l, (6) can be expressed in a structural equation form, similar to the form discussed in Han et al. (2016); Pearl et al. (2000). Given the covariate matrix $\mathbf{z}$, let us define a $p \times p$ coefficient matrix A^l with $\mathrm{A}_{jj}^l = 0, j \in \{1, \ldots, p\}$, and $\mathrm{A}_{jk}^l = \beta_{j,k}^l, j, k \in \{1, 2, , \ldots, p\}$. For $i \in \{1, \ldots, n\}$, let ϵ_i^l be $(p-1)$-variate independent random variables such that given $\mathbf{z}$, $\epsilon_i^l \sim \mathcal{N}(\mu_{\epsilon^l} = \mathbf{0}, \Sigma_i^l)$, where $\Sigma_i^l = \sigma^2 \mathrm{w}_l(\mathbf{z}_i)^{-1} \mathbb{I}_p$. Now, define the $n \times n$ diagonal covariance matrix $\Sigma_0^l = \mathrm{Diag}(\sigma^2 w_l(\mathbf{z}_1)^{-1}, \sigma^2 w_l(\mathbf{z}_2)^{-1}, \ldots, \sigma^2 w_l(\mathbf{z}_n)^{-1})$, and $\epsilon^l = (\epsilon_1^l, \epsilon_2^l, \ldots, \epsilon_n^l)$. Then, it can be shown that given $\mathbf{z}$, $\epsilon^l \sim \mathcal{MN}(\mathbf{0}, \mathbb{I}, \Sigma_0^l)$, where $\mathcal{MN}(\cdot, \cdot, \cdot)$ is the matrix normal distribution with suitably chosen parameters. Thus, conditioned on the covariate $\mathbf{z}$, we can express $\mathbf{X}^{\mathrm{T}}$ as $\mathbf{X}^{\mathrm{T}} = \mathrm{A}^l \mathbf{X}^{\mathrm{T}} + \epsilon^l$. Then, for $\Xi^l = (\mathbb{I} - \mathrm{A}^l)^{-1}$, we have $\mathbf{X}^{\mathrm{T}} = \Xi^l \epsilon^l$, and given $\mathbf{z}$ and Ξ^l,

$$\mathbf{X}^{\mathrm{T}} \sim \mathcal{MN}(0, \Xi^l \Xi^{l^{\mathrm{T}}}, \Sigma_0^l).$$

Thus, estimating the graph G^l is equivalent to estimating the coefficient matrix A^l, with Σ_0^l being the nuisance parameter, and then setting $\mathrm{G}_{ij}^l = \mathbb{I}\{\mathrm{A}_{ij}^l \neq 0\}$, similar to the literature on directed acyclic graphs. This graph estimate $\hat{\mathrm{G}}^l$ given the covariates $\mathbf{z}$ is not a proper undirected graph, and one needs to perform appropriate post-processing to obtain a proper undirected graph.

In this paper, we compare the covariance matrix $\Sigma_0^l = \mathrm{Diag}(\sigma^2 w_l(\mathbf{z}_1)^{-1}, \sigma^2 w_l(\mathbf{z}_2)^{-1}, \ldots, \sigma^2 w_l(\mathbf{z}_n)^{-1})$ of ϵ^l in the proposed setup with the covariate-independent setup in traditional DAG literature, where $\Sigma_0 \equiv \sigma^2 \mathbb{I}_n$, for all individuals. The borrowing of information in standard DAG literature is uniform across all observations resulting in a common graph structure for all individuals. However in the proposed setup, the amount of information borrowed from the i-th observation is covariate-dependent, via the associated weights $\mathrm{w}_l(\mathbf{z}_i)$, resulting in different graph estimates for different individuals. Recall that $\mathrm{w}_l(\mathbf{z}_i) = \phi_\tau(\|\mathbf{z}_i - \mathbf{z}_l\|)$, where ϕ_τ is the Gaussian density with mean 0 and variance τ^2. So, the amount of borrowing is controlled by the bandwidth parameter τ. As $\tau \to \infty$, the weights become equal for all the observations $i \in \{1, \ldots, n\}$, and thus the amount of information borrowed becomes uniform across observations. In this situation, conditioned on $\mathbf{z}$ and Ξ^l, $\mathbf{X}^{\mathrm{T}} \xrightarrow{D} \mathcal{MN}(0, \Xi^l \Xi^{l^T}, \Sigma_0)$ for $l = \{1, 2, \ldots, n\}$. As a result, the graph estimates G^l would be the same for all individuals in the study, as is assumed in the classic DAG literature. On the other hand, when $\tau \to 0$, we have $\mathrm{w}_l(\mathbf{z}_i) \to c_0 \mathbb{I}\{i = l\}$ (for some constant c_0). In this scenario, $(\Sigma_0^l)^{-1}$ reduces to $\sigma^{-2}\mathrm{Diag}(\mathbb{I}\{1 = l\}, \mathbb{I}\{2 = l\}, \ldots, \mathbb{I}\{n = l\})$. When the covariates have a discrete distribution, this results in a separate estimation algorithm where one estimates the underlying graphs corresponding to the different covariate levels separately with no information shared across different covariate levels. For practical experiments, the covariates vary across different observations, and the choice of the bandwidth parameter τ used for defining the weights becomes important for efficient borrowing of information. Ideally, we want to obtain a bandwidth estimate $\hat{\tau}$ such that $\hat{\tau}$ is larger for individuals when there are relatively few remaining individuals with similar covariates (sparse region in the support of the covariates). Conversely, we want $\hat{\tau}$ to be smaller for individuals when there are several other individuals with similar covariates. Towards this end, we follow the estimate discussed in Dasgupta et al. (2020); Abramson (1982); Van Kerm (2003) and others. Specifically, if $k(\mathbf{z})$ is a kernel density estimate of the covariate $\mathbf{z}$ with bandwidth parameter h, then $\hat{\tau}(\mathbf{z}_0) = h/\sqrt{k(\mathbf{z}_0)}$ is the adaptive bandwidth parameter at covariate value $\mathbf{z}_0$. When the covariates are multi-dimensional, the bandwidth

parameter h can be replaced by the harmonic mean of the bandwidths in each direction. If the dimension of the associated covariates is low, one can efficiently estimate the probability density function of the covariates and obtain a variable kernel bandwidth. In our simulation study, we follow this approach to select bandwidth hyperparameters. To select the prior residual variance σ^2, the prior slab variance σ_θ^2 and the prior inclusion probability π, we use a hybrid of grid search and model averaging, which is discussed in greater detail in Supplement D.

Note that, for $l_1 \neq l_2$, our (weighted) likelihood in (7) for the parameter B^{l_1} is different from the (weighted) likelihood of the parameter B^{l_2} because of the associated weights, that is, we do not have a *single, coherent* probability model consisting of the parameters $\mathrm{B}^l, l \in \{1, \ldots, n\}$. This step is crucial in ensuring that different individuals have potentially different underlying graph structures, depending on the covariate values. Indeed, if we have $\mathbf{z}_{l_1} = \mathbf{z}_{l_2}$, then B^{l_1} and B^{l_2} do have the same probability model.

Using a standard hierarchical model approach, we would put a common prior structure on the parameters B^ls to facilitate the borrowing of information across different individuals. However, in our proposed approach, we allow the sharing of information to come directly from the observations, effectively borrowing information while simultaneously allowing independent estimation of each of the parameters B^l. Thus, this approach allows us to define an empirical distribution on the parameters B^l, and hence the underlying graph G^l, given the covariates $\mathbf{z}$.

Next, we specify the prior distribution for the coefficient parameters corresponding to the regression problem introduced in (6). Fix an observation $l \in \{1, \ldots, n\}$, and a variable $j \in \{1, \ldots, p\}$. Note that, a significantly non-zero regression coefficient corresponds to an edge in the underlying graph structure. With this goal in mind, we use a spike-and-slab prior on the parameter β_j^l. That is, for $k \in \{1, \ldots, p\}$, $\beta_{j,k}^l$ is assumed to come from a zero-mean Gaussian density with a variance component $\sigma^2\sigma_\beta^2$ ("slab" density) with a probability π, and equals zero ("spike" density) with probability $1 - \pi$. Let us define $\gamma_{j,k}^l = \mathbb{I}\{\beta_{j,k}^l \neq 0\}$ which can be treated as Bernoulli random variables with a common probability of success π. Define the row vector

$\gamma_j^l = (\gamma_{j,1}^l, \gamma_{j,2}^l, \ldots, \gamma_{j,p}^l)$, and $\Gamma^l = \{\gamma_j^l, j = 1, 2, \ldots, p\}$. Then, we use the following prior distribution for (β_j^l, γ_j^l) given by

$$p_0\left(\beta_j^l, \gamma_j^l\right) = \prod_{k=1, k\neq j}^{p} \delta_{\{0\}}\left(\beta_{j,k}^l\right)^{1-\gamma_{j,k}^l} \mathcal{N}\left(\beta_{j,k}^l; 0, \sigma^2\sigma_\beta^2\right)^{\gamma_{j,k}^l} \prod_{k=1, k\neq j}^{p} \pi^{\gamma_{j,k}^l}(1-\pi)^{(1-\gamma_{j,k}^l)}.$$

Using the data model as described in (6) for an individual l, we obtain the following posterior distribution for (β_j^l, γ_j^l) as

$$p(\beta_j^l, \gamma_j^l \mid \mathbf{X}) \propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x_{ij} - \sum_{k\neq j, k=1}^{p} x_{ik}\beta_{j,k}^l \right)^2 \mathrm{w}_l(\mathbf{z}_i) \right\} p_0(\beta_j^l, \gamma_j^l).$$

Note that the posterior distributions of B^{l_1} and B^{l_2} are independent, but are similar depending on the covariates through the associated weights. This notion allows independent and fast estimation of the parameters $\mathrm{B}^l, l = 1, 2, \ldots, n$, while ensuring that subjects with similar covariates have similar coefficient parameter estimates. This also allows us to effectively gauge the variability of the graph structure across subjects. However, since the posterior distribution does not have a closed-form solution, we would require MCMC samples in order to obtain posterior samples for the parameters. This can be very time consuming when p is large, especially since we have to essentially compute np such distributions. In the following, we develop an efficient parallelized mean-field variational inference to approximate the posterior distribution.

3 An efficient parallelized block mean-field variational inference

Variational Bayes approximations are deterministic approaches where instead of finding the posterior probability distributions, we aim to find an approximation of them by first introducing a class of approximating distributions and then finding the distribution that best approximates the posterior obtained through some optimizing criterion over the aforesaid class. See, for instance,

Jordan et al. (1999); Wainwright & Jordan (2008); Ormerod & Wand (2010); Blei et al. (2017), only to name a few. In this section, we adopt the block-mean-field approach proposed by Carbonetto et al. (2012) in the context of Bayesian variable selection with spike-and-slab priors in high dimensional regression problems.

Suppose we have a parameter of interest ξ with an intractable posterior distribution $p(\xi)$, an observed data vector y, and the variational tractable family of densities $q(\xi)$. Let $\mathrm{D}_{\mathrm{KL}}(\cdot\|\cdot)$ denote the Kullback-Leibler divergence between two density functions. Then the "best approximating density" over a tractable family of densities Γ is a density $q^*(\xi)$ such that

$$q^*(\xi) = \underset{q\in\Gamma}{\operatorname{argmin}}\, \mathrm{D}_{\mathrm{KL}}(q \,\|\, p(\xi \mid y)).$$

Since $\mathrm{D}_{\mathrm{KL}}(q\|p(\cdot|y)) \geq 0$, we have $\log p(y) \geq$ ELBO, where ELBO$= \int q(\xi) \log\{p(y,\xi)/q(\xi)\}\, d\xi$ is the evidence-lower bound.

In the current study, the parameter of interest is $\xi = (\beta_j^l, \gamma_j^l)$. Here, we adopt the block mean-field approach for the variational approximation considered in Carbonetto et al. (2012) given by

$$q(\beta_j^l, \gamma_j^l; \phi_j^l) = \prod_{k=1, k\neq j}^{p} q_k(\beta_{j,k}^l, \gamma_{j,k}^l; \phi_{j,k}^l).$$

In the above expression, ϕ_j^l's are free parameters corresponding to the l-th individual, and the separate factors q_k have the following form:

$$q_k(\beta_j^l, \gamma_j^l; \phi_j^l) = \mathcal{N}\left(\beta_{j,k}^l; \mu_{j,k}^l, (s_{j,k}^l)^2\right)^{\gamma_{j,k}^l} \delta_0\left(\beta_{j,k}^l\right)^{1-\gamma_{j,k}^l} \left(\alpha_{j,k}^l\right)^{\gamma_{j,k}^l} \left(1-\alpha_{j,k}^l\right)^{1-\gamma_{j,k}^l}$$

where $\phi_{j,k}^l = (\alpha_{j,k}^l, \mu_{j,k}^l, (s_{j,k}^l)^2), k = \{1, \ldots, p, k \neq j\}$ are the free parameters, and δ_0 is the "spike" density degenerate at zero. Thus, the individual factors q_k are independent spike-and-slab densities. The parameter $\beta_{j,k}^l$ comes from a Gaussian density with mean $\mu_{j,k}^l$ and standard deviation $s_{j,k}^l$ (the "slab" part) with probability $\alpha_{j,k}^l$, and is zero (the "spike" part) with probability

$1 - \alpha^l_{j,k}$. Therefore, we can write the variational density family as

$$q(\beta^l_j, \gamma^l_j) = \prod_{k=1}^{p-1} \mathcal{N}\left(\beta^l_{j,k}; \mu^l_{j,k}, \left(s^l_{j,k}\right)^2\right)^{\gamma^l_{j,k}} \delta_0\left(\beta^l_{j,k}\right)^{1-\gamma^l_{j,k}} \left(\alpha^l_{j,k}\right)^{\gamma^l_{j,k}} \left(1-\alpha^l_{j,k}\right)^{1-\gamma^l_{j,k}},$$

where the "best approximating density" $q^*(\beta^l_j, \gamma^l_j)$ can be obtained by optimizing over the free parameters $\phi^l_{j,k} = (\alpha^l_{j,k}, \mu^l_{j,k}, \left(s^l_{j,k}\right)^2)$ in order to maximize the ELBO over the variational family.

The coordinate descent updates for the variational parameters can be obtained by taking partial derivatives of the ELBO, setting them to zero, and solving for $\alpha^l_{j,k}, \mu^l_{j,k}$ and $\left(s^l_{j,k}\right)^2$. Carbonetto et al. (2012) proposed a component-wise algorithm where one iterates between updating $\mu^l_{j,k}$ and $\alpha^l_{j,k}$ for a fixed j, and then updating $j \in \{1, 2, \ldots, p\}$. Huang et al. (2016) proposes a batch-wise updating scheme where one iterates between updating $\left(s^l_{j,k}\right)^2, j = 1, 2, \ldots, p$, in a batch, followed by updating $\mu^l_{j,k}, j = 1, 2, \ldots, p$, in a batch, and subsequently by updating $\alpha^l_{j,k}, j = 1, 2, \ldots, p$, in a batch. Huang et al. (2016) argued that the component-wise update scheme in high-dimensional settings might lead to noise accumulation, and can cause the variational estimates to move away from the true parameter value. Huang et al. (2016) also asserted that the batch-wise updating algorithm achieves frequentist as well as Bayesian consistency even when the dimension p diverges to infinity at an exponential rate as the sample size grows to infinity. We closely follow the batch-wise updating algorithm discussed in Huang et al. (2016), which leads to the following variational parameter updates.

$$\begin{aligned}
\left(s^l_{j,k}\right)^2 &= \frac{\sigma^2}{(1/\sigma^2_\beta + \sum_{i=1}^n x^2_{ik} \mathrm{w}_l(\mathbf{z}_i))}; \quad \mathrm{logit}(\alpha^l_{j,k}) = \mathrm{logit}(\pi) + \frac{(\mu^l_{j,k})^2}{2{s^l_{j,k}}^2} + \log \frac{s^l_{j,k}}{\sigma\sigma_\beta}; \\
\mu^l_{j,k} &= \frac{\left(s^l_{j,k}\right)^2}{\sigma^2} \sum_{i=1}^n \left\{ \mathrm{w}_l(\mathbf{z}_i) x_{ik} \left(x_{ij} - \sum_{m \neq j,k} x_{im} \mu^l_{j,m} \alpha^l_{j,m} \right) \right\}.
\end{aligned}$$

Note that, the two-step approach proposed in this paper might not result in a proper undirected graph as the posterior inclusion probability estimates $\hat{\alpha}^l_{j,k}$ and $\hat{\alpha}^l_{k,j}$ might not be the same. Hence, we perform post-processing steps in order to obtain a bonafide undirected graph estimate in

practice. For our purposes, we set $\tilde{\alpha}^l_{j,k} = \tilde{\alpha}^l_{k,j} = (\hat{\alpha}^l_{j,k} + \hat{\alpha}^l_{k,j})/2$ in order to symmetrize the inclusion probabilities and obtain a proper undirected graph estimate.

4 Variational risk bounds

In this section, we derive risk bounds for our variational estimates and demonstrate the efficiency of the weighted pseudo-likelihood model over a model that treats the covariate levels independently. To that end, we start by defining a few notations. First, it is well-known that the conditional distribution of all variables $x_j, j \in \{1, \ldots, p\}$, given the remaining variables x_{-j}, is

$$p(x_{i,j} \mid \mathbf{X}_{i,-j}, \mathbf{z}_i) \sim \mathcal{N}\left(-\sum_{k \neq j} \frac{\Omega^*_{kj}(\mathbf{z}_i)}{\Omega^*_{jj}(\mathbf{z}_i)} x_{i,k}, \frac{1}{\Omega^*_{jj}(\mathbf{z}_i)} \right), i \in \{1, \ldots, n\}. \tag{8}$$

Let $\Gamma^*(\mathrm{z})$ represent the $p \times (p-1)$ matrix of parameters controlling the latent indicator variables corresponding to the sparsity structure of $\Omega^*(\mathrm{z})$. $\gamma^*_j(\mathrm{z})$, the j-th row of $\Gamma^*(\mathrm{z})$, is the indicator variable corresponding to the j-th variable. Let $\mathrm{B}^*(\mathrm{z})$ denote the true coefficient parameter matrix given covariate z, with j-th row $\beta^*_{j,k}(\mathrm{z}) = \Omega^*_{kj}(\mathrm{z})/\Omega^*_{jj}(\mathrm{z}), k \in \{1, \ldots, p\}$. Here $\gamma^*_j(\mathrm{z})$ is the indicator variable associated with the truth $\beta^*_j(\mathrm{z})$. For simplicity of presentation, we assume that the true variance parameter σ^2_* is correctly specified in the model. Let $\Theta^l(\mathbf{z}) = (\mathrm{B}^l(\mathbf{z}), \Gamma^l(\mathbf{z}))$, where $\mathrm{B}^l(\mathbf{z})$ represents the $p \times (p-1)$ coefficient matrix with $\beta^l_j(\mathbf{z})$ as the j-th row corresponding to the j-th variable as a function of covariates $\mathbf{z}$. Denote by $\theta^l_j(\mathbf{z}) = (\beta^l_j(\mathbf{z}), \gamma^l_j(\mathbf{z}))$ the parameter associated with the j-th variable given the covariates $\mathbf{z}$. Let $p_{\theta^l_j} = p_{\gamma^l_j} p_{\beta^l_j|\gamma^l_j}$ denote the spike-and-slab prior distribution of θ^l_j used in the analysis, Γ denote the variational family of distributions $q(\theta^l_j)$ for the parameter θ^l_j and Λ denote the parameter space for the coefficient parameter B, where Λ^l_j denotes the parameter space for β^l_j for an individual $l \in \{1, ..., n\}$. Then, $\theta^l_j(\mathbf{z})$ denotes the parameter associated with the conditional distribution of the j-th variable given the other variables for the l-th individual. For the estimation of the parameters associated with the l-th individual, we assign weights $\mathrm{W}_l = \mathrm{Diag}\{w_l(\mathbf{z}_1), w_l(\mathbf{z}_2), \ldots, w_l(\mathbf{z}_n)\}$ to the likelihood

contribution of the n individuals under study, depending on the covariates associated with the individuals. In what follows, the results are derived for a fixed individual, and is valid for each individual graph parameter $\Theta^l(\mathbf{z})$ in the study.

We derive risk bounds for the weighted pseudo-likelihood model by casting it as a misspecified model, as in Kleijn et al. (2006). Following (2), define the misspecified conditional distribution function $p^w(x_j \mid \mathbf{X}_{-j}, \theta_j^l(\mathbf{z}), \mathbf{z})$ as

$$p^w(x_j \mid \theta_j^l(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z}) = \left(\prod_{i=1}^{n} \frac{\sqrt{\mathrm{w}_l(\mathbf{z}_i)}}{\sqrt{2\pi}\sigma_*} \right) \exp\left\{ - \frac{(x_j - \mathbf{X}_{-j}\beta_j^l(\mathbf{z}))^{\mathrm{T}} \mathrm{W}_l (x_j - \mathbf{X}_{-j}\beta_j^l(\mathbf{z}))}{2\sigma_*^2} \right\}. \quad (9)$$

Let $p(x_j \mid \mathbf{X}_{-j}, \theta_j^*(\mathbf{z}_l))$ denote the true well-specified conditional distribution with respect to the true parameter $\theta_j^*(\mathbf{z}_l)$. Misspecified models as treated in Kleijn et al. (2006) gives rise to Kullback-Leibler balls centered around $\tilde{\theta}_j^l(\mathbf{z})$ where

$$\tilde{\theta}_j^l(\mathbf{z}) = \underset{\theta_j(\mathbf{z}) \in \mathbb{R}^{(p-1)} \times \{0,1\}^{(p-1)}}{\operatorname{argmin}} \int p(x_j \mid \mathbf{X}_{-j}, \theta_j^*(\mathbf{z}_l)) \log \frac{p(x_j \mid \mathbf{X}_{-j}, \theta_j^*(\mathbf{z}_l))}{p^w(x_j \mid \theta_j^l(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})} dx_j. \quad (10)$$

For $\tilde{\theta}_j^l(\mathbf{z})$ as defined in (10) and the true parameter value $\theta_j^*(\mathbf{z}_l)$, and $\alpha \in (0, 1)$, we measure closeness between $\tilde{\theta}_j^l(\mathbf{z})$ and a value $\theta_j^l(\mathbf{z})$ using the divergence

$$D_{\theta_j^*(\mathbf{z}_l),\alpha}(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z}) \mid \mathbf{X}_{-j}, \mathbf{z}) = \frac{1}{\alpha - 1} \log \int \left\{ \frac{p^w(x_j \mid \mathbf{X}_{-j}, \tilde{\theta}_j^l(\mathbf{z}), \mathbf{z})}{p^w(x_j \mid \mathbf{X}_{-j}, \theta_j^l(\mathbf{z}), \mathbf{z})} \right\}^{\alpha} p(x_j \mid \mathbf{X}_{-j}, \theta_j^*(\mathbf{z}_l)) dx_j.$$

Here, $D_{\theta_j^*(\mathbf{z}_l),\alpha}(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z}) \mid \mathbf{X}_{-j}, \mathbf{z})$ is the α-Rényi divergence measure between the Kullback-Leibler minimizer $\tilde{\theta}_j^l(\mathbf{z})$ and a candidate parameter value $\theta_j^l(\mathbf{z})$ with respect to the true underlying distribution, conditioned on $\mathbf{X}_{-j}$ and covariates $\mathbf{z}$. We define a weighted version of our divergence measure between $\tilde{\Theta}^l(\mathbf{z})$ and $\Theta^l(\mathbf{z})$ as

$$d_{\Theta^*(\mathbf{z}),\alpha}(\Theta^l(\mathbf{z}), \tilde{\Theta}^l(\mathbf{z})) = \max_j d_{\theta_j^*(\mathbf{z}_l),\alpha}(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z})), \quad (11)$$

$$d_{\theta_j^*(\mathbf{z}_l),\alpha}(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z})) = -\frac{1}{1-\alpha} \log \mathbf{E}_{-j} \exp\left\{ -(1-\alpha) D_{\theta_j^*(\mathbf{z}_l),\alpha}(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z}) | \mathbf{X}_{-j}, \mathbf{z}) \right\} \quad (12)$$

Since

$$0 < \exp\left\{-(1-\alpha)D_{\theta_j^*(\mathbf{z}_l),\alpha}(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z}) \mid \mathbf{X}_{-j}, \mathbf{z})\right\} \leq 1,$$

with the right equality holds if and only if $\theta_j^l(\mathbf{z}) = \tilde{\theta}_j^l(\mathbf{z})$, $d_{\theta_j^*(\mathbf{z}_l),\alpha}(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z}))$ is a valid divergence measure. Let $\mathrm{D}_{\mathrm{KL}}(q_\theta \,\|\, p_\theta)$ denote the Kullback-Leibler divergence. Let $\hat{q}_{\theta_j}(\theta_j^l(\mathbf{z}))$ be the α-variational estimate associated with the fractional posterior distribution of $\theta_j^l(\mathbf{z})$ as in Yang et al. (2020).

$$\hat{q}_{\theta_j}(\theta_j^l(\mathbf{z})) = \underset{q_{\beta_j,\gamma_j}=\prod_{k=1}^p q_{\beta_{jk},\gamma_{jk}}}{\operatorname{argmin}} \left\{ -\int_{\Lambda_j} \sum_{\delta\in\{0,1\}^{p-1}} R d\theta_j^l(\mathbf{z}) + \alpha^{-1}\mathrm{D}_{\mathrm{KL}}(q_{\theta_j^l(\mathbf{z})} \| p_{\theta_j}) \right\}, \tag{13}$$

where $R = \log \frac{p^w(x_j|\beta_j^l(\mathbf{z}),\gamma_j^l(\mathbf{z}),\mathbf{X}_{-j},\mathbf{z})}{p^w(x_j|\beta_j^*(\mathbf{z}_l),\gamma_j^*(\mathbf{z}_l),\mathbf{X}_{-j},\mathbf{z})} q_{\theta_j}(\theta_j^l(\mathbf{z}))$. Let $\|\mathrm{A}\|_2 = \sqrt{\tau_{\max}(\mathrm{A}^{\mathrm{T}}\mathrm{A})}$ denote the operator norm of a matrix A, where $\tau_{\max}(\mathrm{A})$ is the maximum eigenvalue of A. Let $\|a\|_\infty = \max_i a_i, \|a\|_2 = \sqrt{a^{\mathrm{T}}a}, \|a\|_0 = \sum_i \mathbb{1}\{a_i \neq 0\}$ be the corresponding ℓ_∞, ℓ_2 and ℓ_0 norms of a vector a. For any $\epsilon \in (0,1)$, let

$$\underline{r}(n,\epsilon) = \frac{\alpha\epsilon^2}{(1-\alpha)} + \frac{s^* \log p}{n(1-\alpha)}. \tag{14}$$

Although the methodology in §3 corresponds to $\alpha = 1$, we present the risk bounds for $\alpha \in (0,1)$ as it greatly simplifies the technicalities without deviating from the main idea. We now adapt Yang et al. (2020) to develop risk bounds for this variational estimate in three situations. First, we examine the case in which the true covariates are continuously distributed, and then we discuss the case in which they are discrete. Finally, we consider the situation where the underlying true distribution is independent of the available covariates. We list our assumptions of theoretical analysis in Supplement A. All proofs are deferred to Supplement B with auxiliary results in Supplement C.

4.1 Continuous covariate-dependent model

In this subsection, we consider the case where the covariates z are drawn from a density which is absolutely continuous with respect to the Lebesgue measure. For simplicity, we consider z as a scalar. Define $\beta_j^*(\mathrm{z})$ as the coefficient corresponding to the j-th variable as a function of the covariate value z. Define $\dot{\beta}_j^*(\mathrm{z})$ and $\ddot{\beta}_j^*(\mathrm{z})$ as the first and second order point-wise derivatives respectively of $\beta_j^*(\mathrm{z})$ as a function of z. Similarly, define $\Sigma(\mathrm{z})$ as the covariance matrix as a function of z, with $\dot{\Sigma}(\mathrm{z})$, $\ddot{\Sigma}(\mathrm{z})$ as the point-wise first and second order derivatives with respect to z.

We consider the predefined misspecified weighted pseudo-likelihood with the parameter space: $\|\beta_j^l(\mathbf{z})\|_0 \leq C_0 s_j^*$, $l = 1, ..., n, j = 1, ..., p$ for constant $C_0 \geq 1$. Then for a subject l, the KL divergence between the truth and weighted pseudo-likelihood is

$$\int p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1), ..., \Theta^*(\mathbf{z}_n)) \log \frac{p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1), ..., \Theta^*(\mathbf{z}_n))}{\prod_{j=1}^p p^{w_l}(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^l(\mathbf{z}))} d\mathbf{X}, \tag{15}$$

where the true model $p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1), ..., \Theta^*(\mathbf{z}_n))$ is induced from equation (5) and $\Theta_{-j}^l(\mathbf{z})$ is the parameter of interest given the constraint that $\|\beta_j^l(\mathbf{z})\|_0 \leq C_0 s_j^*$ with $C_0 \geq 1$. We choose $w_l(\mathbf{z}_k) = c_l K\left((\mathbf{z}_k - \mathbf{z}_l)/\tau\right)/\tau$, where c_l is a subject-specific constant. Then the following lemma characterizes the property of the KL minimizer.

Lemma 1. *Under* **Assumptions K, T1- T4** *in Supplement A, let $\tilde{\beta}_j^l(\mathbf{z})$ be the minimizer of the KL divergence* (15) *under the constraint $\|\beta_j^l(\mathbf{z})\|_0 \leq C_0 s_j^*$ for constant $C_0 \geq 1$. If $\tau \to 0$ and $n\tau \to \infty$, then we have for $j = 1, .., p$ and $l = 1, ..., n$*

$$\mathbf{E}_{\mathbf{z}} \|\tilde{\beta}_j^l(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\|^2 = O\left(s_j^* \tau^4 + \frac{s_j^*}{n\tau}\right).$$

Therefore, by balancing the bias $O(\tau^4)$ and variance $O(1/(n\tau))$, one can achieve an MSE of order $n^{-4/5} s_j^*$ corresponding to the optimal bandwidth $\tau = n^{-1/5}$. It is important to note that the optimal bandwidth has the same optimal rate as kernel density estimation so that we

can borrow the tuning strategy from these problems in practice. There could be multiple KL minimizers due to the non-convexity of the ℓ_0 constraint, however, all the KL minimizers have the same convergence behavior depicted in Lemma 1. The following theorem characterizes the convergence rate of the obtained estimator towards the KL minimizers.

Theorem 1. *Under* **Assumptions T1-T5** *and* **P, K** *in Supplement A, let* $\hat{q}^l(\theta_j^l(\mathbf{z}))$ *be the variational estimate of the* j*-th graph coefficients for subject* l*. Suppose that* $\tau = n^{-1/5}$ *and* $\|\tilde{\beta}_j^l(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\| \leq c_1 n^{-4/5} s_j^*$ *for* $\tilde{\beta}_j^l(\mathbf{z})$ *specified in Lemma 1 for* $l = 1, ..., n$ *and* $j = 1, ..., p$*. Then with probability at least* $1 - c_2 n \exp(-c_3 n) - c_4 n/p^{c_0} - e^{-(c_5 s^* - 1)\log(np)}$*, we have*

$$\max_{l=1,...,n} \max_{j=1,...,p} \int \frac{1}{n} d_\alpha(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z})) \hat{q}_{\theta_j^l(\mathbf{z})}(\theta_j^l(\mathbf{z})) d\theta_j^l(\mathbf{z}) \leq C \frac{1+\alpha}{1-\alpha} \left(\frac{s^* \log(np)}{n} + \frac{s^*}{n^{\frac{3}{5}}} \right),$$

for positive constants $c_0, c_1, c_2, c_3, c_4, c_5, C > 0$.

While the first term in the upper bound of the risk is due to the model selection associated with estimating a sparse precision matrix and cannot be improved, observe that the second term is primarily due to the misspecification error under the true data generating distribution (5). Note that this term ($s^* n^{-3/5}$) is in contrast with the convergence rate in Lemma 1, which is upper bounded by $s_j^* n^{-4/5}$. The extra factor of $n^{1/5}$ is due to the discrepancy in the Euclidean norm and the size of the misspecified Kullback-Leibler neighborhood, which is upper bounded by $\|\mathrm{W}_l^{1/2} \mathbf{X}_{-j} (\tilde{\beta}_j^l(\mathbf{z}) - \beta_j^*(\mathbf{z}_l))\|_2^2$. Since, $\|\mathrm{W}_l\|_2 \leq c \max_x K(x)/\tau \leq c/\tau \leq c n^{1/5}$ and the rate of convergence of $d_\alpha(\theta_j^l(\mathbf{z}), \tilde{\theta}_j^l(\mathbf{z}))$ is associated with the size of the misspecified KL ball, the second term is slowed down by a factor of $n^{1/5}$. We conjecture that this cannot be improved unless one considers the discrete covariate setting as we discuss below.

According to the Theorem 1, graph estimation can be carried out consistently for each observation under proper smoothness assumptions. Note that due to the continuous covariate structure, one cannot perform a valid separate estimation across different covariate values. We also compare our Theorem 1 with Theorem 3.4 of Qiu et al. (2016) who considered joint estimation of multiple graphical models. Our result is non-asymptotic and considers joint risk across all subjects

$l = 1, ..., n$, while Qiu et al. (2016) obtained risk bound for estimating a single graph.

4.2 Discrete covariate-dependent underlying graph

Next, we consider the scenario where there are K covariate levels $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K$ corresponding to K different populations with parameters $\Theta^*(\mathbf{z}_l)$ corresponding to the l-th covariate level. Let $n_l,\ l \in \{1, \ldots, K\}$, be the number of sample observations corresponding to the l-th group, with $n = \sum_{l=1}^{K} n_l$ being the total number of observations. Let $\mathbf{X}_{\mathbf{z}_l}$ denote the $n_l \times p$ data matrix corresponding to the covariate $\mathbf{z}_l$. We assume that the true data generating distribution $p(\mathbf{X}_{i,\mathbf{z}_l})$ for the rows of $\mathbf{X}_{\mathbf{z}_l}$ is a zero-mean multivariate Gaussian with a sparse precision matrix $\Omega^*(\mathbf{z}_l), l = 1, \ldots, K$. In this case, the borrowing of information in the misspecified weighted pseudo-likelihood approach causes the Kullback-Leibler minimizer $\tilde{\theta}_j^l(\mathbf{z})$ as defined in (10) to be different from $\theta_j^*(\mathbf{z}_l)$, where $\theta_j^*(\mathbf{z}_l) = (\beta_j^*(\mathbf{z}_l), \delta_j^*(\mathbf{z}_l))$ is the j-th row of $\Theta^*(\mathbf{z}_l)$.

Lemma 2. *If* **Assumption W** *in Supplement A is satisfied, let* $\tilde{\beta}_j^l(\mathbf{z})$ *be the minimizer of the KL divergence* (10) *under the constrained* $\|\beta_j^l(\mathbf{z})\|_0 \leq C_0 s_j^*$ *for constant* $C_0 \geq 1$. *Then we have for* $j = 1, ..., p$ *and* $l = 1, ..., n$*:*

$$\|\tilde{\beta}_j^l(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\|_2^2 \leq c \exp\left(-2c_l^2/\tau^2 + 2\log(n/n_l - 1)\right) s_j^*, \tag{16}$$

for some constant $c > 0$.

Note that when $\tau < \min_l c_l/\sqrt{\log n}$, the convergence rate of the KL minimizer towards the truth is faster than s_j^*/n. **Assumption W** implies that the weighted estimator will finally converge to the separate estimation, so that our method is no worse than separate estimation asymptotically.

Assume that $s^* \log(np)/n \to 0$. Then we have the following theorem where the risk bound is derived for an individual with covariate value $\mathbf{z}_l$.

Theorem 2. *For any* $\zeta \in (0, 1)$ *and* $\epsilon \in (0, 1)$ *and* $\underline{r}(n, \epsilon)$ *as in* (14)*, if* **Assumptions T, W, P** *in*

Supplement A are satisfied, we have with probability at least $1 - \zeta - c_2 p^{-c_1} - p\exp\{-a_2 n_l\}$,

$$\int \frac{1}{n} d_{\Theta^*(\mathbf{z}_l),\alpha}(\Theta^l(\mathbf{z}), \tilde{\Theta}^l(\mathbf{z}))\hat{q}_{\Theta^l(\mathbf{z})}(\Theta^l(\mathbf{z}))d\Theta^l(\mathbf{z}) \leq \underline{r}(n,\epsilon) + \frac{s^*}{n(1-\alpha)}\log\left(\frac{s^*\sqrt{n}}{\epsilon\sqrt{n_l}}\right) + \frac{\log(p/\zeta)}{n(1-\alpha)},$$

for some positive constants a_1, a_2, c_1, c_2 *and* D.

The optimal error rate is obtained by balancing $\underline{r}(n,\epsilon)$ and $\log(1/\epsilon)$. The proof of the above theorem is similar to that of Theorem 3 and Corollary 1 below, and is therefore omitted.

4.3 Covariate-independent underlying graph

Here, we consider the situation where the underlying structure is independent of the covariate levels. In this case, we assume that the underlying graph structure is homogeneous, as described in (4). Thus there is a common true graph parameter $\Theta^* = (\mathrm{B}^*, \Gamma^*)$ associated with every individual in the study. Then we have the following lemma.

Lemma 3. *For the weighted likelihood model, when the true conditional distribution is covariate-independent, we have* $\tilde{\beta}_j^l(\mathbf{z}) = \beta_j^*$ *for* $l = 1, ..., K$.

Theorem 3. *Under Assumptions* **P, W, A** *in Supplement A and data generating process* (4)*, for any* $\zeta \in (0,1)$ *and* $\epsilon \in (0,1)$ *and* $\underline{r}(n,\epsilon)$ *as in* (14)*, for any* $l = 1, ..., K$*, we have with probability at least* $1 - \zeta - c_2 p^{-c_1} - p\exp(-a_2 n_l)$*, for some positive constants* a_1, a_2, c_1, c_2 *and* D*,*

$$\int \frac{1}{n} d_{\Theta^*,\alpha}(\Theta^l(\mathbf{z}), \Theta^*)\hat{q}_{\Theta^l(\mathbf{z})}(\Theta^l(\mathbf{z}))d\Theta^l(\mathbf{z}) \leq \underline{r}(n,\epsilon) + \frac{s^*}{n(1-\alpha)}\log\left(\frac{s^*\sqrt{n}}{\epsilon\sqrt{n_l}}\right) + \frac{\log(p/\zeta)}{n(1-\alpha)}.$$

The strength of the proposed approach is that the risk bound for the variational estimate is sharper for every individual in the current study as compared to independent modeling of the two groups separately. If one were to perform independent modeling of a *homogeneous* graphical structure for the two covariate levels separately. The risk bound of the parameters of an individual belonging to a group with size n_l would be as below:

Corollary 1. *Under Assumptions* **P, A** *in Supplement A and data generating process* (4)*, for any* $\zeta \in (0,1)$ *and* $\epsilon \in (0,1)$ *and* $\underline{r}(n_1, \epsilon)$ *as in* (14)*, we have with probability at least* $1 - \zeta - c_2 p^{-c_1} - p\exp\{-a_2 n_l\}$*,*

$$\int \frac{1}{n_l} d_\alpha(\Theta, \Theta^*) \hat{q}_\Theta(\Theta) d\Theta \leq \underline{r}(n_l, \epsilon) + \frac{s^*}{n_l(1-\alpha)} \log\left(\frac{s^*}{\epsilon}\right) + \frac{1}{n_l(1-\alpha)} \log\left(\frac{p}{\zeta}\right),$$

for some positive constants a_1, a_2, c_1, c_2 *and* D*.*

If one of the groups has a sample size $n_l = \mathcal{O}(n^h)$ with $h < 1$, then the risk bounds for that group will increase compared to the proposed model because the information is borrowed from every subject in the study for the proposed approach.

5 Simulation Study

We begin our simulation study with a setting defined by a unidimensional covariate before considering a multidimensional covariate. In both cases, the covariate $\mathbf{z}$ is randomly drawn from a uniform distribution. To generate the data for each of the settings, we first define the precision matrix Ω_i for the i-th individual as a function of the covariate $\mathbf{z}_i$. We then generate the observation $\mathbf{X}_i$ for the i-th individual according to the population model (5) from a mean-zero $(p+1)$-dimensional normal distribution with precision matrix Ω_i. Finally, we apply W-PL to estimate the graphs $\hat{\mathrm{G}}^i$ describing the sparsity structure of Ω_i. In addition to varying the dimensionality of the covariate, we also perform experiments in each setting with varying data dimensionality, examining performance for $p \in \{10, 30, 50\}$.

We perform 50 trials for each experiment, repeating the data generation process in each trial. In each trial, we first select an individual-specific bandwidth hyperparameter τ for W-PL using a two-step kernel density estimation technique. We then average over a grid of π candidates using the exponentiated ELBO as our unnormalized model averaging weights, and conduct a two-dimensional grid search to select σ^2 and σ_β^2 for each π candidate, using the ELBO as our

grid search objective function. More details on this hyperparameter specification scheme are included in Supplement D. These hyperparameters are used in our final variational estimate to the posterior inclusion probabilities $\alpha_{j,k}$, which we symmetrize as $\tilde{\alpha}_{j,k} = (\alpha_{j,k} + \alpha_{k,j})/2$. Finally, we threshold the symmetrized probabilities at 0.5 as $\hat{\mathrm{G}}^i_{j,k} = \mathbb{1}\{\tilde{\alpha}_{j,k} > 0.5\}$ to construct the final graph estimates $\hat{\mathrm{G}}^i$. To evaluate the performance of W-PL, we compute the sensitivity and specificity of these estimates compared to to the ground-truth precision structure G^*, where these metrics are defined as:

$$\text{sensitivity} = \frac{\#\{(j,k) : (\mathrm{G}^*_{jk} = 1) \cap (\hat{\mathrm{G}}^i_{jk} = 1)\}}{\#\{(j,k) : (\mathrm{G}^*_{jk} = 1)\}}, \quad \text{specificity} = \frac{\#\{(j,k) : (\mathrm{G}^*_{jk} = 0) \cap (\hat{\mathrm{G}}^i_{jk} = 0)\}}{\#\{(j,k) : (\mathrm{G}^*_{jk} = 0)\}}.$$

We consider two competitors for W-PL in these experiments. The first is a time-varying graphical model from Haslbeck & Waldorp (2020) that uses kernel smoothing and elastic net regularization (mgm). The second is also a time-varying graphical model from Yang & Peng (2020) that uses a local group LASSO penalty (loggle). In both cases, we select hyperparameters using cross-validation.

We consider experiments where the covariate is discrete in Supplement E and where the data distribution departs from Gaussian in Supplement F, as well as a comparison to the method of Qiu et al. (2016) in Supplement G.

5.1 Unidimensional Covariate

We first consider a unidimensional covariate $\mathbf{z}_i \in [-3, 3]$ and define the j, k entry of the ground-truth precision matrices as $\Omega^i_{j,k} = 2$ if $j = k$, $\Omega^i_{j,k} = 1$ if $(j,k) \in \{(2,3),(3,2)$, $\Omega^i_{j,k} = \mathbb{1}\{\mathbf{z}_i < 1\} \cdot \min(1, \frac{1}{2} - \frac{1}{2}\mathbf{z}_i)$ if $(j,k) \in \{(1,2),(2,1)\}$ and $\Omega^i_{j,k} = \mathbb{1}\{\mathbf{z}_i > -1\} \cdot \min(1, \frac{1}{2} + \frac{1}{2}\mathbf{z}_i)$ if $(j,k) \in \{(1,3),(3,1)$. The ground truth precision structures are given in Supplement H.1. To generate the covariate, we sample from uniform distributions on $[-3,-1], [-1,1]$, and $[1,3]$ 50 times each. Thus, in this experiment, $n = 150$.

We present the results for these experiments in Table 1. At each of the considered dimensionalities, W-PL outperforms loggle and mgm in terms of sensitivity. mgm consistently offers the lowest false positive rate, although W-PL remains competitive in this metric. Further, although the sensitivity differential between W-PL and loggle is roughly constant across the different dimensionalities, as p get larger, the performance of mgm relative to W-PL substantially decreases.

p	Method	Sensitivity($\uparrow$)	Specificity($\uparrow$)
10	W-PL	$\mathbf{0.8382}(0.0743)$	$0.9951(0.0057)$
	loggle	$0.7802(0.0707)$	$0.9926(0.0071)$
	mgm	$0.7057(0.0953)$	$\mathbf{0.9991}(0.0018)$
30	W-PL	$\mathbf{0.7758}(0.1068)$	$0.9977(0.0014)$
	loggle	$0.7211(0.0935)$	$0.9981(0.0009)$
	mgm	$0.5894(0.1112)$	$\mathbf{0.9999}(0.0002)$
50	W-PL	$\mathbf{0.7387}(0.0907)$	$0.9984(0.0009)$
	loggle	$0.6982(0.0895)$	$0.9984(0.0005)$
	mgm	$0.5149(0.0761)$	$\mathbf{1.0000}(0.0000)$

Table 1: *Results for 1-dimensional continuous covariate-dependent setting*

In order to gauge the practical performance of the proposed method, we look at the estimated inclusion probability, specifically α_{12} for the edge between x_1 and x_2, and α_{13} for the edge between x_1 and x_3. To gauge the variability in the estimates, we study not only the mean posterior inclusion probability across the trials, but also the 5-th and 95-th quantiles. Figure 1 illustrates the true precision value between the edges and the corresponding mean inclusion probability. This figure shows that the presence (or absence) of an edge between pairs of variables is almost always correctly recovered for the first and third clusters. The behavior of the inclusion probability completely mimics the behavior of the true precision value across individuals, and the variability is naturally most apparent in the middle cluster where the precision matrix varies with the covariate. Note that the dependence structure for variables where the corresponding entry in the precision matrix does not change across subjects is correctly recovered for all subjects across all trials.

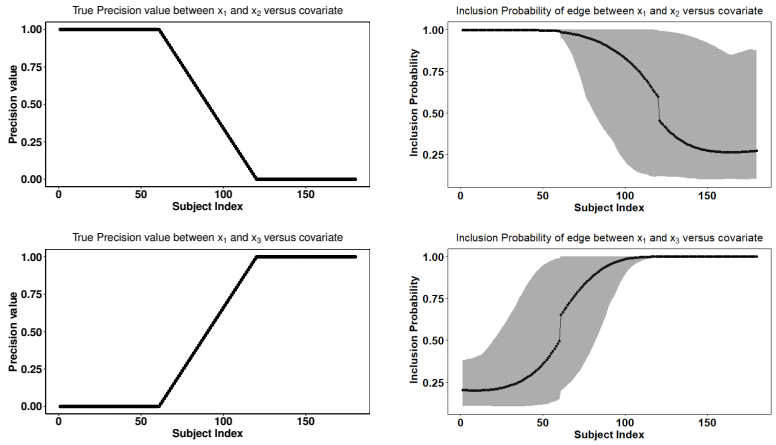

Figure 1: *Left: True precision value for the edge between Variable 1 and 2 (top panel); and Variable 1 and 3 (bottom panel). Right: Corresponding mean inclusion probabilities across 50 simulations, and* 95% *confidence interval of the probabilities.*

5.2 Multidimensional Covariate

We next consider a 2-dimensional covariate $\mathbf{z} \in [-3, 3] \times [-3, 3]$. We define the j, k entry of the ground-truth precision matrices similar to the 1-dimensional case as $\Omega^i_{j,k} = 2$ if $j = k$, $\Omega^i_{j,k} = 1$ if $(j, k) \in \{(2, 3), (3, 2)$, $\Omega^i_{j,k} = \mathbb{1}\{\mathbf{z}_{i1} < 1\} \cdot \min(1, \frac{1}{2} - \frac{1}{2}\mathbf{z}_i)$ if $(j, k) \in \{(1, 2), (2, 1)\}$ and $\Omega^i_{j,k} = \mathbb{1}\{\mathbf{z}_{i2} > -1\} \cdot \min(1, \frac{1}{2} + \frac{1}{2}\mathbf{z}_i)$ if $(j, k) \in \{(1, 3), (3, 1)$. The ground truth precision structures are given in Supplement H.2. We generate a sample of size $n = 225$ by sampling uniformly 25 times from each of the 9 sets generated by taking the Cartesian product of the intervals resulting from partitioning the horizontal and vertical axes of the covariate space into intervals of length 2.

In the unidimensional continuous covariate setting, $\mathbf{z}$ may be thought of as indexing time. Thus, mgm and loggle can both be directly compared to W-PL. However, to include these methods in our multidimensional covariate experiments, a reduction to the dimensionality of the covariate is necessary, as neither model can directly handle a multidimensional extraneous covariate. To do this, we apply a greedy sorting algorithm that re-indexes $\mathbf{z}_1, ..., \mathbf{z}_n$ to $\mathbf{z}_{(1)}, ..., \mathbf{z}_{(n)}$. First, we set $\mathbf{z}_{(1)} = \mathbf{z}_1$. Then, at the t-th step of the algorithm, $t > 1$, we de-

fine $\mathcal{S}_t = \{\mathbf{z}_1, ..., \mathbf{z}_n\} \setminus \{\mathbf{z}_{(1)}, ..., \mathbf{z}_{(t-1)}\}$ as the covariates that have not yet been sorted and set

$$\mathbf{z}_{(t)} = \arg\min_{\mathbf{z} \in \mathcal{S}_t} \|\mathbf{z} - \mathbf{z}_{(t-1)}\|$$

This gives us a bijection ξ mapping from $z_1, ..., z_n$ to $z_{(1)}, ..., z_{(n)}$. We use this mapping to define the 1-dimensional covariate $\mathbf{v} \in 1, ..., n$ mimicking a time index for loggle and mgm, where $\mathbf{v}_l = l'$ if, and only if, $\xi(\mathbf{z}_l) = \mathbf{z}_{(l')}$, i.e., the l'-th timepoint is the individual whose covariate was sorted to the l'-th position. To demonstrate the fairness of this reduction of the covariate, we apply W-PL both to the original covariate $\mathbf{z}$, as well as to the time-indexing covariate $\mathbf{v}$. We refer to the results from the latter as time-varying W-PL (tv W-PL).

We present the results from this experiment in Table 2. W-PL has the best sensitivity of the 4 considered methods across all of the considered dimensionalities. The sensitivity of tv W-PL is less than that of W-PL, but still greater than loggle and mgm in all of the experiments, which is expected, given the results of the previous experiments. Although the differential between W-PL and tv W-PL is only about 0.05 in each experiment, this demonstrates the importance of utilizing covariate information fully in achieving optimal performance. Thus, in addition to the ability to model the precision matrix as varying continuously, another key attribute of W-PL is its ability to directly incorporate a multidimensional covariate into the estimation procedure.

6 Real data analysis

The notion of *non-homogeneous* underlying graphical structure is particularly significant in the field of cancer research, because it is well known that cancer initiates and evolves through co-ordinated changes across multiple molecular levels, networks and pathways. This causes the underlying graph to vary across individuals depending on demographics, genetic markers, and other biological factors(Bolli et al. (2014); Lohr et al. (2014)). These factors can be looked at as extraneous covariates which contain valuable information about how the underlying graph

p	Method	Sensitivity(↑)	Specificity(↑)
10	W-PL	$\mathbf{0.8890}(0.1023)$	$\mathbf{0.9968}(0.0035)$
	tv W-PL	$0.8363(0.1221)$	$0.9906(0.0063)$
	loggle	$0.6360(0.1101)$	$0.9917(0.0076)$
	mgm	$0.6273(0.2031)$	$0.9963(0.0047)$
30	W-PL	$\mathbf{0.8216}(0.1250)$	$0.9995(0.0004)$
	tv W-PL	$0.7689(0.1399)$	$0.9976(0.0012)$
	loggle	$0.5322(0.1159)$	$0.9996(0.0004)$
	mgm	$0.5173(0.1422)$	$\mathbf{0.9998}(0.0003)$
50	W-PL	$\mathbf{0.8399}(0.1173)$	$0.9997(0.0002)$
	tv W-PL	$0.7886(0.1211)$	$0.9980(0.0007)$
	loggle	$0.4809(0.1054)$	$0.9997(0.0002)$
	mgm	$0.4792(0.0749)$	$\mathbf{1.0000}(0.0000)$

Table 2: *Results for the continuous multidimensional covariate-dependent setting*

structure varies across the individuals.

We use data on patients with Breast Invasive Carcinoma (BRCA) from The Cancer Genome Atlas (TCGA) program website at `http://www.compgenome.org/TCGA-Assembler/`. We consider 70 patients with Breast Invasive Carcinoma and 30 patients with normal cells. "FOXC2" is a gene that is well known to be associated with breast cancer, as discussed in Mani et al. (2007). We use the unnormalized copy number variation (`cnv`) values of the gene as our choice of the covariate. We notice that the `cnv` values were very similar among the normal cells and were concentrated in the range of 1.83 to 2.06. However, the values were much more varied among the cancer cells, as shown in the left panel of Figure 2. We estimate the graph dependence structure among the protein expression values of eight genes corresponding to the individuals in our study by treating the given `cnv` values as continuous associated covariates. The eight genes considered were "CTNNB1","BRCA2", "MET", "E-cadherin", "N-cadherin", "NFkB1", "snail" and "STAT3". Based on the covariate values, we describe the `cnv` to be "under-expressed" if the values are less than 1.6, "normally expressed" if the values are between 1.6 and 2.1, and "over-expressed" if the values are greater than 2.1. As opposed to the hyperparameter specification scheme used in Section 5, here, we utilize the scheme described in Supplement E.4.

Figure 3 shows the estimated dependence structure of three individuals with different levels of "FOXC2" `cnv` expression. There is a visible evolution of the dependence structure as the covariate value changes. In particular, we focus on the edge between "N-Cadherin" and "NFkB1" which is present in the under-expressed "FOXC2" gene, but is otherwise not present. The inclusion probability with covariate value is shown in the right panel of Figure 2. We notice a steady decrease of the inclusion probability as the expression level of the "FOXC2" `cnv` increases. The sharp jump on the right is probably because of the sparsity of data points in that neighborhood resulting in inaccurate estimation. N-cadherin is known to promote breast cancer irrespective of the E-cadherin levels, as discussed in Nieman et al. (1999). However, NFkB1 is known to promote breast cancer by suppression of E-cadherin expression in cells, as discussed in ChuaHL et al. (2007); Criswell & Arteaga (2007) and others. Our study corroborates this observation, as we do notice a significant change in the dependence pattern between "E-cadherin" and "NFkB1" at different expression levels of the "FOXC2" gene. For normally expressed cells there is a significant dependence between the protein expressions of the two genes. However, for under-expressed or over-expressed cells, the dependence is no longer present. This is displayed in the middle panel of Figure 2 where we notice a sharp peak in inclusion probability for the normally expressed cells only, except for the outliers on the right.

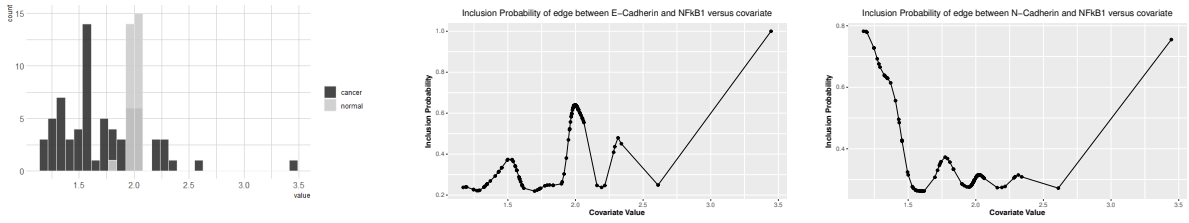

Figure 2: *Left: Histogram of covariate value for normal cells (lighter shade) versus cancer cells (darker shade). Middle: Inclusion Probability between E-Cadherin and NFkB1 versus covariate values. Right: Inclusion Probability between N-Cadherin and NFkB1 versus covariate values.*

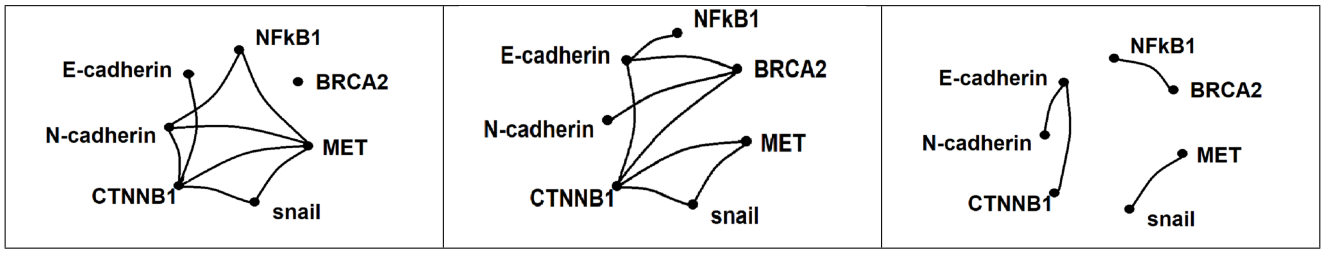

Figure 3: *Predicted network structures for individuals with under expressed (left), normally expressed (middle) and over expressed (right) FOXC2 gene.*

7 Discussion

In this article, we have introduced a novel weighted-pseudo likelihood approach that can provide an estimate of the underlying dependence structure at an individual level using extraneous covariate information. An appealing feature of the proposed approach is that the performance of the estimates does not suffer when the underlying structure does not actually depend on the extraneous covariates, which we demonstrate in Supplement E.1. The variational approach, together with the embarrassingly parallel structure of the parameter estimation avoids the computational complexities associated with running a full-blown Markov chain Monte Carlo. In addition, we also established optimal risk bounds of the proposed method, demonstrating that the approximation through either the variational inference or the pseudo-likelihood framework does not hinder the statistical properties of the method. The theory further demonstrates how borrowing information allows us to obtain a better fit.

Non-Gaussian responses are another direction worth exploring in the future. When the true distribution is non-Gaussian, there is no direct interpretation of the conditional regression coefficients. In contrast, the pseudo-likelihood approach is a practical technique to go beyond the Gaussian assumption by changing the error distribution, see, for instance, Guha et al. (2020). However, it is unclear what true data generation mechanism can be approximated by such a pseudo-likelihood.

Finally, the detection of high-dimensional graphs could be challenging if the SNR is not high enough, which we explore in Supplement E.4. This low SNR issue is more prominent for the

continuous covariate setting when $\Theta^*_{ij}(\mathrm{z})$ is a continuous function of z, and $\Theta^*_{ij}(\mathrm{z})$ takes both zeros and non-zero values. Then by the continuity, $\Theta^*_{ij}(\mathrm{z})$ takes values arbitrarily close to zero, where it is challenging to recover the graphs due to low SNR.

The codes used in our analysis are available online anonymously on Github at `https://anonymous.4open.science/r/covariate-dependent_graphical_modeling/simulation_study_graph_learning/main.R`.

References

ABRAMSON, I. S. (1982). On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics* , 1217–1223.

ATCHADÉ, Y. F. (2019). Quasi-bayesian estimation of large gaussian graphical models. *Journal of Multivariate Analysis* **173**, 656–671.

BESAG, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)* **24**, 179–195.

BESAG, J. (1977). Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika* , 616–618.

BHADRA, A. & MALLICK, B. K. (2013). Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis. *Biometrics* **69**, 447–457.

BLEI, D. M., KUCUKELBIR, A. & MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.

BOLLI, N., AVET-LOISEAU, H., WEDGE, D. C., VAN LOO, P., ALEXANDROV, L. B., MARTINCORENA, I., DAWSON, K. J., IORIO, F., NIK-ZAINAL, S., BIGNELL, G. R. et al. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature communications* **5**, 1–13.

CAI, T. T., LI, H., LIU, W. & XIE, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100**, 139–156.

CARBONETTO, P., STEPHENS, M. et al. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis* **7**, 73–108.

CHUAHL, B., NAKSHATRIP, C. et al. (2007). Nf-kappab represses e-cadherin expression and enhances epithelial to mesenchymal transition of mammary epithelial cells: potential involvement of zeb-1 and zeb-2. *Oncogene* **26**, 711.

CRISWELL, T. L. & ARTEAGA, C. L. (2007). Modulation of nfκb activity and e-cadherin by the type iii transforming growth factor β receptor regulates cell growth and motility. *Journal of Biological Chemistry* **282**, 32491–32500.

CSISZÁR, I. & TALATA, Z. (2006). Consistent estimation of the basic neighborhood of markov random fields. *The Annals of Statistics* , 123–145.

DANAHER, P., WANG, P. & WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **76**, 373.

DASGUPTA, S., PATI, D. & SRIVASTAVA, A. (2020). A two-step geometric framework for density modeling. *Statistica Sinica* **30**, 2155–2177.

FOX, E. B. & DUNSON, D. B. (2015). Bayesian nonparametric covariance regression. *The Journal of Machine Learning Research* **16**, 2501–2542.

FRENO, A., TRENTIN, E. & GORI, M. (2009). Scalable pseudo-likelihood estimation in hybrid random fields. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

GIUDICI, P. & GREEN, P. (1999). Decomposable graphical gaussian model determination. *Biometrika* **86**, 785–801.

GUHA, N., BALADANDAYUTHAPANI, V. & MALLICK, B. K. (2020). Quantile graphical models: a bayesian approach. *The Journal of Machine Learning Research* **21**, 3023–3069.

GUO, J., LEVINA, E., MICHAILIDIS, G. & ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1–15.

HA, M. J., BALADANDAYUTHAPANI, V. & DO, K.-A. (2015). Dingo: differential network analysis in genomics. *Bioinformatics* **31**, 3413–3420.

HAN, S. W., CHEN, G., CHEON, M.-S. & ZHONG, H. (2016). Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *Journal of the American Statistical Association* **111**, 1004–1019.

HASLBECK, J. M. B. & WALDORP, L. J. (2020). mgm: Estimating Time-Varying Mixed Graphical Models in High-Dimensional Data. *Journal of Statistical Software* **93**, 1–46.

HECKERMAN, D., GEIGER, D. & CHICKERING, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning* **20**, 197–243.

HOFF, P. D. & NIU, X. (2012). A covariance regression model. *Statistica Sinica* , 729–753.

HUANG, X., WANG, J. & LIANG, F. (2016). A variational algorithm for bayesian variable selection. *arXiv preprint arXiv:1602.07640* .

JI, C., SEYMOUR, L. et al. (1996). A consistent model selection procedure for markov random fields based on penalized pseudolikelihood. *The annals of applied probability* **6**, 423–443.

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. & SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning* **37**, 183–233.

KLEIJN, B. J., VAN DER VAART, A. W. et al. (2006). Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics* **34**, 837–877.

KOLAR, M., P PARIKH, A. & P XING, E. (2010a). On sparse nonparametric conditional covariance selection .

KOLAR, M., SONG, L., AHMED, A., XING, E. P. et al. (2010b). Estimating time-varying networks. *The Annals of Applied Statistics* **4**, 94–123.

LEE, W. & LIU, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of multivariate analysis* **111**, 241–255.

LIU, H., CHEN, X., WASSERMAN, L. & LAFFERTY, J. D. (2010). Graph-valued regression. In *Advances in Neural Information Processing Systems*.

LOHR, J. G., STOJANOV, P., CARTER, S. L., CRUZ-GORDILLO, P., LAWRENCE, M. S., AUCLAIR, D., SOUGNEZ, C., KNOECHEL, B., GOULD, J., SAKSENA, G. et al. (2014). Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell* **25**, 91–101.

MANI, S. A., YANG, J., BROOKS, M., SCHWANINGER, G., ZHOU, A., MIURA, N., KUTOK, J. L., HARTWELL, K., RICHARDSON, A. L. & WEINBERG, R. A. (2007). Mesenchyme forkhead 1 (foxc2) plays a key role in metastasis and is associated with aggressive basal-like breast cancers. *Proceedings of the National Academy of Sciences* **104**, 10069–10074.

MEINSHAUSEN, N., BÜHLMANN, P. et al. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of statistics* **34**, 1436–1462.

NI, Y., STINGO, F. C. & BALADANDAYUTHAPANI, V. (2019). Bayesian graphical regression. *Journal of the American Statistical Association* **114**, 184–197.

NIEMAN, M. T., PRUDOFF, R. S., JOHNSON, K. R. & WHEELOCK, M. J. (1999). N-cadherin promotes motility in human breast cancer cells regardless of their e-cadherin expression. *The Journal of cell biology* **147**, 631–644.

ORMEROD, J. T. & WAND, M. P. (2010). Explaining variational approximations. *The American Statistician* **64**, 140–153.

PEARL, J. et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* .

PENSAR, J., NYMAN, H., NIIRANEN, J., CORANDER, J. et al. (2017). Marginal pseudo-likelihood learning of discrete markov network structures. *Bayesian analysis* **12**, 1195–1215.

PETERSON, C., STINGO, F. C. & VANNUCCI, M. (2015). Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association* **110**, 159–174.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–690.

POURAHMADI, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425–435.

POURAHMADI, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*, vol. 882. John Wiley & Sons.

QIU, H., HAN, F., LIU, H. & CAFFO, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 487–504.

REN, M., ZHANG, S., ZHANG, Q. & MA, S. (2022). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics* **78**, 524–535.

VAN KERM, P. (2003). Adaptive kernel density estimation. *The Stata Journal* **3**, 148–156.

WAINWRIGHT, M. J. & JORDAN, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

WANG, J. & KOLAR, M. (2014). Inference for sparse conditional precision matrices. *arXiv preprint arXiv:1412.7638* .

YANG, J. & PENG, J. (2020). Estimating Time-Varying Graphical Models. *Journal of Computational and Graphical Statistics* **29**, 191–202.

YANG, Y., PATI, D., BHATTACHARYA, A. et al. (2020). α-variational inference with statistical guarantees. *Annals of Statistics* **48**, 886–905.

YIN, J. & LI, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics* **5**, 2630.

YUAN, M. & LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94**, 19–35.

ZHANG, W. & LENG, C. (2012). A moving average cholesky factor model in covariance modelling for longitudinal data. *Biometrika* **99**, 141–150.

ZHOU, S., LAFFERTY, J. & WASSERMAN, L. (2010). Time varying undirected graphs. *Machine Learning* **80**, 295–319.