Statistical Methods & Applications (2022) 31:241–251
https://doi.org/10.1007/s10260-021-00605-2

COMMENT

Extending graphical models for applications: on covariates, missingness and normality

Luigi Augugliaro[1] · Veronica Vinciotti[2] · Ernst C. Wit[3]

Accepted: 12 October 2021 / Published online: 28 October 2021

Abstract

The authors of the paper "Bayesian Graphical Models for Modern Biological Applications" have put forward an important framework for making graphical models more useful in applied settings. In this discussion paper, we give a number of suggestions for making this framework even more suitable for practical scenarios. Firstly, we show that an alternative and simplified definition of covariate might make the framework more manageable in high-dimensional settings. Secondly, we point out that the inclusion of missing variables is important for practical data analysis. Finally, we comment on the effect that the Gaussianity assumption has in identifying the underlying conditional independence graph and how this can be circumvented. The Bayesian framework proposed by the authors is flexible enough to accommodate extensions that can deal with these aspects, which are often encountered in real data analyses such as the complex modern applications considered by the authors.

Keywords Conditional graphical models · Copula graphical models · Missing data · Sparse inference

✉ Luigi Augugliaro
luigi.augugliaro@unipa.it

Veronica Vinciotti
veronica.vinciotti@unitn.it

Ernst C. Wit
wite@usi.ch

[1] University of Palermo, Viale delle Scienze - Building 13, 90128 Palermo, Italy

[2] University of Trento, Via Sommarive, 14, 38123 Povo, Italy

[3] Università della Svizzera italiana, Via G. Buffi 13, 6900 Lugano, Switzerland

1 Introduction

The authors of the discussion paper should be congratulated with the considerable effort in making graphical models more suitable for real world applications. Graphical models are the archetypal way of studying complex systems in an integrated probabilistic way. The factorization of the likelihood in combination with the graphical representation of the system make graphical models both formally tractable and directly interpretable. However, until recently, graphical models were rarely used in practice in conjunction with substantial covariate information. This has limited the application of graphical models in experimental settings, where often the aim is to describe the effect of one or more factors on the behaviour of the system of interest. The aim of the discussion paper is to extend the graphical model for those situations. In our discussion we focus on a number of aspects that deserve further attention.

In particular, in Sect. 2 we discuss in what ways covariate information can be included in a graphical model. Whereas the authors of the discussion paper have chosen a particular approach, this is clearly not the only way one can include dependent variables. We aim to clarify the various options that are available, what their strengths and limitations are and how they are related. Furthermore, we describe a related conditional graphical model approach that is particularly useful in high-dimensional settings.

In Sect. 3 we discuss the situation of missing and censored data. Whereas we do not intend to repeat the discussion that missing data can be the Achilles heel of any statistical analysis, we do believe that especially in practical scenarios, where missingness and other artifacts such as censoring or saturation are common, it is crucial to include as many samples as possible in the analysis. Simply discarding incomplete data can be disastrous, especially in the graphical model setting. Although in principle the Bayesian framework should be particularly suitable for dealing with data that is missing at random (or completely at random), the authors unfortunately only dedicate a single line to the matter.

Finally, in Sect. 4 we return to the age-old issue of assuming normality. Although we are all for making convenient and practically workable assumptions, it is important to realize that in graphical model settings this tends to have disproportionate effect on the structural inference. Practically, this can affect the conclusions, such as in the myeloma network analysis, where various inferred networks are compared with each other. There are simple and practical ways to mitigate these issues and we describe this in the Gaussian copula approach.

All our suggestions built forth on the very useful developments on practical and accessible graphical modelling presented in the discussion paper. By making graphical models more flexible, they can become a standard tool in the applied data analyst's tool box. A golden age of graphical models is ahead of us.

2 What is a covariate?

It is instructive to go back to basics from time to time to take stock of what has actually been achieved. Ni et al. (2021) introduce two types of graphical models that depend on covariates and that might be called *conditional graphical models*. It does raise the question: what exactly *is* a covariate? In short, *covariates* are observations conditional on whom another observation, typically referred to as the response, has a particular distribution. In a parametric setting, this means that X is a covariate relative to Y if

$$Y|X = x \sim \mathcal{L}(\theta(x)),$$

for some parametric distribution $\mathcal{L}$, whose parameters θ are a function of x. Given that a Gaussian graphical model is defined relative to its mean μ, its precision matrix Ω and its conditional independence graph G, a *conditional Gaussian graphical model* (Y, X, μ, Ω, G) is defined as

$$Y|X = x \sim N(\mu(x), \Omega(x)) \text{ relative to a conditional independence graph } G(x).$$

This definition is general and holds both for undirected and directed Gaussian graphical models. It is easy to see that both the *Bayesian multiple graph* model of Sect. 3.1, its dynamic extension in Sect. 3.3 and the *covariate-dependent graph* models of Sect. 4 of the discussion paper are covered by this definition. For directed conditional Gaussian graphical models, such as the graphical regression models in Sect. 4.1, the functional space of precision matrix functions $\Omega(x)$ should be constrained to satisfy that the resulting graphical structure $G(x)$ is a Directed Acyclic Graph. The authors do not explicitly constrain their inference in this way, but undoubtedly they do some a posteriori sanity checks to make sure that no cycles occur.

The conditional graphical models described in Ni et al. (2021) focus heavily on the precision matrix $\Omega(x)$. In fact, the multiple graph models in Sect. 3 consider a single categorical covariate with K levels and conditional on the level, say k, the precision matrix is freely specified as Ω_k. The graphical regression considers Q continuous covariates $\mathbf{x}$ and defines the entries $\Omega_{jk}(\mathbf{x})$ of the precision matrix as a thresholded function f_{jk} (for example defined as a sum of Q univariate b-splines),

$$\Omega_{jk}(\mathbf{x}) = f_{jk}(\mathbf{x}) 1_{\{|f_{jk}(\mathbf{x})| > t_{jk}\}}.$$

This is a very useful class of models and it has the ability to show how certain factors affect the strength of various interactions. Given the detailed nature of the comparisons in the multiple graphs model and interpretation of the effect shapes in the graphical regression model, this method comes to the fore best in low-dimensional problems.

Given the fundamental linear structure of a Gaussian graphical model in the first place, it seems not unreasonable to consider only linear functions with additional hierarchical constraints, say,

$$\Omega_{jk}(\mathbf{x}) = \theta^0_{jk} + \sum_{i=1}^{Q} \theta^i_{jk} x_i, \text{ such that } \forall i : \theta^0_{jk} = 0 \Longrightarrow \theta^i_{jk} = 0.$$

The hierarchical constraints impose that an intercept is always included. Furthermore, for undirected graphical models we impose $\theta^i_{jk} = \theta^i_{kj}$, whereas for directed graphical models we only consider $\Omega(\mathbf{x})$ that satisfy the DAG structure of $G(\mathbf{x})$. Due to the hierarchical constraints, this means that we only have to put the DAG constraints on the intercepts $\{\theta^0_{jk}\}$.

A simpler, but still interesting class of models are the conditional Gaussian graphical models with a model structure on the mean parameters, such as

$$\mu(\mathbf{x}) = B\mathbf{x}.$$

This class of models can be studied easily in high-dimensional settings, given the straightforward nature of the interpretation of the parameters.

So far, we have not spoken about inference. Whereas the authors suggest a Bayesian implementation, which can indeed be suitable for low- to medium-dimensional settings, for high-dimensional settings, penalized likelihood approaches can be particularly useful.

2.1 Conditional graphical lasso

In this section, we consider a recent development of Gaussian graphical modelling approaches that allow for dependency of the mean on the covariates. These models are generally referred to as conditional Gaussian graphical models, also known as covariate adjusted Gaussian graphical models, and are a class of conditional probabilistic graphical models used to encode the dependence structure among the elements of a set of random variables conditional on a second set of random variables (Lafferty et al. 2001). Formally, let $y = (y_1, \ldots, y_p)^\top$ and $x = (x_1, \ldots, x_q)^\top$ be p- and q-dimensional random vectors, respectively, and let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \ldots, p\}$, indexing only the entries in y, and edge set $E \subseteq V \times V$, where $(h, k) \in E$ iff there is a directed edge from the vertex h to k in G. Suppose that the distribution of y conditional on x is a multivariate Gaussian distribution with probability density function defined as follows:

$$\phi(y|x; B, \Omega) = (2\pi)^{-\frac{p}{2}} |\Omega|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}(y - B^\top x)^\top \Omega (y - B^\top x) \right\}, \tag{1}$$

where, with a little abuse of notation, we let $x = (1, x^\top)^\top$ be the vector of predictors and $B = (\beta_0, \beta^\top)^\top$ the $(q + 1) \times p$ regression coefficient matrix. The model in (1) assumes that the predictors affect the distribution of the response variables y only via the p conditional expected values and through a linear function, that is:

$$E(y|x) = B^\top x, \quad V(y|x) = \Sigma.$$

The inverse of the variance matrix, denoted as before by $\Omega = (\omega_{hk})$, is called the precision matrix and its entries have a one-to-one correspondence with the partial correlation coefficients. Using standard results about the multivariate Gaussian distribution, it is possible to show that y_h and y_k are conditionally independent given x and all the remaining variables in y iff the corresponding partial correlation coefficient is zero (Lauritzen 1996). This remarkable property of the multivariate Gaussian distribution gives rise in a natural way to the notion of conditional Gaussian graphical model, which is based on the idea of relating the factorization of the density (1) to the topological structure of the undirected graph G.

Inference of a conditional Gaussian graphical model is particularly challenging under censoring and missing-at-random structures, which occur frequently in real data and which raise computational challenges already for moderate sized datasets (Augugliaro et al. 2020). In order to see this, consider a set of n independent observations denoted by (y_i, x_i) with $i = 1, \ldots, n$. For observations $i \in \mathcal{O}$, in which the observation vector y_i is fully observed, the contribution to the log-likelihood is

$$\ell_i(B, \Omega) = \log \phi(y_i|x_i; B, \Omega).$$

However, for observations $i \in \mathcal{C}$, in which some entries j of y_i are censored, $j \in c_i$, either from below or from above, the contribution to the likelihood is given by the multi-dimensional integral across the censored variables,

$$\ell_i(B, \Omega) = \log \int_{D_{c_i}} \phi(y_i|x_i; B, \Omega) dy_{ic_i},$$

where the region $D_{c_i} = \prod_{j \in c_i} D_{ij}$ is the censoring region, with $D_{ij} = (-\infty, l_j)$ if $y_{ij} \leq l_j$ (censored from below) or $D_{ij} = (u_j, \infty)$ if $y_{ij} \geq u_j$ (censored from above). Finally, if some entries y_{ij} are missing-at-random, then it is possible to extend the definition of the censoring region to encompass such missingness in the likelihood. In particular, $D_{ij} = \mathbf{R}$ for these cases. Considering all possible cases, the relevant average observed log-likelihood function is given by

$$\bar{\ell}(B, \Omega) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(B, \Omega). \tag{2}$$

Under a high-dimensional setting, that is $\min\{p, q\} > n$, inference about B and Ω can be carried out under the assumption that these matrices have a sparse structure, i.e., only a few regression coefficients and partial correlation coefficients are different from zero. To this end, Augugliaro et al. (2020) propose to estimate the parameters of a conditional Gaussian graphical model by maximizing a new objective function whereby two specific lasso-type penalty functions are added to the average observed log-likelihood. The resulting estimator is defined as follows:

$$\{\widehat{B}, \widehat{\Omega}\} = \arg\max \bar{\ell}(B, \Omega) - \lambda \sum_{k=1}^{p} \omega_{kk} \|\beta_k\|_1 - \rho \|\Omega\|_1^- \tag{3}$$

where β_k denotes the kth column of β, $\|\beta_k\|_1 = \sum_{h,k} |\beta_{hk}|$ and $\|\Omega\|_1^- = \sum_{h \neq k} |\omega_{hk}|$. Like in the standard penalized inference approaches, the tuning parameter λ is used to control the amount of sparsity in the estimated regression coefficient matrix whereas ρ is devoted to control the sparsity in $\widehat{\Omega} = (\hat{\omega}_{hk})$ and, consequently, in the corresponding estimated conditional independence graph $\widehat{G} = \{V, \widehat{E}\}$, where $\widehat{E} = \{(h,k) : \hat{\omega}_{hk} \neq 0\}$. When ρ is sufficiently large, some $\hat{\omega}_{hk}$ are shrunken to zero resulting in the removal of the corresponding link in $\widehat{G}$; on the other hand, when ρ is equal to zero and the sample size is large enough the estimator $\widehat{\Omega}$ coincides with the maximum likelihood estimator of the precision matrix, which implies a fully connected conditional independence graph.

2.2 Computational time

Augugliaro et al. (2020) propose a unifying algorithm for inference of a sparse conditional Gaussian graphical model that can accommodate both the case of censoring (Augugliaro et al. 2020), missingness-at-random (Städler and Bühlmann 2012) as well as the high-dimensionality of the data. Inference is based on an Expectation-Maximization (EM) algorithm for maximizing the penalized log-likelihood and is efficiently implemented in the `R` package `cglasso`.

In general, the EM algorithm is based on the idea of repeating the expectation and maximization steps, until a convergence criterion is met. For the sake of simplicity, in the remaining part of this section, we use $\vartheta = \{B, \Theta\}$ for the parameters and $\hat{\vartheta}$ to denote their current estimates inside the EM. Moreover, r_{ik} indicates whether y_{ik} is observed ($r_{ik} = 0$) or not ($r_{ik} \neq 0$), with the latter case including both censoring and missingness.

Since the complete probability density function is a member of the regular exponential family, the E-step consists in computing two quantities. First, the imputed response matrix $\widehat{Y} = (\hat{y}_{i,k})$ is obtained, whose entries are defined as:

$$\hat{y}_{i,k} = \begin{cases} y_{ik} & \text{if} \quad r_{ik} = 0 \\ E(y_{ik} \mid y_{ic_i} \in D_{c_i}, x_i; \hat{\vartheta}) & \text{otherwise} \end{cases},$$

where $E(\cdot \mid y_{ic_i} \in D_{c_i}, x_i; \hat{\vartheta})$ denotes the expected value operator computed with respect to the conditional Gaussian distribution of y_{ic_i} given $\{x_i, y_{io_i}\}$ and truncated over the region D_{c_i}. Secondly, it involves the matrix $\widehat{C}_{yy} = \sum_{i=1}^{n} \widehat{C}_i$, whose components have entries:

$$\widehat{C}_{i,hk} = \begin{cases} y_{ih}y_{ik} & \text{if} \quad r_{ih} = 0 \text{ and } r_{ik} = 0 \\ y_{ih}E(y_{ik} \mid y_{ic_i} \in D_{c_i}, x_i; \hat{\vartheta}) & \text{if} \quad r_{ih} = 0 \text{ and } r_{ik} \neq 0 \\ E(y_{ih} \mid y_{ic_i} \in D_{c_i}, x_i; \hat{\vartheta})y_{ik} & \text{if} \quad r_{ih} \neq 0 \text{ and } r_{ik} = 0 \\ E(y_{ih}y_{ik} \mid y_{ic_i} \in D_{c_i}, x_i; \hat{\vartheta}) & \text{if} \quad r_{ih} \neq 0 \text{ and } r_{ik} \neq 0. \end{cases}$$

From this, the working empirical covariance matrix is given by:

$$\widehat{S}_{y|x}(B) = n^{-1}\{\widehat{C}_{yy} - \widehat{Y}^{\top}XB - (XB)^{\top}\widehat{Y} + X^{\top}XB\}, \tag{4}$$

where X denotes the design matrix. Given the matrix (4), the M-step involves solving a new maximization problem obtained by replacing the objective function in definition (3), with the so-called penalized Q-function:

$$Q(B, \Theta) = \log\det\Theta - \mathrm{tr}\{\Theta\widehat{S}_{y|x}(B)\} - \lambda\sum_{k=1}^{p}\theta_{kk}\|\beta_k\|_1 - \rho\|\Theta\|_1^{-}. \tag{5}$$

Since, for a fixed $\hat{\vartheta}$, the penalized Q-function in (5) is a bi-convex function in B and Θ, its maximization can be obtained by repeating two sub-steps until a convergence criterion is met. Given the current estimate of the precision matrix $\widehat{\Theta}$, the first sub-step consists in estimating the regression coefficient matrix by solving the following maximization problem:

$$\min_{B}\mathrm{tr}\{\widehat{\Theta}\widehat{S}_{y|x}(B)\} + \lambda\sum_{k=1}^{p}\hat{\theta}_{kk}\|\beta_k\|_1, \tag{6}$$

whereas, in the second sub-step, given $\widehat{B}$, the precision matrix is estimated by solving the sub-problem:

$$\max_{\Theta\succ 0}\log\det\Theta - \mathrm{tr}\{\Theta\widehat{S}_{y|x}(\widehat{B})\} - \rho\|\Theta\|_1^{-}. \tag{7}$$

While problem (7) is a standard graphical lasso problem that can be efficiently solved using, for example, a block-coordinate descent algorithm (Friedman et al. 2008), problem (6) is similar to that studied by Rothman et al. (2010) and Yin and Li (2011) in the case of no censoring. However, instead of solving this problem through a cyclic coordinate descent algorithm, Augugliaro et al. (2020) use a more efficient and easy-to-implement block-coordinate descent algorithm.

3 Who is afraid of missing data?

In the paper, the authors study an application of their methods to gene expressions of 48 genes. Just like in many multivariate applications, the chances that one of those 48 measurements is missing or corrupted in some way can be considerable. In fact, the original study reports 414 multiple myeloma samples (Chapman et al. 2011), whereas the authors only consider 154 samples without missing values. This will clearly have a downstream impact on the analysis.

As an illustrative example, we consider the subset of 304 samples from the same study, provided to us by the authors. Figure 1 (left) shows the regulatory network of the same 48 genes, inferred from the samples in the first stage of myeloma, defined according to the serum beta-2 microglobulin and serum albumin prognostic factors as in the paper. We use a Gaussian graphical model and perform L_1 penalised inference accounting for missing data (`cglasso` R package Augugliaro et al. 2020). The red edges are those that would be inferred using only the 35 samples that are fully observed, clearly providing only a partial view of the underlying network.

The right figure shows the network inferred from samples classified to belong to the three stages of the disease (168 samples, including 1.6% of missing data). Here we decide to include three covariates in the mean of the model, namely the classification of the samples in the three stages, the gender of the patient and whether they were treated or not. Using again a penalised inference approach, both on the precision matrix and on the regression coefficients (`cglasso` R package Augugliaro et al. 2020), the figure shows the reconstructed network, after accounting for a difference in mean across the three stages and the treatment status for two of the genes. Although this analysis does not have the same objective as the analysis conducted in the paper and it is much simpler in many respects, it is worth pointing out how it took 0.25 s across a default grid of 100 values for the two tuning parameters. A significant difference to the 47 hours needed for the richer analysis presented in the paper. We hope to see further computational developments in Bayesian inferential procedures, in order to allow for more in-depth fine tuning of individual analyses as well as to avoid unnecessary selections of the variables (nodes) to be investigated.

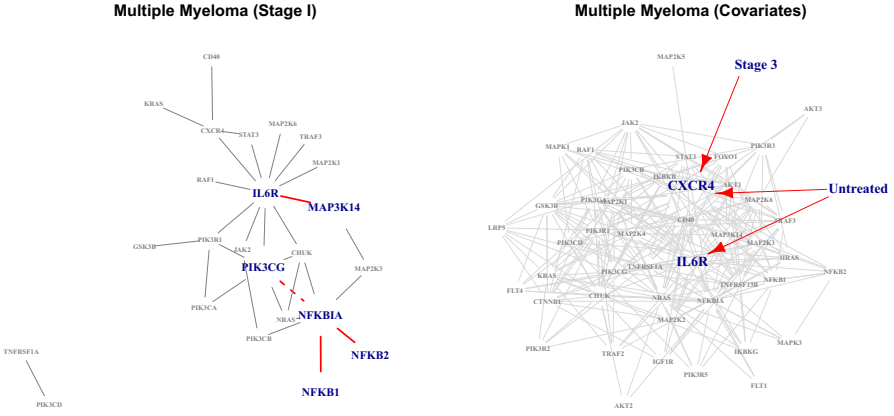

Fig. 1 Regulatory network in multiple myeloma samples. Left: The eBIC selected network, inferred from the 83 samples from Stage I myeloma. The red edges are those inferred from the 35 fully observed samples (solid: in common, dashed: additional). Right: The eBIC-optimal network from 168 samples (Stage I, II and III), including a mean dependency on three covariates (Stage, Gender and Treatment)

4 Is there life beyond Gauss?

Normality is celebrated and despised in equal measure around the scientific community. Whereas it is the workhorse of many elegant theoretical results as well as many practical implementations, there is a vocal opposition that questions its uncritical use. As any statistician knows—but which sometimes gets forgotten by overly cautious practitioners—it is important to draw a distinction between requiring the data to be normal or assuming that the estimated parameters are normal. The former is typically not necessary for the latter.

4.1 Spurious links: model selection issues

The case considered here, however, is fundamentally different from typical discussions of normality. Here the normality assumption of the data is crucial in order to be able to connect zeroes in the precision matrix with the absence of a link in the conditional independence graph, as the authors state early on in Sect. 2.1. So what happens if normality is violated? In general, as shown e.g. in Abegaz and Wit (2015), conditional independence does not correspond anymore to zeroes in the precision matrix. Therefore, any Bayesian or frequentist method that aims to identify zeroes in the precision matrix will result in incorrect identification of the conditional independence graphs. In a small simulation study, we found that any deviation from normality leads to lower true positive and true negative rates, and that this is particularly pronounced for skewed distributions.

In short, deviations from normality lead to the identification of spurious links as well as not being able to detect true edges.

4.2 Gaussian copula graphical models

There have been a number of suggestions to deal with non-Gaussian graphical models. In fact, there is a rich literature of graphical models for discrete data (Madigan et al. 1995). More recently, the nonparanormal graphical model was developed for dealing with arbitrary distributions (Liu et al. 2009). Although the original approach was entirely based on a frequentist approach, the method can be formalized more generally via a Gaussian copula. The beauty of this method is that it can model arbitrary marginal distributions, while retaining a simple Gaussian covariance structure. Particularly simple is the case of absolute continuous Gaussian copula random variable Y, whose components Y_i have arbitrary marginal distributions F_i and the joint distribution of $(\Phi^{-1}(F_1(Y_1)), \ldots, \Phi^{-1}(F_d(Y_d)))$ is multivariate normally distributed. This approach can also easily be incorporated in the framework proposed by Ni and co-authors.

5 Conclusions

It has been an absolute pleasure and honour to be commenting on the excellent discussion paper presented here. The authors have put forward an important framework to make graphical models more useful in applied settings. They should be congratulated for this. In this discussion paper, we have merely been giving a number of suggestions for making this framework even more suitable for practical scenarios.

In particular, firstly, we showed that an alternative and simplified definition of covariate might make the framework more manageable in high-dimensional settings. Secondly, we pointed out that the inclusion of missing variables is important for practical data analysis. Fortunately, the Bayesian framework is particularly suited for this and so, conceptually, including this feature in the method should be straightforward. Finally, we commented on the fact that the Gaussianity assumption is clearly a crucial assumption for identifying the underlying conditional independence graph. However, recent ideas concerning gaussian Copula graphical models should also allow the current framework to be extended to deal with more general distributions.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abegaz F, Wit E (2015) Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. Stat Neerl 69(4):419

Augugliaro L, Abbruzzo A, Vinciotti V (2020) l 1 penalized censored Gaussian graphical model. Biostatistics 21(2):e1–e16. https://doi.org/10.1093/biostatistics/kxy043

Augugliaro L, Sottile G, Vinciotti V (2020) The conditional censored graphical lasso estimator. Stat Comput 30:1273

Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M et al (2011) Initial genome sequencing and analysis of multiple myeloma. Nature 471(7339):467

Friedman JH, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432

Lafferty J, McCallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning 2001 (ICML 2001), pp. 282–289. https://doi.org/10.5555/645530.655813

Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford

Liu H, Lafferty J, Wasserman L (2009) The nonparanormal: semiparametric estimation of high dimensional undirected graphs. J Mach Learn Res 10(10):229

Madigan D, York J, Allard D (1995) Bayesian graphical models for discrete data. Int Stat Rev 63(2):215

Ni Y, Baladandayuthapani V, Vannucci M, Stingo FC (2021) Bayesian graphical models for modern biological applications. Stat Methods Appl. https://doi.org/10.1007/s10260-021-00572-8

Rothman AJ, Levina E, Zhu J (2010) Sparse multivariate regression with covariance estimation. J Comput Graph Stat 19(4):947. https://doi.org/10.1198/jcgs.2010.09188

Städler N, Bühlmann P (2012) Missing values: sparse inverse covariance estimation and an extension to sparse regression. Stat Comput 22(1):219. https://doi.org/10.1007/s11222-010-9219-7
Yin J, Li H (2011) A sparse conditional Gaussian graphical model for analysis of genetical genomics data. Ann Appl Stat 5(4):2630. https://doi.org/10.1214/11-AOAS494