Structure learning of undirected graphical models for count data

Nguyen Thi Kim Hue
Department of Statistical Sciences
University of Padova
Via C. Battisti, 241 - 35121 Padova, Italy

Monica Chiogna
Department of Statistical Sciences
University of Padova
Via C. Battisti, 241 - 35121 Padova, Italy

Abstract

Biological processes underlying the basic functions of a cell involve complex interactions between genes. From a technical point of view, these interactions can be represented through a graph where genes and their connections are, respectively, nodes and edges. The main objective of this paper is to develop a statistical framework for modelling the interactions between genes when the activity of genes is measured on a discrete scale. In detail, we define a new algorithm for learning the structure of undirected graphs, PC-LPGM, proving its theoretical consistence in the limit of infinite observations. The proposed algorithm shows promising results when applied to simulated data as well as to real data.

Keywords: Graphical models, Undirected graphs, Structure learning, Sparsity, Conditional independence tests.

1. Introduction

Current demand for modelling complex interactions between genes, combined with the greater availability of high-dimensional discrete data, possibly showing a large number of zeros and measured on a small number of units, has led to an increased focus on structure learning for discrete data in high dimensional settings.

Various solutions are nowadays available in the literature for learning (sparse) graphical models for discrete data. Höfling and Tibshirani (2009) consider the problem of estimating the parameters as well as the structure of binary-valued Markov networks; Ravikumar et al. (2010) consider the problem of estimating the graph associated with a binary Ising Markov random field; Jalali et al. (2011) consider learning general discrete graphical models, where each variable can take a multiplicity of possible values, and factors can be of order higher than two and Allen and Liu (2013) consider learning graphical models for Poisson counts. To deal with high dimensionality, most methods resort on penalization, which simultaneously performs parameter estimation and model selection.

In this paper, we concentrate on count data and introduce a simple algorithm for structure learning of undirected graphical models, called PC-LPGM, particularly useful when
sparse graphs are under consideration. The algorithm stems from the conditional approach of Allen and Liu (2013), where the neighbourhood of each node is estimated in turn by solving a lasso penalized regression problem and the resulting local structures stitched together to form the global graph. We propose to substitute penalized estimation with a testing procedure on the parameters of the local regressions following the lines of the PC algorithm, see Spirtes et al. (2000). This solution is particularly attractive, since it inherits the potential of the PC algorithm to estimate a sparse graph even if \( p \), the number of variables, is in the hundreds or thousands.

We give a theoretical proof of convergence of PC-LPGM that shows that the proposed algorithm consistently estimates the edges of the underlying (sparse) undirected graph, as the sample size \( n \to \infty \). For such proof to be developed, a joint distribution must exist, a condition which might be questionable when relying on a conditional model specification such as the one behind a neighbourhood approach. If one assumes that each variable conditioned on all other variables follows a Poisson distribution, for example, a unique joint distribution compatible with the given conditionals exists provided that conditional dependencies are all negative. As this condition, known as “competitive relationship” among variables, highly limits attractiveness of such specification in applications, we have chosen to develop statistical guarantees for PC-LPGM under the assumption that conditional distributions follow a truncated Poisson law. Such choice admits dependencies richer than those under competitive relationship; see, however, Yang et al. (2013) for a discussion about its limitations. For the truncated Poisson model, under mild assumptions on the expected Fisher information matrix, and fixing the truncation point \( R > 0 \), convergence is guaranteed for \( n > O_p (p^2 \log p) \) (or \( n > \max \{O_p (m \log p), O_p (m^2 \log m)\} \)), where \( m \) is the maximum neighbourhood size).

To explore whether it is reasonable to extend the desirable properties of PC-LPGM to the case of conditional Poisson distributions with unrestricted conditional dependencies, extensive simulations studies are conducted to empirically evaluate statistical properties of the algorithm in such cases.

The paper is organized as follows. After reviewing some essential concepts on undirected graphical models and Truncated Poisson models in Section 2, we introduce PC-LPGM algorithm in Section 3. We then provide statistical guarantees in Section 4. Properties of the algorithm in the setting of conditional Poisson distributions with unrestricted conditional dependencies are explored, also relative to various alternatives, in Section 5. A validation of the algorithm on a real case is offered in Section 6 and some concluding remarks are presented in Section 7.

2. A quick review on truncated Poisson undirected graphical models

In this section, we review some essential concepts on undirected graphical models and introduce truncated Poisson undirected graphical models.

Consider a \( p \)-dimensional random vector \( \mathbf{X} = (X_1, \ldots, X_p) \) such that each random variable \( X_t \) corresponds to a node of a graph \( G = (V, E) \) with index set \( V = \{1, 2, \ldots, p\} \). An edge between two nodes \( s \) and \( t \) will be denoted by \( (s, t) \). The neighbourhood of a node \( s \in V \) is defined to be the set \( N(s) = \{t \in V : (s, t) \in E\} \) consisting of all nodes connected

\[ \text{NGUYEN AND CHIOGNA} \]
to $s$. The random vector $\mathbf{X}$ satisfies the pairwise Markov property with respect to $G$ if

$$X_s \perp \!\!\!\!\perp X_t | \mathbf{X}_{V \setminus \{s,t\}}$$

whenever $(s, t) \notin E$. When all variables $X_s, s \in V$, are discrete with positive joint probabilities, as in the case under consideration, the pairwise Markov property coincides with the local and global Markov property, according to which, respectively,

$$X_s \perp \!\!\!\!\perp \mathbf{X}_{V \setminus \{N(s) \cup \{s\}\}} | \mathbf{X}_{N(s)}$$

for every $s \in V$, and

$$\mathbf{X}_A \perp \!\!\!\!\perp \mathbf{X}_B | \mathbf{X}_C,$$

for any triple of pairwise disjoint subsets $A, B, C \subset V$ such that $C$ separates $A$ and $B$ in $G$, that is, every path between a node in $A$ and a node in $B$ contains a node in $C$.

To specify a probabilistic model for $\mathbf{X}$, we take a conditional approach. Assume that each conditional distribution of node $X_s$ given other variables $\mathbf{X}_{V \setminus \{s\}}$ follows a Poisson distribution truncated at $R$, $R > 0$, written as $X_s | \mathbf{X}_{V \setminus \{s\}} \sim TP(\exp(\theta_s + \sum_{t \neq s} \theta_{st} x_t))$, with node conditional distribution

$$P(x_s | \mathbf{x}_{V \setminus \{s\}}) = \frac{\exp \{\theta_s x_s + \sum_{t \neq s} \theta_{st} x_t x_s - \log x_s!\}}{\sum_{k=0}^{R} \exp \{\theta_s k + k \sum_{t \neq s} \theta_{st} x_t - \log k!\}}$$

$$= \frac{\exp \{\theta_s x_s + x_s \langle \theta_s, \mathbf{x}_{V \setminus \{s\}} \rangle - \log x_s!\}}{\sum_{k=0}^{R} \exp \{\theta_s k + k \langle \theta_s, \mathbf{x}_{V \setminus \{s\}} \rangle - \log k!\}}$$

$$= \exp \{\theta_s x_s + x_s \langle \theta_s, \mathbf{x}_{V \setminus \{s\}} \rangle - \log x_s! - D(\langle \theta_s, \mathbf{x}_{V \setminus \{s\}} \rangle)\},$$

(1)

where $\theta_s = \{\theta_{st}, t \in V, t \neq s\}$ denotes the set of conditional dependence parameters, $\langle \cdot, \cdot \rangle$ denotes the inner product, and $D(\langle \theta_s, \mathbf{x}_{V \setminus \{s\}} \rangle) = \log (\sum_{k=0}^{R} \exp \{\theta_s k + k \langle \theta_s, \mathbf{x}_{V \setminus \{s\}} \rangle - \log k!\})$.

An application of Proposition 1 in Yang et al. (2015) shows that a valid joint probability distribution function from the above given set of specified conditional distributions can be constructed. By Assumption 1 and Assumption 2 in Section 4.1 of Besag (1974), such distribution defines an undirected graph $G = (V, E)$ in which a missing edge between node $s$ and node $t$ corresponds to the condition $\theta_{st} = \theta_{ts} = 0$. On the other side, one edge between node $s$ and node $t$ implies $\theta_{st} \equiv \theta_{ts}$.

The existence of a joint distribution suggests that the structure of the network might be recovered from observed data within a likelihood approach by mean of a set of statistical tests. Indeed, in an undirected graphical model, the pairwise Markov property infers a collection of full conditional independences encoded in absent edges. For this reason, performing $\binom{|V|}{2}$ pairwise full conditional independence tests yields a method to estimate the graph $G$. However, such an approach might be impractical even for modestly sized graphs. The existence of the maximum likelihood estimates is, in general, not guaranteed if the number of observations is small, the basic problem being that the number of parameters in $\theta$ is of the order $p^2$. Hence, the sample size is often not large enough to obtain a good estimator. Moreover, it requires computing complex normalization constants and combinatorial searches through the space of graph structures. For this reason, in what follows, we
will exploit the local Markov property, according to which every variable is conditionally independent of the remaining ones given its neighbours. This property suggests that each variable \( X_s, s \in V \) can be optimally predicted from its neighbour \( X_{N(s)} \).

3. The PC-LPGM algorithm

We will work within the neighbourhood selection approach. The analysis of this setting is related to the concept of pseudo-likelihood,

\[
P_L(\theta) = \prod_{s \in V} \mathbb{P}(x_s | \mathbf{x}_{V \setminus \{s\}}),
\]

where \( \mathbb{P}(x_s | \mathbf{x}_{V \setminus \{s\}}) \) is the distribution function of each node conditional distribution. Standard model specifications treat different conditional distributions \( \mathbb{P}(x_s | \mathbf{x}_{V \setminus \{s\}}) \) as unrelated. In other words, the symmetry of interaction parameters \( \theta_{st} \) and \( \theta_{ts} \) is usually not explicitly taken into account (see, however, Peng et al. (2009) for a solution that takes the natural symmetry of coefficients into account in the Gaussian setting).

In this setting, structure learning usually proceeds by disjointly maximizing the single factors in \( P_L(\theta) \). In high-dimensional sparse settings, many up-to-date algorithms are based on solving local convex optimization problems, typically formed by the sum of a loss function, such as the local negative log likelihood, with a sparsity inducing penalty function. Each local penalized estimate \( \hat{\theta}_s \) is then combined into a single non-degenerate global estimate, possibly employing consensus operators aimed at solving inconsistencies with respect to parameters shared between factors (see, for example, Mizrahi et al., 2014). From empirical studies, it is in most cases easy to check that such algorithms converge, sometimes also reasonably quickly thanks to the possibility of distributing the various maximization tasks. However, it is not immediately clear if convergence can be established theoretically, so that it cannot be given for granted that such algorithms ultimately yield correct graphs.

Our proposal, called PC-LPGM, is a pseudo-likelihood based algorithm that stems from current neighbourhood selection methods for count data (see Allen and Liu, 2013), but substitutes penalization with hypothesis testing. In Section 4, it is rigorously demonstrated that the sequence of tests does indeed converge to the true structure in the limit of infinite observations, regardless of the dimension of the problem.

We consider the same model specification as in (1). In detail, we assume that each node conditional distribution follows a truncated Poisson distribution. As we are only interested in the structure of graph \( G \), without loss of generality we can assume \( \theta_s = 0, \ s \in V \). In line with the most common solutions, we also treat the conditional distributions \( \mathbb{P}_{\theta_s}(x_s | \mathbf{x}_{V \setminus \{s\}}) \) as unrelated.

In PC-LPGM, neighbours are identified by mean of conditional independence tests built from the conditional models and aimed at identifying the set of non-zero conditional dependence parameters. Tests are based on Wald type statistics built on exploiting the asymptotic normality of the local maximum likelihood estimators. To face the high computational complexity related to the testing procedure, we employ the PC algorithm, which relies on controlling the number of variables in the conditional sets, a strategy particularly effective when sparse graphs are under consideration.
In what follows, let \( \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)} \) be \( n \) independent \( p \)-random vectors drawn from \( \mathbf{X} \), where \( \mathbf{X}^{(i)} = (X_{i1}, \ldots, X_{ip}) \); and \( \mathbb{X} = \{ \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \} \) be the collection of \( n \) samples drawn from the random vectors \( \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)} \), with \( \mathbf{x}^{(i)} = (x_{i1}, \ldots, x_{ip}), \quad i = 1, \ldots, n \). For each \( U \subset V \), let \( \mathbb{X}_U \) be the set of \( n \) samples of the \( |U| \)-random vector \( \mathbf{X}_U = (X_i : i \in U) \), with \( \mathbf{x}^{(i)}_U = (x_{ij})_{j \in U}, \quad i = 1, \ldots, n \). Starting from the complete graph, for each \( s \) and \( t \in V \setminus \{s\} \) and for any set of variables \( S \subset \{1, \ldots, p\} \setminus \{s, t\} \), we test, at some pre-specified significance level, the null hypothesis \( H_0 : \theta_{st|K} = 0 \), with \( K = S \cup \{s, t\} \). In other words, we test if data support existence of the conditional independence relation \( X_s \independent X_t | X_S \). If the null hypothesis is not rejected, the edge \((s, t)\) is considered to be absent from the graph.

A control is operated on the cardinality of the set \( S \) of conditioning variables, which is progressively increased from 0 to \( p - 2 \) or to \( m, \quad m < (p - 2) \).

Assume

\[
X_s | X_{K\setminus\{s\}} \sim \text{TP}\left( \exp \left\{ \sum_{t \in K \setminus \{s\}} \theta_{st|K} x_t \right\} \right), \quad \forall s \in K \subset \{1, \ldots, p\},
\]

and denote \( \mathbf{\theta}_{s|K} = \{ \theta_{st|K} : t \in K \setminus \{s\} \} \). A rescaled negative node conditional log-likelihood given the conditioning variables \( X_{K\setminus\{s\}} = (X_k : k \in K \setminus \{s\}) \) can be written as

\[
l(\mathbf{\theta}_{s|K}, X_s; X_{K\setminus\{s\}}) = -\frac{1}{n} \log \prod_{i=1}^{n} \mathbb{P}_{\mathbf{\theta}_{s|K}}(x_{is} | X_{K\setminus\{s\}}) = \frac{1}{n} \sum_{i=1}^{n} \left[ -x_{is}(\mathbf{\theta}_{s|K}, X_{K\setminus\{s\}}) + \log x_{is} + D((\mathbf{\theta}_{s|K}, X_{K\setminus\{s\}})) \right],
\]

where the scaling factor is taken for later mathematical convenience. The estimate \( \hat{\mathbf{\theta}}_{s|K} \) of the parameter \( \mathbf{\theta}_{s|K} \) is determined by minimizing the rescaled negative conditional log-likelihood given in Equation (3), i.e.,

\[\hat{\mathbf{\theta}}_{s|K} = \arg\min_{\mathbf{\theta}_{s|K} \in \mathbb{R}^{|K|-1}} l(\mathbf{\theta}_{s|K}, X_s; X_{K\setminus\{s\}}).\]

A Wald-type test statistic for the hypothesis \( H_0 : \theta_{st|K} = 0 \) can be obtained from asymptotic normality of \( \hat{\mathbf{\theta}}_{s|K} \),

\[\sqrt{n}(\mathbf{\theta}_{s|K} - \hat{\mathbf{\theta}}_{s|K}) \xrightarrow{d} N(0, I(\mathbf{\theta}_{s|K})^{-1}),\]

where \( I(\mathbf{\theta}_{s|K}) \) denotes the expected Fisher information matrix,

\[I(\mathbf{\theta}_{s|K}) = \mathbb{E}_{\theta_s} \left[ \frac{\partial^2 l(\mathbf{\theta}_{s|K}, X_s; X_{K\setminus\{s\}})}{\partial^2 \theta_{s|K}} \right],\]

which holds under fairly general regularity conditions. The test statistic for the null hypothesis \( H_0 : \theta_{st|K} = 0 \) can be obtained on exploiting the marginal asymptotic normality of the component \( \theta_{s|K} \).
In practice, the observed information $J(\theta_{s|K}) = n \frac{\partial^2 l(\theta_{s|K}; X_{\{s\}}; X_{K\{s\}})}{\partial^2 \theta_{s|K}}$, that is, the second derivative of the negative log-likelihood function, is more conveniently used evaluated at $\hat{\theta}_{s|K}$ as variance estimate of maximum likelihood quantities instead of the expected Fisher information matrix, a modification which comes from the use of an appropriately conditioned sampling distribution for the maximum likelihood estimators. Following this line, the test statistic for the hypothesis $H_0 : \theta_{st|K} = 0$ is given by

$$Z_{st|K} = \frac{\sqrt{n}\hat{\theta}_{st|K}}{\sqrt{[J(\hat{\theta}_{s|K})^{-1}]}_{tt}}, \quad (4)$$

where $[A]_{jj}$ denotes the element in position $(j,j)$ of matrix $A$. It is readily available that $Z_{st|K}$ is asymptotically standard normally distributed under the null hypothesis, provided that some general regularity conditions hold (Lehmann, 1986, page 185). Possible inconsistencies with respect to parameters shared between local conditional models are solved by removing edge $(s,t)$ if either $H_0 : \theta_{st|K} = 0$ or $H_0 : \theta_{ts|K} = 0$ is not rejected.

The conditional independence tests are prone to mistakes. Moreover, incorrectly deleting or retaining an edge would result in changes in the neighbour sets of other nodes, as the graph is updated dynamically. Therefore, the resulting graph is dependent on the order in which the conditional independence tests are performed. To avoid this problem, we employ the solution in Colombo and Maathuis (2014), who developed a modification of the PC algorithm that removes the order-dependence, called PC-stable. In this modification, the neighbours of all nodes are searched for and kept unchanged at each particular cardinality $l$ of the set $K_s$. As a result, an edge deletion at one level does not affect the conditioning sets of the other nodes, and thus the output is independent on the variable ordering.

The pseudo-code of our algorithm is illustrated in Algorithm 1, where $\text{adj}(\hat{G}, s) = \{t \in V : (s,t) \in \hat{G}\}$ denotes the estimated set of all nodes that are adjacent to $s$ on the graph $\hat{G}$. We note that the pseudo-code is identical to Algorithm 4.1 in Colombo and Maathuis (2014). Indeed, the difference lies in the statistical procedure used to test the hypothesis at line 15.

4. Statistical Guarantees

In this section, we address the property of statistical consistency of our algorithm. In detail, we study the limiting behaviour of our estimation procedure as the sample size $n$, and the model size $p$ go to infinity. In what follows, we derive uniform consistency of our distributed estimators explicitly as a function of the sample size, $n$, the number of nodes, $p$, the truncation point $R$. Moreover, we prove consistency of the graph estimator as a function of the previous quantities and of the maximum number of neighbours, $m$, by assuming that the true distribution is faithful to the graph. We acknowledge that our results are based on the work of Yang et al. (2012) for exponential family models, combined with ideas coming from Kalisch and Bühlmann (2007). In detail, we borrowed some ideas from the proof of consistency of estimators in $l_1$ regularized local models given in Yang et al. (2012) and we adapted to our setting the ideas of Kalisch and Bühlmann (2007) for proving consistency of the graph estimator.
For the readers’ convenience, before stating the main result, we summarize some notation that will be used throughout this proof. Given a vector $v \in \mathbb{R}^p$, and a parameter $q \in [0, \infty]$, we write $\|u\|_q$ to denote the usual $l_q$ norm. Given a matrix $A \in \mathbb{R}^{p \times p}$, denote the largest and smallest eigenvalues as $\Lambda_{\text{max}}(A)$, $\Lambda_{\text{min}}(A)$, respectively. We use $|||A|||_2 = \sqrt{\Lambda_{\text{max}}(A^T A)}$ to denote the spectral norm, corresponding to the largest singular value of $A$, and the $l_\infty$ matrix norm is defined as $|||A|||_\infty = \max_{i=1,...,a} \sum_{j=1}^{a} |A_{i,j}|$.

**Algorithm 1** The PC-LPGM algorithm.

1. **Input:** $n$ independent realizations of the $p$-random vector $X; x^{(1)}, x^{(2)}, \ldots, x^{(n)}$; an ordering $\text{order}(V)$ on the variables, (and a stopping level $m$).

2. **Output:** An estimated undirected graph $\hat{G}$.

3. Form the complete undirected graph $\tilde{G}$ on the vertex set $V$.

4. $l = -1$; $\hat{G} = \tilde{G}$

5. repeat

6. $l = l + 1$

7. for all vertices $s \in V$, do

8. let $K_s = \text{adj}(\hat{G}, s)$

9. end for

10. repeat

11. Select a (new) ordered pair of nodes $s, t$ that are adjacent in $\hat{G}$ such that $|K_s \{t\}| \geq l$, using $\text{order}(V)$.

12. repeat

13. choose a (new) set $S \subset K_s \{t\}$ with $|S| = l$, using $\text{order}(V)$.

14. if $H_0 : \theta_{st|S} = 0$ not rejected

15. delete edge $(s, t)$ from $\hat{G}$

16. end if

17. until edge $(s, t)$ is deleted or all $S \subset K_s \{t\}$ with $|S| = l$ have been considered.

18. until all ordered pair of adjacent variables $s$ and $t$ such that $|K_s \{t\}| \geq l$ and $S \subset K_s \{t\}$ with $|S| = l$ have been tested for conditional independence.

19. until $l = m$ or for each ordered pair of adjacent nodes $s, t$: $|\text{adj}(\hat{G}, s) \{t\}| < l$.

4.1 Assumptions

We will begin by stating the assumptions that underlie our analysis, and then give a precise statement of the main result.

Denote the population Fisher information and the sample Fisher information matrix corresponding to the covariates in model (2) with $K = V$ as follows

$$I_s(\theta_s) = -\mathbb{E}_{\theta} \left( \nabla^2 \log \left( \mathbb{P}_{\theta_s}(X_s|X_{V\{s\}}) \right) \right),$$

and

$$Q_s(\theta_s) = \nabla^2 \ell(\theta_s, X_s; X_{V\{s\}}).$$

We note that we will consider the problem of maximum likelihood on a closed and bounded dish $\Theta \subset \mathbb{R}^{(p-1)}$. For $\theta_s|K \in \mathbb{R}^{|K|-1}$, we can immerse $\theta_s|K$ into $\Theta \subset \mathbb{R}^{(p-1)}$ by zero-pad $\theta_s|K$ to include zero weights over $\{V\setminus K\}$.  

7
Assumption 4.1 The coefficients $\theta_{s|K} \in \Theta$ for all sets $K \subset V$ and all $s \in K$ have an upper bound norm, $\max_{s,t,K} |\theta_{st|K}| \leq M$, $\forall \theta_{st|K} \neq 0$, and a lower bound norm, $\min_{s,t,K} |\theta_{st|K}| \geq c$, $\forall \theta_{st|K} \neq 0$, where $t \in K$.

Assumption 4.2 The Fisher information matrix corresponding to the covariates in model (2) with $K = V$ has bounded eigenvalues; that is, there exists a constant $\lambda_{\min} > 0$ such that

$$\Lambda_{\min}(I_s(\theta_s)) \geq \lambda_{\min}, \forall \theta_s \in \Theta.$$ 

Moreover, we require that

$$\Lambda_{\max}\left(\mathbb{E}_\theta \left(X_{V \setminus \{s\}}^T X_{V \setminus \{s\}}\right)\right) \leq \lambda_{\max}, \forall s \in V, \forall \theta \in \Theta,$$

where $\lambda_{\max}$ is some constant such that $\lambda_{\max} < \infty$.

The first assumption simply bounds the effects of covariates in all local models. In other words, we consider parameters $\theta_{st|K}$ belong to a compact set bounded by $M$. Being the expected value of the rescaled negative log-likelihood twice differentiable, the lower bound on the eigenvalues of the Fisher information matrix in the second assumption guarantees strong convexity in all partial models. Condition on the upper eigenvalue of the covariance matrix guarantees that the relevant covariates do not become overly dependent, a requirement which is commonly adopted in these settings.

4.2 Convergence guarantees of local estimators

We are now ready to consider the question of whether convergence guarantees can be proved in the setting of our interest. Before proving our main theorem, we show some intermediate results of independent interest (see Appendix A for related proofs).

Proposition 4.3 Assume 4.1-4.2 and let $K \subset V$. Then, for all $s \in K$ and any $\delta > 0$

$$\mathbb{P}_\theta(\|\nabla l(\theta_{s|K}, X_{\{s\}}; X_{K \setminus \{s\}})\|_\infty \geq \delta) \leq \exp\{-c_1 n\},$$

$\forall \theta_{s|K} \in \Theta$, when $n \to \infty$.

Theorem 4.4 Assume 4.1-4.2 and let $K \subset V$. Then there exists a non-negative decreasing sequence $\delta_n \to 0$, such that

$$\mathbb{P}_\theta(\|\hat{\theta}_{s|K} - \theta_{s|K}\|_2 \leq \delta_n) \geq 1 - \exp\{-cn\}, \forall s \in K, \theta \in \Theta,$$

when $n \to \infty$.

We now proceed to consider uniform consistency of the local estimators. Let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_p)^T$ be the array of rowwise local estimators $\hat{\theta}_{s|K}$ with $K = V$. We can state the following theorem, which extends Theorem 4.4 without any additional conditions.

Theorem 4.5 (uniform consistency) Assume 4.1-4.2. Then, $\hat{\theta}$ converges in probability to $\theta$, the true value, as $n$ increases, uniformly in $\theta$. 

8
Proof We have to show that given \( \epsilon > 0, \mu > 0 \), there exists an integer \( n_0 \) dependent on \( \epsilon \) and \( \mu \) but not on \( \theta \), such that for all \( n > n_0 \),
\[
P_{\theta}(||\theta - \hat{\theta}||_2 \leq \epsilon) \geq 1 - \mu.
\]

Take \( \frac{\epsilon}{p} \) as the number \( \delta_n \), and the \( \frac{\mu}{p} \) to be the \( \exp\{ -cn \} \) in Theorem A.3. Then, for each \( s \in V \), there exist \( n_s \), such that for all \( n > n_s \)
\[
P_{\theta} \left( \| \hat{\theta}_s - \theta_s \|_2 \leq \frac{\epsilon}{p} \right) \geq 1 - \frac{\mu}{p}.
\]

Let \( \Omega_s \) be the space such that for all \( X \in \Omega_s \),
\[
\| \hat{\theta}_s - \theta_s \|_2 \leq \frac{\epsilon}{p},
\]
and \( P_{\theta}(\Omega_s) \geq 1 - \frac{\mu}{p} \). Define \( n_0 = \max_{s \in V} \{ n_s \} \) and \( \Omega = \cap_{s \in V} \Omega_s \). Then, for all \( X \in \Omega \),
\[
||| \theta - \hat{\theta} |||_2 \leq \sqrt{ \sum_{s=1}^{p} \sum_{t \neq s} | \theta_{st} - \hat{\theta}_{st} |^2 }
\]
\[
= \sqrt{ \sum_{s=1}^{p} \| \theta_s - \hat{\theta}_s \|_2^2 }
\]
\[
\leq \sqrt{ p \left( \frac{\epsilon}{p} \right)^2 }
\]
\[
\leq \epsilon.
\]

Moreover, it is easy to prove by induction that
\[
P_{\theta}(\Omega) \geq 1 - p \frac{\mu}{p} = 1 - \mu.
\]

Hence, for all \( X \in \Omega \), we have \( ||| \theta - \hat{\theta} |||_2 \leq \epsilon \), and \( P_{\theta}(\Omega) \geq 1 - \mu \). In other words, we have
\[
P_{\theta}(||| \theta - \hat{\theta} |||_2 \leq \epsilon) \geq 1 - \mu.
\]

\[\blacksquare\]

Remark 1 With suitable modifications, uniform consistency can be proved in the case of Poisson node conditional distributions with “competitive relationships” between variables, that is, with only negative conditional interaction parameters. Analogously, it can be extended to other distributions for count data belonging to the exponential family, such as the Negative Binomial distribution, provided that a joint distribution compatible with the conditional specifications can be constructed.
Remark 2 Convergence of the pseudo likelihood estimator $\hat{\theta}$ might also have been proved by characterizing its asymptotic behaviour in terms of law of large numbers. Indeed, the pseudo likelihood estimator $\hat{\theta}$ can be proved to converge to the true parameter value when some conditions on the parameter space $\theta$ and moments of the variables $X$ are satisfied (see, for example, Theorem 5.7 from Van der Vaart, 2000). It is worth noting that our proof allows to highlight the relative scaling of $n$, $p$ and $R$ needed to reach convergence.

4.3 Consistency of the graph estimator

In what follows, we assume faithfulness of the truncated Poisson node conditional distributions to the graph $G$. We restrict the parameter space $\Theta$ to the subspace, $\Omega(\Theta)$ say, on which the faithfulness condition is guaranteed. We recall that a distribution $P_X$ is said to be faithful to the graph $G$ if

$$X_A \perp \perp X_B | X_C \Rightarrow A \perp \perp G B | C,$$

for all disjoint vertex sets $A, B, C$. It is worth noting that faithfulness of the local distributions guarantees faithfulness of the joint distributions, thanks to the equivalence between local and global Markov property.

Now we state the main result of this work for the consistency of the graph estimate. We note that PC-LPGM employs a modification of the PC algorithm, PC-stable. However, the proof of consistency of the algorithm in Kalisch and Bühlmann (2007) is unchanged.

**Theorem 4.6** Assume 4.1- 4.2. Denote by $\hat{G}(\alpha_n)$ the estimator resulting from from Algorithm 1, and by $G$ the true graph. Then, there exists a numerical sequence $\alpha_n \to 0$, such that

$$\mathbb{P}_{\theta}(\hat{G}(\alpha_n) = G) = 1, \forall \theta \in \Omega(\Theta),$$

when $n \to \infty$.

**Proof** Let $\hat{\theta}_{st|K}$ and $\theta^*_{st|K}$ denote the estimated and true partial weights between $X_s$ and $X_t$ given $X_r, r \in S$, where $S = K \setminus \{s, t\} \subset \{1, \ldots, p\} \setminus \{s, t\}$. Many partial weights are tested for being zero during the run of the PC-procedure. For a fixed ordered pair of nodes $s, t$, the conditioning sets are elements of

$$K^m_{st} = \{S \subset \{1, \ldots, p\} \setminus \{s, t\} : |S| \leq m\}.$$

The cardinality is bounded by

$$|K^m_{st}| \leq C p^m, \text{ for some } 0 < C < \infty.$$

Let $E_{st|K}$ denote type I or type II errors occurring when testing $H_0 : \theta_{st|K} = 0$. Thus

$$E_{st|K} = E_{st|K}^I \cup E_{st|K}^{II},$$

in which, for $n$ large enough

- type I error $E_{st|K}^I$: $Z_{st|K} > \Phi^{-1}(1 - \alpha/2)$ and $\theta^*_{st|K} = 0;$

10
type II error $E^I_{st|\mathbf{K}}$: $Z_{st|\mathbf{K}} \leq \Phi^{-1}(1 - \alpha/2)$ and $\theta^*_{st|\mathbf{K}} \neq 0$;

where $Z_{st|\mathbf{K}}$ was defined in (4), and $\alpha$ is a chosen significance level. Consider an arbitrary matrix $\theta_{st|\mathbf{K}} = \{\theta_{st|\mathbf{K}}\}^T_{s,t \in \mathbf{K}} \in \Omega(\Theta)$, such that $|\theta_{st|\mathbf{K}}| \geq \delta$, for some $\delta > 0$. Let $\theta^0_{st|\mathbf{K}}$ be the matrix that has the same elements as $\theta_{st|\mathbf{K}}$ except $\theta^0_{st|\mathbf{K}} = 0$. Choose $\alpha_n = 2(1 - \Phi(n^d))$, with $0 < d < 1/2$, then

$$
\sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}} (E^I_{st|\mathbf{K}}) = \sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}}(\hat{\theta}_{st|\mathbf{K}} > n^{d-1/2} \sqrt{[J(\hat{\theta}_{st|\mathbf{K}})^{-1}]_{tt}}) \\
= \sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}}(\hat{\theta}_{st|\mathbf{K}} - \theta^0_{st|\mathbf{K}} > n^{d-1/2} \sqrt{[J(\hat{\theta}_{st|\mathbf{K}})^{-1}]_{tt}}) \\
\leq \exp\{-cn\},
$$

(6)

using Theorem A.3 and the fact that $n^{d-1/2} \sqrt{[J(\hat{\theta}_{st|\mathbf{K}})^{-1}]_{tt}} \rightarrow 0$ as $n \rightarrow \infty$. Furthermore, with the choice of $\alpha_n$ above, and $\delta \geq 2n^{d-1/2} \sqrt{[J(\hat{\theta}_{st|\mathbf{K}})^{-1}]_{tt}}$,

$$
\sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}} (E^I_{st|\mathbf{K}}) = \sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}}(\hat{\theta}_{st|\mathbf{K}} \leq n^{d-1/2} \sqrt{[J(\hat{\theta}_{st|\mathbf{K}})^{-1}]_{tt}}) \\
= \sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}}(\theta_{st|\mathbf{K}} - \hat{\theta}_{st|\mathbf{K}} \geq |\theta_{st|\mathbf{K}}| - n^{d-1/2} \sqrt{[J(\hat{\theta}_{st|\mathbf{K}})^{-1}]_{tt}}) \\
\leq \sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}}(\theta_{st|\mathbf{K}} - \hat{\theta}_{st|\mathbf{K}} \geq |\theta_{st|\mathbf{K}}| - n^{d-1/2} \sqrt{[J(\hat{\theta}_{st|\mathbf{K}})^{-1}]_{tt}}) \\
\leq \sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}}(\hat{\theta}_{st|\mathbf{K}} - \theta_{st|\mathbf{K}} \geq n^{d-1/2} \sqrt{[J(\hat{\theta}_{st|\mathbf{K}})^{-1}]_{tt}}),
$$

Finally, by Theorem 4.4, we then obtain

$$
\sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}} (E^I_{st|\mathbf{K}}) \leq \exp\{-cn\},
$$

(7)

as $n \rightarrow \infty$. Now, by (5)-(7), we get

$$
\mathbb{P}_{\theta}(\text{a type I or II error occurs in testing procedure}) \leq \mathbb{P}_{\theta}(\bigcup_{s,t,\mathbf{K} \in K^m_{st}} E_{st|\mathbf{K}}) \\
\leq O(p^{m+2}) \sup_{s,t,\mathbf{K} \in K^m_{st}} \mathbb{P}_{\theta^0_{st|\mathbf{K}}} (E_{st|\mathbf{K}}) \\
\leq O(p^{m+2}) \exp\{-cn\} \\
\rightarrow 0.
$$

(8)

as $n \rightarrow \infty$.  \blacksquare
5. Unrestricted Poisson graphical models

It is interesting to ask if consistency of PC-LPGM holds also in the case of Poisson node conditional distributions with unrestricted conditional interaction parameters, although a theoretical proof is still an unsolved question.

We devote this section to an empirical study of consistency of our proposed algorithm in this setting. We aim to measure the ability of PC-LPGM to recover the true structure of the graphs, also in situations where relatively moderate sample sizes are available. As measure of ability, we adopt two measures: PPV that stands for Positive Predictive Value and is defined as TP/(TP+FP); and Sensitivity (Se), defined as TP/(TP+FN), where TP (true positive), FP (false positive), and FN (false negative) refer to the inferred edges.

In doing these studies, we also aim to compare PC-LPGM to a number of popular structure learning algorithms. We therefore consider Local Poisson Graphical Models (LPGM) (Allen and Liu, 2013) and Poisson dependency networks (PDNs) (Hadiji et al., 2015). It is worth remembering that structure learning for discrete undirected graphical models is usually performed by employing methods for continuous data after proper data transformation. We therefore consider two representatives of approaches based on the Gaussian assumption, variable selection with lasso (VSL) (Meinshausen and Bühlmann, 2006), and graphical lasso algorithm (GLASSO) (Friedman et al., 2008). Moreover, we consider two structure learning methods dealing with the class of nonparanormal distributions, the nonparanormal-Copula algorithm (NPN-Copula) (Liu et al., 2009), and the nonparanormal-SKEPTIC algorithm (NPN-Skeptic) (Liu et al., 2012).

5.1 Data generation

For two different cardinalities, \( p = 10 \) and \( p = 100 \), we consider three graphs of different structure: (i) a scale-free graph, in which the node degree distribution follows a powerlaw; (ii) a hub graph, where each node is connected to one of the hub nodes; (iii) a random graph, where presence of edges are independent and identically distributed Bernoulli random variables. To construct the scale-free and hub networks, we employed the \( R \) package \texttt{XMRF}. For the scale-free network, we assumed a power law with parameter 0.01 for the node degree distribution. For the hub network, we assumed two hub nodes for \( p = 10 \), and 5 hub nodes for \( p = 100 \). To construct the random network, we employed the \( R \) package \texttt{igraph} with edge probability 0.2 for \( p = 10 \), and 0.02 for \( p = 100 \). See Figure 1 and 2 for a plot of the three chosen graphs for \( p = 10 \) and \( p = 100 \), respectively.

For each graph, 500 datasets were sampled for three sample sizes, \( n = 200, 1000, 2000 \). To generate the data, we followed the approach in Allen and Liu (2013). Let \( X \in \mathbb{R}^{n \times p} \) be the set of \( n \) independent observations of random vector \( X \). Then, \( X \) is obtained from the following model \( X = YW + \epsilon \), where \( Y = (y_{st}) \) is an \( n \times (p+p(p-2)/2) \) matrix whose entries \( y_{st} \) are realizations of independent random variables \( Y_{st} \sim \text{Pois}(\lambda_{\text{true}}) \) and \( \epsilon = (e_{st}) \) is an \( n \times p \) matrix with entries \( e_{st} \) which are realizations of random variables \( E_{st} \sim \text{Pois}(\lambda_{\text{noise}}) \). Let \( W \) be the adjacency matrix of a given true graph, then the adjacency matrix is encoded by matrix \( W \) as \( W = [I_p, P \odot (1_p\text{tri}(W)^T)]^T \). Here, \( P \) is a \( p \times (p(p-1)/2) \) pairwise permutation matrix, \( \odot \) denotes the element-wise product, and \( \text{tri}(W) \) is the \( (p(p-1)/2) \times 1 \) vectorized upper triangular part of \( W \). As in Allen and Liu (2013), we simulated data at two signal-
to-noise ratio (SNR) levels. We set $\lambda_{\text{true}} = 1$ with $\lambda_{\text{noise}} = 5$ for the low SNR level, and $\lambda_{\text{noise}} = 0.5$ for the high SNR level.

Figure 2: The graph structures for $p = 100$ employed in the simulation studies: (a) scale-free; (b) hub; (c) random graph.

5.2 Results

The considered algorithms are listed below, along with specifications, if needed, of tuning parameters. Algorithms for Gaussian data have been used on log transformed data shifted by 1. Whenever a regularization parameter $\lambda$ had to be chosen, the StARS algorithm (Liu et al., 2010) was employed, which aims to seek the value of $\lambda \in (\lambda_{\text{min}}, \lambda_{\text{max}})$, $\lambda_{\text{opt}}$ say, leading to the most stable set of edges. We refer the reader to Appendix C, for details on the StARS algorithm and its tuning parameters, in particular the variability threshold $\beta$ and the number of subsamplings $B$. It is worth noting that, whenever the graph corresponding to $\lambda_{\text{opt}}$ was empty, we shifted to the first nonempty graph (if it existed) in the decreasing regularization path. We therefore considered:

- **PC-LPGM**: level of significance of tests 1%;
- **LPGM:** $\beta = 0.05; \lambda = 20; \frac{\lambda_{\text{min}}}{\lambda_{\text{max}}} = 0.01; \gamma = 0.001$
- **VSL:** $\beta = 0.1; B = 20$
- **GLASSO:** $\beta = 0.1; B = 20$
- **NPN-Copula:** $\beta = 0.1; B = 20$
- **NPN-Skeptic:** $\beta = 0.1; B = 20$

For the two considered vertex cardinalities, $p = 10, 100$, and for the chosen sample sizes $n = 200, 1000, 2000$, Table 1 and Table 2 report, respectively, Monte Carlo means of TP, FP, FN, PPV and Se for each of considered method at low ($\lambda_{\text{noise}} = 5$) and high ($\lambda_{\text{noise}} = 0.5$) SNR levels. Each value is computed as an average of the 1500 values obtained by simulating 500 samples for each of the three networks. Monte Carlo means (and standard deviations) of the same quantities disaggregated by network type are given in Appendix D, Tables 3 – 6. These results indicate that the PC-LPGM algorithm is consistent and outperforms, on average, Gaussian-based competitors (VSL, GLASSO), nonparanormal-based competitors (NPN-Copula, NPN-Skeptic) as well as the state-of-the-art algorithms that are designed specifically for Poisson graphical models (LPGM, PDN) on average in terms of reconstructing the structure from given data.

![Figure 3](image-url)  
**Figure 3:** Number of TP edges recovered by PC-LPGM; LPGM; PDN; VSL; GLASSO; NPN-Copula; NPN-Skeptic for networks in Figure 1 ($p = 10$) and sample sizes $n = 200, 1000, 2000$. First panel row corresponds to high SNR level ($\lambda_{\text{noise}} = 0.5$); second panel row corresponds to low SNR level ($\lambda_{\text{noise}} = 5$).

When $p = 10$, the PC-LPGM algorithm reaches the highest TP value, followed by the PDN and the LPGM algorithms. When $n \geq 1000$, PC-LPGM recovers almost all edges for both low and high SNR levels, see Figure 3. A closer look at the PPV and Se plot (see
Figure 4 and Figure 5) provides further insight of the behaviour of considered methods. Among the algorithms with highest PPV, PC-LPGM shows a sensitivity approaching 1 already at the sample size $n = 1000$ for both a high and a low SNR level (Figure 4). It is worth noting that, LPGM algorithm was successful only for a high SNR level ($\lambda_{\text{noise}} = 0.5$).

It is interesting to note that the performance of the PC-LPGM algorithm is far better than that of the competing algorithms employing the Poisson assumption, PDN and LPGM. This might be explained in terms of difference between penalization and restriction of the conditional sets. In the LPGM algorithm, as well as in the PDN algorithm, a prediction model is fitted locally on all other variables, by mean of a series of independent penalized regressions. In the PC-LPGM algorithm, the number of variables in the conditional sets is controlled and progressively increased from 0 to $p - 2$ (or to the maximum number of neighbours $m$). In our simulations, this second strategy appears to be more powerful in the network reconstruction.

![Figure 4](image-url)  
**Figure 4:** PPV (first panel row) and Se (second panel row) for PC-LPGM; LPGM; PDN; VSL; GLASSO; NPN-Copula; NPN-Skeptic for networks in Figure 1 ($p = 10$), sample sizes $n = 200, 1000, 2000$ and $\lambda_{\text{noise}} = 0.5$.

The Gaussian based methods (VSL, GLASSO) perform reasonably well, with an inferior score with respect to the leading threesome only for the hub graph at high SNR level. It is worth noting that sophisticated techniques that replace the Gaussian distribution with a more flexible continuous distribution such as the nonparanormal distribution, for example, NPN-Copula, NPN-Skeptic show slight gains in accuracy over the naive analysis.

Results for the high dimensional setting ($p = 100$) are somehow comparable, as it can be seen in Figures 6, 7 and 8. The PC-LPGM outperforms all competing methods, and differences among algorithms are more evident. The TP score of PC-LPGM becomes already reasonable when $n$ approaches 2000 observations. It is worth noting that performances of
methods based on $l_1$-regularized regression are overall less accurate and more variable in this scenario. For example, the number of recovered edges with LPGM is almost comparable to an empty graph in a number of cases, a result possibly related to the levels of $\beta$ chosen in the exercise. To ascertain such explanation, we run some simulations with higher variability threshold levels, $\beta = 0.5$ and 0.3 for LPGM (results not reported here). Although the TP scores improved, they were still unable to compete with the best performing algorithms.

Overall, results seem to demonstrate the good performances of PC-LPGM algorithm in all considered situations.
Figure 6: Number of TP edges recovered by PC-LPGM; LPGM; PDN; VSL; GLASSO; NPN-Copula; NPN-Skeptic for networks in Figure 2 (\(p = 100\)) and sample sizes \(n = 200, 1000, 2000\). First panel row corresponds to high SNR level (\(\lambda_{\text{noise}} = 0.5\)); second panel row corresponds to low SNR level (\(\lambda_{\text{noise}} = 5\)).

Figure 7: PPV (first panel row) and Se (second panel row) for PC-LPGM; LPGM; PDN; VSL; GLASSO; NPN-Copula; NPN-Skeptic for networks in Figure 2 (\(p = 100\)), sample sizes \(n = 200, 1000, 2000\) and \(\lambda_{\text{noise}} = 0.5\).
Figure 8: PPV (first panel row) and Se (second panel row) for PC-LPGM; LPGM; PDN; VSL; GLASSO; NPN-Copula; NPN-Skeptic for networks in Figure 2 \((p = 100)\), sample sizes \(n = 200, 1000, 2000\) and \(\lambda_{\text{noise}} = 5\).

Table 1: Monte Carlo marginal means of TP, FP, FN, PPV, Se obtained by simulating 500 samples from each of the three networks shown in Figure 1 \((p = 10)\).

| \(\lambda_{\text{noise}}\) | \(n\) | Algorithm | TP   | FP   | FN   | PPV  | Se   |
|-------------------------|------|-----------|------|------|------|------|------|
| 200                     |      | PC-LPGM   | 6.336| 0.067| 2.023| 0.991| 0.758|
|                         |      | LPGM      | 3.768| 0.255| 4.565| 0.955| 0.449|
|                         |      | PDN       | 5.784| 1.035| 2.549| 0.858| 0.696|
|                         |      | VSL       | 4.169| 0.033| 4.190| 0.995| 0.498|
|                         |      | GLASSO    | 4.076| 0.026| 4.283| 0.996| 0.487|
|                         |      | NPN-Copula| 4.568| 0.029| 3.791| 0.996| 0.546|
|                         |      | NPN-Skeptic| 4.476| 0.034| 3.883| 0.995| 0.534|
| 1000                    |      | PC-LPGM   | 8.359| 0.090| 0.000| 0.990| 1.000|
|                         |      | LPGM      | 5.307| 1.909| 3.027| 0.869| 0.637|
|                         |      | PDN       | 5.991| 0.721| 2.342| 0.901| 0.722|
|                         |      | VSL       | 4.694| 0.000| 3.665| 1.000| 0.562|
|                         |      | GLASSO    | 4.624| 0.000| 3.735| 1.000| 0.554|
|                         |      | NPN-Copula| 4.954| 0.000| 3.405| 1.000| 0.592|
|                         |      | NPN-Skeptic| 4.819| 0.000| 3.540| 1.000| 0.576|
| 0.5                     |      | PC-LPGM   | 8.422| 0.090| 0.000| 0.990| 1.000|
|                         |      | LPGM      | 7.132| 4.639| 1.201| 0.690| 0.856|
|                         |      | PDN       | 5.981| 0.694| 2.353| 0.904| 0.721|
|                         |      | VSL       | 5.657| 0.000| 2.765| 1.000| 0.675|
|                         |      | GLASSO    | 5.620| 0.000| 2.802| 1.000| 0.670|
|                         |      | NPN-Copula| 5.901| 0.000| 2.521| 1.000| 0.702|
|                         |      | NPN-Skeptic| 5.779| 0.000| 2.643| 1.000| 0.688|
| 2000                    |      | PC-LPGM   | 8.422| 0.090| 0.000| 0.990| 1.000|
|                         |      | LPGM      | 7.132| 4.639| 1.201| 0.690| 0.856|
|                         |      | PDN       | 5.981| 0.694| 2.353| 0.904| 0.721|
|                         |      | VSL       | 5.657| 0.000| 2.765| 1.000| 0.675|
|                         |      | GLASSO    | 5.620| 0.000| 2.802| 1.000| 0.670|
|                         |      | NPN-Copula| 5.901| 0.000| 2.521| 1.000| 0.702|
|                         |      | NPN-Skeptic| 5.779| 0.000| 2.643| 1.000| 0.688|
| 200                     |      | PC-LPGM   | 2.059| 0.704| 6.357| 0.755| 0.245|
Table 1 – continued from previous page

| λ_{noise} | n  | Algorithm       | TP    | FP    | FN    | PPV   | Se   |
|-----------|----|-----------------|-------|-------|-------|-------|------|
| 1000      | 1000 | PC-LPGM         | 7.889 | 1.063 | 0.444 | 0.890 | 0.946 |
|           |     | LPGM            | 4.115 | 2.176 | 4.219 | 0.686 | 0.494 |
|           |     | PDN             | 5.853 | 1.249 | 2.481 | 0.833 | 0.703 |
|           | 5   | VSL             | 3.135 | 0.012 | 5.198 | 0.998 | 0.377 |
|           |     | GLASSO          | 3.118 | 0.012 | 5.215 | 0.998 | 0.375 |
|           |     | NPN-Copula      | 3.211 | 0.006 | 5.122 | 0.999 | 0.386 |
|           |     | NPN-Skeptic     | 3.007 | 0.008 | 5.327 | 0.998 | 0.362 |
| 2000      | 2000 | PC-LPGM         | 8.355 | 1.056 | 0.002 | 0.897 | 1.000 |
|           |     | LPGM            | 4.337 | 2.151 | 3.996 | 0.703 | 0.566 |
|           |     | PDN             | 6.153 | 0.805 | 5.180 | 0.892 | 0.740 |
|           | 0.5 | VSL             | 3.954 | 0.000 | 4.404 | 1.000 | 0.473 |
|           |     | GLASSO          | 3.931 | 0.000 | 4.426 | 1.000 | 0.470 |
|           |     | NPN-Copula      | 4.094 | 0.000 | 4.264 | 1.000 | 0.490 |
|           |     | NPN-Skeptic     | 3.863 | 0.000 | 4.494 | 1.000 | 0.462 |

Table 2: Monte Carlo marginal means of TP, FP, FN, PPV, Se obtained by simulating 500 samples from each of the three networks shown in Figure 2 (p = 100).

| λ_{noise} | n  | Algorithm       | TP    | FP    | FN    | PPV   | Se   |
|-----------|----|-----------------|-------|-------|-------|-------|------|
| 200       | 200 | PC-LPGM         | 54.780| 9.430 | 46.913| 0.822 | 0.535 |
|           |     | LPGM            | 5.879 | 2.414 | 94.550| 0.786 | 0.058 |
|           |     | PDN             | 41.476| 47.640| 59.524| 0.493 | 0.406 |
|           |     | VSL             | 57.990| 24.512| 43.703| 0.703 | 0.566 |
|           |     | GLASSO          | 56.531| 23.983| 45.161| 0.703 | 0.552 |
|           |     | NPN-Copula      | 60.315| 21.202| 41.378| 0.737 | 0.589 |
|           |     | NPN-Skeptic     | 58.967| 26.466| 42.726| 0.695 | 0.576 |
| 1000      | 1000 | PC-LPGM         | 98.693| 13.201| 4.398 | 0.882 | 0.956 |
|           |     | LPGM            | 34.694| 0.377 | 66.152| 0.929 | 0.339 |
|           |     | PDN             | 68.954| 9.723 | 32.046| 0.890 | 0.688 |
|           |     | VSL             | 81.930| 0.107 | 21.160| 0.999 | 0.782 |
|           |     | GLASSO          | 81.316| 0.129 | 21.775| 0.998 | 0.776 |
|           |     | NPN-Copula      | 85.150| 0.078 | 17.941| 0.999 | 0.814 |
|           |     | NPN-Skeptic     | 84.277| 0.160 | 18.814| 0.998 | 0.806 |
| 5         | 5   | PC-LPGM         | 101.508| 14.405| 1.114 | 0.879 | 0.990 |
|           |     | LPGM            | 43.743| 0.305 | 57.257| 0.872 | 0.421 |
|           |     | PDN             | 73.431| 3.448 | 27.569| 0.953 | 0.736 |
|           |     | VSL             | 93.355| 0.004 | 9.266 | 1.000 | 0.904 |
|           |     | GLASSO          | 93.127| 0.004 | 9.494 | 1.000 | 0.902 |
|           |     | NPN-Copula      | 96.317| 0.000 | 6.305 | 1.000 | 0.935 |
|           |     | NPN-Skeptic     | 95.303| 0.006 | 7.319 | 1.000 | 0.924 |

| 200       | 200 | PC-LPGM         | 6.170 | 14.292| 94.830| 0.288 | 0.060 |
|           |     | LPGM            | 7.075 | 56.433| 93.925| 0.124 | 0.068 |
|           |     | PDN             | 11.220| 97.543| 89.780| 0.104 | 0.110 |
|           |     | VSL             | 7.752 | 23.011| 93.248| 0.276 | 0.076 |
6. Real data analysis: inferring networks from next generation sequencing data

To make our evaluation of PC-LPGM stronger, we perform some biological validation by applying the new algorithm to level III breast cancer microRNAs (miRNAs) expression, retrieved from the Cancer Genome Atlas. Here, we expect to obtain results coherent with the current biological knowledge.

miRNAs are non-coding RNAs that are transcribed but do not encode proteins. miRNAs have been reported to play a pivotal role in regulating key biological processes, for example, post-transcriptional modifications and translation processes. Some studies revealed that some disease-related miRNAs can indirectly regulate the function of other miRNAs associated with the same phenotype. In this perspective, studying the features of the interaction pattern of miRNAs in some conditions might help understand complex phenotype conditions.

Here, we consider level III breast cancer. Our interest lies in the pattern of interactions among miRNAs, with a particular focus on the existence of hubs. In fact, nodes with atypically high numbers of connections represent sites of signalling convergence with potentially large explanatory power for network behaviour or utility for clinical prognosis and therapy. By applying our algorithm, we expect to obtain results in line with known associations between miRNAs and breast cancer, and possibly gain more understanding of the nature of their effect on other genes. In other words, we expect some miRNAs associated with this phenotype to be the hubs of our estimated structure.

miRNAs expression, obtained by high-throughput sequencing, was downloaded from The Cancer Genome Atlas (TCGA) portal (https://tcga-data.nci.nih.gov/docs/publications/brca_2012/). The raw count data set consisted of 544 patients and 1046 miRNAs. As measurements were zero-inflated and highly skewed, with total count volumes depending on experimental condition, standard preprocessing was applied to the data (see Allen and
Liu, 2013). In particular: we normalized the data by the 75% quantile matching (Bullard et al., 2010); selected top 25% most variable miRNAs across the data; used a power transform $X^\alpha$ for $\alpha \in [0, 1]$ with $\alpha$ chosen via the minimum Kolmogorov-Smirnov statistic (Li et al., 2012). The miRNAs with little variation across the samples were filtered out, leaving 544 patients ($n = 544$) and 261 miRNA ($p = 261$). The effect of preprocessing on four prototype miRNA are shown in Figure 9.

Normalized data was used as input to PC-LPGM. A significance level of 10% resulted in a spare graph is shown in Figure 10.

![Figure 9: Distribution of four miRNA-Seq: raw data (top), normalized data (bottom).](image)

![Figure 10: Breast cancer miRNA network estimated by the PC-LPGM algorithm (hub nodes coloured red).](image)
We identified ten hub nodes in the network, miR-10b, -30a, -143, -375, -145, -210, -139, -934, -190b, -590. Almost all of them are known to be related to breast cancer (Volinia et al., 2012), providing a biological validation of the potential of the algorithm to recover the sites of the network with high explanatory power. In particular, miR-10b and -210 highly express in breast cancer, when high expression is related to poor prognosis; miR-30a, -143 and -145 appear to be inhibitors of progression, and should therefore be low in patients with good survival (Zhang et al., 2014; Yan et al., 2014). These results play the role of a biological validation of the ability of PC-LPGM to retrieve structures reflecting existing relations among variables.

7. Discussion

The main contribution of this paper is a careful analysis of the numerical and statistical efficiency of PC-LPGM, a simple method for structure learning of undirected graphical models for count data. A key strategy of our approach is controlling the number of variables in the conditional sets, as done in the PC algorithm. In this way, we control problems of estimation when the number of random variables $p$ is large possibly goes to infinity.

Our main theoretical result on truncated Poisson counts provides sufficient conditions on the set $(n, p, m, R)$ and on the model parameters for the method to succeed in consistently estimating the neighbours of every node in the graph. Precisely, Theorem 4.6 not only specifies sufficient conditions but also provides the probability with which the method recovers the true edge set. Indeed, Equation (8) shows that

$$\Pr_{\theta}(\text{a type I or II error occurs in testing procedure}) \leq O(p^{m+2}) \exp\{-cn\}$$

Hence, the right hand sight of the Equation will tend to 0 if $p^{m+2} \exp\{-cn\} \to 0$, equivalent to $n > O_p(m \log p)$. Moreover, Proposition A.2, and Lemma A.1 require

$$n > \max\{O_p(R^2\kappa_1 \log p), O_p(\kappa_1^2 R^4 p^2 \log p)\} = O_p(\kappa_1^2 R^4 p^2 \log p)$$

Thus, the sufficient condition becomes

$$n > \max\{O_p(m \log p), O_p(\kappa_1^2 R^4 p^2 \log p)\} = O_p(\kappa_1^2 R^4 p^2 \log p)$$

to guarantee the convergences. Appendix B shows that $\kappa_1 \leq O_p(R^2)$. Hence, to have consistency of PC-LPGM with exponentially decaying error, it is sufficient to have $n > O_p(R^8 p^2 \log p)$.

When $R$ is fixed, the condition reduces to $n > O_p(p^2 \log p)$. However, it is worth remembering that when the maximum number of neighbours that one node is allowed to have is fixed to $m$, a limitation is operated on the cardinalities $m + 1$ of the sets $K$. In this situation, the condition for convergence is relaxed to $n > \max\{O_p(m \log p), O_p(m^2 \log m)\}$ (see Note A.4 and A.5 for details).

Our simulation results show that the algorithm perform well also when Poisson conditional distributions with no constraints on the interaction parameters are taken as starting point for model specification. The empirical comparison shows that the algorithm outperforms its natural competitors.
Acknowledgments

We would like to acknowledge support for this project from the research grant BIRD172830 of University of Padova, Italy.
Appendix A. Proofs

In this section, we provide proofs of Proposition 4.3 and Theorem 4.4 stated in Section 4 of the main paper. We begin by introducing results for the case $K = V$. Then, the same results for general case $K \subset V$ are deduced.

Before going into details, we first prove the following Lemma, used in the proof of Theorem A.3.

**Lemma A.1** Assume 4.2. Then, for any $\delta > 0$, we have

$$
\mathbb{P}_\theta \left( \Lambda_{\text{max}} \left( \frac{1}{n} \sum_{i=1}^{n} (X^{(i)}_{V \setminus \{s\}})^T X^{(i)}_{V \setminus \{s\}} \right) \leq \lambda_{\text{max}} + \delta \right) \geq 1 - \exp \left( -c_2 n \right)
$$

$$
\mathbb{P}_\theta \left( \Lambda_{\text{min}} (Q_s(\theta_s)) \geq \lambda_{\text{min}} - \delta \right) \geq 1 - \exp \left( -c_2 n \right).
$$

**Proof** The $(j, k)$ element of the matrix $Z^n = Q_s(\theta_s) - I_s(\theta_s)$ can be written as

$$
Z_{jk}^n(\theta_s) = \frac{1}{n} \sum_{i=1}^{n} D''((\theta_s, X^{(i)}_{V \setminus \{s\}})) X_{ij} X_{ik} - \mathbb{E}_\theta \left( D''((\theta_s, X_{V \setminus \{s\}})) X_j X_k \right)
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} Y_i - \mathbb{E}_\theta \left( \frac{1}{n} \sum_{i=1}^{n} Y_i \right),
$$

where $Y_i = D''((\theta_s, X^{(i)}_{V \setminus \{s\}})) X_{ij} X_{ik}$, $i = 1, \ldots, n$ are independent and bounded by $|Y_i| \leq \kappa_1 R^2$.

By the Azuma-Hoeffding inequality (Theorem 2 in Hoeffding, 1963), for any $\epsilon > 0$, we have

$$
\mathbb{P}_\theta \left( (Z_{ij}^n)^2 \geq \epsilon^2 \right) = \mathbb{P}_\theta \left( |Z_{ij}^n| \geq \epsilon \right) \leq 2 \exp \left( -\frac{c_2^2 n}{2 \kappa_1^2 R^4} \right).
$$

Moreover,

$$
\Lambda_{\text{min}}(I_s(\theta_s)) = \min_{\|y\|_2 = 1} y I_s(\theta_s) y^T
$$

$$
= \min_{\|y\|_2 = 1} \left\{ y Q_s(\theta_s) y^T + y (I_s(\theta_s) - Q_s(\theta_s)) y^T \right\}
$$

$$
\leq y Q_s(\theta_s) y^T + y (I_s(\theta_s) - Q_s(\theta_s)) y^T,
$$

where $y \in \mathbb{R}^{p-1}$ is an arbitrary vector with unit norm. Hence,

$$
\Lambda_{\text{min}}(Q_s(\theta_s)) \geq \Lambda_{\text{min}}(I_s(\theta_s)) - \max_{\|y\|_2 = 1} y (I_s(\theta_s) - Q_s(\theta_s)) y^T \geq \lambda_{\text{min}} - \||I_s(\theta_s) - Q_s(\theta_s)||_2\|.
$$

We now derive a bound on the spectral norm $||I_s(\theta_s) - Q_s(\theta_s)||_2$. Let $\epsilon = \delta/p$, then

$$
\mathbb{P}_\theta \left( ||I_s(\theta_s) - Q_s(\theta_s)||_2 \geq \delta \right) \leq \mathbb{P}_\theta \left( \left( \sum_{j,k \neq s} (Z_{jk}^n)^2 \right)^{1/2} \geq \delta \right)
$$

$$
\leq 2p^2 \exp \left\{ -\frac{\delta^2 n}{2p^2 \kappa_1^2 R^4} \right\}
$$

$$
\leq \exp\{-c_2 n\}. \quad (9)
$$
Form Equation (9) and (10), we have
\[ \mathbb{P}_\theta (\Lambda_{\min}(Q_s(\theta_s)) \geq \lambda_{\min} - \delta) \geq 1 - \exp\{-c_2n\}. \]

Similarly, we have
\[ \mathbb{P}_\theta \left( \Lambda_{\max} \left[ \frac{1}{n} \sum_{i=1}^{n} (X^{(i)}_{V \setminus \{s\}})^T X^{(i)}_{V \setminus \{s\}} \right] \leq \lambda_{\max} + \delta \right) \geq 1 - \exp\{-c_2n\}. \]

We now introduce results for the case \( K = V \).

**Proposition A.2** Assume 4.1-4.2. Then, for any \( \delta > 0 \)
\[ \mathbb{P}_\theta (\|\nabla l(\theta_s, X_s; X_{V \setminus \{s\}})\|_\infty \geq \delta) \leq \exp\{-c_1n\}, \forall \theta \in \Theta, \]
when \( n \to \infty \).

**Proof** A rescaled negative node conditional log-likelihood can be written as
\[ l(\theta_s, X_s; X_{V \setminus \{s\}}) = -\frac{1}{n} \log \prod_{i=1}^{n} \mathbb{P}_{\theta_s}(x_{is}|X_{V \setminus \{s\}}) \]
\[ = \frac{1}{n} \sum_{i=1}^{n} [-x_{is}(\theta_s, X_{V \setminus \{s\}}) + \log x_{is}! + D((\theta_s, X_{V \setminus \{s\}}))] \]
The \( t \)-partial derivative of the node conditional log-likelihood \( l(\theta_s, X_s; X_{V \setminus \{s\}}) \) is:
\[ W_t = \nabla l(\theta_s, X_s; X_{V \setminus \{s\}}) = \frac{1}{n} \sum_{i=1}^{n} [-x_{is}x_{it} + x_{it}D'(\theta_s, X_{V \setminus \{s\}})) \]
Let \( V_{is}(t) = X_{is}X_{it} - X_{it}D'(\theta_s, X_{V \setminus \{s\}})) \). We have,
\[ \mathbb{P}_\theta (\|W\|_\infty \geq \delta) = \mathbb{P}_\theta (\max_{t \in V \setminus \{s\}} |\nabla l(\theta_s, X_s; X_{V \setminus \{s\}})| \geq \delta) \]
\[ = \mathbb{P}_\theta \left( \max_{t \in V \setminus \{s\}} \left| \frac{1}{n} \sum_{i=1}^{n} V_{is}(t) \right| \geq \delta \right) \]
\[ \leq p \left[ \mathbb{P}_\theta \left( \frac{1}{n} \sum_{i=1}^{n} V_{is}(t) \geq \delta \right) + \mathbb{P}_\theta \left( -\frac{1}{n} \sum_{i=1}^{n} V_{is}(t) \geq \delta \right) \right] \]
\[ \leq p \left[ \frac{\mathbb{E}_\theta \left[ \prod_{i=1}^{n} \exp \{hV_{is}(t)\} \right]}{\exp\{nh\delta\}} + \frac{\mathbb{E}_\theta \left[ \prod_{i=1}^{n} \exp \{-hV_{is}(t)\} \right]}{\exp\{nh\delta\}} \right] \]
\[ = p \left[ \prod_{i=1}^{n} \mathbb{E}_\theta \left[ \exp \{hV_{is}(t)\} \right] + \prod_{i=1}^{n} \mathbb{E}_\theta \left[ \exp \{-hV_{is}(t)\} \right] \right] \frac{\exp\{nh\delta\}}{\exp\{nh\delta\}} \]
\[ = p \left[ \exp \left\{ \sum_{i=1}^{n} \log \mathbb{E}_\theta \left[ \exp \{hV_{is}(t)\} \right] \right\} - nh\delta \right] \]
\[ + \exp \left\{ \sum_{i=1}^{n} \log \mathbb{E}_\theta \left[ \exp \{-hV_{is}(t)\} \right] - nh\delta \right\} \],
\[ \text{(11)} \]
for some $h > 0$. We therefore need to compute

$$
\sum_{i=1}^{n} \log \mathbb{E}_{\theta} \left[ \exp \left\{ hV_{is}(t) \right\} \right],
$$

and

$$
\sum_{i=1}^{n} \log \mathbb{E}_{\theta} \left[ \exp \left\{ -hV_{is}(t) \right\} \right].
$$

First, we have

$$
\mathbb{E}_{\theta} \left[ \exp \left\{ hV_{is}(t) \right\} | x^{(i)}_{V \setminus \{s\}} \right] = \sum_{x_{is}=0}^{R} \exp \left\{ h[x_{is} x_{it} - x_{it} D'(\langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle)] + x_{is} \{ \theta_s, x^{(i)}_{V \setminus \{s\}} \} - \log x_{is}! - D(\langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) \right\}
$$

$$
= \sum_{x_{is}=0}^{R} \exp \left\{ x_{is} [h x_{it} + \langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle] - \log x_{is}! - h x_{it} D'(\langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) - D(\langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) \right\}
$$

$$
= \exp \left\{ D(\langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) - D(\langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) \right\}
$$

$$
= \exp \left\{ \frac{h^2}{2} (x_{it})^2 D''(v h x_{it} + \langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) \right\},
$$

for some $v \in [0, 1]$, where we move from line 2 to line 3 by applying $\sum_{x_{is}=0}^{R} \exp \left\{ x_{is} [h x_{it} + \langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle] - \log x_{is}! - D(\langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) \right\} = 1$, and from line 3 to line 4 by using a Taylor expansion for function $D(.)$ at $\langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle$.

Therefore,

$$
\sum_{i=1}^{n} \log \mathbb{E}_{\theta} \left[ \exp \left\{ hV_{is}(t) \right\} \right] = \sum_{i=1}^{n} \log \mathbb{E}_{\theta^{V \setminus \{s\}}} \left[ \mathbb{E}_{\theta_s} \left[ \exp \left\{ hV_{is}(t) \right\} | x^{(i)}_{V \setminus \{s\}} \right] \right]
$$

$$
= \sum_{i=1}^{n} \log \mathbb{E}_{\theta^{V \setminus \{s\}}} \left[ \exp \left\{ \frac{h^2}{2} (x_{it})^2 D''(v h x_{it} + \langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) \right\} \right]
$$

$$
\leq n \frac{h^2}{2} R^2 \kappa_1,
$$

(12)

where $D''(v h x_{it} + \langle \theta_s, X^{(i)}_{V \setminus \{s\}} \rangle) < \kappa_1$, $\forall \theta_s \in \Omega(\Theta)$ (since $D''(.)$ is a continuous function, and $\Omega(\Theta)$ is bounded, see Appendix B for details). Similarly,

$$
\mathbb{E}_{\theta_s} \left[ \exp \left\{ -hV_{is}(t) \right\} | x^{(i)}_{V \setminus \{s\}} \right] = \exp \left\{ \frac{h^2}{2} (x_{it})^2 D''(-v h x_{it} + \langle \theta_s, x^{(i)}_{V \setminus \{s\}} \rangle) \right\},
$$
Therefore,
\[
\sum_{i=1}^{n} \log E_{\theta} [\exp \{-hV_{is}(t)\}] = \sum_{i=1}^{n} \log E_{\theta_{V\{s\}}} [E_{\theta_{s}} [\exp\{-hV_{is}(t)|X_{V\{s\}}^{(i)}\}]] \\
= \sum_{i=1}^{n} \log E_{\theta_{V\{s\}}} \left[ \exp \left\{ \frac{h^2}{2} (X_{it})^2 D''(-vX_{it} + \langle \theta_{s}, X_{V\{s\}}^{(i)} \rangle) \right\} \right] \\
\leq n \frac{h^2}{2} R^2 \kappa_{1}. \tag{13}
\]

Let \( h = \frac{\delta}{R^2 \kappa_{1}} \), from (11)–(13), we have
\[
P_{\theta}(\|W\|_{\infty} \geq \delta) \leq p \left[ \exp \left\{ \frac{n}{2} h^2 R^2 \kappa_{1} - nh\delta \right\} + \exp \left\{ \frac{n}{2} h^2 R^2 \kappa_{1} - nh\delta \right\} \right] \\
= 2p \left[ \exp \left\{ \frac{-n\delta^2}{2R^2 \kappa_{1}} \right\} \right] \\
\leq \exp \left\{ -c_{1} n \right\},
\]
provided that \( p < \frac{1}{2} \exp \left\{ \frac{n\delta^2}{4R^2 \kappa_{1}} \right\} \).

\[\square\]

**Theorem A.3** Assume 4.1–4.2. Then, there exists a non-negative decreasing sequence \( \delta_{n} \to 0 \), such that
\[
P_{\theta}(\|\hat{\theta}_{V\{s\}} - \theta_{V\{s\}}\|_{2} \geq \delta_{n}) \geq 1 - \exp \left\{ -c_{n} \right\}, \forall \theta \in \Theta,
\]
when \( n \to \infty \).

**Proof** For a fixed design \( X \), define \( G : \mathbb{R}^{p-1} \to \mathbb{R} \) as
\[
G(u, X_{s}; X_{V\{s\}}) = l(\theta_{s} + u, X_{s}; X_{V\{s\}}) - l(\theta_{s}, X_{s}; X_{V\{s\}}).
\]

Then, \( G(0, X_{s}; X_{V\{s\}}) = 0 \). Moreover, let \( \hat{\theta} = \theta_{s} - \theta_{s} \), we have \( G(\hat{u}, X_{s}; X_{V\{s\}}) \leq 0 \).

Given a value \( \epsilon > 0 \), if \( G(u, X_{s}; X_{V\{s\}}) > 0 \), \( \forall u \in \mathbb{R}^{p-1} \) such that \( \|u\|_{2} = \epsilon \), then \( \|\hat{u}\|_{2} \leq \epsilon \), since \( G(., X_{s}; X_{V\{s\}}) \) is a convex function. Therefore,
\[
P_{\theta} (\|\hat{\theta} - \theta_{s}\|_{2} \leq \epsilon) \geq P_{\theta} (G(u, X_{s}; X_{V\{s\}}) > 0), \forall u \in \mathbb{R}^{p-1} \) such that \( \|u\|_{2} = \epsilon \).

A Taylor expansion of the rescaled negative node conditional log-likelihood at \( \theta_{s} \) yields
\[
G(u, X_{s}; X_{V\{s\}}) = l(\theta_{s} + u, X_{s}; X_{V\{s\}}) - l(\theta_{s}, X_{s}; X_{V\{s\}}) \\
= \nabla l(\theta_{s}, X_{s}; X_{V\{s\}})u^{T} + \frac{1}{2} u^{T} \nabla^{2} l(\theta_{s} + vu, X_{s}; X_{V\{s\}})u^{T}.
\]
for some \( v \in [0, 1] \). Let
\[
q = \Lambda_{\min}(\nabla^2(l(\theta_s + vu, X_s; X_{V\setminus\{s\}}))) \\
\geq \min_{v \in [0, 1]} \Lambda_{\min}(\nabla^2(l(\theta_s + vu, X_s; X_{V\setminus\{s\}}))) \\
= \min_{v \in [0, 1]} \Lambda_{\min} \left[ \frac{1}{n} \sum_{i=1}^{n} D''(\langle \theta_s + vu, X_{iV\setminus\{s\}}^{(i)} \rangle)(X_{iV\setminus\{s\}}^{(i)})^T X_{iV\setminus\{s\}}^{(i)} \right].
\]

By using Taylor expansion for \( D''(\langle \theta_s + vu, X_{iV\setminus\{s\}}^{(i)} \rangle) \) at \( \langle \theta_s, X_{iV\setminus\{s\}}^{(i)} \rangle \), we have
\[
\frac{1}{n} \sum_{i=1}^{n} D''(\langle \theta_s + vu, X_{iV\setminus\{s\}}^{(i)} \rangle)(X_{iV\setminus\{s\}}^{(i)})^T X_{iV\setminus\{s\}}^{(i)} \\
= \frac{1}{n} \sum_{i=1}^{n} D''(\langle \theta_s, X_{iV\setminus\{s\}}^{(i)} \rangle)(X_{iV\setminus\{s\}}^{(i)})^T X_{iV\setminus\{s\}}^{(i)} + \\
\frac{1}{n} \sum_{i=1}^{n} D''(\langle \theta_s + v'u, X_{iV\setminus\{s\}}^{(i)} \rangle)[vu(X_{iV\setminus\{s\}}^{(i)})^T][X_{iV\setminus\{s\}}^{(i)}]^T X_{iV\setminus\{s\}}^{(i)},
\]
for some \( v' \in [0, 1] \). Fixed \( \delta = \frac{\lambda_{\min}}{8} \) in Lemma A.1. We have
\[
q \geq \Lambda_{\min} \left[ \frac{1}{n} \sum_{i=1}^{n} D''(\langle \theta_s, X_{iV\setminus\{s\}}^{(i)} \rangle)(X_{iV\setminus\{s\}}^{(i)})^T X_{iV\setminus\{s\}}^{(i)} \right] \\
- \max_{v' \in [0, 1]} \Lambda_{\max} \left[ \frac{1}{n} \sum_{i=1}^{n} |D''(\langle \theta_s + v'u, X_{iV\setminus\{s\}}^{(i)} \rangle)|[vu(X_{iV\setminus\{s\}}^{(i)})^T][X_{iV\setminus\{s\}}^{(i)}]^T X_{iV\setminus\{s\}}^{(i)} \right] \\
\geq \lambda_{\min} - \delta - \max_{v' \in [0, 1]} \Lambda_{\max} \left[ \frac{1}{n} \sum_{i=1}^{n} |D''(\langle \theta_s + v'u, X_{iV\setminus\{s\}}^{(i)} \rangle)|[vu(X_{iV\setminus\{s\}}^{(i)})^T][X_{iV\setminus\{s\}}^{(i)}]^T X_{iV\setminus\{s\}}^{(i)} \right] \\
\geq \lambda_{\min} - 2\delta - \kappa_2 R \sqrt{p} \|v\|_2 \lambda_{\max} \\
= \lambda_{\min} - 2\delta - \kappa_2 R \epsilon \lambda_{\max} \\
> \frac{\lambda_{\min}}{2}, \quad \text{provided that } \epsilon < \frac{\lambda_{\min}}{4 \sqrt{p} \lambda_{\max} \kappa_2 R};
\]
with probability at least \( 1 - \exp \{-c_2 n\} \), where \( |D''(\langle \theta_s + v'u, X_{iV\setminus\{s\}}^{(i)} \rangle)| < \kappa_2, \ \forall \ \theta_s \in \Theta \) (since \( D''(\cdot) \) is a continuous function, and \( \Theta \) is bounded, see Appendix B for details).

Let \( \delta = \frac{\lambda_{\min}}{4} \epsilon \) in Proposition A.2. Then, from Proposition A.2, we have
\[
\nabla l(\theta_s, X_s; X_{V\setminus\{s\}}) \geq -\frac{\lambda_{\min}}{4} \epsilon,
\]
for some \( v \in [0, 1] \). Let
with probability at least \( 1 - \exp \{ - c_1 n \} \), provided that \( p < \frac{1}{2} \exp \left( \frac{n \lambda_{\min}^2}{64R^2 \kappa_1} \right) \). Combining with the inequality of \( q \), we have

\[
G(u, X_s; X_{V \setminus \{s\}}) = \nabla l(\theta_s, X_s; X_{V \setminus \{s\}})u^T + \frac{1}{2} u[\nabla^2 l(\theta_s + vu_s, X_s; X_{V \setminus \{s\}})]u^T > -\frac{\lambda_{\min}}{4} \epsilon^2 + \frac{\lambda_{\min}}{4} \epsilon^2 = 0
\]

provided that \( p < \min \left\{ \frac{\lambda_{\min}^2}{16\epsilon^2 \lambda_{\max}^2 R^2}, \frac{1}{2} \exp \left( \frac{n \lambda_{\min}^2}{64R^2 \kappa_1} \right) \right\} \). It means that \( \|u\|_2 < \epsilon \).

When \( n \to \infty \) we can choose a non-negative decreasing sequence \( \delta_n \) such that \( \delta_n < \frac{\lambda_{\min}}{4\sqrt{R}\lambda_{\max} \kappa_2 R} \), then

\[
\mathbb{P}_\theta(\|\hat{\theta}_{V \setminus \{s\}} - \theta_{V \setminus \{s\}}\|_2 \leq \delta_n) \geq 1 - \exp \{ - cn \},
\]

when \( n \to \infty \).

Results for \( K \subset V \) are derived as following.

**Proposition 4.3** Assume \( 4.1 \cdot 4.2 \) and let \( K \subset V \). Then, for all \( s \in K \) and any \( \delta > 0 \)

\[
\mathbb{P}_\theta(\|\nabla l(\theta_{s\mid K}, X_s; X_{K \setminus \{s\}})\|_\infty \geq \delta) \leq \exp\{ -c_1 n \},
\]

\( \forall \, \theta_{s\mid K} \in \Theta \), when \( n \to \infty \).

**Proof** The proof of Proposition 4.3 follows the lines of Proposition A.2. We note that the set of explanatory variables \( X_{K \setminus \{s\}} \) in the generalized linear model \( X_s \) given \( X_{K \setminus \{s\}} \) does not include variables \( X_t \), with \( t \in V \setminus K \). Suppose we zero-pad the true parameter \( \theta_{s\mid K} \in \mathbb{R}^{|K| - 1} \) to include zero weights over \( V \setminus K \), then the resulting parameter would lie in \( \mathbb{R}^{|v\setminus 1|} \).

**Note A.4** When the maximum number of neighbours that one node is allowed to have is fixed, a control is operated on the cardinality of the set \( K, |K| \leq m + 1 \). In this case, parameters \( \theta_{s\mid K} \) are estimated from models that are restricted on subsets of variables with their cardinalities less than or equal to \( m + 1 \). Therefore, \( p \) in Proposition A.2 is replaced by \( m + 1 \). In detail, for all \( s \in K \) and any \( \delta > 0 \)

\[
\mathbb{P}_\theta(\|\nabla l(\theta_{s\mid K}, X_s; X_{K \setminus \{s\}})\|_\infty \geq \delta) \leq \exp\{ -c_1 n \},
\]

\( \forall \, \theta_{s\mid K} \in \Theta \), provided that \( m < \frac{1}{2} \exp \left\{ \frac{n \delta^2}{4R^2 \kappa_1} \right\} \).

We take the same way as in the proof of Theorem A.3 to prove Theorem 4.4.

**Theorem 4.4** Assume \( 4.1 \cdot 4.2 \) and let \( K \subset V \). Then, there exists a non-negative decreasing sequence \( \delta_n \to 0 \), such that

\[
\mathbb{P}_\theta(\|\hat{\theta}_{s\mid K} - \theta_{s\mid K}\|_2 \leq \delta_n) \geq 1 - \exp\{ - cn \}, \forall \, s \in K, \theta \in \Theta,
\]

29
Proof Let \( \hat{u} = \hat{\theta}_{s|K} - \theta_{s|K} \) and define \( G : \mathbb{R}^{|K|-1} \rightarrow \mathbb{R} \) as
\[
G(\hat{u}, X_{\{s\}}; X_{\backslash\{s\}}) = l(\theta_{s|K} + \hat{u}, X_{\{s\}}; X_{\backslash\{s\}}) - l(\theta_{s|K}, X_{\{s\}}; X_{\backslash\{s\}}).
\]

Similar to Theorem A.3, we have
\[
\mathbb{P}_\theta(||\hat{\theta}_{s|K} - \theta_{s|K}||_2 \leq \epsilon) \geq \mathbb{P}_\theta \left( G(u, X_{\{s\}}; X_{\backslash\{s\}}) > 0, \forall u \in \mathbb{R}^{|K|} \text{ such that } ||u||_2 = \epsilon \right).
\]

Recall the conditional rescaled negative log-likelihood function:
\[
l(\theta_{s|K}, X_{\{s\}}; X_{\backslash\{s\}}) = \frac{1}{n} \sum_{i=1}^{n} \left[ -X_{is}(\theta_{s|K}, X_{\{\backslash{s}\}}^{(i)}) + D((\theta_{s|K}, X_{\{\backslash{s}\}}^{(i)})) \right].
\]

By its Taylor expansion at \( \theta_{s|K} \), we have
\[
G(u) = l(\theta_{s|K} + u, X_{\{s\}}; X_{\backslash\{s\}}) - l(\theta_{s|K}, X_{\{s\}}; X_{\backslash\{s\}}) = \nabla l(\theta_{s|K}, X_{\{s\}}; X_{\backslash\{s\}}) u^T + \frac{1}{2} u [\nabla^2 l(\theta_{s|K} + v u, X_{\{s\}}; X_{\backslash\{s\}})] u^T.
\]

Let
\[
q = \Lambda_{\min}(\nabla^2 l(\theta_{s|K} + u, X_{\{s\}}; X_{\backslash\{s\}}))
\geq \min_{v \in [0,1]} \Lambda_{\min}(\nabla^2 l(\theta_{s|K} + v u, X_{\{s\}}; X_{\backslash\{s\}}))
= \min_{v \in [0,1]} \Lambda_{\min} \left[ \frac{1}{n} \sum_{i=1}^{n} D''((\theta_{s|K} + v u, X_{\{\backslash{s}\}}^{(i)})) (X_{\{\backslash{s}\}}^{(i)})^T X_{\{\backslash{s}\}}^{(i)} \right].
\]

By using Taylor expansion of \( D''((\theta_{s|K} + v u, X_{\{\backslash{s}\}}^{(i)})) \) at \( (\theta_{s|K}, X_{\backslash\{s\}}^{(i)}) \), we have
\[
\frac{1}{n} \sum_{i=1}^{n} D''((\theta_{s|K} + v u, X_{\{\backslash{s}\}}^{(i)})) (X_{\{\backslash{s}\}}^{(i)})^T X_{\{\backslash{s}\}}^{(i)}
= \frac{1}{n} \sum_{i=1}^{n} D''((\theta_{s|K}, X_{\{\backslash{s}\}}^{(i)})) (X_{\{\backslash{s}\}}^{(i)})^T X_{\{\backslash{s}\}}^{(i)}
+ \frac{1}{n} \sum_{i=1}^{n} D''((\theta_{s|K} + v' u, X_{\{\backslash{s}\}}^{(i)})) [v u (X_{\{\backslash{s}\}}^{(i)})^T (X_{\{\backslash{s}\}}^{(i)})^T X_{\{\backslash{s}\}}^{(i)}] ...
\]

Hence,
\[
q \geq \Lambda_{\min} \left[ \frac{1}{n} \sum_{i=1}^{n} D''((\theta_{s|K}, X_{\{\backslash{s}\}}^{(i)})) (X_{\{\backslash{s}\}}^{(i)})^T X_{\{\backslash{s}\}}^{(i)} \right]
- \max_{v' \in [0,1]} \Lambda_{\max} \left[ \frac{1}{n} \sum_{i=1}^{n} D''((\theta_{s|K} + v' u, X_{\{\backslash{s}\}}^{(i)})) [v u (X_{\{\backslash{s}\}}^{(i)})^T (X_{\{\backslash{s}\}}^{(i)})^T X_{\{\backslash{s}\}}^{(i)}] \right]
\geq \lambda_{\min} - \delta - \max_{v' \in [0,1]} \Lambda_{\max} \left[ \frac{1}{n} \sum_{i=1}^{n} D''((\theta_{s|K} + v' u, X_{\{\backslash{s}\}}^{(i)})) [v u (X_{\{\backslash{s}\}}^{(i)})^T (X_{\{\backslash{s}\}}^{(i)})^T X_{\{\backslash{s}\}}^{(i)}] \right]
\geq \lambda_{\min} - 2\delta - \kappa_2 \sqrt{p R c} \lambda_{\max} \]
The second and third inequality are due to well-known results on eigenvalue inequalities for a matrix and its submatrix (see, for example, Johnson and Robinson, 1981). Here,

$$Q_s|K(\theta|s|K) = \frac{1}{n} \sum_{i=1}^{n} D'' \left( (\theta|s|K, X^{(i)}_{K\setminus\{s\}}) \right) \left( X^{(i)}_{K\setminus\{s\}} \right)^T X^{(i)}_{K\setminus\{s\}}$$

is a sub-matrix of the Hessian matrix $Q_s(\theta)$). Hence,

$$\Lambda_{\min} \left[ \frac{1}{n} \sum_{i=1}^{n} D'' \left( (\theta|s|K, X^{(i)}_{K\setminus\{s\}}) \right) \left( X^{(i)}_{K\setminus\{s\}} \right)^T X^{(i)}_{K\setminus\{s\}} \right] \geq \Lambda_{\min}(Q_s(\theta_s)) \geq \lambda_{\min} - \delta.$$ 

Similarly, for the matrix $\left( X^{(i)}_{K\setminus\{s\}} \right)^T X^{(i)}_{K\setminus\{s\}}$, we have

$$\max_{v' \in [0,1]} \Lambda_{\max} \left[ \frac{1}{n} \sum_{i=1}^{n} D''' \left( (\theta|s|K + v'u, X^{(i)}_{K\setminus\{s\}}) \right) \left( v'u \left( X^{(i)}_{K\setminus\{s\}} \right)^T \right) \left( X^{(i)}_{K\setminus\{s\}} \right)^T X^{(i)}_{K\setminus\{s\}} \right] \leq \kappa_2 \sqrt{p} \text{Re} \lambda_{\max} + \delta.$$

Then, by performing the same analysis as in the proof of Theorem A.3 and Proposition A.2, we get the result. \(\blacksquare\)

**Note A.5** In the proof of Theorem 4.4, we only require the uniform convergence of a submatrix (restricted on $K$), $Q_s|K(\theta|s|K)$, of the sample Fisher information matrix $Q_s(\theta)$. Therefore, when the maximum neighbourhood size is known, $|K| \leq m + 1$, we have convergence provided that $n > O_p(\kappa_1 R^4 m^2 \log m)$. In detail, let $I_s|K(\theta|s|K)$ be the submatrix of $I_s(\theta)$ indexed in $K$, Equation (10) becomes

$$\mathbb{P}_{\theta} \left( \|I_s|K(\theta|s|K) - Q_s|K(\theta|s|K)\|_2 \geq \delta \right) \leq \mathbb{P}_{\theta} \left( \left( \sum_{j,k \in K\setminus\{s\}} (Z_{jk}^n)^2 \right)^{1/2} \geq \delta \right) \leq 2m^2 \exp \left\{ -\frac{\delta^2 n}{2m^2 \kappa_1^2 R^4} \right\} \leq \exp\{-c_2n\},$$

provided that $n > O_p(\kappa_1 R^4 m^2 \log m)$.

**Appendix B. A bound on the second and third derivative of the log normalizing term $D(.)$**

Here, we derive bounds $\kappa_1$ and $\kappa_2$ for the second and third derivative of the log normalizing term $D(.)$, that is, $D''(vhX_{it} + (\theta, X^{(i)}_{V\setminus\{s\}}))$, and $D'''(vhX_{it} + (\theta, X^{(i)}_{V\setminus\{s\}}))$. For the sake of simplicity, we write

$$D(x) = \log \left( \sum_{k=0}^{R} \exp \left\{ kx - \log k! \right\} \right),$$
which we consider on a compact set $U \subset \mathbb{R}$. The first and second derivative of $D(.)$ is

$$D'(x) = \frac{\sum_{k=0}^{R} \exp \{ kx - \log k! \} k}{\sum_{k=0}^{R} \exp \{ kx - \log k! \}}$$

$$D''(x) = \frac{\sum_{k=0}^{R} \exp \{ kx - \log k! \} k^2 \sum_{k=0}^{R} \exp \{ kx - \log k! \} - \left( \sum_{k=0}^{R} \exp \{ kx - \log k! \} k \right)^2}{\left( \sum_{k=0}^{R} \exp \{ kx - \log k! \} \right)^2}$$

$$= \frac{\sum_{k,h=0}^{R,\infty} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \} (k^2 - kh)}{\sum_{k,h=0}^{R,\infty} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \}}.$$ 

Hence,

$$|D''(x)| = \left| \frac{\sum_{k,h=0}^{R,\infty} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \} (k^2 - kh)}{\sum_{k,h=0}^{R,\infty} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \}} \right|$$

$$\leq \frac{\sum_{k,h=0}^{R,\infty} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \} |k^2 - kh|}{\sum_{k,h=0}^{R,\infty} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \}}$$

$$\leq 2R^2 \frac{\sum_{k,h=0}^{R,\infty} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \}}{\sum_{k,h=0}^{R,\infty} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \}}$$

$$= 2R^2.$$

Therefore, $\kappa_1 \leq O_p(R^2)$. Similarly,

$$D'''(x) = \frac{N(x)}{\left( \sum_{k=0}^{R} \exp \{ kx - \log k! \} \right)^4}$$

where

$$N(x) = \left( \sum_{k=0}^{R} \exp \{ kx - \log k! \} k^3 \sum_{k=0}^{R} \exp \{ kx - \log k! \} + \sum_{k=0}^{R} \exp \{ kx - \log k! \} k^2 \right)$$

$$- \left( \sum_{k=0}^{R} \exp \{ kx - \log k! \} k \right)^2 - \left( \sum_{k=0}^{R} \exp \{ kx - \log k! \} k^2 \sum_{k=0}^{R} \exp \{ kx - \log k! \} \right)$$

$$- \left( \sum_{k=0}^{R} \exp \{ kx - \log k! \} k \right)^2 2 \sum_{k=0}^{R} \exp \{ kx - \log k! \} \sum_{k=0}^{R} \exp \{ kx - \log k! \} k$$

$$= \sum_{k,h,r,t=0}^{R} \exp \{ kx - \log k! \} \exp \{ hx - \log h! \} \exp \{ rx - \log r! \} \exp \{ tx - \log t! \}$$

$$\left( k^3 - kh^2 - 2k^2t + 2kht \right).$$
Hence,

\[
|D''(x)| \leq 6R^3 \sum_{k,h,r,t=0}^R \exp \left\{ \frac{kx - \log k!}{k} + \frac{hx - \log h!}{h} + \frac{rx - \log r!}{r} + \frac{tx - \log t!}{t} \right\} \\
= 6R^3.
\]

Therefore, \( \kappa_2 \leq O_p(R^3) \).

Appendix C. The StARS algorithm

The StARS algorithm introduced in Liu et al. (2010), aims to seek the value of \( \lambda \) leading to the most stable set of edges. More precisely, it considers a range \( \Lambda = \{\lambda_1, \ldots, \lambda_k\} \) of values for \( \lambda \), and fixes a number \( n_B \), \( 1 < n_B < n \) of observations in one sample. Then, \( B \) samples of size \( n_B \), \( S_1, \ldots, S_B \), are generated from \( x_1, \ldots, x_n \). For each \( \lambda \in \Lambda \), the graph is estimated by solving a lasso problem. Let \( A_{\lambda}^{n_B}(S_1), \ldots, A_{\lambda}^{n_B}(S_B) \) be estimated adjacency matrices of the graph in the subsamples. The stability of one edge can be estimated by

\[
\epsilon_{s,t}^{n_B}(\lambda) = 2\psi_{s,t}^{n_B}(\lambda)(1 - \psi_{s,t}^{n_B}(\lambda)),
\]

where \( \psi_{s,t}^{n_B}(\lambda) = \frac{1}{B} \sum_{i=1}^B A_{\lambda}^{n_B}(S_i)_{st} \) is the estimated probability of one edge between nodes \( s \) and \( t \). The optimal value \( \lambda_{opt} \) is defined as the largest value that maximizes the total stability

\[
\bar{D}_{n_B}(\lambda) = \sup_{0 \leq \rho \leq \lambda} \sum_{s<t} \epsilon_{s,t}^{n_B}(\sigma)/\binom{p}{2},
\]

smaller than an upper bound \( \beta \), \( \lambda_{opt} = \sup\{\lambda : \bar{D}_{n_B}(\lambda) \leq \beta\} \).

Appendix D. Simulation study results

Table 3 to Table 6 report TP, FP, FN, PPV and Se for each of methods considered in Section 5 of the main paper. Two different graph dimensions, \( p = 10, 100 \), and three graph structures (see Figure 1 and Figure 2 of the main paper) are considered at one low (\( \lambda_{noise} = 5 \)) and one high (\( \lambda_{noise} = 0.5 \)) SNR levels.

Table 3: Simulation results from 500 replicates of the undirected graphs shown in Figure 1 of the main paper for \( p = 10 \) variables with Poisson node conditional distribution and level of noise \( \lambda_{noise} = 0.5 \). Monte Carlo means (standard deviations) are shown for TP, FP, FN, PPV and Se.

| Graph | n | Algorithm | TP   | FP   | FN   | PPV  | Se  |
|-------|---|-----------|------|------|------|------|-----|
| 200   | 0 | PC-LPGM   | 6.838 (1.152) | 0.048 (0.230) | 2.163 (1.152) | 0.994 (0.208) | 0.760 (0.169) |
|       |   | LPGM      | 4.732 (1.407) | 0.384 (0.644) | 4.268 (1.407) | 0.941 (0.097) | 0.526 (0.156) |
|       |   | PDN       | 5.872 (0.741) | 0.182 (0.430) | 3.128 (0.741) | 0.972 (0.065) | 0.652 (0.082) |
|       |   | VSL       | 4.625 (2.056) | 0.034 (0.181) | 4.375 (2.056) | 0.996 (0.021) | 0.514 (0.228) |
|       |   | GLASSO    | 4.502 (1.961) | 0.023 (0.151) | 4.498 (1.961) | 0.997 (0.018) | 0.500 (0.218) |
|       |   | NPN-Copula| 5.073 (2.169) | 0.034 (0.191) | 3.927 (2.169) | 0.996 (0.023) | 0.564 (0.241) |
|       |   | NPN-Skeptic| 5.030 (2.177) | 0.039 (0.230) | 3.970 (2.177) | 0.994 (0.023) | 0.559 (0.242) |
| 1000  | 0 | PC-LPGM   | 9.000 (0.000) | 0.071 (0.258) | 0.000 (0.000) | 0.993 (0.026) | 1.000 (0.000) |
|       |   | LPGM      | 5.780 (1.253) | 0.692 (2.730) | 3.220 (1.253) | 0.964 (0.135) | 0.642 (0.139) |
| Graph  | n   | Algorithm | TP    | FP    | FN    | PPV    | Se     |
|--------|-----|-----------|-------|-------|-------|--------|--------|
|        |     | PDN       | 5.780 | 0.000 | 3.220 | 1.000  | 0.642  |
|        |     | VSL       | 4.954 | 0.000 | 4.046 | 1.000  | 0.550  |
|        |     | GLASSO    | 4.889 | 0.000 | 4.111 | 1.000  | 0.543  |
|        |     | NPN-Copula| 5.377 | 0.000 | 3.623 | 1.000  | 0.597  |
|        |     | NPN-Skeptic| 5.232 | 0.000 | 3.768 | 1.000  | 0.581  |
| Scale-free | 2000 | PC-LPGM | 9.000 | 0.071 | 0.000 | 0.993  | 1.000  |
|        |     | LPGM      | 7.660 | 5.180 | 1.340 | 0.703  | 0.851  |
|        |     | PDN       | 5.658 | 0.000 | 3.342 | 1.000  | 0.629  |
|        |     | VSL       | 5.566 | 0.000 | 3.434 | 1.000  | 0.618  |
|        |     | GLASSO    | 5.573 | 0.000 | 3.427 | 1.000  | 0.619  |
|        |     | NPN-Copula| 6.055 | 0.000 | 2.945 | 1.000  | 0.673  |
|        |     | NPN-Skeptic| 5.945 | 0.000 | 3.055 | 1.000  | 0.661  |
|        |     | Hub       | 8.000 | 0.122 | 0.000 | 0.987  | 1.000  |
|        |     | LPGM      | 4.392 | 1.452 | 3.608 | 0.885  | 0.549  |
|        |     | PDN       | 7.128 | 0.030 | 3.684 | 0.995  | 0.540  |
|        |     | VSL       | 5.908 | 0.000 | 2.092 | 1.000  | 0.739  |
|        |     | GLASSO    | 5.842 | 0.000 | 2.158 | 1.000  | 0.730  |
|        |     | NPN-Copula| 6.000 | 0.000 | 2.000 | 1.000  | 0.750  |
|        |     | NPN-Skeptic| 5.818 | 0.000 | 2.182 | 1.000  | 0.727  |
|        |     | Random    | 8.000 | 0.078 | 0.000 | 0.991  | 1.000  |
|        |     | LPGM      | 5.748 | 3.584 | 2.252 | 0.758  | 0.633  |
|        |     | PDN       | 5.066 | 0.000 | 0.099 | 1.000  | 0.369  |
|        |     | VSL       | 3.190 | 0.000 | 4.810 | 1.000  | 0.399  |
|        |     | GLASSO    | 3.110 | 0.000 | 4.890 | 1.000  | 0.389  |
|        |     | NPN-Copula| 3.434 | 0.000 | 4.566 | 1.000  | 0.429  |
|        |     | NPN-Skeptic| 3.358 | 0.000 | 4.642 | 1.000  | 0.420  |
|        |     | Random    | 8.000 | 0.048 | 0.000 | 0.995  | 1.000  |
|        |     | LPGM      | 4.748 | 0.256 | 0.516 | 0.576  | 0.936  |
|        |     | PDN       | 5.068 | 0.000 | 2.932 | 0.713  | 0.634  |
|        |     | VSL       | 2.952 | 0.000 | 5.048 | 1.000  | 0.369  |
|        |     | GLASSO    | 2.828 | 0.000 | 5.172 | 1.000  | 0.353  |
| Graph | n  | Algorithm     | TP    | FP    | FN    | PPV   | Se    |
|------|----|---------------|-------|-------|-------|-------|-------|
|      |    | NPN-Copula    | 3.356 | 0.000 | 4.644 | 1.000 | 0.420 |
|      |    | NPN-Skeptic   | 3.384 | 0.000 | 4.616 | 1.000 | 0.423 |

Table 4: Simulation results from 500 replicates of the undirected graphs shown in Figure 1 of the main paper for $p = 10$ variables with Poisson node conditional distribution and level of noise $\lambda_{\text{noise}} = 5$. Monte Carlo means (standard deviations) are shown for TP, FP, FN, PPV and Se.
| Graph | n   | Algorithm | TP     | FP     | FN     | PPV    | Se     |
|-------|-----|-----------|--------|--------|--------|--------|--------|
|       | 200 | PC-LPGM   | 1.685  | 0.740  | 6.315  | 0.716  | 0.211  |
|       |     | LPGM      | 1.552  | 2.264  | 4.796  | 0.513  | 0.194  |
|       |     | PDN       | 3.204  | 4.904  | 6.498  | 0.513  | 0.194  |
|       |     | VSL       | 1.800  | 0.850  | 6.200  | 0.757  | 0.225  |
|       |     | GLASSO    | 1.805  | 0.845  | 6.195  | 0.758  | 0.226  |
|       |     | NPN-Copula| 1.980  | 0.735  | 6.020  | 0.801  | 0.248  |
|       |     | NPN-Skeptic| 1.795 | 0.830  | 6.205  | 0.752  | 0.224  |
|       | 1000| PC-LPGM   | 7.470  | 0.980  | 0.530  | 0.895  | 0.934  |
|       |     | LPGM      | 3.724  | 1.872  | 4.276  | 0.704  | 0.466  |
|       |     | PDN       | 4.816  | 2.600  | 3.184  | 0.653  | 0.602  |
|       |     | VSL       | 3.042  | 0.016  | 4.958  | 0.997  | 0.380  |
|       |     | GLASSO    | 3.018  | 0.016  | 4.982  | 0.997  | 0.378  |
|       |     | NPN-Copula| 3.164  | 0.008  | 4.836  | 0.998  | 0.396  |
|       |     | NPN-Skeptic| 2.972 | 0.010  | 5.028  | 0.998  | 0.372  |
| Random| 200 | PC-LPGM   | 61.585 | 8.490  | 37.415 | 0.880  | 0.622  |
|       |     | LPGM      | 5.564  | 0.824  | 93.436 | 0.985  | 0.056  |
|       |     | PDN       | 53.080 | 26.007 | 47.480 | 0.673  | 0.536  |
|       |     | VSL       | 63.915 | 22.308 | 35.085 | 0.760  | 0.446  |
|       |     | GLASSO    | 62.755 | 22.956 | 36.245 | 0.754  | 0.630  |
|       |     | NPN-Copula| 65.647 | 18.345 | 33.352 | 0.797  | 0.663  |
|       |     | NPN-Skeptic| 64.343| 0.010  | 4.215  | 1.000  | 0.473  |
| Scale-free| 200 | PC-LPGM   | 98.580 | 9.589  | 0.420  | 0.912  | 0.996  |
|       |     | LPGM      | 51.520 | 0.012  | 47.480 | 1.000  | 0.520  |
|       |     | PDN       | 65.357 | 0.050  | 33.643 | 0.999  | 0.660  |
|       |     | VSL       | 94.438 | 0.089  | 4.562  | 0.999  | 0.954  |
|       |     | GLASSO    | 93.830 | 0.161  | 5.170  | 0.998  | 0.948  |
|       |     | NPN-Copula| 94.571 | 0.054  | 4.429  | 1.000  | 0.952  |
|       |     | NPN-Skeptic| 94.277| 0.134  | 4.723  | 1.000  | 0.952  |
|       | 1000| PC-LPGM   | 95.580 | 9.589  | 0.420  | 0.912  | 0.996  |
|       |     | LPGM      | 51.520 | 0.012  | 47.480 | 1.000  | 0.520  |
|       |     | PDN       | 65.357 | 0.050  | 33.643 | 0.999  | 0.660  |
|       |     | VSL       | 94.438 | 0.089  | 4.562  | 0.999  | 0.954  |
|       |     | GLASSO    | 93.830 | 0.161  | 5.170  | 0.998  | 0.948  |
|       |     | NPN-Copula| 94.571 | 0.054  | 4.429  | 1.000  | 0.952  |
|       |     | NPN-Skeptic| 94.277| 0.134  | 4.723  | 1.000  | 0.952  |
|       | 2000| PC-LPGM   | 99.000 | 9.759  | 0.000  | 0.911  | 1.000  |
|       |     | LPGM      | 54.185 | 0.010  | 44.815 | 1.000  | 0.547  |
|       |     | PDN       | 64.370 | 0.000  | 34.630 | 1.000  | 0.650  |
|       |     | VSL       | 96.821 | 0.000  | 2.179  | 1.000  | 0.978  |
|       |     | GLASSO    | 96.518 | 0.000  | 2.482  | 1.000  | 0.975  |
|       |     | NPN-Copula| 97.375 | 0.000  | 1.625  | 1.000  | 0.984  |
|       |     | NPN-Skeptic| 97.214| 0.000  | 1.786  | 1.000  | 0.982  |
### Table 5 – continued from previous page

| Graph | n  | Algorithm       | TP      | FP      | FN      | PPV    | Se     |
|-------|----|-----------------|---------|---------|---------|--------|--------|
|       | 200| PC-LPGM         | 13.393  | 14.518  | 81.607  | 0.486  | 0.141  |
|       |    | LPGM            | 4.344   | 5.840   | 90.656  | 0.426  | 0.046  |
|       |    | PDN             | 19.340  | 84.747  | 75.660  | 0.186  | 0.204  |
|       |    | VSL             | 16.643  | 26.982  | 78.357  | 0.427  | 0.175  |
|       |    | GLASSO          | 15.991  | 25.518  | 76.509  | 0.434  | 0.168  |
|       |    | NPN-Copula      | 18.491  | 26.625  | 76.509  | 0.451  | 0.195  |
|       |    | NPN-Skeptic     | 17.473  | 31.348  | 77.527  | 0.406  | 0.184  |
|       | 1000| PC-LPGM         | 84.794  | 25.238  | 10.206  | 0.772  | 0.893  |
|       |    | LPGM            | 4.555   | 0.910   | 90.445  | 0.792  | 0.048  |
|       |    | PDN             | 78.487  | 19.650  | 16.513  | 0.800  | 0.826  |
|       |    | VSL             | 29.651  | 0.063   | 65.349  | 0.998  | 0.312  |
|       |    | GLASSO          | 29.341  | 0.056   | 65.659  | 0.998  | 0.309  |
|       |    | NPN-Copula      | 37.746  | 0.048   | 57.254  | 0.999  | 0.397  |
|       |    | NPN-Skeptic     | 35.476  | 0.119   | 59.524  | 0.998  | 0.373  |
|       | 2000| PC-LPGM         | 94.949  | 26.942  | 0.051   | 0.781  | 0.999  |
|       |    | LPGM            | 7.145   | 0.625   | 87.855  | 0.620  | 0.775  |
|       |    | PDN             | 93.073  | 1.113   | 1.927   | 0.988  | 0.980  |
|       |    | VSL             | 69.263  | 0.013   | 25.737  | 0.999  | 0.729  |
|       |    | GLASSO          | 68.647  | 0.048   | 57.254  | 0.998  | 0.397  |
|       |    | NPN-Copula      | 77.833  | 0.000   | 17.167  | 1.000  | 0.819  |
|       |    | NPN-Skeptic     | 74.987  | 0.013   | 20.013  | 1.000  | 0.789  |
|       | 200| PC-LPGM         | 62.432  | 8.656   | 46.568  | 0.879  | 0.573  |
|       |    | LPGM            | 8.190   | 0.120   | 100.810 | 0.987  | 0.075  |
|       |    | PDN             | 52.007  | 32.167  | 56.993  | 0.619  | 0.477  |
|       |    | VSL             | 67.032  | 26.932  | 41.968  | 0.735  | 0.615  |
|       |    | GLASSO          | 64.736  | 25.440  | 38.480  | 0.769  | 0.647  |
|       |    | NPN-Copula      | 68.956  | 29.956  | 40.044  | 0.722  | 0.635  |
|       | 1000| PC-LPGM         | 105.748 | 8.752   | 3.252   | 0.924  | 0.970  |
|       |    | LPGM            | 43.800  | 100.810 | 45.980  | 0.870  | 0.578  |
|       |    | PDN             | 63.020  | 9.470   | 45.980  | 0.870  | 0.578  |
|       |    | VSL             | 102.676 | 26.932  | 41.968  | 0.735  | 0.615  |
|       |    | GLASSO          | 101.904 | 25.440  | 38.480  | 0.769  | 0.647  |
|       |    | NPN-Copula      | 104.820 | 29.956  | 40.044  | 0.722  | 0.635  |
|       |    | NPN-Skeptic     | 104.392 | 104.820 | 104.820 | 0.970  | 0.970  |
|       | 2000| PC-LPGM         | 106.724 | 8.664   | 2.276   | 0.925  | 0.979  |
|       |    | LPGM            | 69.900  | 9.470   | 45.980  | 0.870  | 0.578  |
|       |    | PDN             | 62.850  | 26.932  | 41.968  | 0.735  | 0.615  |
|       |    | VSL             | 106.836 | 26.932  | 41.968  | 0.735  | 0.615  |
|       |    | GLASSO          | 106.884 | 26.932  | 41.968  | 0.735  | 0.615  |
|       |    | NPN-Copula      | 107.376 | 26.932  | 41.968  | 0.735  | 0.615  |
|       |    | NPN-Skeptic     | 107.124 | 26.932  | 41.968  | 0.735  | 0.615  |
Table 6: Simulation results from 500 replicates of the undirected graphs shown in Figure 2 of the main paper for $p = 100$ variables with Poisson node conditional distribution and level of noise $\lambda_{\text{noise}} = 5$. Monte Carlo means (standard deviations) are shown for TP, FP, FN, PPV and Se.

| Graph | n | Algorithm   | TP     | FP     | FN     | PPV  | Se     |
|-------|---|-------------|--------|--------|--------|------|--------|
|       |   |             |        |        |        |      |        |
| 200   |   | PC-LPGM     | 7.780  | 14.470 | 91.220 | 0.348| 0.079 |
|       |   | LPGM        | 10.188 | 65.352 | 88.812 | 0.152| 0.107 |
|       |   | PDN         | 13.457 | 94.817 | 85.543 | 0.125| 0.136 |
|       |   | VSL         | 9.316  | 22.496 | 89.684 | 0.332| 0.094 |
|       |   | GLASSO      | 9.052  | 21.372 | 89.948 | 0.336| 0.091 |
|       |   | NPN-Copula  | 10.012 | 21.924 | 88.988 | 0.359| 0.101 |
|       |   | NPN-Skeptic | 9.868  | 27.424 | 89.132 | 0.320| 0.100 |
|       |   | PC-LPGMC    | 75.130 | 24.805 | 23.870 | 0.753| 0.105 |
| 1000  |   | PC-LPGM     | 65.025 | 29.855 | 29.975 | 0.687| 0.075 |
|       |   | LPGM        | 52.827 | 31.153 | 46.173 | 0.630| 0.534 |
|       |   | PDN         | 14.844 | 0.044  | 84.156 | 0.998| 0.150 |
|       |   | VSL         | 14.936 | 0.044  | 84.064 | 0.998| 0.151 |
|       |   | GLASSO      | 17.124 | 0.040  | 81.876 | 0.998| 0.173 |
|       |   | NPN-Copula  | 16.708 | 0.116  | 82.292 | 0.996| 0.169 |
|       |   | NPN-Skeptic | 96.400 | 26.500 | 2.600  | 0.786| 0.974 |
| Scale-free | 2000 | PC-LPGM     | 96.400 | 26.500 | 2.600  | 0.786| 0.974 |
|       |   | PC-LPGMC    | 25.733 | 0.000  | 73.267 | 0.998| 0.264 |
|       |   | NPN-Copula  | 33.672 | 0.000  | 65.328 | 0.100| 0.335 |
|       |   | NPN-Skeptic | 32.267 | 0.000  | 66.733 | 0.100| 0.321 |
|       |   | PC-LPGM     | 2.690  | 13.600 | 92.108 | 0.166| 0.028 |
|       |   | LPGM        | 6.630  | 103.063| 88.370 | 0.060| 0.070 |
|       |   | PDN         | 3.392  | 23.688 | 91.608 | 0.143| 0.036 |
|       |   | VSL         | 3.304  | 22.964 | 91.696 | 0.145| 0.035 |
|       |   | GLASSO      | 3.392  | 21.852 | 91.608 | 0.150| 0.036 |
|       |   | NPN-Copula  | 3.108  | 23.476 | 91.892 | 0.134| 0.033 |
|       |   | NPN-Skeptic | 29.525 | 24.635 | 65.475 | 0.548| 0.311 |
|       |   | PC-LPGM     | 0.892  | 1.076  | 94.108 | 0.429| 0.009 |
|       |   | LPGM        | 7.424  | 84.333 | 71.573 | 0.217| 0.247 |
|       |   | PDN         | 7.364  | 1.424  | 87.576 | 0.884| 0.078 |
|       |   | VSL         | 8.440  | 1.392  | 86.560 | 0.895| 0.089 |
|       |   | GLASSO      | 8.208  | 1.804  | 86.792 | 0.860| 0.086 |
| Hub   | 1000 | PC-LPGM     | 29.025 | 24.635 | 65.475 | 0.548| 0.311 |
|       |   | LPGM        | 0.392  | 1.712  | 94.098 | 0.187| 0.064 |
|       |   | PDN         | 49.100 | 54.997 | 45.900 | 0.472| 0.517 |
|       |   | VSL         | 8.893  | 0.068  | 86.017 | 0.996| 0.095 |
|       |   | GLASSO      | 8.924  | 0.068  | 86.076 | 0.996| 0.094 |
|       |   | NPN-Copula  | 9.797  | 0.042  | 85.203 | 0.998| 0.103 |
|       |   | NPN-Skeptic | 9.305  | 0.068  | 85.695 | 0.995| 0.098 |
|       | 2000 | PC-LPGM     | 65.025 | 29.855 | 29.975 | 0.687| 0.684 |
|       |   | LPGM        | 0.392  | 1.712  | 94.098 | 0.187| 0.064 |
|       |   | PDN         | 49.100 | 54.997 | 45.900 | 0.472| 0.517 |
|       |   | VSL         | 8.893  | 0.068  | 86.017 | 0.996| 0.095 |
|       |   | GLASSO      | 8.924  | 0.068  | 86.076 | 0.996| 0.094 |
|       |   | NPN-Copula  | 9.797  | 0.042  | 85.203 | 0.998| 0.103 |
|       |   | NPN-Skeptic | 9.305  | 0.068  | 85.695 | 0.995| 0.098 |
|       | 200  | PC-LPGM     | 8.040  | 14.805 | 100.960| 0.350| 0.074 |
|       |   | LPGM        | 10.592 | 69.316 | 98.408 | 0.175| 0.097 |
|       |   | PDN         | 13.573 | 94.750 | 95.427 | 0.126| 0.125 |
|       |   | VSL         | 10.548 | 22.848 | 98.452 | 0.353| 0.097 |
|       |   | GLASSO      | 10.160 | 21.460 | 98.840 | 0.358| 0.093 |
|       |   | NPN-Copula  | 11.064 | 22.136 | 97.936 | 0.382| 0.102 |

*Nguyen and Chiogna*
| Graph | n  | Algorithm          | TP     | FP     | FN     | PPV   | Se    |
|-------|----|--------------------|--------|--------|--------|-------|-------|
|       | 1000 | NPN-Skeptic     | 10.648 (6.242) | 26.632 (23.376) | 98.352 (6.642) | 0.341 (0.134) | 0.098 (0.057) |
| Random |     | PC-LPGM           | 81.055 (4.632)  | 23.665 (4.941)  | 27.945 (4.632)  | 0.775 (0.038)  | 0.744 (0.042)  |
|       |     | LPGM              | 1.776 (2.675)   | 3.196 (5.107)   | 107.224 (2.675) | 0.397 (0.401)  | 0.016 (0.025)  |
|       |     | PDN               | 53.207 (3.471)  | 33.383 (10.084) | 55.793 (3.471)  | 0.616 (0.046)  | 0.488 (0.032)  |
|       |     | VSL               | 14.741 (6.294)  | 0.022 (0.148)   | 94.259 (6.294)  | 0.397 (0.401)  | 0.135 (0.058)  |
|       |     | GLASSO            | 14.741 (6.291)  | 0.022 (0.148)   | 94.259 (6.291)  | 0.999 (0.006)  | 0.135 (0.058)  |
|       |     | NPN-Copula        | 16.333 (7.249)  | 0.022 (0.148)   | 92.667 (7.249)  | 0.999 (0.005)  | 0.150 (0.067)  |
|       |     | NPN-Skeptic       | 15.178 (7.307)  | 0.044 (0.296)   | 93.822 (7.307)  | 0.998 (0.011)  | 0.139 (0.067)  |
|       | 2000 | PC-LPGM           | 104.010 (1.992) | 24.370 (4.706)  | 4.990 (1.992)   | 0.811 (0.029)  | 0.954 (0.018)  |
|       |     | LPGM              | 1.995 (1.800)   | 1.260 (1.825)   | 107.005 (1.880) | 0.671 (0.360)  | 0.018 (0.017)  |
|       |     | PDN               | 65.093 (2.892)  | 12.207 (1.837)  | 43.907 (2.892)  | 0.841 (0.021)  | 0.597 (0.027)  |
|       |     | GLASSO            | 26.038 (12.457) | 0.000 (0.000)   | 82.962 (12.457) | 1.000 (0.000)  | 0.239 (0.114)  |
|       |     | NPN-Copula        | 30.340 (14.496) | 0.000 (0.000)   | 78.660 (14.496) | 1.000 (0.000)  | 0.278 (0.133)  |
|       |     | NPN-Skeptic       | 28.474 (14.777) | 0.000 (0.000)   | 80.526 (14.777) | 1.000 (0.000)  | 0.261 (0.136)  |

References

G. Allen and Z. Liu. A local Poisson graphical model for inferring networks from sequencing data. *NanoBioscience, IEEE Transactions on*, 12(3):189–198, 2013.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 2010.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

F. Hadiji, A. Molina, S. Natarajan, and K. Kersting. Poisson dependency networks: Gradient boosted models for multivariate count data. *Machine Learning*, 100(2-3):477–507, 2015.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10(Apr):883–906, 2009.

A. Jalali, P. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 378–387, 2011.
C. R. Johnson and H. A. Robinson. Eigenvalue inequalities for principal submatrices. *Linear Algebra and its Applications*, 37:11–22, 1981.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

E. L. Lehmann. *Testing statistical hypotheses (2nd ed)*. New York, NY: Wiley, 1986.

J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440, 2010.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. The nonparanormal Skeptic. *arXiv preprint arXiv:1206.6488*, 2012.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

Y. D. Mizrahi, M. Denil, and N. de Freitas. Distributed parameter estimation in probabilistic graphical models. In *Advances in Neural Information Processing Systems*, pages 1700–1708, 2014.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using $l_1$-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 06 2010.

P. Spirtes, C. N Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

S. Volinia, M. Galasso, M. E. Sana, T. F. Wise, J. Palatini, K. Huebner, and C. M. Croce. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. *Proceedings of the National Academy of Sciences*, 109(8):3024–3029, 2012.

X. Yan, X. Chen, H. Liang, T. Deng, W. Chen, S. Zhang, M. Liu, X. Gao, Y. Liu, and C. Zhao. miR-143 and miR-145 synergistically regulate ERBB3 to suppress cell proliferation and invasion in breast cancer. *Molecular Cancer*, 13(1):220, 2014.
E. Yang, G. Allen, Z. Liu, and P. K. Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366, 2012.

E. Yang, P. Ravikumar, G. Allen, and Z. Liu. On Poisson graphical models. In *Advances in Neural Information Processing Systems*, pages 1718–1726, 2013.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847, 2015.

N. Zhang, X. Wang, Q. Huo, M. Sun, C. Cai, Z. Liu, G. Hu, and Q. Yang. MicroRNA-30a suppresses breast tumor growth and metastasis by targeting metadherin. *Oncogene*, 33 (24):3119–3128, 2014.