Published in final edited form as:
Biometrika. 2015 March ; 102(1): 47–64. doi:10.1093/biomet/asu051.

Selection and estimation for mixed graphical models

Shizhe Chen, **Daniela M. Witten**, and **Ali shojaie**
Department of Biostatistics, University of Washington, Box 357232, Seattle, Washington 98195, U.S.A
Shizhe Chen: szchen@uw.edu; Daniela M. Witten: dwitten@uw.edu; Ali shojaie: ashojaie@uw.edu

Summary

We consider the problem of estimating the parameters in a pairwise graphical model in which the distribution of each node, conditioned on the others, may have a different exponential family form. We identify restrictions on the parameter space required for the existence of a well-defined joint density, and establish the consistency of the neighbourhood selection approach for graph reconstruction in high dimensions when the true underlying graph is sparse. Motivated by our theoretical results, we investigate the selection of edges between nodes whose conditional distributions take different parametric forms, and show that efficiency can be gained if edge estimates obtained from the regressions of particular nodes are used to reconstruct the graph. These results are illustrated with examples of Gaussian, Bernoulli, Poisson and exponential distributions. Our theoretical findings are corroborated by evidence from simulation studies.

Keywords

Compatibility; Conditional likelihood; Exponential family; High dimensionality; Model selection consistency; Neighbourhood selection; Pairwise Markov random field

1. Introduction

In this paper, we consider the task of learning the structure of an undirected graphical model that encodes pairwise conditional dependence relationships among random variables. Specifically, suppose that we have p random variables represented as nodes of the graph $G = (V, E)$ with vertex set $V = \{1, \dots, p\}$ and edge set $E \subseteq V \times V$. An edge in the graph indicates a pair of random variables that are conditionally dependent given all the other variables. The problem of reconstructing the graph from a set of n observations has attracted much interest in recent years, especially when $p > n$ and $p(p - 1)/2$ edges must be estimated from n observations.

Many authors have studied the estimation of high-dimensional undirected graphical models in the setting where the distribution of each node, conditioned on all the other nodes, has the same parametric form. In particular, Gaussian graphical models have been studied

Supplementary material
Supplementary material available at *Biometrika* online includes proofs of the theoretical results, additional details about the numerical studies, and R code for replicating the numerical results.

extensively (see, e.g., Meinshausen & Bühlmann, 2006; Yuan & Lin, 2007; Friedman et al., 2008; Rothman et al., 2008;Wainwright & Jordan, 2008; Peng et al., 2009; Ravikumar et al., 2011) and generalized to account for nonnormality and outliers (see, e.g., Miyamura & Kano, 2006; Finegold & Drton, 2011; Vogel & Fried, 2011; Sun & Li, 2012). Other authors have considered the setting in which all node-conditional distributions are Bernoulli (Lee et al., 2007; Höfling & Tibshirani, 2009; Ravikumar et al., 2010), multinomial (Jalali et al., 2011), Poisson (Allen & Liu, 2012), or any univariate distribution in the exponential family (Yang et al., 2012). An extended version of Yang et al. (2012) is available as an unpublished technical report.

In this paper, we seek to estimate a graphical model in which the variables are of different types. Here, the type of a node refers to the parametric form of its distribution, conditioned on all the other nodes. For instance, the variables might include DNA nucleotides, taking binary values, and gene expression levels measured using RNA sequencing, taking nonnegative integer values. We could model the first set of nodes as Bernoulli, which means that each of their distributions, conditional on the other nodes, is Bernoulli; similarly, we could model the second set as Poisson. We assume that the type of each node is known a priori, and refer to this set-up as a mixed graphical model.

In the low-dimensional setting, Lauritzen (1996) studied a special case of the mixed graphical model, known as the conditional Gaussian model, in which each node is either Gaussian or Bernoulli. More recent work has focused on the high-dimensional setting. Lee & Hastie (2015) proposed two algorithms for reconstructing conditional Gaussian models using a group lasso penalty. In a 2013 unpublished technical report (arXiv:1304.2810), J. Cheng, E. Levina and J. Zhu modified this approach by using a weighted ℓ_1 penalty.

A related line of research considers semiparametric or nonparametric approaches to estimating conditional dependence relationships (Liu et al., 2009; Xue & Zou, 2012; Fellinghauer et al., 2013; Voorman et al., 2014); of these methods, that of Fellinghauer et al. (2013) is specifically proposed for mixed graphical models. However, despite their flexibility, these nonparametric methods are often less efficient than their parametric counterparts, if the type of each node is known.

In this paper, we propose an estimator and develop theory for the parametric mixed graphical model, under a much more general setting than existing approaches (e.g., Lee & Hastie, 2015). We allow the conditional distribution of each node to belong to the exponential family. Unlike Yang et al. (2012), nodes may be of different types. For instance, within a single graph, some nodes may be Bernoulli, some Poisson, and some exponential.

In parallel efforts, Yang et al. (2014) recently presented general results on strong compatibility for mixed graphical models for which the node-conditional distributions belong to the exponential family and the graph contains only two types of nodes. We instead consider the setting where the graph can contain more than two types of nodes, and provide specific requirements for strong compatibility for some common distributions.

2. A model for mixed data

2·1. Conditionally specified models for mixed data

We consider the pairwise graphical model (Wainwright et al., 2007), which takes the form

$$p(x) \propto \exp\left\{\sum_{s=1}^{p} f_s(x_s) + \sum_{s=2}^{p}\sum_{t<s} f_{ts}(x_s, x_t)\right\}, \quad (1)$$

where $x = (x_1, \ldots, x_p)^{\mathrm{T}}$ and $f_{ts} = 0$ for $\{t, s\} \notin E$. Here, $f_s(x_s)$ is the node potential function and $f_{st}(x_s, x_t)$ is the edge potential function. We further simplify the pairwise interactions by assuming that $f_{st}(x_s, x_t) = \theta_{st}x_s x_t = \theta_{ts}x_s x_t$, so that we can write the parameters associated with edges in a symmetric square matrix $\Theta = (\theta_{st})_{p\times p}$ in which the diagonal elements equal zero. The joint density can then be written as

$$p(x) = \exp\left\{\sum_{s=1}^{p} f_s(x_s) + \frac{1}{2}\sum_{s=1}^{p}\sum_{t\neq s} \theta_{ts}x_s x_t - A(\Theta, \alpha)\right\}, \quad (2)$$

where $A(\Theta, \alpha)$ is the log-partition function, a function of Θ and α. Here α is a $K \times p$ matrix, where K is some known integer, of parameters involved in the node potential functions; that is, $f_s(x_s)$ involves α_s, the sth column of α. For $\{s, t\} \notin E$, the edge potentials satisfy $\theta_{st} = \theta_{ts} = 0$. We define the neighbours of the sth node by $N(x_s) = \{t : \theta_{st} = \theta_{ts} \neq 0\}$.

In principle, given a parametric form for the joint density (2), we can estimate the conditional dependence relationships among the p variables, and hence the edges in the graph. But this approach requires calculation of the log-partition function $A(\Theta, \alpha)$, which is often intractable. To overcome this difficulty, we instead use the framework of conditionally specified models (Besag, 1974): we specify the distribution of each node conditional on the others, and then combine the p conditional distributions to form a single graphical model. This approach has been widely used in estimating high-dimensional graphical models where all nodes are of the same type (Meinshausen & Bühlmann, 2006; Ravikumar et al., 2010; Allen & Liu, 2012; Yang et al., 2012). However, as we will discuss in § 2·2, a conditionally specified model may not correspond to a valid joint distribution.

Define $x_{-s} = (x_1, \ldots, x_{s-1}, x_{s+1}, \ldots, x_p)^{\mathrm{T}}$. We consider conditional densities of the form

$$p(x_s \mid x_{-s}) = \exp\left\{f_s(x_s) + \sum_{t\neq s} \theta_{ts}x_t x_s - D_s(\eta_s)\right\}, \quad (3)$$

where $\eta_s = \eta_s(\Theta_s, x_{-s}, \alpha_s)$ is a function of α_s, x_{-s} and Θ_s, with Θ_s being the sth column of Θ without the diagonal element. Suppose that $f_s(x_s) = \alpha_{1s}x_s + \alpha_{2s}x_s^2/2 + \sum_{k=3}^{K} \alpha_{ks}B_{ks}(x_s)$, where each α_{ks} is a parameter, which could be zero, and $B_{ks}(x_s)$ is a known function for $k = 3, \ldots, K$. Under this assumption, (3) belongs to the exponential family.

The assumed form of $f_s(x_s)$ is quite general. We now consider some special cases of (3) corresponding to commonly used distributions in the exponential family, for which $f_s(x_s)$ takes a very simple form. In the following examples, we assume that $\eta_s(\Theta_s, x_{-s}, \alpha_s) = \alpha_{1s} + \sum_{t:\, t \neq s} \theta_{ts} x_t$.

Example 1—The conditional density is Gaussian and $\alpha_{2s} = -1$:

$$p(x_s|x_{-s})=\exp\left\{-\frac{1}{2}x_s^2+\eta_s x_s-\frac{1}{2}\eta_s^2-\frac{1}{2}\log(2\pi)\right\}\quad(x_s\in\mathbb{R}), \qquad (4)$$

where $f_s(x_s)=\alpha_{1s}x_s-x_s^2/2$ and $D_s(\eta_s)=\eta_s^2/2+\log(2\pi)/2$.

Example 2—The conditional density is Bernoulli and, instead of coding x_s as $\{0, 1\}$, we code x_s as $\{-1, 1\}$; this yields the conditional density

$$p(x_s|x_{-s})=\exp\{\eta_s x_s-D_s(\eta_s)\}\ (x_s\in\{-1,1\}), \qquad (5)$$

where $f_s(x_s) = \alpha_{1s}x_s$ and $D_s(\eta_s) = \log\{\exp(\eta_s) + \exp(-\eta_s)\}$.

Example 3—The conditional density is Poisson:

$$p(x_s|x_{-s})=\exp\{\eta_s x_s-\log(x_s!)-D_s(\eta_s)\}\ (x_s\in\{0,1,\ldots\}), \qquad (6)$$

where $f_s(x_s) = \alpha_{1s}x_s - \log(x_s!)$ and $D_s(\eta_s) = \exp(\eta_s)$.

Example 4—The conditional density is exponential:

$$p(x_s|x_{-s})=\exp\{\eta_s x_s-D_s(\eta_s)\}\ (x_s\in\mathbb{R}^+), \qquad (7)$$

where $f_s(x_s) = \alpha_{1s}x_s$ and $D_s(\eta_s) = -\log(-\eta_s)$.

These four examples have been studied in the context of conditionally specified graphical models in which all nodes are of the same type (Besag, 1974; Meinshausen & Bühlmann, 2006; Ravikumar et al., 2010; Allen & Liu, 2012; Yang et al., 2012).

In what follows, we will consider the conditionally specified mixed graphical model, with conditional distributions given by (3), in which each node can be of a different type. This class of mixed graphical models is not closed under marginalization; for instance, given a graph composed of Gaussian and Bernoulli nodes, integrating out the Bernoulli nodes leads to a conditional density that is a mixture of Gaussians, which does not belong to the exponential family.

2·2. Compatibility of conditionally specified models

Under what circumstances does the conditionally specified model with node-conditional distributions given in (3) correspond to a well-defined joint distribution? We first adapt and restate a definition from Wang & Ip (2008), which applies to any conditional density.

Definition 1—*A nonnegative function g is capable of generating a conditional density function* $p(y \mid x)$ *if*

$$p(y|x)=\frac{g(y,x)}{\int g(y,x)\mathrm{d}y}.$$

Two conditional densities are said to be compatible if there exists a function g that is capable of generating both conditional densities. When g is a density, the conditional densities are said to be strongly compatible.

The following proposition relates Definition 1 to the conditional density in (3). Its proof, as well as the proofs of other statements in this paper, is given in the Supplementary Material.

Proposition 1—*Let* $x = (x_1, \ldots, x_p)^{\mathrm{T}}$ *be a random vector. Suppose that for each* x_s*, the conditional density takes the form* (3). *If* $\theta_{st} = \theta_{ts}$*, then the conditional densities are compatible. Furthermore, any function g that is capable of generating the conditional densities is of the form*

$$g(x) \propto \exp\left\{\sum_{s=1}^{p} f_s(x_s)+\frac{1}{2}\sum_{s=1}^{p}\sum_{t\neq s}\theta_{ts}x_s x_t\right\}. \qquad (8)$$

Under the conditions of Proposition 1, if we further assume that *g* in (8) is integrable, then by Definition 1 the conditional densities of the form (3) are strongly compatible. Proposition 1 indicates that, provided (2) is a valid joint distribution, we can arrive at it via the conditional densities in (3). This justifies the conditionally specified modelling approach taken in this paper. Proposition 1 is closely related to § 4.3 in Besag (1974) and Proposition 1 of Yang et al. (2012), with small modifications. More general theory is developed in Wang & Ip (2008).

We now return to the four examples (4)–(7). Lemma 1 summarizes the conditions under which a conditionally specified model with nondegenerate conditional distributions of the form (4)–(7) leads to a valid joint distribution.

Lemma 1—*If* $\theta_{st} = \theta_{ts}$*, then the subset of conditions with a dagger* (†) *in* Table 1 *is necessary and sufficient for the conditional densities in* (4)–(7) *to be compatible. Moreover, the complete set of conditions in* Table 1 *is necessary and sufficient for the conditional densities in* (4)–(7) *to be strongly compatible.*

To simplify the presentation of the conditions for the Gaussian nodes, in Table 1 it is assumed that J is the index set of the Gaussian nodes. Without loss of generality, we further assume that the nodes are ordered such that $J = \{1, \ldots, m\}$, and define

$$\Theta_{JJ} = \begin{pmatrix} \alpha_{21} & \theta_{12} & \cdots & \theta_{1m} \\ \theta_{21} & \alpha_{22} & \cdots & \theta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{m1} & \theta_{m2} & \cdots & \alpha_{2m} \end{pmatrix}. \qquad (9)$$

Table 1 displays the set of restrictions on the parameter space that must hold in order for the conditional densities in (4)–(7) to be compatible or strongly compatible. The diagonal entries of this table were previously studied in Besag (1974). In general, strong compatibility imposes more restrictions on the parameter space than does compatibility. For instance, compatibility does not place any restrictions on the edges between two Poisson nodes, but for strong compatibility the edge potentials must be negative. Compatibility and strong compatibility even restrict the relationships that can be modelled using the conditional densities (4)–(7); for instance, no edges are possible between Gaussian and exponential nodes, or between Gaussian and Poisson nodes.

To summarize, given conditional densities of the form (4)–(7), existence of a joint density imposes substantial constraints on the parameter space, and thus limits the flexibility of the corresponding graph. However, we will see in § 5 that it is possible to consistently estimate the structure of a graph even when the requirements for compatibility or strong compatibility are violated, i.e., even in the absence of a joint density.

While Table 1 examines only conditionally specified models composed of the conditional densities in (4)–(7), the estimator proposed in § 3 and the theory developed in § § 4 and 5 apply to other types of conditional densities of the form (3).

3. Estimation via neighbourhood selection

3·1. Estimation

We now present a neighbourhood selection approach for recovering the structure of a mixed graphical model, by maximizing penalized conditional likelihoods node by node. A similar approach has been studied when all nodes in the graph are of the same type (Meinshausen & Bühlmann, 2006; Ravikumar et al., 2010; Allen & Liu, 2012, Yang et al., 2012).

Recall from § 2·1 that $f_s(x_s)=\alpha_{1s}x_s+\alpha_{2s}x_s^2/2+\sum_{k=3}^{K}\alpha_{ks}B_{ks}(x_s)$. We now simplify the problem by assuming that the α_{ks} are known, and possibly zero, for $k \geq 2$. Let X denote an $n \times p$ data matrix, with the ith row given by $x^{(i)}$. From now on, we will use an asterisk to indicate the true parameter values. We estimate Θ_s^* and α_{1s}^*, the parameters for the sth node, as

$$\underset{\Theta_s \in \mathbb{R}^{p-1}, \alpha_{1s} \in \mathbb{R}}{\arg\min} - \ell_s(\Theta_s, \alpha_{1s}; X) + \lambda_n \|\Theta_s\|_1, \qquad (10)$$

where $\ell_s(\Theta_s, \alpha_{1s}; X) = \sum_{i=1}^{n} \log p(x_s^{(i)} | x_{-s}^{(i)})/n$; recall that the conditional density $p(x_s^{(i)} | x_{-s}^{(i)})$ was defined in (3). Finally, we define the estimated neighbourhood of x_s to be $\hat{N}(x_s) = \{t : \hat{\theta}_{ts} \neq 0\}$, where $\hat{\Theta}_s$ solves (10) and $\hat{\theta}_{ts}$ is the element corresponding to an edge with the tth node.

In practice, to avoid a situation where variables of different types are on different scales, we may wish to modify (10) in order to allow a different weight for the ℓ_1 penalty on each coefficient. We define a weight vector w equal to the empirical standard errors of the corresponding variables: $w = (\hat{\sigma}_1, \ldots, \hat{\sigma}_{s-1}, \hat{\sigma}_{s+1}, \ldots, \hat{\sigma}_p)^T$. Then (10) can be replaced with

$$\underset{\Theta_s \in \mathbb{R}^{p-1}, \alpha_{1s} \in \mathbb{R}}{\arg\min} - \ell_s(\Theta_s, \alpha_{1s}; X) + \lambda_n \|\mathrm{diag}(w)\Theta_s\|_1. \qquad (11)$$

The analysis in § § 4 and 5 uses (10) for simplicity, but could be generalized to (11) with additional bookkeeping. Both (10) and (11) can be easily solved (see, e.g., Friedman et al., 2010).

In the joint density (2), the parameter matrix Θ is symmetric, i.e., $\theta_{st} = \theta_{ts}$, but the neighbourhood selection method does not guarantee symmetric estimates; for instance, it could happen that $\hat{\theta}_{st} = 0$ but $\hat{\theta}_{ts} = \neq 0$. Our analysis in § 4·2 shows that we can exploit the asymmetry in $\hat{\theta}_{st}$ and $\hat{\theta}_{ts}$ when x_s and x_t are of different types in order to obtain more efficient edge estimates.

3·2. Tuning

In order to select the value of the tuning parameter λ_n in (10), we use the Bayesian information criterion (Zou et al., 2007; Peng et al., 2009; Voorman et al., 2014), which takes the form

$$\mathrm{BIC}_s(\lambda_n) = -2n\ell_s(\hat{\Theta}_s, \hat{\alpha}_{1s}; X) + \log(n)\|\hat{\Theta}_s\|_0.$$

where $\|\hat{\Theta}_s\|_0$ is the number of nonzero elements in $\hat{\Theta}_s$ for a given value of λ_n. We allow a different value of λ_n for each node type. For instance, to select λ_n for the Poisson nodes, we choose the value of λ_n such that $\mathrm{BIC}_s(\lambda_n)$, summed over the Poisson nodes, is minimized. We evaluate the performance of this method for tuning parameter selection in § 6·3.

4. Recovery with strongly compatible conditional distributions

4·1. Neighbourhood recovery

In this subsection we show that if the conditional distributions in (3) are strongly compatible, as they will be under the conditions discussed in § 2·2, then under some additional assumptions, the true neighbourhood of each node is consistently selected using the neighbourhood selection approach proposed in § 3·1. Here we rely heavily on results from Yang et al. (2012), who consider a related problem in which all nodes are of the same type.

In the following discussion, we assume $p > n$ for simplicity. For any s, let Δ_s denote the set of indices for elements of $(\Theta_s^{\mathrm{T}}, \alpha_{1s})^{\mathrm{T}}$ that correspond to non-neighbours of the sth node, and let $Q_s^* = -\nabla^2 \ell_s(\Theta_s^*, \alpha_{1s}^*; X)$ be the negative Hessian of $\ell_s(\Theta_s, \alpha_{1s}; X)$ with respect to $(\Theta_s^{\mathrm{T}}, \alpha_{1s})^{\mathrm{T}}$, evaluated at the true values of the parameters. Below we suppress the subscript s for simplicity, and we remind the reader that all quantities are related to the conditional density of the sth node. We express Q^* in blocks as

$$Q^* = \begin{pmatrix} Q^*_{\Delta^c\Delta^c} & Q^*_{\Delta^c\Delta} \\ Q^*_{\Delta\Delta^c} & Q^*_{\Delta\Delta} \end{pmatrix}.$$

Assumption 1—There exists a positive number a such that

$$\max_{l \in \Delta} \| Q^*_{l\Delta^c} (Q^*_{\Delta^c\Delta^c})^{-1} \|_1 \leq 1 - a.$$

Assumption 1 limits the association between neighbours and non-neighbours of the sth node: if the association is too high, then it is not possible to select the correct neighbourhood. This type of assumption is standard for variable selection consistency of ℓ_1-penalized estimators (see, e.g., Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Wainwright, 2009; Ravikumar et al., 2010, 2011; Yang et al., 2012; Lee et al., 2013).

Assumption 2—There exists $\Lambda_1 > 0$ such that the smallest eigenvalue of $Q^*_{\Delta^c\Delta^c}$, $\Lambda_{\min}(Q^*_{\Delta^c\Delta^c})$, is greater than or equal to Λ_1. Also, there exists $\Lambda_2 < \infty$ such that the largest eigenvalue of $\sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} / n$, $\Lambda_{\max}\{\sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} / n\}$, is less than or equal to Λ_2, where $x_0 = (x_{-s}^{\mathrm{T}}, 1)^{\mathrm{T}}$.

The lower bound in Assumption 2 is needed to prevent singularity among the true neighbours, which would prevent neighbourhood recovery. The bound on the largest eigenvalue of the sample covariance matrix is needed to prevent a situation where most of the variance in the data is due to a single feature. Similar assumptions were made in Meinshausen & Bühlmann (2006), Zhao & Yu (2006), Wainwright (2009), Ravikumar et al. (2010) and Yang et al. (2012).

Assumption 3—The log-partition function $D(\cdot)$ of the conditional density $p(x_s \mid x_{-s})$ is third-order differentiable, and there exist κ_2 and κ_3 such that $|D''(y)| \leq \kappa_2$ and $|D'''(y)| \leq \kappa_3$ for $y \in \{y : y \in \mathscr{D}, |y| \leq M\delta_1 \log p\}$, where $\mathscr{D}$ is the support of $D(\cdot)$.

Remark 1: The two quantities κ_2 and κ_3 in Assumption 3 are functions of p. The quantity δ_1 is a constant to be chosen in Proposition 2. The constant M is a sufficiently large constant that plays a role in Assumption 6.

Assumption 3 controls the smoothness of the log-partition function $D(\cdot)$ for conditional densities of the form (3). Recall from § 2·1 that the log-partition function of the node x_s is $D(\eta_s)$, where η_s equals $\alpha_{1s} + \sum_{t\neq s}\theta_{ts}x_t$. To apply Assumption 3 to $D(\eta_s)$, we will need to bound $\sum_{t\neq s}\theta_{ts}x_t$, so that $|\eta_s| \le M\delta_1 \log(p)$.

In order to obtain such a bound, we need another assumption.

Assumption 4—For $t = 1, \ldots, p$:

- **i.** $|E(x_t)| \le \kappa_m$;
- **ii.** $E(x_t^2) \le \kappa_v$;
- **iii.** $\max_{u:|u|\le 1}(\partial^2 A/\partial \alpha_{1t}^2)\Big|_{\alpha_{1t}^*+u} \le \kappa_h$ and $\max_{u:|u|\le 1}(\partial^2 A/\partial \alpha_{2t}^2)\Big|_{\alpha_{2t}^*+u} \le \kappa_h$.

Assumption 4 controls the moments of each node, as well as the local smoothness of the log-partition function A in (2). Given Assumption 4, the following propositions on the marginal behaviour of random variables hold; see Propositions 3 and 4 in Yang et al. (2012).

Proposition 2—*Define the event*

$$\xi_1 = \left(\max_{i\in\{1,\ldots,n\};t\in\{1,\ldots,p\}} |x_t^{(i)}| < \delta_1 \log p\right).$$

Then, assuming $p > n$, $\mathrm{pr}(\xi_1) \ge 1 - c_1 p^{-\delta_1+2}$ *where* $c_1 = \exp(\kappa_m + \kappa_h/2)$.

Proposition 3—*Define the event*

$$\xi_2 = \left[\max_{t\in\{1,\ldots,p\}} \left\{\frac{1}{n}\sum_{i=1}^{n}\left(x_t^{(i)}\right)^2\right\} < \delta_2\right],$$

where $\delta_2 \ge 1$. *If* $\delta_2 \le \min(2\kappa_\upsilon/3, \kappa_h + \kappa_\upsilon)$ *and* $n \ge 8\kappa_h^2 \log p/\delta_2^2$, *then* $\mathrm{pr}(\xi_2) \ge 1 - \exp(-c_2\delta_2^2 n)$ *where* $c_2 = 1/(4\kappa_h^2)$.

We now present three additional assumptions that relate to the node-wise regression in (10).

Assumption 5—The minimum of the edge potentials related to node x_s, $\min_{t\in N(x_s)}|\theta_{ts}|$, is larger than $10(d+1)^{1/2}\lambda_n/\Lambda_1$, where d is the number of neighbours of x_s.

Assumption 6—The tuning parameter λ_n is in the range

$$\left[\frac{8(2-a)}{a}\left\{\delta_2\kappa_2\frac{\log(2p)}{n}\right\}^{1/2},\ \min\left\{\frac{2(2-a)}{a}\kappa_2\delta_2 M,\frac{a\Lambda_1^2(d+1)^{-1}}{288(2-a)\kappa_2\Lambda_2},\frac{\Lambda_1^2(d+1)^{-1}}{12\Lambda_2\kappa_3\delta_1\ \log\ p}\right\}\right].$$

(12)

Remark 2: Of the three quantities in the upper bound for λ_n, $\Lambda_1^2/\{12\Lambda_2(d+1)\kappa_3\delta_1\ \log\ p\}$ is usually the smallest because of the log p in the denominator.

Assumption 7—The sample size n is no smaller than $8\kappa_h^2\ \log\ p/\delta_2^2$; also, the range of feasible λ_n in Assumption 6 is nonempty, i.e.,

$$n\geq\frac{96^2(2-a)^2\Lambda_2^2}{a^2\Lambda_1^4}(d+1)^2\kappa_2\kappa_3^2\delta_1^2\delta_2\ \log(2p)(\log\ p)^2. \qquad (13)$$

Assumptions 5, 6 and 7 specify the minimum edge potential, the range of the tuning parameter and the minimum sample size required for Theorem 1 to hold, i.e., for our neighbourhood selection approach (10) to achieve model selection consistency. Similar assumptions are made in related work (e.g., Yang et al., 2012).

Remark 3: Suppose that $n = \Omega\{(d+1)^2 \log^{3+\varepsilon}(p)\}$ for $\varepsilon > 0$, $\lambda_n = c\{\log(p)/n\}^{1/2}$ for some constant c, and κ_2 and κ_3 are $O(1)$. Then Assumptions 6 and 7 are satisfied asymptotically as n and p tend to infinity. Similar rates appear in Meinshausen & Bühlmann (2006), Ravikumar et al. (2010) and Yang et al. (2012).

Theorem 1—*Suppose that the joint density* (2) *exists and that Assumptions* 1–7 *hold for the sth node. Then, with probability at least* $1-c_1p^{-\delta_1+2}-\exp(-c_2\delta_2^2n)-\exp(-c_3\delta_3n)$, *for some constants* c_1, c_2, c_3, $\delta_2 \le \min(2\kappa_\upsilon/3, \kappa_h + \kappa_\upsilon)$ *and* $\delta_3 = 1/(\kappa_2\delta_2)$, *the estimator from* (10) *recovers the true neighbourhood of* x_s *exactly, so that* $\hat{N}(x_s) = N(x_s)$.

Theorem 1 shows that the probability of successful recovery converges to unity exponentially fast with the sample size n. We note that the number of neighbours d appears in Assumptions 5–7. As d increases, the minimum edge potential for each neighbour increases, the upper range for λ_n decreases, and the required sample size increases. Therefore, we need the true graph G to be sparse, $d = o(n)$, in order for Theorem 1 to be meaningful.

The quantities $\delta_2\kappa_2$ and $\delta_1\kappa_3$ appear in the upper bound for λ_n, in (12), and the minimum sample size, (13). The fact that κ_2 and δ_2 appear together in a product implies that we can relax the restriction on δ_2 if κ_2 is small. The same applies to δ_1 and κ_3.

For certain types of nodes, Theorem 1 holds with a less stringent set of assumptions. For a Gaussian node, the second- and higher-order derivatives of $D(\cdot)$ are always bounded, i.e., $\kappa_2 = 1$ and $\kappa_3 = 0$. This has profound effects on the theory, as illustrated in Corollary 1.

Corollary 1—*Suppose that the joint density* (2) *exists and that Assumptions* 1–5 *hold for a Gaussian node x_s. If*

$$\lambda_n \in \left[\frac{8(2-a)}{a}\left\{\delta_2 \frac{\log(2p)}{n}\right\}^{1/2}, \frac{2(2-a)}{a}\delta_2 M\right],\ n \geq \frac{8\kappa_h^2 \log p}{\delta_2^2},$$

then with probability at least $1-\exp(-c_2\delta_2^2 n)-\exp(-c_3\delta_3 n)$, *for some constants* c_2, c_3, $\delta_2 \leq \min(2\kappa_\upsilon/3, \kappa_h + \kappa_\upsilon)$ *and* $\delta_3 = \delta_2$, *the estimator from* (10) *recovers the true neighbourhood of x_s exactly, so that $\hat{N}(x_s) = N(x_s)$.*

4·2. Combining neighbourhoods to estimate the edge set

The neighbourhood selection approach may give asymmetric estimates, in the sense that $t \in \hat{N}(x_s)$ but $s \notin \hat{N}(x_t)$. To deal with this discrepancy, two strategies for estimating a single edge set were proposed by Meinshausen & Bühlmann (2006) and adapted in other work:

$$\hat{E}_{\text{and}}=\left\{(s,t):s \in \hat{N}(x_t) \text{ and } t \in \hat{N}(x_s)\right\},\ \hat{E}_{\text{or}}=\left\{(s,t):s \in \hat{N}(x_t) \text{ or } t \in \hat{N}(x_s)\right\}.$$

When the sth and tth nodes are of the same type, there is no clear reason to prefer the edge estimate from $\hat{N}(x_s)$ over the one from $\hat{N}(x_t)$, and so the choice between the intersection rule, $\hat{E}_{\text{and}}$, and the union rule, $\hat{E}_{\text{or}}$, is not crucial (Meinshausen & Bühlmann, 2006).

When the sth and tth nodes are of different types, however, the choice of neighbourhood matters. We now take a closer look at this situation with examples of Gaussian, Bernoulli, exponential and Poisson nodes as in (4)–(7). The quantities c_1, c_2 and c_3 in Theorem 1 are the same regardless of the node type, while the values of κ_2 and κ_3 depend on the type of node being regressed on the others in (10). We fix $B_1 = \kappa_3\delta_1$ for Bernoulli, Poisson and exponential nodes. For a Gaussian node, this quantity will always equal zero, since $D(\eta_s)=\eta_s^2/2+\log(2\pi)/2$ and hence $D'''(\eta_s) = 0 = \kappa_3$. Furthermore, we fix $B_2 = 1/\delta_3 = \delta_2\kappa_2$ for all four types of nodes. With B_1 and B_2 fixed, the minimum sample size and the feasible range of the tuning parameter for Bernoulli, Poisson and exponential nodes are exactly the same, as these quantities involve only B_1 and B_2. In particular, from Assumption 6, the range of feasible λ_n is $[8(2-a)\{\log(2p)B_2/n\}^{1/2}/a, \Lambda_1^2/\{12\Lambda_2(d+1)B_1 \log p\}]$, and from Assumption 7, the minimum sample size is $96^2(2-a)^2\Lambda_2^2(d+1)^2 B_2 B_1^2 \log(2p)(\log p)^2/(a^2\Lambda_1^4)$. These bounds are more restrictive than the corresponding bounds for Gaussian nodes in Corollary 1. We now derive a lower bound on the probability of successful neighbourhood recovery for each node type.

Example 5—If x_s is a Gaussian node, then the log-partition function is $D(\eta_s)=\eta_s^2/2+\log(2\pi)/2$. It follows that $D''(\eta_s) = 1 = \kappa_2$. Thus $\delta_2 = B_2$. By Corollary 1, a lower bound for the probability of successful neighbourhood recovery is

$$\text{pr}\{\hat{N}(x_s)=N(x_s)\} \geq 1-\exp(-c_2B_2^2n)-\exp(-c_3n/B_2).$$

Example 6—If x_s is a Bernoulli node, then the log-partition function is $D(\eta_s) = \log\{\exp(-\eta_s) + \exp(\eta_s)\}$, so that $|D''(\eta_s)| \leq 1$ and $|D'''(\eta_s)| \leq 2$. Consequently, $\delta_2 = B_2$ and $\delta_1 = B_1/\kappa_3 = B_1/2$. By Theorem 1, a lower bound for the probability of successful neighbourhood recovery is

$$\text{pr}\{\hat{N}(x_s)=N(x_s)\} \geq 1-c_1p^{-B_1/2+2}-\exp(-c_2B_2^2n)-\exp(-c_3n/B_2). \quad (14)$$

Example 7—If x_s is a Poisson node, then the log-partition function is $D(\eta_s) = \exp(\eta_s)$, so $D''(\eta_s) = D'''(\eta_s) = \exp(\eta_s)$. To bound $D''(\eta_s)$ and $D'''(\eta_s)$, we need to bound $\exp(\eta_s)$. Recall from Table 1 that strong compatibility requires that $\theta_{ts}x_t \leq 0$ when x_t is Gaussian, Poisson or exponential. Therefore, an upper bound for $\exp(\eta_s)$ is

$$\exp(\eta_s) \leq \exp\left(\alpha_{1s}+\sum_{t\in I}|\theta_{ts}|\right) \equiv b_{\mathrm{P}}, \quad (15)$$

with I being the set of Bernoulli nodes. Therefore $\kappa_2 = \kappa_3 = b_{\mathrm{P}}$, and so $\delta_2 = B_2/b_{\mathrm{P}}$ and $\delta_1 = B_1/b_{\mathrm{P}}$. By Theorem 1, a lower bound on the probability of successful neighbourhood recovery is

$$\text{pr}\{\hat{N}(x_s)=N(x_s)\} \geq 1-c_1p^{-B_1/b_{\mathrm{P}}+2}-\exp(-c_2B_2^2n/b_{\mathrm{P}}^2)-\exp(-c_3n/B_2). \quad (16)$$

Example 8—If x_s is an exponential node, then the log-partition function is $D(\eta_s) = -\log(-\eta_s)$, so $D''(\eta_s)=\eta_s^{-2}$ and $D'''(\eta_s)=-2\eta_s^{-3}$. Furthermore,

$$\eta_s=\alpha_{1s}+\sum_{t\neq s}\theta_{ts}x_t \leq \alpha_{1s}+\sum_{t\in I}\theta_{ts}x_t \leq \alpha_{1s}+\sum_{t\in I}|\theta_{ts}|<0, \quad (17)$$

with I being the set of Bernoulli nodes. In (17), the first inequality follows from the requirement for compatibility from Table 1 that $\theta_{ts}x_t \leq 0$ when x_t is Gaussian, Poisson or exponential; the second inequality follows from the fact that Bernoulli nodes are coded as +1 and −1; and the third inequality follows from the Bernoulli-exponential entry in Table 1. Therefore,

$$|\eta_s| \geq \left|\alpha_{1s}+\sum_{t\in I}|\theta_{ts}|\right| \geq |\alpha_{1s}| - \sum_{t\in I}|\theta_{ts}| \equiv b_{\mathrm{E}}. \qquad (18)$$

As a result, $|D''(\eta_s)|$ and $|D'''(\eta_s)|$ are bounded by $\kappa_2 = b_{\mathrm{E}}^{-2}$ and $\kappa_3 = 2b_{\mathrm{E}}^{-3}$, respectively. For fixed B_1 and B_2, we have $\delta_2 = b_{\mathrm{E}}^2 B_2$ and $\delta_1 = B_1 b_{\mathrm{E}}^3/2$. By Theorem 1, a lower bound for the probability of successful neighbourhood recovery is

$$\mathrm{pr}\{\hat{N}(x_s)=N(x_s)\} \geq 1 - c_1 p^{-b_{\mathrm{E}}^3 B_1/2+2} - \exp(-c_2 b_{\mathrm{E}}^4 B_2^2 n) - \exp(-c_3 n/B_2). \qquad (19)$$

Examples 5–8 reveal that the neighbourhood of a Gaussian node is easier to recover than the neighbourhoods of the other three types of nodes: the former requires a smaller minimum sample size when p is large, allows for a wider range of feasible tuning parameters, and has in general a higher probability of success. As a result, the neighbourhood of the Gaussian node should be used when estimating an edge between a Gaussian node and a non-Gaussian node.

Which neighbourhood should we use to estimate an edge between two non-Gaussian nodes? There are no clear winners: while (14) can be evaluated given knowledge of c_1, c_2 and c_3, (16) and (19) also require knowledge of the unknown quantities b_{E} and b_{P}, which are functions of unknown quantities Θ_s and α_{1s} in (15) and (18). One possibility is to substitute a consistent estimator for these parameters (see, e.g., Bunea, 2008; van de Geer, 2008) to obtain a consistent estimator for b_{P} or b_{E}. This leads to the following lemma.

LEMMA 2—*Suppose that* $\tilde{\Theta}_s$ *and* $\tilde{\alpha}_{1s}$ *are consistent estimators of the true parameters in the conditional densities* (6) *and* (7). *Let I be the index set of the Bernoulli nodes.*

i. *If* x_s *is a Poisson node and* $\tilde{b}_{\mathrm{P}} = \exp(\tilde{\alpha}_{1s} + \sum_{t\in I}|\tilde{\theta}_{ts}|)$, *then*

$$1 - c_1 p^{-B_1/\tilde{b}_{\mathrm{P}}+2} - \exp(-c_2 B_2^2 n/\tilde{b}_{p}^{2}) - \exp(-c_3 n/B_2) \qquad (20)$$

is a consistent estimator of a lower bound for $\mathrm{pr}\{\hat{N}(x_s) = N(x_s)\}$.

ii. *If* x_s *is an exponential node and* $\tilde{b}_{\mathrm{E}} = |\alpha_{1s}| - \sum_{t\in I}|\tilde{\theta}_{ts}|$, *then*

$$1 - c_1 p^{-\tilde{b}_{\mathrm{E}}^3 B_1/2+2} - \exp(-c_2 \tilde{b}_{\mathrm{E}}^4 B_2^2 n) - \exp(-c_3 n/B_2) \qquad (21)$$

is a consistent estimator of a lower bound for $\mathrm{pr}\{\hat{N}(x_s) = N(x_s)\}$.

Therefore, by substituting consistent estimators of Θ_s and α_{1s} into (15) or (18), we can reconstruct an edge by choosing the estimate with the highest probability of correct recovery according to (14), (20) and (21). The rules are summarized in Table 2. The results in this section illustrate a worst-case scenario for recovery of each neighbourhood, in that Theorem 1 provides a lower bound for the probability of successful neighbourhood recovery.

5. Recovery in partially specified models

In § 4, we showed that the neighbourhood selection approach of § 3·1 can recover the true graph when each node's conditional distribution is of the form (3), provided that the conditions for strong compatibility are satisfied. In this section, we consider a partially specified model in which some of the nodes are assumed to have conditional distributions of the form (3), and we make no assumption on the conditional distributions of the remaining nodes. We will show that in this setting, neighbourhoods of the nodes with conditional distributions of the form (3) can still be recovered.

Here the neighbourhood of x_s is defined based on its conditional density, (3), as $N^0(x_s) = \{t : \theta_{ts} \neq 0\}$. Assumption 4 in § 4·1 is inappropriate since we no longer assume that all p nodes have conditional densities of the form (3), and consequently we are not assuming a particular form for the joint density. Therefore, we make the following assumption to replace Propositions 2 and 3.

Assumption 8

Let (i) $\text{pr}(\xi_1) \geq 1 - c_1 p^{-\delta_1+2}$; (ii) $\text{pr}(\xi_2) \geq 1 - \exp(-c_2\delta_2^2 n)$.

Theorem 2

Suppose that the sth node has conditional density (3) *and that Assumptions* 1–3 *and* 5–8 *hold. Then, with probability at least* $1 - c_1 p^{-\delta_1+2} - \exp(-c_2\delta_2^2 n) - \exp(-c_3\delta_3 n)$, *for some constants* c_1, c_2, c_3 *and* $\delta_3 = 1/(\kappa_2\delta_2)$, *the estimator from* (10) *recovers the true neighbourhood of* x_s *exactly, so that* $\hat{N}(x_s) = N^0(x_s)$.

The proof of Theorem 2 is similar to that of Theorem 1, and is thus omitted. Theorem 2 indicates that our neighbourhood selection approach can recover the neighbourhood of any node for which the conditional density is of the form (3), provided that Assumption 8 holds. This means that in order to recover an edge between two nodes using our neighbourhood selection approach, it suffices for one of the two nodes to have conditional density of the form (3). Consequently, we can model relationships that are far more flexible than those outlined in Table 1, such as an edge between a Poisson node and a node that takes values on the whole real line.

Although Theorem 2 allows us to go beyond some of the restrictions in Table 1, it is still restricted in that it only guarantees recovery of an edge between two nodes for which at least one of the node-conditional densities is exactly of the form (3). In future work, we could generalize Theorem 2 to the case where (3) is simply an approximation to the true node-conditional distribution.

6. Numerical studies

6·1. Data generation

We consider mixed graphical models with two types of nodes and $m = p/2$ nodes per type, for Gaussian-Bernoulli and Poisson-Bernoulli models. We order the nodes so that the Gaussian or Poisson nodes precede the Bernoulli nodes.

For both models, we construct a graph in which the jth node ($j = 1, \ldots, m$) is connected with the adjacent nodes of the same type, as well as the $(m + j)$th node of the other type, as shown in Fig. 1. This encodes the edge set E. For $(i, j) \in E$ and $i < j$, we generate the edge potentials θ_{ij} and θ_{ji} as

$$\theta_{ij}=\theta_{ji}=y_{ij}r_{ij},\ \mathrm{pr}(y_{ij}=1)=\mathrm{pr}(y_{ij}=-1)=0\cdot 5,\ r_{ij}\sim \mathrm{Un}(a,b). \quad (22)$$

We set $\theta_{ij} = \theta_{ji} = 0$ if $(i, j) \notin E$. Additional steps to ensure strong compatibility of the conditional distributions are discussed in the Supplementary Material. The values of a and b in (22), as well as the parameters of $f_s(x_s)$ in the conditional density (3), are specified in § § 6·2–6·4.

To sample from the joint density $p(x)$ in (2) without calculating the log-partition function A, we employ a Gibbs sampler as in Lee & Hastie (2015). Briefly, we iterate through the nodes and sample from each node's conditional distribution. To ensure independence, after a burn-in period of 3000 iterations, we select samples from the chain which are 500 iterations apart from each other.

6·2. Probability of successful neighbourhood recovery

In § 4·1 we saw that the probability of successful neighbourhood recovery for neighbourhood selection converges to unity exponentially fast with the sample size, and in § 4·2 we saw that the estimates from the Gaussian nodes are superior to those from the Bernoulli nodes, in the sense that a smaller sample size is needed in order to achieve a given probability of successful recovery. We now verify these findings empirically. Here, successful neighbourhood recovery is defined to mean that the estimated and true edge sets of a graph or a subgraph are identical.

We set $a = b = 0·3$ in (22) so that Assumption 5 is satisfied, and generate one Gaussian-Bernoulli graph for each of $p = 60$, 120 and 240. We set $\alpha_{1s} = 0$ and $\alpha_{2s} = -1$ in (4) for Gaussian nodes, and set $\alpha_{1s} = 0$ in (5) for Bernoulli nodes. For each graph, 100 independent datasets are drawn from the Gibbs sampler. We perform neighbourhood selection using the estimator from (11), with the tuning parameter λ_n set to be a constant c times $\{\log(p)/n\}^{1/2}$, so that it is on the scale required by Assumption 6, as illustrated in Remark 3.

In order to achieve successful neighbourhood recovery as the sample size increases, the value of c must be in a range matching the requirement of Assumption 6. We explored a range of values of c, and in Fig. 2 we show the probability of successful neighbourhood recovery for $c = 2·6$. For ease of viewing, we display separate empirical probability curves for the Gaussian-Gaussian, Bernoulli-Bernoulli and Bernoulli-Gaussian subgraphs. Figs. 2(a) and (b) show estimates obtained by regressing the Gaussian nodes onto the others, while panels (c) and (d) are the estimates from regressing the Bernoulli nodes onto the others. We see that the probability of successful recovery increases to unity once the scaled sample size exceeds the threshold required in Assumption 7 and Corollary 1. Furthermore, Figs. 2(b) and (c) agree with the conclusions of § 4·2: neighbourhood recovery using regression of a Gaussian node onto the others requires fewer samples than does recovery using regression of a Bernoulli node onto the others.

6·3. Comparison with competing approaches

In this section, we compare the proposed method with alternative approaches on a Gaussian-Bernoulli graph. We limit the number of nodes to $p = 40$ in order to facilitate comparison with the computationally intensive approach of Lee & Hastie (2015). We generate 100 random graphs with $a = 0{\cdot}3$ and $b = 0{\cdot}6$ in (22), and we set $\alpha_{1s} = 0$ and $\alpha_{2s} = -1$ in (4) for Gaussian nodes and $\alpha_{1s} = 0$ in (5) for Bernoulli nodes. Twenty independent samples of $n =$ 200 observations are generated from each graph. We evaluate the performance of each approach by computing the number of correctly estimated edges as a function of the number of estimated edges in the graph. Results are averaged over 20 datasets from each of 100 random graphs, for a total of 2000 simulated datasets.

Seven approaches are compared in this study: (i) our proposed method for neighbourhood selection in the mixed graphical model; (ii) penalized maximum likelihood estimation in the mixed graphical model (Lee et al., 2013); (iii) weighted ℓ_1-penalized regression in the mixed graphical model, as proposed by Cheng et al. in their unpublished technical report (arXiv: 1304.2810); (iv) graphical random forests (Fellinghauer et al., 2013); (v) neighbourhood selection in the Gaussian graphical model (Meinshausen & Bühlmann, 2006), where we use an ℓ_1-penalized linear regression to estimate the neighbourhoods of all nodes; (vi) the graphical lasso (Friedman et al., 2008), which treats all features as Gaussian; and (vii) neighbourhood selection in the Ising model (Ravikumar et al., 2010), where we use ℓ_1-penalized logistic regression on all nodes after dichotomizing the Gaussian nodes by their means. The first four methods are designed for mixed graphical models, with the methods of Lee & Hastie (2015) and Cheng et al. proposed specifically for Gaussian-Bernoulli networks. In contrast, the last three methods ignore the presence of mixed node types. For methods based on neighbourhood selection, we use the union rule of Meinshausen & Bühlmann (2006) to reconstruct the edge set from the estimated neighbourhoods, with one exception: to estimate the Gaussian-Bernoulli edges for our proposed method, we use the estimates from the Gaussian nodes, as suggested by the theory developed in § 4·2.

Owing to its high computational cost, th emethod of Lee & Hastie (2015) is run on 250 datasets from 50 graphs rather than 2000 datasets from 100 graphs.

Figure 3(a) displays results for Bernoulli-Bernoulli and Gaussian-Gaussian edges, while Fig. 3(b) displays results for edges between Gaussian and Bernoulli nodes.

The curves in Fig. 3 correspond to the estimated graphs as the tuning parameter for each method is varied. Recall from § 3·2 that our proposed method involves a tuning parameter

λ_n^{G} for the ℓ_1-penalized linear regressions of the Gaussian nodes onto the others, as well as a tuning parameter λ_n^{B} for the ℓ_1-penalized logistic regressions of the Bernoulli nodes onto the others. The triangle in each panel of Fig. 3 shows the average performance of our proposed method with the tuning parameters $\hat{\lambda}_n^{\mathrm{B}}$ and $\hat{\lambda}_n^{\mathrm{G}}$ selected using BIC summed over the Bernoulli and Gaussian nodes, respectively, as described in § 3·2. This choice yields good precision (52%) and recall (95%) for edge recovery in the graph. To obtain the curves in Fig. 3, we set $\lambda_n^{\mathrm{B}}=(\hat{\lambda}_n^{\mathrm{B}}/\hat{\lambda}_n^{\mathrm{G}})\lambda_n^{\mathrm{G}}$ and varied the value of λ_n^{G}.

In general, our proposed method outperforms the competitors, which is expected since it assumes the correct model. Although the approaches of Lee & Hastie (2015) and Cheng et al. are intended for a Gaussian-Bernoulli graph, they attempt to capture more complicated relationships than in (2), and so they perform worse than our method. On the other hand, the graphical random forest of Fellinghauer et al. (2013) performs reasonably well, despite the fact that it is a nonparametric approach. Neighbourhood selection in the Gaussian graphical model performs closest to the proposed method in terms of edge selection. The Ising model suffers substantially due to dichotomization of the Gaussian variables. The graphical lasso algorithm experiences serious violations to its multivariate Gaussian assumption, leading to poor performance.

6·4. Application of selection rules to mixed graphical models

In § 6·3, in keeping with the results of § 4·2, we always used the estimates from the Gaussian nodes in estimating an edge between a Bernoulli node and a Gaussian node. Here we consider a mixed graphical model of Poisson and Bernoulli nodes. In this case, the selection rules in § 4·2 are more complex, and whether it is better to use a Poisson node or a Bernoulli node to estimate a Bernoulli-Poisson edge depends on the true parameter values in Table 2.

We generate a graph with $p = 80$ nodes as follows: we take $a = 0{\cdot}8$ and $b = 1$ in (22), $\alpha_{1s} = -3$ for $s = 1, \ldots, 20$ and $\alpha_{1s} = 0$ for $s = 21, \ldots, 40$ for the Poisson nodes, and $\alpha_{1s} = 0$ for the Bernoulli nodes. This guarantees that b_{P} in (15) is smaller than 1 for the first half of the Poisson nodes, and is larger than 2 for the second half, due to the structure of the graph in Fig. 1. In order to estimate a Bernoulli-Poisson edge, we will use the estimates from the Poisson nodes if $b_{\mathrm{P}} < 1$ and the estimates from the Bernoulli nodes if $b_{\mathrm{P}} > 2$, according to the selection rules in Table 2.

We compare the performances of: our proposed approach using the selection rules in Table 2, with the true or estimated parameters; our proposed approach using the union and intersection rules presented in § 4·2; and the graphical random forest method of Fellinghauer et al. (2013). To prevent overshrinkage of the parameters for estimation of b_{P} in (15), we set λ_n in (10) to equal 0·5 times the value from the Bayesian information criterion for each node type. We present only the results for Poisson-Bernoulli edges, as the selection rules in § 4·2 apply to edges between nodes of different types.

The results are shown in Fig. 4, averaged over 20 samples from each of 25 random graphs. The selection rules proposed in § 4·2 clearly outperform the commonly used union and intersection rules. The curve for the selection rule from § 4·2 using the estimated parameter values is almost identical to the curve using the true parameter values, which indicates that in this case the quantity b_{P} is accurately estimated for each node. The graphical random forest approach of Fellinghauer et al. (2013) slightly outperforms our proposed method when few edges are estimated, but performs worse when the estimated graph includes more edges. This may indicate that as the graph becomes less sparse, the nonparametric graphical random forest approach suffers from insufficient sample size.

7. Discussion

In § 2·2 we saw that a stringent set of restrictions is required for compatibility or strong compatibility of the node-conditional distributions given in (4)–(7). These restrictions limit the theoretical flexibility of the conditionally specified mixed graphical model, especially when modelling unbounded variables. It is possible that by truncating unbounded variables, we may be able to circumvent some of these restrictions. Furthermore, the model (2) assumes pairwise interactions in the form of $x_s x_t$, which can be seen as a second-order approximation of the true edge potentials in (1). We can relax this assumption by fitting nonlinear edge potentials using semiparametric penalized regressions, as in Voorman et al. (2014).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank two reviewers and an associate editor for helpful comments, as well as Jie Cheng, Bernd Fellinghauer and Jason Lee for providing code and responding to our inquiries. Witten and Shojaie were partially supported by the U.S. National Science Foundation and National Institutes of Health. Witten also acknowledges support from a Sloan Fellowship.

References

Allen, GI.; Liu, Z. Proc. IEEE Int. Conf. Bioinfo. Biomed. 2012. New York: Curran Associates; 2012. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data; p. 1-6.

Besag JE. Spatial interaction and the statistical analysis of lattice systems (with Discussion). J. R. Statist. Soc. B. 1974; 36:192–236.

Bunea F. Honest variable selection in linear and logistic regression models via ℓ_1 and + ℓ_2 penalization. Electron. J. Statist. 2008; 2:1153–1194.

Fellinghauer B, Bühlmann P, Ryffel M, von Rhein M, Reinhardt JD. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. Comp. Statist. Data Anal. 2013; 64:132–142.

Finegold M, Drton M. Robust graphical modeling of gene networks using classical and alternative *t*-distributions. Ann. Appl. Statist. 2011; 5:1057–1080.

Friedman JH, Hastie TJ, Tibshirani RJ. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]

Friedman JH, Hastie TJ, Tibshirani RJ. Regularization paths for generalized linear models via coordinate descent. J. Statist. Software. 2010; 33:1–22.

Höfling H, Tibshirani RJ. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. J. Mach. Learn. Res. 2009; 10:883–906. [PubMed: 21857799]

Jalali, A.; Ravikumar, PK.; Vasuki, V.; Sanghavi, S. Proc. 13th Int. Conf. Artif. Intel. Statist., 2010. New York: Association for Computing Machinery; 2011. On learning discrete graphical models using group-sparse regularization; p. 378-387.

Lauritzen, SL. Graphical Models. New York: Oxford University Press; 1996.

Lee JD, Hastie TJ. Learning the structure of mixed graphical models. J. Comp. Graph. Statist. 2015 to appear.

Lee, JD.; Sun, Y.; Taylor, JE. On model selection consistency of penalized M-estimators: A geometric theory. In: Burges, CJC.; Bottou, L.; Welling, M.; Ghahramani, Z.; Weinberger, KQ., editors. Adv. Neural Info. Proces. Syst. Vol. 26. New York: Curran Associates; 2013. p. 342-350.

Lee, S-I.; Ganapathi, V.; Koller, D. Efficient structure learning of Markov networks usingℓ_1-regularization. In: Schölkopf, B.; Platt, JC.; Hoffman, T., editors. Adv. Neural Info. Proces. Syst. Vol. 19. Cambridge, Massachusetts: MIT Press; 2007. p. 817-824.

Liu H, Lafferty JD, Wasserman LA. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. J. Mach. Learn. Res. 2009; 10:2295–2328.

Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Ann. Statist. 2006; 34:1436–1462.

Miyamura M, Kano Y. Robust Gaussian graphical modeling. J. Mult. Anal. 2006; 97:1525–1550.

Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. J. Am. Statist. Assoc. 2009; 104:735–746.

Ravikumar PK, Wainwright MJ, Lafferty JD. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. Ann. Statist. 2010; 38:1287–1319.

Ravikumar PK, Wainwright MJ, Raskutti G, Yu B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electron. J. Statist. 2011; 5:935–980.

Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electron. J. Statist. 2008; 2:494–515.

Sun H, Li H. Robust Gaussian graphical modeling via ℓ_1 penalization. Biometrics. 2012; 68:1197–1206. [PubMed: 23020775]

van de Geer SA. High-dimensional generalized linear models and the lasso. Ann. Statist. 2008; 36:614–645.

Vogel D, Fried R. Elliptical graphical modelling. Biometrika. 2011; 98:935–951.

Voorman AL, Shojaie A, Witten DM. Graph estimation with joint additive models. Biometrika. 2014; 101:85–101. [PubMed: 25013234]

Wainwright MJ. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). IEEE Trans. Info. Theory. 2009; 55:2183–2202.

Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Foundat. Trends Mach. Learn. 2008; 1:1–305.

Wainwright, MJ.; Lafferty, JD.; Ravikumar, PK. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In: Schölkopf, B.; Platt, JC.; Hoffman, T., editors. Adv. Neural Info. Proces. Syst. Vol. 19. Cambridge, Massachusetts: MIT Press; 2007. p. 1465-1472.

Wang YJ, Ip EH. Conditionally specified continuous distributions. Biometrika. 2008; 95:735–746.

Xue L, Zou H. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. Ann. Statist. 2012; 40:2541–2571.

Yang, E.; Allen, GI.; Liu, Z.; Ravikumar, PK. Graphical models via generalized linear models. In: Bartlett, P.; Pereira, F.; Burges, C.; Bottou, L.; Weinberger, K., editors. Adv. Neural Info. Proces. Syst. Vol. 25. New York: Curran Associates; 2012. p. 1367-1375.

Yang E, Baker Y, Ravikumar PK, Allen GI, Liu Z. Mixed graphical models via exponential families. Proc. 17th Int. Conf. Artif. Intel. Statist., 2014. JMLR Workshop and Conference Proceedings. 2014; 33:1042–1050.

Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007; 94:19–35.

Zhao P, Yu B. On model selection consistency of lasso. J.Mach. Learn. Res. 2006; 7:2541–2563.

Zou H, Hastie TJ, Tibshirani RJ. On the "degrees of freedom" of the lasso. Ann. Statist. 2007; 35:2173–2192.

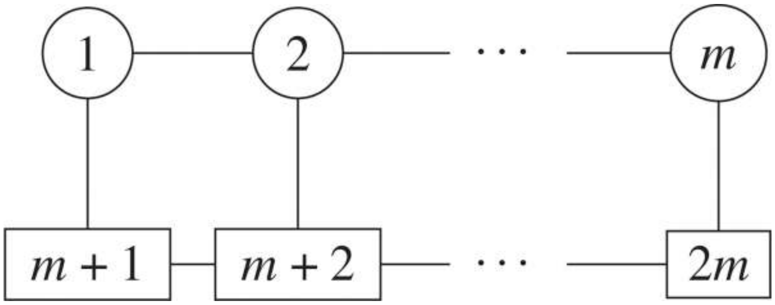

Fig. 1.

The graph used to generate the data in § § 6·2–6·4, consisting of $m = p/2$ Gaussian or Poisson nodes, shown as circles, and $m = p/2$ Bernoulli nodes, shown as rectangles.

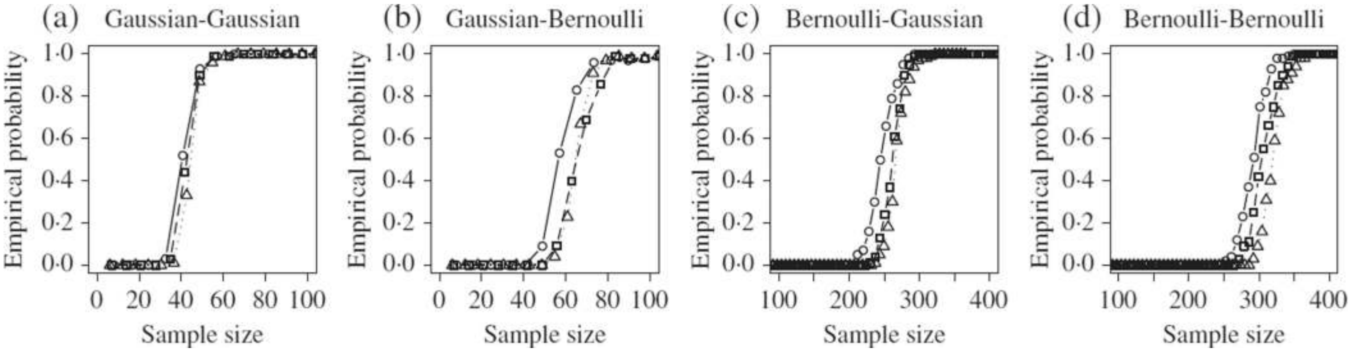

Fig. 2.

Probability of successful neighbourhood recovery plotted as a function of scaled sample size $n/\{3 \log(p)\}$, for the set-up of § 6·2. The curves are empirical probabilities of successful neighbourhood recovery for graphs with 60 (∘—∘), 120 (□- -□) or 240 nodes (△···△), averaged over 100 independent datasets. The tuning parameter is set to $2{\cdot}6\{\log(p)/n\}^{1/2}$. The title above each panel indicates the subgraph for which the recovery probability is displayed, and the first word in the title indicates the node type that was regressed in order to obtain the subgraph estimate. For instance, panel (b) displays probability curves for edges between Gaussian and Bernoulli nodes that are estimated from the ℓ_1-penalized linear regression of Gaussian nodes; panel (c) displays the same quantity, but estimated via an ℓ_1-penalized logistic regression of the Bernoulli nodes.

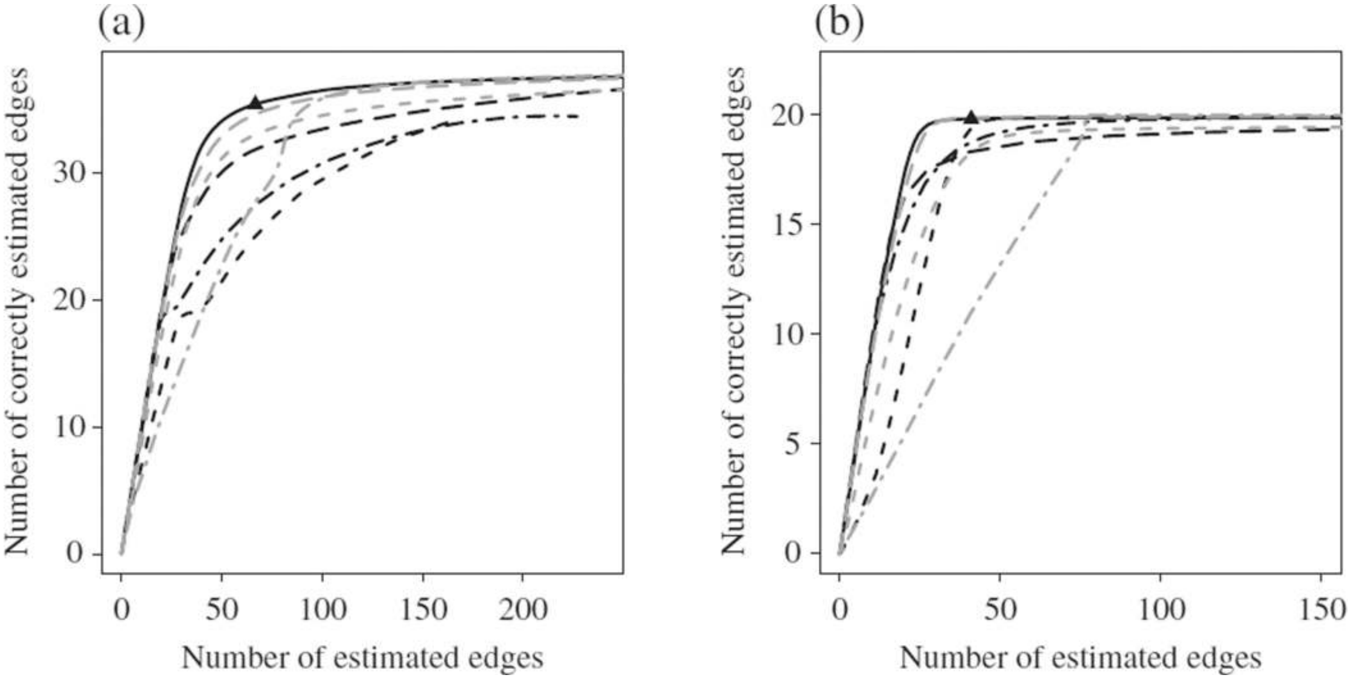

Fig. 3.

Simulation results for the Gaussian-Bernoulli graph, as described in § 6·3. The number of correctly estimated edges is displayed as a function of the number of estimated edges, for a range of tuning parameter values in a graph with $p = 40$ and $n = 200$: (a) edges between nodes of the same type, Bernoulli-Bernoulli or Gaussian-Gaussian; (b) edges between Gaussian and Bernoulli nodes. In each panel the different curves represent the methods of the present paper (solid), Lee & Hastie (2015) (short-dashed), Cheng et al. (long-dashed), Fellinghauer et al. (2013) (dot-dashed), neighbourhood selection in the Gaussian graphical model (grey long-dashed), neighbourhood selection in the Ising model (grey short-dashed), and the graphical lasso (grey dot-dashed). The black triangle shows the average performance of our proposed method with tuning parameter selected by the Bayesian information criterion (see § 3·2).

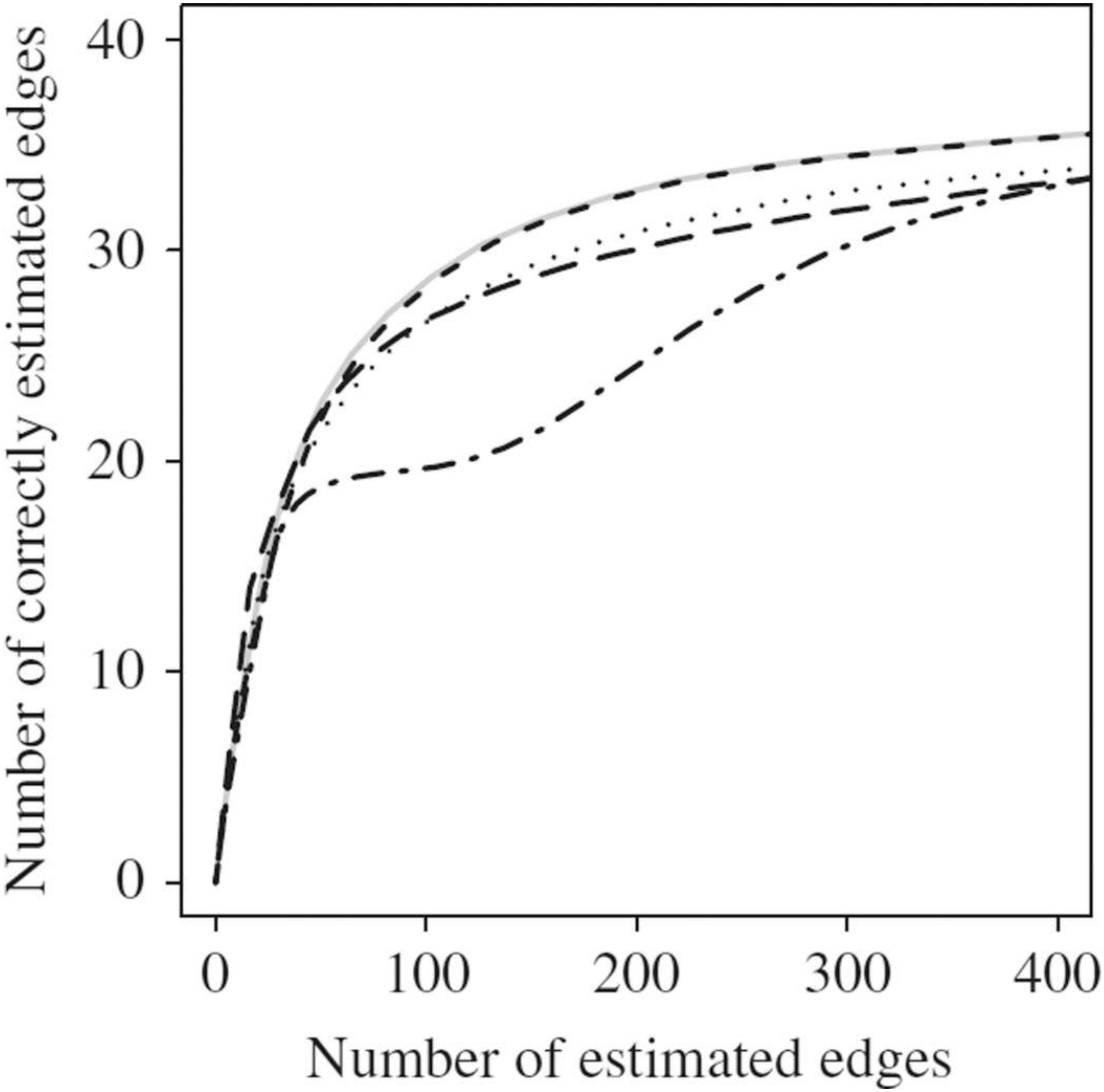

Fig. 4.

Summary of the simulation results for the Poisson-Bernoulli graph, as described in § 6·4. The number of correctly estimated edges is displayed as a function of the number of estimated edges, for a range of tuning parameter values in a graph with $p = 80$ nodes from $n = 200$ observations. The different curves represent the selection rule from § 4·2 with the true parameters (grey solid), the selection rule from § 4·2 with estimated parameters (short-dashed), the union rule (dot-dashed), the intersection rule (dotted), and the graphical random forest method of Fellinghauer et al. (2013) (long-dashed).

Table 1

Restrictions on the parameter space required for compatibility or strong compatibility of the conditional densities in (4)–(7)

	Gaussian	**Poisson**	**Exponential**	**Bernoulli**
Gaussian	$\Theta_{JJ} \prec 0$	$\theta_{ts} = 0$	$\theta_{ts} = 0^{\dagger}$	$\theta_{ts} \in \mathbb{R}^{\dagger}$
Poisson		$\theta_{ts} \leq 0$	$\theta_{ts} \leq 0^{\dagger}$	$\theta_{ts} \in \mathbb{R}^{\dagger}$
Exponential			$\theta_{ts} \leq 0^{\dagger}$	$\sum_{s\in I}\lvert\theta_{st}\rvert < -\alpha_{1t}^{\dagger}$
Bernoulli				$\theta_{ts} \in \mathbb{R}^{\dagger}$

Each column specifies the type of the sth node, and each row specifies the type of the *t*th node; conditions marked with a dagger, †, are necessary and sufficient for the conditional densities in (4)–(7) to be compatible, and the complete set of conditions is necessary and sufficient for the conditional densities to be strongly compatible. For compatibility to hold for a Gaussian node x_S, $\alpha_{2S} < 0$ is also required. Here Θ_{JJ} is as defined in (9), and I denotes the set of Bernoulli nodes.

Table 2

Neighbourhood to use in estimating an edge between two non-Gaussian nodes of different types; when the conditions in this table are not met, there is no clear preference in terms of which neighbourhood to use

Pair of nodes	**Selection rule**
Poisson & exponential	Choose Poisson if $\tilde{b}_{\mathrm{E}}^{2}\tilde{b}_{\mathrm{P}} < 1$ and $\tilde{b}_{\mathrm{E}}^{3}\tilde{b}_{\mathrm{P}} < 2$. Choose exponential if $\tilde{b}_{\mathrm{E}}^{2}\tilde{b}_{\mathrm{P}} > 1$ and $\tilde{b}_{\mathrm{E}}^{3}\tilde{b}_{\mathrm{P}} > 2$.
Poisson & Bernoulli	Choose Poisson if $\tilde{b}_{\mathrm{P}} < 1$. Choose Bernoulli if $\tilde{b}_{\mathrm{P}} > 2$.
Exponential & Bernoulli	Choose exponential if $\tilde{b}_{\mathrm{E}} \geq 1$. Choose Bernoulli if $\tilde{b}_{\mathrm{E}} < 1$.