Published in final edited form as:
Biometrika. 2014 March 1; 101(1): 85–101. doi:10.1093/biomet/ast053.

Graph Estimation with Joint Additive Models

Arend Voorman, **Ali Shojaie**, and **Daniela Witten**
Department of Biostatistics, University of Washington, Seattle, Washington, 98195-7232, U.S.A.

Abstract

In recent years, there has been considerable interest in estimating conditional independence graphs in high dimensions. Most previous work has assumed that the variables are multivariate Gaussian, or that the conditional means of the variables are linear; in fact, these two assumptions are nearly equivalent. Unfortunately, if these assumptions are violated, the resulting conditional independence estimates can be inaccurate. We propose a semi-parametric method, graph estimation with joint additive models, which allows the conditional means of the features to take on an arbitrary additive form. We present an efficient algorithm for our estimator's computation, and prove that it is consistent. We extend our method to estimation of directed graphs with known causal ordering. Using simulated data, we show that our method performs better than existing methods when there are non-linear relationships among the features, and is comparable to methods that assume multivariate normality when the conditional means are linear. We illustrate our method on a cell-signaling data set.

Keywords

Conditional independence; Graphical model; Lasso; Nonlinearity; Non-Gaussianity; Sparse additive model; Sparsity

1. Introduction

In recent years, there has been considerable interest in developing methods to estimate the joint pattern of association within a set of random variables. The relationships between d random variables can be summarized with an undirected graph $\Gamma = (V, S)$ in which the random variables are represented by the vertices $V = \{1, \ldots, d\}$ and the conditional dependencies between pairs of variables are represented by edges $S \subset V \times V$. That is, for each $j \in V$, we want to determine a minimal set of variables on which the conditional densities $p_j(\{X_j \mid \{X_k, k \neq j\})$ depend,

$$p_j\left(X_j \middle| \{X_k, k \neq j\}\right) = p_j\left(X_j \middle| \{X_k : (k, j) \in S\}\right).$$

There has also been considerable work in estimating marginal associations between a set of random variables (see e.g. Basso et al., 2005; Meyer et al., 2008; Liang & Wang, 2008;

voorma@uw.edu ashojaie@uw.edu dwitten@uw.edu

Hausser & Strimmer, 2009; Chen et al., 2010). But in this paper we focus on conditional dependencies, which, unlike marginal dependencies, cannot be explained by the other variables measured.

Estimating the conditional independence graph Γ based on a set of *n* observations is an old problem (Dempster, 1972). In the case of high-dimensional continuous data, most previous work has assumed either multivariate Gaussianity (see e.g. Friedman et al., 2008; Rothman et al., 2008; Yuan & Lin, 2007; Banerjee et al., 2008) or linear conditional means (see e.g. Meinshausen & Bühlmann, 2006; Peng et al., 2009) for the features. However, as we will see, these two assumptions are essentially equivalent. Some recently proposed methods relax the multivariate Gaussian assumption using univariate transformations (Liu et al., 2009, 2012; Xue & Zou, 2012; Dobra & Lenkoski, 2011), restrictions on the graph structure (Liu et al., 2011), or flexible random forests (Fellinghauer et al., 2013). However, we will see that these methods may not capture realistic departures from multivariate Gaussianity.

For illustration, consider the cell signaling data set from Sachs et al. (2005), which consists of protein concentrations measured under a set of perturbations. We analyze the data in more detail in Section 5·3. Pairwise scatterplots of three of the variables are given in Fig. 1 (a)-(c) for one of 14 experiments. The data have been transformed to be marginally normal, as suggested by Liu et al. (2009), but the transformed data clearly are not multivariate normal, as confirmed by a Shapiro–Wilk test, which yields a p-value less than 2×10^{-16}.

Can the data in Fig. 1 be well-represented by linear relationships? In Fig. 1 (d), we see strong evidence that the conditional mean of the protein P38 given PKC and PJNK is nonlinear. This is corroborated by the fact that the p-value for including quadratic terms in the linear regression of P38 onto PKC and PJNK is less than 2×10^{-16}. Therefore in this data set, the features are not multivariate Gaussian, and marginal transformations do not remedy the problem.

In order to model this type of data, we could specify a more flexible joint distribution. However, joint distributions are difficult to construct and computationally challenging to fit, and the resulting conditional models need not be easy to obtain or interpret. Alternatively, we can specify the conditional distributions directly; this has the advantage of simpler interpretation and greater computational tractability. In this paper, we will model the conditional means of non-Gaussian random variables with generalized additive models (Hastie & Tibshirani, 1990), and will use these in order to construct conditional independence graphs.

Throughout this paper, we will assume that we are given *n* independent and identically distributed observations from a *d*-dimensional random vector $X = (X_1, \ldots, X_d) \sim \mathscr{P}$. Our observed data can be written as $x = (x_1, \ldots, x_d) \in \mathbb{R}^{n \times d}$.

2. MODELING CONDITIONAL DEPENDENCE RELATIONSHIPS

Suppose we are interested in estimating the conditional independence graph for a random Γ vector $X \in \mathbb{R}^d$. If the joint distribution is known up to some parameter θ, it suffices to estimate θ via e.g. maximum likelihood. One practical difficulty is specification of a

plausible joint distribution. Specifying a conditional distribution, such as in a regression model, is typically much less daunting. We therefore consider pseudo-likelihoods (Besag, 1974, 1975) of the form

$$log\left\{p_{_{PL}}\left(x;\theta\right)\right\}=\sum_{j=1}^{d}log\left\{p_{j}\left(x_{j}|\left\{x_{k}:\left(j,k\right)\in S\right\};\theta\right)\right\}.$$

For a set of arbitrary conditional distributions, there need not be a compatible joint distribution (Wang & Ip, 2008). However, the conditionally specified graphical model has an appealing the-oretical justification, in that it minimizes the Kullback–Leibler distances to the conditional distributions (Varin & Vidoni, 2005). Furthermore, in estimating conditional independence graphs, our scientific interest is in the conditional independence relationships rather than in the joint distribution. So in a sense, modeling the conditional distribution rather than the joint distribution is a more direct approach to graph estimation. We therefore advocate an approach for non-Gaussian graphical modeling based on conditionally specified models (Varin et al., 2011).

3. PREVIOUS WORK

3·1. Estimating graphs with Gaussian data

Suppose for now that X has a joint Gaussian distribution with mean 0 and precision matrix Θ. The negative log-likelihood function evaluated at Θ, up to constants, is

$$-log\,det\left(\Theta\right)+tr\left(x^{T}x\Theta\right)/n. \quad (1)$$

In this case, the conditional relationships are linear,

$$X_{j}|\left\{X_{k},k\neq j\right\}=\sum_{k\neq j}\beta_{jk}X_{k}+\epsilon_{j}\quad\left(j=1,\ldots,d\right), \quad (2)$$

where $\beta_{jk} = -\Theta_{jk}/\Theta_{kk}$ and $\in_j \sim N_1(0, 1/\Theta_{jj})$. To estimate the graph Γ, we must determine which β_{jk} are zero in (2), or equivalently, which Θ_{jk} are 0 in (1). This is simple when $n > > d$.

In the high-dimensional setting, when the maximum likelihood estimate is unstable or undefined, a number of approaches to estimate the conditional independence graph Γ have been proposed. Meinshausen & Bühlmann (2006) proposed fitting (2) using an l_1-penalized regression. This is referred to as neighborhood selection:

$$\left\{\hat{\beta}_{jk}:1\leq j,k\leq d\right\}=\underset{\beta_{jk}:1\leq j,k\leq d}{arg\,min}\left\{\frac{1}{2}\sum_{j=1}^{d}\left\|x_{j}-\sum_{k\neq j}x_{k}\beta_{jk}\right\|^{2}+\lambda\sum_{j=1}\sum_{k\neq j}\left|\beta_{jk}\right|\right\}. \quad (3)$$

Here λ is a nonnegative tuning parameter that encourages sparsity in the coefficient estimates. Peng et al. (2009) improved upon the neighborhood selection approach by applying l_1 penalties to the partial correlations; this is known as sparse partial correlation estimation.

As an alternative to (3), many authors have considered estimating Θ by maximizing an l_1-penalized joint log likelihood (see e.g. Yuan & Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008). This amounts to the optimization problem

$$\hat{\Theta}=\underset{W\succ 0}{arg\,min}\left\{-log\,det\,(W)+tr\left(x^{T}xW\right)/n+\lambda\|W\|_{1}\right\}, \quad (4)$$

known as the graphical lasso. Here, $W \succ 0$ 0 indicates that *W* must be positive definite. The sparsity pattern of the solution $\hat{\Theta}$ to (4) serves as an estimate of Γ.

At first glance, neighborhood selection and sparse partial correlation estimation may seem semi-parametric: a linear model may hold in the absence of multivariate normality. However, while (2) can accurately model each conditional dependence relationship semi-parametrically, the accumulation of these specifications sharply restricts the joint distribution: Khatri & Rao (1976) proved that if (2) holds, along with some mild assumptions, then the joint distribution must be multivariate normal, regardless of the distribution of the errors ε_j in (2). In other words, even though (3) does not explicitly involve the multivariate normal likelihood, normality is implicitly assumed. Thus, if we wish to model non-normal data, non-linear conditional models are necessary.

3·2. Estimating graphs with non-Gaussian data

We now briefly review three existing methods for modeling conditional independence graphs with non-Gaussian data. The normal copula or nonparanormal model (Liu et al. 2009, Liu et al. 2012, Xue & Zou 2012, studied in the Bayesian context by Dobra & Lenkoski 2011) assumes that *X* has a nonparanormal distribution: that is, $\{h_1(X_1), \ldots, h_d(X_d)\} \sim N_d(0, \Theta)$ for monotone functions $h_1, \ldots, h_d$. After these are estimated, one can apply any of the methods mentioned in Section 3·1 to the transformed data. The conditional model implicit in this approach is

$$h_j\,(X_j)\,|\,\{X_k, k \neq j\}=\sum_{k\neq j}\beta_{jk}h_k\,(X_k)+\epsilon_j \quad (j=1,\ldots,d)\,. \quad (5)$$

This restrictive assumption may not hold, as seen in Fig. 1.

Forest density estimation (Liu et al., 2011) replaces the need for distributional assumptions with graphical assumptions: the underlying graph is assumed to be a forest, and bivariate densities are estimated non-parametrically. Unfortunately, the restriction to acyclic graphs may be inappropriate in applications, and maximizing over all possible forests is infeasible.

The graphical random forests (Fellinghauer et al., 2013) approach uses random forests to flexibly model conditional means, and allows for interaction terms. However, random forests do not correspond to a well-defined statistical model or optimization problem, and results on its feature selection consistency are in general unavailable. In contrast, our proposed method corresponds to a statistical model for which we can prove results on edge selection consistency.

4. METHOD

4·1. Jointly additive models

In order to estimate a conditional independence graph using pseudolikelihood, we must estimate the variables on which the conditional distributions $p_j(\cdot)$ depend. However, since density estimation is generally a challenging task, especially in high dimensions, we focus on the simpler problem of estimating the conditional mean $E_{x_j}(X_j \mid \{X_k : (j, k) \in S\})$, under the assumption that the conditional distribution and the conditional mean depend on the same set of variables. Thus, we seek to estimate the conditional mean $f_j(\cdot)$ in the regression model $X_j \mid \{X_k, k \neq j\} = f_j(X_k, k \neq j) + \in_j) + \in_j$ where $\in_j$ is a mean-zero error term. Since estimating arbitrary functions $f_j(\cdot)$ is infeasible in high dimensions, we restrict ourselves to additive models

$$X_j \mid \{X_k, k \neq j\} = \sum_{k \neq j} f_{jk}(X_k) + \epsilon_j, \quad (6)$$

where $f_{jk}(\cdot)$ lies in some space of functions $\mathscr{F}$. This amounts to modeling each variable using a generalized additive model (Hastie & Tibshirani, 1990). Unlike Fellinghauer et al. (2013), we do not assume that the errors $\in_j$ are independent of the additive components $f_{jk}(\cdot)$, but merely that the conditional independence structure can be recovered from the functions $f_{jk}(\cdot)$.

4·2. Estimation

Since we believe that the conditional independence graph is sparse, we fit (6) using a penalty that performs simultaneous estimation and selection of the $f_{jk}(\cdot)$. Specifically, we link together d sparse additive models (Ravikumar et al., 2009) using a penalty that groups the parameters corresponding to a single edge in the graph. This results in the problem

$$\underset{f_{jk} \in \mathscr{F}, 1 \leq j,k \leq d}{minimize} \left[\frac{1}{2n} \sum_{j=1}^{d} \| x_j - \sum_{k \neq j} f_{jk}(x_k) \|_2^2 + \lambda \sum_{k>j} \left\{ \| f_{jk}(x_k) \|_2^2 + \| f_{kj}(x_j) \|_2^2 \right\}^{1/2} \right]. \quad (7)$$

We consider $f_{jk}(x_k) = \mathbf{\Psi}_{jk} \beta_{jk}$, where $\mathbf{\Psi}_{jk}$ is a $n \times r$ matrix whose columns are basis functions used to model the additive components f_{jk}, and β_{jk} is an r-vector containing the associated coefficients. For instance, if we use a linear basis function, $\mathbf{\Psi}_{jk} = x_k$, then $r = 1$ and we model only linear conditional means, as in Meinshausen & Bühlmann (2006). Higher-order terms allow us to model more complex dependencies. The standardized group lasso penalty (Simon & Tibshirani, 2012) encourages sparsity and ensures that the estimates of $f_{jk}(\cdot)$ and $f_{kj}(\cdot)$ will be simultaneously zero or non-zero. Problem (7) is an extension of sparse additive modeling (Ravikumar et al., 2009) to graphs, and generalizes neighborhood selection (Meinshausen & Bühlmann, 2006) and sparse partial correlation (Peng et al., 2009) to allow for flexible conditional means.

Algorithm 1. Given initial values for the $\hat{\beta}_{jk}$, perform Steps 1–3 for $(j, k) \in V \times V$, and repeat until convergence.

Step 1. Calculate the vector of errors for the jth and kth variables:

$$r_{jk} \leftarrow x_j - \sum_{i \neq j,k} \Psi_{ji} \hat{\beta}_{ji}, \quad r_{kj} \leftarrow x_k - \sum_{i \neq j,k} \Psi_{ki} \hat{\beta}_{ki}.$$

Step 2. Regress the errors on the specified basis functions:

$$\hat{\beta}_{jk} \leftarrow \left(\Psi_{jk}^T \Psi_{jk}\right)^{-1} \Psi_{jk}^T r_{jk}, \quad \hat{\beta}_{kj} \leftarrow \left(\Psi_{kj}^T \Psi_{kj}\right)^{-1} \Psi_{kj}^T r_{jk}.$$

Step 3. Threshold:

$$\hat{\beta}_{jk} \leftarrow \left\{1 - n\lambda\left(\|\Psi_{jk}\hat{\beta}_{jk}\|_2^2 + \|\Psi_{kj}\hat{\beta}_{kj}\|_2^2\right)^{-1/2}\right\}_+ \hat{\beta}_{jk},$$
$$\hat{\beta}_{kj} \leftarrow \left\{1 - n\lambda\left(\|\Psi_{jk}\hat{\beta}_{jk}\|_2^2 + \|\Psi_{kj}\hat{\beta}_{kj}\|_2^2\right)^{-1/2}\right\}_+ \hat{\beta}_{kj}.$$

Algorithm 1 uses block coordinate descent to achieve the global minimum of the convex problem (7) (Tseng, 2001). Performing Step 2 requires an $r \times r$ matrix inversion; this must be performed only twice per pair of variables. Estimating 30 conditional independence graphs with $r = 3$ on a simulated data set with $n = 50$ and $d = 100$ takes 1.1 seconds on a 2.8 GHz Intel Core i7 Macbook Pro. The R package spacejam, available at cran.r-project.org/package=spacejam, implements the proposed approach.

4·3. Tuning

A number of options for tuning parameter selection are available, such as generalized crossvalidation (Tibshirani, 1996), the Bayesian information criterion (Zou et al., 2007), and stability selection (Meinshausen & Bühlmann, 2010). We take an approach motivated by the Bayesian information criterion, as in Peng et al. (2009). For the *j*th variable, the criterion is

$$BIC_j(\lambda) = n \quad log\left(\left\|x_j - \sum_{k \neq j} \Psi_{jk} \hat{\beta}_{jk}^{(\lambda)}\right\|_2^2\right) + log(n)\, DF_j(\lambda), \quad (8)$$

where $DF_j(\lambda)$ is the degrees of freedom used in this regression. We seek the value of λ that minimizes $\sum_{j=1}^{d} BIC_j(\lambda)$. When a single basis function is used, we can approximate the degrees of freedom by the number of non-zero parameters in the regression (Zou et al., 2007; Peng et al., 2009). But with $r > 1$ basis functions, we use

$$DF_j(\lambda) = |S_j^{(\lambda)}| + (r-1) \sum_k \frac{\|\Psi_{jk}\hat{\beta}_{jk}^{(\lambda)}\|_2^2}{\|\Psi_{jk}\hat{\beta}_{jk}^{(\lambda)}\|_2^2 + \lambda}, \quad (9)$$

where $S_j^{(\lambda)} = \left\{k : \|\hat{\beta}_{jk}^{(\lambda)}\| \neq 0\right\}$. Although (9) was derived under the assumption of an orthogonal design matrix, it is a good approximation for the non-orthogonal case (Yuan & Lin, 2006). Chen & Chen (2008) proposed modifications of the Bayesian information

criterion for high-dimensional regression, which Gao & Song (2010) extended to the psuedo-likelihood setting. We leave evaluation of these alternatives for future work.

In order to perform Algorithm 1, we must select a set of basis functions. Domain knowledge or experience with similar data may guide basis choice. In the absence of domain knowledge we use cubic polynomials, which can approximate a wide range of functions. In Section 5·1, we use several different bases, and find that even misspecified sets of functions can provide more accurate graph estimates than methods that assume linearity.

5. NUMERICAL EXPERIMENTS

5·1. Simulation setup

As discussed in Section 2, it can be difficult to specify flexible non-Gaussian distributions for continuous variables. However, construction of multivariate distributions via conditional distributions is straightforward when the variables can be represented with a directed acyclic graph. The distribution of variables in a directed acyclic graph can be decomposed as $p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p_j(x_j | \{x_k : (k, j) \in S_D\})$, where S_D denotes the directed edge set of the graph. This is a valid joint distribution regardless of the choice of conditional distributions $p_j(x_j \mid \{x_k : (k, j) \in S_D)$ (Pearl, 2000, Chapter 1.4). We chose structural equations of the form

$$X_j | \{X_k : (k, j) \in S_D\} = \sum_{(k,j) \in S_D} f_{jk}(X_k) + \epsilon_j, \quad (10)$$

with $\epsilon_j \sim N(0, 1)$. If the f_{jk} are chosen to be linear, then the data are multivariate normal, and if the f_{jk} are non-linear, then the data will typically not correspond to a well-known multivariate distribution. We moralized the directed graph in order to obtain the conditional independence graph (Cowell et al., 2007, Chapter 3.2). Here we have used directed acyclic graphs simply as a tool to generate non-Gaussian data; the full conditional distributions of the random variables created using this approach are not necessarily additive.

We first generated a directed acyclic graph with $d = 100$ nodes and 80 edges chosen at random from the 4950 possible edges. We used two schemes to construct a distribution on this graph. In the first setting, we chose $f_{jk}(x_k) = b_{jk1}x_k + b_{jk2}x_k^2 + b_{jk3}x_k^3$, where the b_{jk1}, b_{jk2}, and b_{jk3} are independent and normally distributed with mean zero and variance 1, 0.5, and 0.5, respectively. In the second case, we chose $f_{jk}(x_k) = x_k$, resulting in multivariate normal data. In both cases, we scaled the $f_{jk}(x_k)$ to have unit variance, and generated $n = 50$ observations. We compared our method to sparse partial correlation (Peng et al., 2009, R package space), graphical lasso (Yuan & Lin, 2007, R package glasso), neighborhood selection (Meinshausen & Bühlmann, 2006, R package glasso), nonparanormal (Liu et al., 2012, R package huge), forest density estimation (Liu et al., 2011, code provided by authors), the method of Basso et al. (2005, R package minet), and graphical random forests (Fellinghauer et al., 2013, code provided by authors). In performing neighborhood selection, we declared an edge between the jth and kth variables if $\hat{\beta}_{jk} \neq 0$ or $\hat{\beta}_{kj} \neq 0$. We performed

our method using three sets of basis functions: $\Psi_{jk} = \left(x_k, x_k^2\right)$, $\Psi_{jk} = \left(x_k, x_k^3\right)$, $\Psi_{jk} = \left(x_k, x_k^2, x_k^3\right)$.

5·2. Simulation results

Figure 2 summarizes the results of our simulations. For each method, the numbers of correctly and incorrectly estimated edges were averaged over 100 simulated data sets for a range of 100 tuning parameter values. When the $f_{jk}(\cdot)$ are non-linear, our method with the basis $\Psi_{jk} = \left(x_k, x_k^2, x_k^3\right)$ dominates the basis sets $\Psi_{jk} = \left(x_k, x_k^2\right)$ or $\left(x_k, x_k^3\right)$, which in turn tend to enjoy superior performance relative to all other methods, as seen in Fig. 2 (a).

Furthermore, even though the basis sets $\Psi_{jk} = \left(x_k, x_k^2\right)$ and $\left(x_k, x_k^3\right)$ do not entirely capture the functional forms of the data-generating mechanism, they still outperform methods that assume linearity, as well as competitors intended to model non-linear relationships.

When the conditional means are linear and the number of estimated edges is small, all methods perform roughly equally, as seen in Fig. 2 (b). As the number of estimated edges increases, sparse partial correlation performs best, while the graphical lasso, the nonparanormal and the forest-based methods perform worse. This agrees with the observations of Peng et al. (2009) that sparse partial correlation and neighborhood selection tend to outperform the graphical lasso. In this setting, since non-linear terms are not needed to model the conditional dependence relationships, sparse partial correlation outperforms our method with two basis functions, which performs better than our method with three basis functions. Nonetheless, the loss in accuracy due to the inclusion of non-linear basis functions is not dramatic, and our method still tends to outperform other methods for non-Gaussian data, as well as the graphical lasso.

5·3. Application to cell signaling data

We apply our method to a data set consisting of measurements for 11 proteins involved in cell signaling, under 14 different perturbations (Sachs et al., 2005). To begin, we consider data from one of the 14 perturbations with $n = 911$, and compare our method using cubic polynomials to neighborhood selection, the nonparanormal, and graphical random forests with stability selection. Minimizing BIC(λ) for our method yields a graph with 16 edges. We compare our method to competing methods, selecting tuning parameters such that each graph estimate contains 16 edges, as well as 10 and 20 edges, for the sake of comparison. Figure 3 displays the estimated graphs, along with the directed graph presented in Sachs et al. (2005).

The graphs estimated using different methods are qualitatively different. If we treat the directed graph from Sachs et al. (2005) as the ground truth, then our method with 16 edges correctly identifies 12 of the edges, compared to 11, 9, and 8 using sparse partial correlation, the nonparanormal, and graphical random forests, respectively.

Next, we examined the other 13 perturbations, and found that for graphs with 16 edges, our method chooses on average 0.93, 0.64 and 0.2 more correct edges than sparse partial

correlation, nonparanormal, and graphical random forests, respectively, yielding $p = 0.001$, 0.19 and 0.68 using the paired t-test. Since graphical random forests does not permit arbitrary specification of graph size, when graphs with 16 edges could not be obtained, we used the next largest graph, resulting in a larger number of correct edges for their method.

In Section 1, we showed that these data are not well-represented by linear models even after the nonparanormal transformation. The superior performance of our method in this section confirms this observation. The qualitative differences between our method and graphical random forests indicate that the approach taken for modeling non-linearity does affect the results obtained.

6. EXTENSION TO DIRECTED GRAPHS

In certain applications, it can be of interest to estimate the causal relationships underlying a set of features, typically represented as a directed acyclic graph. Though directed acyclic graph estimation is in general NP-hard, it is computationally tractable when a causal ordering is known. In fact, in this case, a modification of neighborhood selection is equivalent to the graphical lasso (Shojaie & Michailidis, 2010b). We extend the penalized likelihood framework of Shojaie & Michailidis (2010b) to non-linear additive models by solving

$$\underset{\beta_{jk},2\leq j\leq p,k\prec j}{minimize}\left\{\frac{1}{2n}\|x_j-\sum_{k\prec j}\Psi_{jk}\beta_{jk}\|_2^2+\lambda\sum_{k\prec j}\|\Psi_{jk}\beta_{jk}\|_2\right\},$$

where $k \prec j$ indicates that k precedes j in the causal ordering. When $\mathbf{\Psi}_{jk} = x_k$, the model is exactly the penalized Gaussian likelihood approach of Shojaie & Michailidis (2010b).

Figure 4 displays the same simulation scenario as Section 5·1, but with the directed graph estimated using the known causal ordering. Results are compared to the penalized Gaussian likelihood approach of Shojaie & Michailidis (2010b). Our method performs best when the true relationships are non-linear, and performs competitively when the relationships are linear.

7. THEORETICAL RESULTS

In this section, we establish consistency of our method for undirected graphs. Similar results hold for directed graphs, but we omit them due to space considerations. The theoretical development follows the sparsistency results for sparse additive models (Ravikumar et al., 2009).

First, we must define the graph for which our method is consistent. Recall that we have the random vector $X=(X_1,\ldots,X_d)\sim\mathscr{P}$, and $x=(x_1,\ldots,x_d)\in\mathbb{R}^{n\times d}$ is a matrix where each row is an independent draw from $\mathscr{P}$. For each $(j,k)\in V\times V$ consider the orthogonal set of basis functions $\psi_{jkt}(\cdot)$, $t\in\mathbb{N}$. Define the population level parameters $\beta^*_{jk}\in\mathbb{R}^\infty$ as

$$\left\{\beta_{jk}^{*}:k=1,\ldots,d\right\}\equiv\underset{\beta_{jk}:k=1,\ldots,d}{arg\,min}\left\{E|X_j-\sum_{k\neq j}\sum_{t-1}^{\infty}\psi_{jkt}\left(X_k\right)\beta_{jkt}|^{2}\right\}\left(j=1,\ldots,d\right).$$

Let $S_j=\left\{k:\|\beta_{jk}^{*}\|\neq 0\right\}$, $s_j = |S_j|$, and $f_{jk}=\sum_{t=1}^{\infty}\psi_{jkt}\beta_{jkt}^{*}\in\mathscr{F}$. Then

$$X_j=\sum_{k\in S_j}f_{jk}\left(X_k\right)+\epsilon_j\quad\left(j=1,\ldots,d\right),$$

where $\in_1, \ldots, \in_d$ are errors, and $\Sigma_{k\in S_j} f_{jk}(X_k)$ is the best additive approximation to $E(X_j \mid X_k : k \neq j)$, in the least-squares sense. We wish to determine which of the $f_{jk}(\cdot)$ are zero.

On observed data, we use a finite set of basis functions to model the $f_{jk}(\cdot)$. Denote the set of r orthogonal basis functions used in the regression of x_j on x_k by $\mathbf{\Psi}_{jk} = \{\boldsymbol{\psi}_{jk1}(x_k),..., \psi_{jkr}(x_k)\}$, a matrix of dimension $n \times r$ such that $\Psi_{jk}^{T}\Psi_{jk}/n=I_r$. Let $\beta_{jk}^{*(r)}=\left(\beta_{jk1}^{*},\ldots,\beta_{jkr}^{*}\right)^{T}$ denote the first r components of β_{jk}^{*}. Further, let $\Psi_{S_j}\in\mathbb{R}^{n\times s_j r}$ be the concatenated basis functions in $\{\mathbf{\Psi}_{jk} : k \in S_j\}$ and β_{Sj} be the corresponding coefficients. Also let $\Sigma_{S_j,S_j}=n^{-1}\Psi_{S_j}^{T}\Psi_{S_j}$, $\Sigma_{jk,S_j}=n^{-1}\Psi_{jk}^{T}\Psi_{S_j}$, and $\Lambda_{min}\left(\Sigma_{S_j,S_j}\right)$, and $\Lambda_{\min}(\Sigma_{S_j,S_j})$ be the minimum eigenvalue of Σ_{S_j,S_j}. Define the subgradient of the penalty in (7) with respect to β_{jk} as $g_{jk}(\beta)$. On the set S_j, we write the concatenated subgradients as gS_j, a vector of length $s_j r$.

Let $\hat{\beta}$ be the parameter estimates from solving (7), let $\hat{S}_n=\left\{(j,k):\|\hat{\beta}_{jk}\|_2^2+\|\hat{\beta}_{kj}\|_2^2\neq 0\right\}$ be the corresponding estimated edge set, and let $S^* = \{(j,k) : k \in S_j \text{ or } j \in S_k\}$ be the graph obtained from the population level parameters. In Theorem 1, proved in the Appendix, we give precise conditions under which $pr\left(\hat{S}_n=S^*\right)\rightarrow 1$ as $n \rightarrow \infty$.

THEOREM 1. Let the functions f_{jk} be sufficently smooth, in thesense that if $f_{jk}^{(r)}=\sum_{t=1}^{r}\psi_{jkt}\left(X_k\right)\beta_{jkt}^{*}$, *then* $|f_{jk}^{(r)}\left(X_k\right)-f_{jk}\left(X_k\right)|=O_p\left(1/r^m\right)$, *uniformly in* $(j,k)\in V\times V$ *for some* $m\in\mathbb{N}$. For j = 1,..., d, Assume the that the basis functions satisfy $\Lambda_{\min}(\Sigma_{S_j, S_j}) \geq C_{min} > 0$ with probality tending to 1. Assume the irrrepresentability condition holds with probality

$$\|\Sigma_{jk,S_j}\Sigma_{S_j,S_j}^{-1}\hat{g}_{S_j}\|_2^2+\|\Sigma_{kj,S_k}\Sigma_{S_k,S_k}^{-1}\hat{g}_{S_k}\|_2^2\leq 1-\delta,\quad(11)$$

for $(j,k)\notin S^*$ *and some* $\delta > 0$, *where* $\hat{g}_{S_j}=g_{S_j}\left(\hat{\beta}\right)$. Assume the followint conditions on the number of edges $|S^*|$, *the neighborhood size* s_j, *the regularization parameter* λ, *and the truncation dimension r:*

$$\frac{r\,log(r|S^{*c}|)}{\lambda^2 n} \to 0, \quad \max_j \frac{rs_j\,log(r|S^*|)}{\lambda^2 n} \to 0, \quad \max_j \frac{s_j}{r^m \lambda} \to 0,$$
$$\frac{1}{\rho^*} \max_j \left[\left\{ \frac{s_j r\,log(r|S^*|)}{n} \right\}^{1/2} + \frac{s_j}{r^m} + \lambda (rs_j)^{1/2} \right] \to 0,$$

where $\rho^* = min_j\, min_{k \in S_j} \|\beta_{jk}^{*(r)}\|_\infty$. *Further, assume the variables* $\xi_{jkt} \equiv \psi_{jkt}(X_k) \in_j$ *have exponential tails, that is* $pr(|\xi_{jkt}| > z) \le ae^{-bz2}$ *for some* $a,b > 0$.

Then, the graph estimated using our method is consistent: $pr\left(\hat{S}_n = S^*\right) \to 1$ as $n \to \infty$.

Remark 1. The conditions on $|S^*|$, s_j, λ, and r hold if, for instance, $\lambda \propto n^{-\gamma\lambda}$, $d \propto \exp(n^{\gamma d})$, $r \propto n^{\gamma r}$, $\max_j S_j \propto n^{\gamma s}$, $m = 2$, and $\rho^* > \delta > 0$ for positive constants $\gamma\lambda$, γd, γ_r, γ_s, and δ, while $\gamma_r + \gamma_s < 2\gamma\lambda < 1 - \gamma_r - \gamma_s - \gamma d$, $2\gamma_s + \gamma\lambda$ and $n \to \infty$.

8. EXTENSION TO HIGH DIMENSIONS

In this section, we propose an approximation to our method that can speed up computations in high dimensions. Our proposal is motivated by recent work in the Gaussian setting (Witten et al., 2011; Mazumder & Hastie, 2012): the connected components of the conditional independence graph estimated using the graphical lasso (4) are precisely the connected components of the marginal independence graph estimated by thresholding the empirical covariance matrix. Consequently, one can obtain the exact solution to the graphical lasso problem in substantially reduced computational time by identifying the connected components of the marginal independence graph, and solving the graphical lasso optimization problem for the variables within each connected component.

We now apply the same principle to our method in order to quickly approximate the solution to (7). Let $\rho_m^{(jk)} = sup_{f,g \in \mathscr{F}} \rho\{f(X_k), g(X_j)\}$ be the maximal correlation between X_j and X_k over the univariate functions in $\mathscr{F}$ such that $f(X_k)$ and $g(X_j)$ have finite variance. Define the marginal dependence graph $\Gamma_M = (V, S_M)$, where $(j, k) \in S_M$ when $\rho_m^{(jk)} \neq 0$. If the jth and kth variables are in different connected components of Γ_M, then they must be conditionally independent. Theorem 2, proved in the Appendix, makes this assertion precise.

THEOREM 2. *Let* $C_1, \ldots, C_l$ be the connected components of Γ_M. Suppose the space of functions $\mathscr{F}$ contains linear functions. If $j \in C_u$ and $k \notin C_u$ *for some* $1 \le u \le l$, *then* $(j, k) \notin S^*$.

Theorem 2 forms the basis for Algorithm 2. There, we approximate $\rho_m^{(jk)}$ using the canonical correlation (Mardia et al., 1980) between the basis expansions $\mathbf{\Psi}_{kj}$ and $\mathbf{\Psi}_{jk}$:

$\hat{\rho}_m^{(jk)} = max_{v,w \in \mathbb{R}^r} \rho(\Psi_{jk} v, \Psi_{jk} w)$.

Algorithm 2. Given λ_1 and λ_2, perform Steps 1–4.

Step 1. For $(j, k) \in V \times V$, calculate $\rho_m^{(jk)}$, the canonical correlation between $\mathbf{\Psi}_{kj}$ and $\mathbf{\Psi}_{jk}$.

Step 2. Construct the marginal independence graph $\hat{\Gamma}_M : (j,k) \in \hat{\Gamma}_M$ when $|\rho_m^{(jk)}| \geq \lambda_2$.

Step 3. Find the connected components $C_1, \ldots, C_l$ of $\hat{\Gamma}_M$.

Step 4. Perform Algorithm 1 on each connected component with $\lambda = \lambda_1$.

In order to show that (i) Algorithm 2 provides an accurate approximation to the original algorithm, (ii) the resulting estimator outperforms methods that rely on Gaussian assumptions when those assumptions are violated, and (iii) Algorithm 2 is indeed faster than Algorithm 1, we replicated the graph used in Section 5·1 five times. This gives $d = 500$ variables, broken into five components. We took $n = 250$, and set $\Psi_{jk} = \left(x_k, x_k^2, x_k^3\right)$.

In Fig. 5, we see that when λ_2 is small, there is little loss in statistical efficiency relative to Algorithm 1, which is a special case of Algorithm 2 with $\lambda_2 = 0$. Further, we see that our method outperforms neighborhood selection even when λ_2 is large. Using Algorithm 2 with $\lambda_2 = 0.5$ and $\lambda_2 = 0.63$ led to a reduction in computation time by 25% and 70%, respectively.

Theorem 2 continues to hold if maximal correlation $\rho_m^{(jk)}$ is replaced with some other measure of marginal association $\rho_*^{(jk)}$, provided that $\rho_*^{(jk)}$ dominates maximal correlation in the sense that $\rho_*^{(jk)} = 0$ implies that $\rho_m^{(jk)} = 0$. That is, any measure of marginal association, such as mutual information, which detects the same associations as maximal correlation (i.e. $\rho_*^{(jk)} \neq 0$ if $\rho_m^{(jk)} \neq 0$) can be used in Algorithm 2.

9. DISCUSSION

A possible extension to this work involves accommodating temporal information. We could take advantage of the natural ordering induced by time, as considered by Shojaie & Michailidis (2010a), and apply our method for directed graphs. We leave this to future work.

Acknowledgments

We thank two anonymous reviewers and an associate editor for helpful comments, Thomas Lumley for valuable insights, and Han Liu and Bernd Fellinghauer for providing R code. D.W. and A.S. are partially supported by National Science Foundation grants, D.W. is partially supported by a National Institutes of Health grant.

APPENDIX 1: TECHNICAL PROOFS

A·1. Proof of Theorem 1

First, we restate a theorem which will be useful in the proof of the main result.

Theorem A1. (Kuelbs & Vidyashankar, 2010) Let $\{\xi_{n,j,i} : i = 1, \ldots, n; j \in A_n\}$ be a set of random variables such that $\xi_{n,j,i}$ is independent of $\xi_{n,j,i'}$ for $i \neq i'$. That is, $\xi_{n,j,i}$, $i = 1, \ldots, n$ denotes independent observations of feature j, and the features are indexed by some finite set A_n. Assume $E(\xi_{n,j,i}) = 0$, and there exist constants $a > 1$ and $b > 0$ such that $\text{pr}(|\xi_{n,j,i}| \geq x) \leq$

ae^{-bx2} for all $x > 0$. Further, assume that $|A_n| < \infty$ for all x all n and that $|A_n| \to \infty$ as $n \to \infty$. Denote $z_{n,j} = \sum_{i=1}^{n} \xi_{n,j,i}$. Then

$$\frac{max_{j \in A_n} |z_{n,j}|}{n} = O_p \left[\left\{ \frac{log\,(|A_n|)}{n} \right\}^{1/2} \right].$$

We now prove Theorem 1 of Section 7.

Proof of Theorem 1. First, $\hat{\beta}$ is a solution to (7) if and only if

$$-\frac{1}{n} \Psi_{jk}^T \left(x_j - \sum_{l \neq j} \Psi_{jl} \hat{\beta}_{jl} \right) + \lambda g_{jk} \left(\hat{\beta} \right) = 0 \quad \{(j,k) \in V \times V\}, \quad \text{(A1)}$$

where $g_{jk}\left(\hat{\beta}\right)$ is the subgradient vector satisfying $\|g_{jk}\left(\hat{\beta}\right)\|_2^2 + \|g_{kj}\left(\hat{\beta}\right)\|_2^2 \leq 1$ when $\|\hat{\beta}_{jk}\|_2 + \|\hat{\beta}_{kj}\|_2 = 0$, otherwise $g_{jk}\left(\hat{\beta}\right) = \Psi_{jk}\hat{\beta}_{jk}\left(\|\Psi_{jk}\hat{\beta}_{jk}\|_2^2 + \|\Psi_{kj}\hat{\beta}_{kj}\|_2^2\right)^{-1/2}$.

We base our proof on the primal-dual witness method of Wainwright (2009). That is, we construct a coefficient-subgradient pair $\left(\hat{\beta}, \hat{g}\right)$ and show that they solve (7) and produce the correct sparsity pattern, with probability tending to 1. For $(j, k) \in S^*$, we construct $\hat{\beta}_{jk}$ and the corresponding subgradients $\hat{g}_{jk}$ using our method, restricted to edges in S^* :

$$\underset{\beta_{jk}:(j,k)\in S^*}{arg\,min} \left\{ \frac{1}{2n} \sum_{j=1}^{d} \|x_j - \sum_{k \in S_j} \Psi_{jk}\beta_{jk}\|_2^2 + \lambda \sum_{(j,k)\in S^*} \left(\|\Psi_{jk}\beta_{jk}\|_2^2 + \|\Psi_{kj}\beta_{jk}\|_2^2 \right)^{1/2} \right\}. \quad \text{(A2)}$$

For $(j, k) \in S^{*c}$, we set $\hat{\beta}_{jk} = 0$, and use (A1) to solve for the remaining $\hat{g}_{jk}$ when $k \notin S_j$. Now, $\hat{\beta}$ is a solution to (7) if

$$\|g_{jk}\left(\hat{\beta}_{jk}\right)\|_2^2 + \|g_{kj}\left(\hat{\beta}_{kj}\right)\|_2^2 \leq 1 \qquad \{(j,k) \in S^{*c}\}. \quad \text{(A3)}$$

In addition, $\hat{S}_n = S^*$ provided that

$$\hat{\beta}_{S_j} \neq 0 \quad (j = 1, \ldots, d). \quad \text{(A4)}$$

Thus, it suffices to show that that conditions (A3) and (A4) hold with high probability.

We start with the condition (A4). The stationary condition for $\hat{\beta}_{S_j}$ is given by

$$-\frac{1}{n} \Psi_{S_j}^T \left(x_j - \Psi_{S_j} \hat{\beta} S_j \right) + \lambda \hat{g}_{S_j} = 0.$$

Denote by $\sum_{k\in S_j}\left\{f_{jk}(x_k)-f_{jk}^{(r)}(x_k)\right\}=w_j$ the truncation error from including only r basis terms. We can write $x_j=\Psi_{S_j}\beta_{S_j}^{*(r)}+w_j+\epsilon_j$. And so

$$\frac{1}{n}\Psi_{S_j}^T\left\{\Psi_{S_j}\left(\hat{\beta}_{S_j}-\beta_{S_j}^{*(r)}\right)-w_j-\epsilon_j\right\}+\lambda\hat{g}_{S_j}=0,$$

or

$$\left(\hat{\beta}_{S_j}-\beta_{S_j}^{*(r)}\right)=\left(\frac{1}{n}\Psi_{S_j}^T\Psi_{S_j}\right)^{-1}\left(\frac{1}{n}\Psi_{S_j}^T w_j+\frac{1}{n}\Psi_{S_j}^T\epsilon_j-\lambda\hat{g}_{S_j}\right), \quad \text{(A5)}$$

using the assumption that $\Psi_{S_j}^T\Psi_{S_j}$ is invertible. We will now show that the inequality

$$\max_j\|\hat{\beta}_{S_j}-\beta_{S_j}^{*(r)}\|_\infty<\min_j\min_{k\in S_j}\|\beta_{jk}^{*(r)}\|_\infty/2\equiv\rho^*/2 \quad \text{(A6)}$$

holds with high probability. This implies that $\|\hat{\beta}_{jk}\|_2\neq 0$ if $\|\beta_{jk}^{*(r)}\|_2\neq 0$.

From (A5) we have that

$$\max_j\|\hat{\beta}_{S_j}-\beta_{S_j}^{*(r)}\|_\infty\leq\max_j\left\|\Sigma_{S_j,S_j}^{-1}\frac{1}{n}\Psi_{S_j}^T w_j\right\|_\infty+\max_j\left\|\Sigma_{S_j,S_j}^{-1}\frac{1}{n}\Psi_{S_j}^T\epsilon_j\right\|_\infty+\max_j\lambda\left\|\Sigma_{S_j,S_j}^{-1}\hat{g}_{S_j}\right\|_\infty\equiv T_1+T_2+T_3.$$

Thus, to show (A6) it suffices to bound T_1, T_2, and T_3.

We first bound T_1. By assumption, $|f_{jk}^{(r)}(X_k)-f_{jk}(X_k)|=O_p(1/r^m)$ uniformly in k. Thus, $n^{-1/2}\|w_j\|_2=n^{-1/2}\|\sum_{k\in S_j}\left\{f_{jk}^{(r)}(x_k)-f_{jk}(x_k)\right\}\|_2=O_p(s_j/r^m)$ uniformly in j.

This implies that

$$T_1\leq\max_j\left\|\Sigma_{S_j,S_j}^{-1}\frac{1}{n}\Psi_{S_j}^T w_j\right\|_2\leq\max_j\left\|\Sigma_{S_j,S_j}^{-1}\frac{1}{\sqrt{n}}\Psi_{S_j}^T\right\|_2\frac{1}{\sqrt{n}}\|w_j\|_2\leq C_{min}^{-1/2}\max_j O_p(s_j/r^m)=O_p\left(\frac{\max_j s_j}{r^m}\right).$$

In the above, we used that $\Lambda_{max}\left(\sum_{S_j,S_j}^{-1}\Psi_{S_j}^T/\sqrt{n}\right)=\left\{\Lambda_{min}\left(\sum_{S_j,S_j}\right)\right\}^{-1/2}$.

We now bound T_2. Here, we use Theorem A1 which bounds the l_∞ norm of the average of high-dimensional independent vectors. First, by the definition of $\in_j$ we must have that $E\{\psi_{jkt}(X_k)\in_j\}=0$, i.e. the errors are uncorrelated with the covariates.

Let $z_{jkt}\equiv\psi_{jkt}(x_k)^{\mathrm{T}}\in_j$, which is the sum of n independent random variables with exponential tails. We have that

$$\max_j \|\Psi_{S_j}^T \epsilon_j\|_\infty / n = \max_j \max_{k \in S_j} \max_{t=1,\dots,r} |z_{jkt}|/n \leq \max_{(j,k) \in S^*} \max_{t=1,\dots,r} \left(|z_{jkt}| \vee |z_{kjt}|\right)/n,$$

the maximum of $2r|S^*|$ elements. We can thus apply Theorem A1, with A_n indexing the $2r|S^*|$ elements above, to obtain

$$T_2 = \max_j \left\| \Sigma^{-1}_{S_j,S_j} \frac{1}{n} \Psi^T_{S_j} \epsilon_j \right\|_\infty \leq \max_j \left\| \Sigma^{-1}_{S_j,S_j} \right\|_\infty \left\| \frac{1}{n} \Psi^T_{S_j} \epsilon_j \right\|_\infty \leq \max_j (rs_j)^{1/2} C_{min}^{-1} O_p \left[\left\{ \frac{\log(2r|S^*|)}{n} \right\}^{1/2} \right]$$
$$= O_p \left[\max_j \left\{ \frac{s_j r \log(r|S^*|)}{n} \right\}^{1/2} \right].$$

We now bound T_3. We have that $\|\hat{g}_{jk}\|_2^2 \leq 1$ for $(j, k) \in S^*$, so

$$T_3 \leq \lambda \max_j \|\Sigma^{-1}_{S_j,S_j}\|_\infty \leq \lambda \max_j (rs_j)^{1/2} \|\Sigma^{-1}_{S_j,S_j}\|_2 \leq \lambda \max_j \frac{(rs_j)^{1/2}}{C_{min}}.$$

Altogether, we have shown that

$$\max_j \|\hat{\beta}_{S_j} - \beta^{*(r)}_{S_j}\|_\infty \leq O_p\left(\frac{\max_j s_j}{r^m}\right) + O_p \left[\max_j \left\{ \frac{s_j r \log(r|S^*|)}{n} \right\}^{1/2} \right] + \lambda \max_j \frac{(rs_j)^{1/2}}{C_{min}}.$$

By assumption,

$$\frac{1}{\rho^*} \max_j \left[\left\{ \frac{s_j r \log(r|S^*|)}{n} \right\}^{1/2} + \frac{s_j}{r^m} + \lambda (rs_j)^{1/2} \right] \to 0$$

which implies that $\max_j \|\hat{\beta}_{S_j} - \beta^{*(r)}_{S_j}\|_\infty < \rho^*/2$ with probability tending to 1 as $n \to \infty$.

We now consider the dual problem, condition (A3). We must show that $\|\hat{g}_{jk}\|_2 + \|\hat{g}_{kj}\|_2 \leq 1$ for each $(j, k) \notin S^*$. From the discussion of condition (A4), we know that

$$\begin{aligned} -\hat{g}_{jk} &= \frac{1}{\lambda n} \Psi^T_{jk} \left\{ \Psi_{S_j} \left(\hat{\beta}_{S_j} - \beta^{*(r)}_{S_j} \right) - w_j - \epsilon_j \right\} \\ &= \frac{1}{\lambda n} \Psi^T_{jk} \left\{ \Psi_{S_j} \Sigma^{-1}_{S_j,S_j} \left(\frac{1}{n} \Psi^T_{S_j} w_j + \frac{1}{n} \Psi^T_{S_j} \epsilon_j - \lambda \hat{g}_{S_j} \right) - w_j - \epsilon_j \right\} \\ &= -\frac{1}{\lambda n} \Psi^T_{jk} \left(I - \frac{1}{n} \Psi_{S_j} \Sigma^{-1}_{S_j,S_j} \Psi^T_{S_j} \right) w_j - \frac{1}{\lambda n} \Psi^T_{jk} \left(I - \frac{1}{n} \Psi_{S_j} \Sigma^{-1}_{S_j,S_j} \Psi^T_{S_j} \right) \epsilon_j - \frac{1}{n} \Psi^T_{jk} \Psi_{S_j} \Sigma^{-1}_{S_k,S_j} \hat{g}_{S_j} \\ &\equiv M_1^{jk} + M_2^{jk} + M_3^{jk}. \end{aligned}$$

We will proceed by bounding $\|M_1^{jk}\|_2 + \|M_1^{kj}\|_2$, $\|M_2^{jk}\|_2 + \|M_2^{kj}\|_2$ and $\|M_3^{jk}\|_2 + \|M_3^{kj}\|_2$, which will a bound for the quantity of interest, $\|\hat{g}_{jk}\|_2 + \|\hat{g}_{kj}\|_2$.

We first bound M_1. When bounding T_1 earlier, we saw that $n^{-1/2}\|w_j\|_2 = O_p(s_j/r^m)$. Now $I - \Psi_{S_j}\sum^{-1}_{S_j,S_j}\Psi^T_{S_j}/n$ is a projection matrix with eigenvalues equal to 1, and by design $n^{-1/2}\boldsymbol{\Psi}_{jk}$ is orthogonal, so that all the singular values of $n^{-1/2}\boldsymbol{\Psi}_{jk}$ are 1. Therefore

$$\|M_1^{jk}\|_2 \leq \frac{1}{\lambda} n^{-1/2}\|\Psi_{jk}\|_2 n^{-1/2}\|w_j\|_2 = O_p\left(\frac{s_j}{\lambda r^m}\right),$$

and

$$\|M_1^{jk}\|_2 + \|M_1^{kj}\|_2 \leq O_p\left(\frac{s_j \vee s_k}{\lambda r^m}\right),$$

which tends to zero because $s_j(\lambda r^m)^{-1} \to 0$ uniformly in j.

We now bound M_2. First, note that

$$\begin{aligned}\lambda\|M_2^{kl}\|_2 &\leq n^{-1}\|\Psi_{jk}^T\epsilon_j\|_2 + n^{-1/2}\|\Psi_{jk}\|_2\left\|n^{-1/2}\Psi_{S_j}\Sigma^{-1}_{S_j,S_j}\right\|_2\|\Psi^T_{S_j}\epsilon_j\|_2/n \\ &\leq n^{-1}\|\Psi_{jk}^T\epsilon_j\|_2 + C_{min}^{-1/2}\|\Psi^T_{S_j}\epsilon_j\|_2/n \\ &\leq \sqrt{r}\|\Psi_{jk}^T\epsilon_j\|_\infty/n + (rs_j/C_{min})^{1/2}\|\Psi^T_{S_j}\epsilon_j\|_\infty/n.\end{aligned}$$

Then, applying Theorem A1, as in the bound for T_2, we get

$$\lambda \max_{(j,k)\in S^{*c}}\|M_2^{jk}\|_2 \leq O_p\left[\left\{\frac{r\,log\,(r|S^{*c}|)}{n}\right\}^{1/2}\right] + O_p\left[\left\{\frac{r\,max_j s_j\,log\,(r|S^*|)}{n}\right\}^{1/2}\right].$$

Thus, $max_{(j,k)\in S^{*c}}\left(\|M_2^{jk}\|_2 + \|M_2^{jk}\|_2\right) \to 0$ when

$$\frac{r\,log\,(r|S^{*c}|)}{\lambda^2 n} \to 0 \quad \text{and} \quad \max_j \frac{rs_j\,log\,(r|S^*|)}{\lambda^2 n} \to 0.$$

We now bound M_3. By the irrepresentability assumption, we have that $\|M_3^{jk}\|_2^2 + \|M_3^{kj}\|_2^2 \leq 1 - \delta$ with probability tending to 1.

Thus, since $\|M_1^{jk}\|_2 + \|M_1^{kj}\|_2 + \|M_2^{jk}\|_2 + \|M_2^{kj}\|_2 = o_p(1)$, we have that for $(j, k) \in S^{*c}$

$$\max_{(j,k)\in S^{*c}}\left(\|\hat{g}_{jk}\|_2 + \|\hat{g}_{kj}\|_2\right) \leq 1 - \delta$$

with probability tending to 1. Further, since we have strict dual feasibility, i.e, $\|\hat{g}_{jk}\|_2 + \|\hat{g}_{kj}\|_2 < 1$ for $(j, k) \in S^{*c}$, with probability tending to 1, the estimated graph is unique.

A·2. Proof of Theorem 2

We now prove Theorem 2 of Section 8.

Proof of Theorem 2. Consider a variable $j \in C_u$. Our large-sample model minimizes $E|X_j - \Sigma_{k\neq j} f_{jk}(X_k)|^2$ over functions $f_{jk} \in \mathscr{F}$. We have that

$$\begin{aligned}
&E|X_j - \sum_{k\neq j} f_{jk}(X_k)|^2 \\
&= EX_j^2 - 2\sum_{k\neq j} E\{X_j f_{jk}(X_k)\} \\
&+ \sum_{k\neq j}\sum_{l\neq j} E\{f_{jk}(X_k) f_{jt}(X_t)\} \\
&= EX_j^2 - 2\sum_{k\in C_u\setminus j} E\{X_j f_{jk}(X_k)\} \\
&- 2\sum_{k\notin C_u} E\{X_j f_{jk}(X_k)\} \\
&+ \sum_{k\in C_u\setminus j}\sum_{l\in C_u\setminus j} E\{f_{jk}(X_k) f_{jl}(X_l)\} \\
&+ \sum_{k\notin C_u}\sum_{l\notin C_u} E\{f_{jk}(X_k) f_{jl}(X_l)\} \\
&+ 2\sum_{k\notin C_u}\sum_{l\in C_u\setminus j} E\{f_{jk}(X_k) f_{jl}(X_l)\}.
\end{aligned}$$

By assumption, $\sum_{k\notin C_u} E\{X_j f_{jk}(X_k)\} = \sum_{k\notin C_u}\sum_{l\in C_u\setminus j} E\{f_{jk}(X_k) f_{jt}(X_t)\} = 0$. Thus, collecting terms, get

$$E|X_j = \sum_{k\neq j} f_{jk}(X_k)|^2 = E|X_j - \sum_{k\in C_u\setminus j} f_{jk}(X_k)|^2 + E|\sum_{k\notin C_u} f_{jk}(X_k)|^2.$$

Minimization of this quantity with respect to $\{f_{jk} \in \mathscr{F} : k \notin C_u\}$ only involves the last term, which achieves its minimum at zero when $f_{jk}(\cdot) = 0$ almost everywhere for each $k \notin C_u$.

REFERENCES

Banerjee O, El Ghaoui L, D'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res. 2008; 9:485–516.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nature Genet. 2005; 37:382–390. [PubMed: 15778709]

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

Besag JE. Statistical analysis of non-lattice data. The Statistician. 1975; 24:179–195.

Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008; 95:759–771.

Chen YA, Almeida JS, Richards AJ, Müller P, Carroll RJ, Rohrer B. A nonparametric approach to detect nonlinear correlation in gene expression. J. Comp. Graph. Stat. 2010; 19:552–568.

Cowell, RG.; Dawid, P.; Lauritzen, SL.; Spiegelhalter, DJ. Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks. Springer; New York: 2007. chap. 3.2.1.

Dempster AP. Covariance selection. Biometrics. 1972; 28:157–175.

Dobra A, Lenkoski A. Copula Gaussian graphical models and their application to modeling functional disability data. Ann. App. Statist. 2011; 5:969–993.

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. Stat. & Data An. 2013; 64:132–152.

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

Gao X, Song PX-K. Composite likelihood Bayesian information criteria for model selection in high-dimensional data. J. Am. Statist. Assoc. 2010; 105:1531–1540.

Hastie, T.; Tibshirani, RJ. Generalized Additive Models. Chapman & Hall/CRC; Boca Raton: 1990.

Hausser J, Strimmer K. Entropy inference and the james–stein estimator, with application to nonlinear gene association networks. J. of Mach. Learn. Res. 2009; 10:1469–1484.

Khatri CG, Rao CR. Characterizations of multivariate normality. I. through independence of some statistics. J. Mult. An. 1976; 6:81–94.

Kuelbs J, Vidyashankar A. Asymptotic inference for high-dimensional data. Ann. Statist. 2010; 38:836–869.

Liang K-C, Wang X. Gene regulatory network reconstruction using conditional mutual information. EURASIP J. Bioinf. and Syst. Biol. 2008:253894–253894.

Liu H, Han F, Yuan M, Lafferty J, Wasserman LA. High-dimensional semiparametric Gaussian copula graphical models. Ann. Statist. 2012; 40:2293–2326.

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

Liu H, Xu M, Gu H, Gupta A, Lafferty J, Wasserman LA. Forest density estimation. J. Mach. Learn. Res. 2011; 12:907–951.

Mardia, K.; Kent, J.; Bibby, J. Multivariate Analysis. Academic press; Waltham: 1980.

Mazumder R, Hastie TJ. Exact covariance thresholding into connected components for large-scale graphical lasso. J. of Mach. Learn. Res. 2012; 13:781–794.

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

Meinshausen N, Bühlmann P. Stability selection. J. R. Statist. Soc. B. 2010; 72:417–473.

Meyer PE, Lafitte F, Bontempi G. Minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics. 2008; 9:461–471. [PubMed: 18959772]

Pearl, J. Causality: Models, Reasoning, and Inference. Vol. 47. Functional Causal Models. Cambridge Univ Press; 2000. p. 27-38.chap. 1.4

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 P, Lafferty J, Liu H, Wasserman LA. Sparse additive models. J. R. Statist. Soc. B. 2009; 71:1009–1030.

Rothman A, Bickel P, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:494–515.

Sachs K, Perez O, Pe'er D, Lauffenburger D, Nolan G. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005; 308:523–529. [PubMed: 15845847]

Shojaie A, Michailidis G. Discovering graphical Granger causality using the truncating lasso penalty. Bioinformatics. 2010a; 26:517–523.

Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika. 2010b; 97:519–538. [PubMed: 22434937]

Simon N, Tibshirani RJ. Standardization and the group lasso penalty. Statistica Sinica. 2012; 22:983–1001.

Tibshirani RJ. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B. 1996; 58:267–288.

Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. J. Opt. Theo. App. 2001; 109:475–494.

Varin C, Reid N, Firth D. An overview of composite likelihood methods. Statistica Sinica. 2011; 21:5–42.

Varin C, Vidoni P. A note on composite likelihood inference and model selection. Biometrika. 2005; 92:519–528.

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

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

Witten DM, Friedman J, SIMON N. New insights and faster computations for the graphical lasso. J. Comp. Graph. Stat. 2011; 20:892–900.

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

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J. R. Statist. Soc. B. 2006; 68:49–67.

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

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

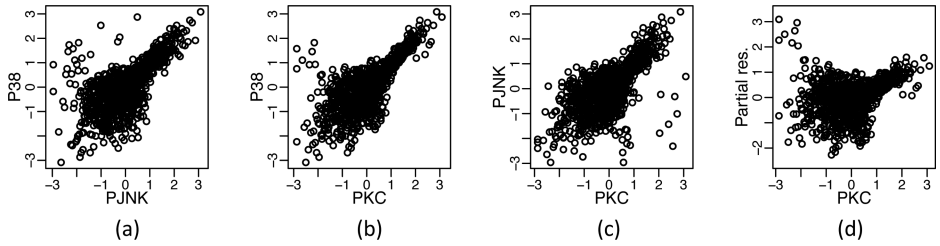

Fig. 1.

Cell signaling data from Sachs et al. (2005). (a)–(c) Pairwise scatterplots for PKC, P38 and PJNK. (d) Partial residuals from the linear regression of P38 on PKC and PJNK. The data are transformed to have normal marginal distributions, but are clearly not multivariate normal.

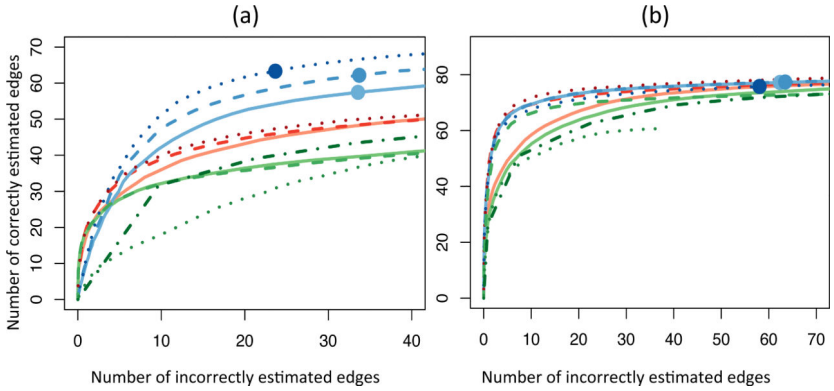

Fig. 2.

Summary of simulation study. The number of correctly estimated edges is displayed as a function of incorrectly estimated edges, for a range of tuning parameter values, in the (a) non-linear and (b) Gaussian set-ups, averaged over 100 simulated data sets. Dots indicate the average model size chosen by minimizing BIC(λ). The lines display our method with $\Psi_{jk} = \left(x_k, x_k^2\right)$, ( ), $\Psi_{jk} = \left(x_k, x_k^3\right)$ ( ), and $\Psi_{jk} = \left(x_k, x_k^2, x_k^3\right)$ ( ), as well as the methods of Liu et al. (2012) ( ), Basso et al. (2005) ( ), Liu et al. (2011) ( ), Fellinghauer et al. (2013) ( ), Yuan & Lin (2007) ( ), Meinshausen & Bühlmann (2006) ( ), and Peng et al. (2009) ( ).

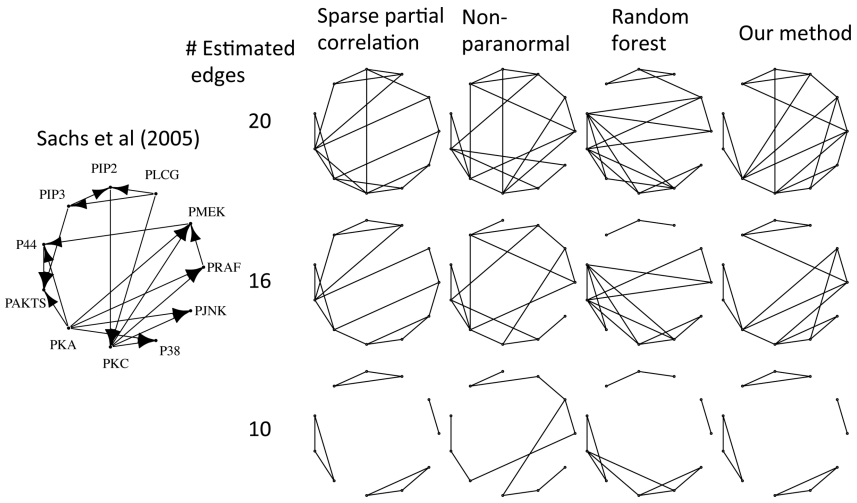

Fig. 3.

Cell signaling data set; graph reported in Sachs et al. (2005) is shown on the left. On the right, graphs were estimated using data from one perturbation of the data set. From top to bottom, panels contain graphs with 20, 16 and 10 edges. From left to right, comparisons are to Peng et al. (2009); Liu et al. (2012); Fellinghauer et al. (2013). We cannot specify an arbitrary graph size using graphical random forests, so graph sizes for that approach do not match exactly.

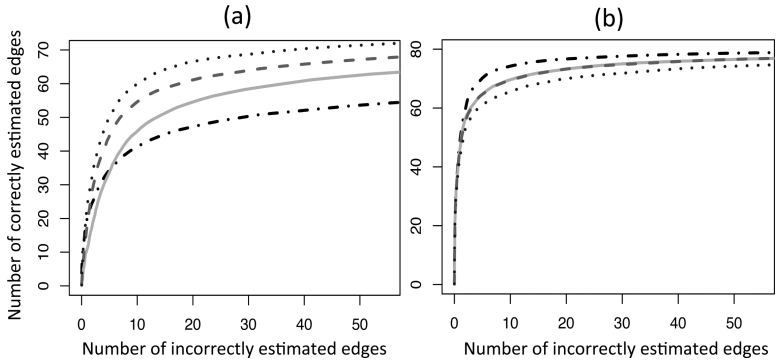

Fig. 4.

Summary of the directed acyclic graph simulation. The simulation is exactly as in Section 5·1 and Fig. 2. Again, (a) contains the non-linear simulation and (b) contains the Gaussian simulation. For each method, the number of correctly and incorrectly estimated edges are averaged over 100 simulated data sets, for a range of 100 tuning parameter values. The curves displayed are those of our method with $\Psi_{jk} = \left(x_k, x_k^2\right)$ (), $\Psi_{jk} = \left(x_k, x_k^3\right)$ (), and $\Psi_{jk} = \left(x_k, x_k^2, x_k^3\right)$ (), as well as the method of Shojaie & Michailidis (2010b) ().

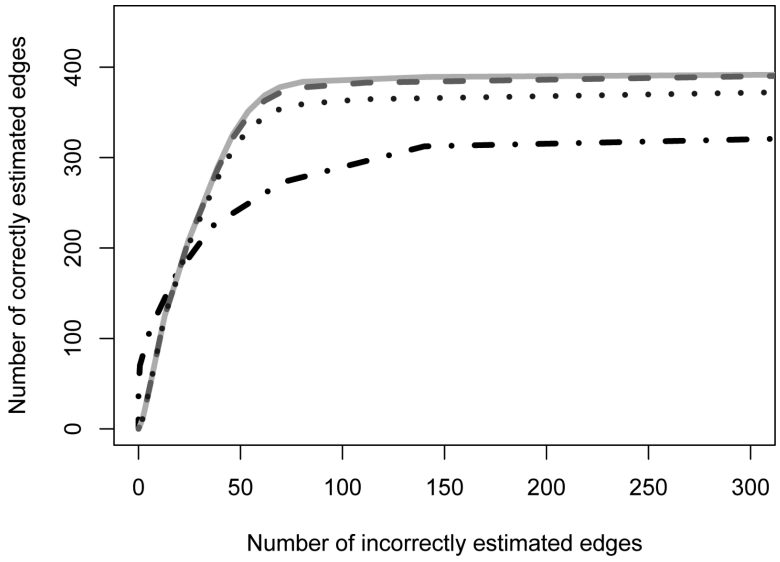

Fig. 5.

Performance of Algorithm 2. The number of correctly and incorrectly estimated edges are averaged over 100 simulated data sets, for each of 100 tuning parameter values. The curves displayed are from our method with $\lambda_2 = 0$ (), $\lambda_2 = 0.5$ () and $\lambda_2 = 0.63$ (), as well as the method of Meinshausen & Bühlmann (2006) ().