Learning Nonparametric Forest Graphical Models with Prior Information

Yuancheng Zhu
University of Pennsylvania

Zhe Liu
University of Chicago

Siqi Sun
Toyota Technological Institute at Chicago

Abstract

We present a framework for incorporating prior information into nonparametric estimation of graphical models. To avoid distributional assumptions, we restrict the graph to be a forest and build on the work of forest density estimation (FDE). We reformulate the FDE approach from a Bayesian perspective, and introduce prior distributions on the graphs. As two concrete examples, we apply this framework to estimating scale-free graphs and learning multiple graphs with similar structures. The resulting algorithms are equivalent to finding a maximum spanning tree of a weighted graph with a penalty term on the connectivity pattern of the graph. We solve the optimization problem via a minorize-maximization procedure with Kruskal's algorithm. Simulations show that the proposed methods outperform competing parametric methods, and are robust to the true data distribution. They also lead to improvement in predictive power and interpretability in two real data sets.

1 Introduction

Graphical models are widely used to encode the conditional independence relationships between random variables. In particular, a random vector $X = (X_1, \ldots, X_d)$ is represented by an undirected graph $G = (V, E)$ with $d = |V|$ vertices and missing edges $(i, j) \notin E$ whenever X_i and X_j are conditionally independent given the other variables. One major statistical task is to learn the graph from n i.i.d. copies of the random vector.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

Existing approaches for estimating graphical models make assumptions on either the underlying distribution or the graphical structure. Currently the most popular method, called *graphical lasso* (Friedman et al., 2008), assumes that the random vector follows a multivariate Gaussian distribution. In this way, learning the graph is equivalent to estimating the precision matrix Ω, since the conditional independence of a Gaussian random vector is entirely determined by the sparsity pattern of Ω. The graphical lasso finds a sparse estimate of Ω by maximizing the ℓ_1-regularized log-likelihood. On the other hand, we can make no distributional assumptions but restrict the graph to be a forest instead. Under this structural constraint, there exists a factorization of the density function involving only the univariate and bivariate marginal densities, which makes nonparametric estimation tractable in high dimensions. In this case, estimating the graph amounts to finding the maximum spanning tree of a weighted graph; see, for example, Chow and Liu (1968); Liu et al. (2011) for details.

Oftentimes, additional information on the structure of a graph is available *a priori*, which could be utilized to assist the estimation task (Koivisto and Sood, 2004). For example, a wide variety of the networks in recent literature, such as protein, gene, and social networks, are reported to be scale-free. That is, the degree distribution of the vertices follows a power law: $p(\text{degree} = k) \propto k^{-\alpha}$ for some $\alpha > 1$. In such scale-free networks, some vertices have many more connections than others, and these highest-degree vertices are usually called hubs and serve significant roles in their networks. As another example of prior information, consider the applications where we believe that several networks share similar but not necessarily identical structures. This phenomenon is not unusual when we have multiple sets of data across distinct classes or units, such as gene expression measurements collected on a set of normal tissue samples and a set of cancer tissue samples. It is thus natural to ask whether such prior information can be integrated to improve estimation.

Various approaches have been proposed to incorporate the prior belief of the underlying graphs, for example, Defazio and Caetano (2012); Liu and Ihler (2011); Tan et al. (2014); Tang et al. (2015) for learning scale-free graphical models, and Guo et al. (2011); Danaher et al. (2014); Peterson et al. (2015); Zhu and Barber (2015) for joint estimation of multiple graphical models. Nevertheless, to the best of our knowledge, all the existing methods assume some parametric distribution on the data, mostly multivariate Gaussian. Such distributional assumptions can be quite unrealistic and unnecessary in many applications. Even though the marginal distribution of each variable can be transformed to approximately Gaussian, which allows arbitrary univariate distributions, the joint dependence is still restricted under the Gaussian assumption.

In this paper, we relax such distributional assumptions and estimate graphical models nonparametrically. We build on the forest density estimation (FDE) method introduced in Liu et al. (2011). In particular, we reformulate the FDE approach from a Bayesian perspective, and encode the prior information by putting some prior distribution on the graphs, which favors those that are more consistent with our prior belief. We further show that for the scale-free-graph case and the multiple-graph case, such an approach amounts to finding a maximum spanning tree of a weighted graph with a penalty term on the connection pattern of the nodes. We then devise an algorithm based on a minorize-maximization procedure and Kruskal's algorithm (Kruskal, 1956) to find a local optimal solution.

The rest of the paper is organized as follows. In the following section, we give background on forest density estimation. In Section 3, we first give a general framework on how to incorporate prior information to nonparametric forest-based graphical model estimation, and then illustrate how the framework can be specialized to model scale-free graphical models and jointly estimate multiple graphical models with similar structure. We present theoretical results of the algorithm in Section 4. In Section 5, we provide a brief review on the related work. Experimental results on synthetic data sets and real applications are presented in Section 6, followed by a conclusion in Section 7. Proofs and some additional experimental results are collected in the supplementary material.

2 Forest density estimation

We say an undirected graph is a forest if it is acyclic. Let $F = (V_F, E_F)$ be a forest with vertices $V_F = \{1, \ldots, d\}$ and edge set $E_F \in V_F \times V_F$. Let $X = (X_1, \ldots, X_d)$ be a d-dimensional random vector with density $p(x) > 0$. We say that X, or equivalently, its density p, is Markov to F if X_i and X_j are conditionally independent given the other random variables whenever edge (i, j) is missing in E_F. A density p that is Markov to F has the following factorization

$$p(x) = \prod_{(i,j)\in E_F} \frac{p_{ij}(x_i, x_j)}{p_i(x_i)p_j(x_j)} \prod_{\ell \in V_F} p_\ell(x_\ell), \qquad (1)$$

where each $p_{ij}(x_i, x_j)$ is a bivariate density and each $p_l(x_l)$ is a univariate density. With this factorization, we can write the expected log-likelihood as

$$\mathbb{E} \log p(X) = \sum_{(i,j)\in E_F} I(X_i; X_j) - \sum_{\ell \in V_F} H(X_\ell), \qquad (2)$$

where $I(X_i; X_j) = \int p_{ij}(x_i, x_j) \log \frac{p_{ij}(x_i,x_j)}{p_i(x_i)p_j(x_j)} dx_i dx_j$ is the mutual information between X_i and X_j, and $H(X_\ell) = -\int p_\ell(x_\ell) \log p_\ell(x_\ell) dx_\ell$ is the entropy of X_ℓ. We maximize the right hand side of (2) to find the optimal forest F

$$\widehat{F} = \arg\max_{F \in \mathcal{F}_d} \sum_{(i,j)\in E_F} I(X_i; X_j), \qquad (3)$$

where $\mathcal{F}_d$ is the collection of spanning trees on vertices $\{1, \ldots, d\}$. We let $\mathcal{F}_d$ contain only spanning trees because there is always a spanning tree that solves the problem (3). This problem can be recast as the problem of finding a maximum spanning tree for a weighted graph, where the weight w_{ij} of the edge between nodes i and j is $I(X_i; X_j)$. Kruskal's algorithm (Kruskal, 1956) is a greedy algorithm that is guaranteed to find an optimal solution, while Chow and Liu (1968) propose the procedure in the setting of discrete random variables. The method is described in Algorithm 1. See Zhou (2011) for a discussion of a number of greedy algorithms for graphical model selection.

Algorithm 1 Kruskal's (Chow-Liu) algorithm

Input Weight matrix $W = (w_{ij})_{d\times d}$
Initialize $E^{(0)} \leftarrow \emptyset$
for $\ell = 1, \ldots, d-1$ **do**
  $(i^{(\ell)}, j^{(\ell)}) \leftarrow \arg\max_{(i,j)} w_{ij}$ such that $E^{(\ell-1)} \cup \{(i^{(\ell)}, j^{(\ell)})\}$ doesn't contain a cycle
  $E^{(\ell)} \leftarrow E^{(\ell-1)} \cup \{(i^{(\ell)}, j^{(\ell)})\}$
end for
Output The final edge set $E^{(d-1)}$

However, this procedure is not practical since the true density p is unknown. Suppose instead that we have $X_{1,1:d}, \ldots, X_{n,1:d}$, which are n i.i.d. copies of the random vector X. We replace the population mutual information by the estimates

$$\widehat{I}(X_i; X_j) = \int \widehat{p}_{ij}(x_i, x_j) \log \frac{\widehat{p}_{ij}(x_i, x_j)}{\widehat{p}_i(x_i)\widehat{p}_j(x_j)} dx_i dx_j,$$

where $\widehat{p}_{ij}(x_i, x_j)$ and $\widehat{p}_\ell(x_\ell)$ are kernel density estimators of the bivariate and univariate marginal densities

$$\widehat{p}_{ij}(x_i, x_j) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{h_2^2} K\left(\frac{X_{ti} - x_i}{h_2}\right) K\left(\frac{X_{tj} - x_j}{h_2}\right),$$
$$\widehat{p}_\ell(x_\ell) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{h_1} K\left(\frac{X_{t\ell} - x_\ell}{h_1}\right)$$

with a kernel function K and bandwidths h_2 and h_1. The resulting estimator of the graph becomes

$$\tilde{F} = \arg\max_{F \in \mathcal{F}_d} \sum_{(i,j)\in E_F} \widehat{I}(X_i; X_j). \tag{4}$$

A held-out set is usually used to prune the spanning tree $\tilde{F}$ by stopping early in Algorithm 1 when the likelihood on the held-out set is maximized. Thus we obtain a forest estimate of the graph.

3 Learning forest graphical model with prior knowledge

3.1 A Bayesian framework

Sometimes we have some prior information about the structure of the underlying graphical models, and would like to incorporate that to assist the estimation. One way to realize that is to encode the prior knowledge into prior distributions on the spanning trees. Let $\pi(F)$ be a prior distribution on $\mathcal{F}_d$, the set of the spanning trees with d nodes. Given the data $X_{1,1:d}, \ldots, X_{n,1:d}$ and assuming the density p is known and Markov to the spanning tree F, we can write the likelihood as

$$p(X|F) = \prod_{t=1}^{n} \left(\prod_{(i,j)\in E_F} \frac{p_{ij}(X_{ti}, X_{tj})}{p_i(X_{ti}) p_j(X_{tj})} \prod_{\ell \in V_F} p_\ell(X_{t\ell}) \right).$$

Then the posterior probability of F is

$$p(F|X) \propto p(X|F)\pi(F)$$
$$\propto \prod_{t=1}^{n} \left(\prod_{(i,j)\in E_F} \frac{p_{ij}(X_{ti}, X_{tj})}{p_i(X_{ti}) p_j(X_{tj})} \prod_{k \in V_F} p_k(X_{tk}) \right) \cdot \pi(F). \tag{5}$$

The maximum a posteriori (MAP) estimate is given by

$$\widehat{F}_{\text{MAP}} = \arg\max_{F \in \mathcal{F}_d} \left\{ \sum_{(i,j)\in E_F} \sum_{t=1}^{n} \frac{1}{n} \log \frac{p_{ij}(X_{ti}, X_{tj})}{p_i(X_{ti}) p_j(X_{tj})} + \frac{1}{n}\log \pi(F) \right\}. \tag{6}$$

Since we do not know the true density p in practice, we can plug in the estimator (4) and obtain

$$\tilde{F}_\pi = \arg\max_{F \in \mathcal{F}_d} \left\{ \sum_{(i,j)\in E_F} \widehat{I}(X_i; X_j) + \frac{1}{n}\log \pi(F) \right\} \tag{7}$$

as an approximation of $\widehat{F}_{\text{MAP}}$. In fact, $\tilde{F}_\pi$ is obtained by replacing the true marginal densities and the empirical distributions in (6) by their corresponding density estimates. It can also be viewed as a penalized version of the estimator (4).

The penalty term $\frac{1}{n}\log \pi(F)$, which is sometimes combinatorial, could make the optimization problem extremely hard to solve. However, when $\log \pi(F)$ is convex with respect to the entries of the adjacency matrix of F, we can adopt a minorize-maximization algorithm (Hunter and Lange, 2004) to find a local optimal solution. In fact, given the convexity of $\log \pi(F)$, the objective function adopts a linear lower bound at any current estimates. This linear lower bound can be then decomposed into a sum of weights over the edges, and we can apply the Kruskal's algorithm to update our estimate. We shall see in details in the following two concrete examples how this can be carried out.

3.2 Scale-free graphs

Now suppose that we have reasons to believe that the graph is scale-free, or more generally, that the graph consists of several nodes that have dominating degrees compared to the rest. Let $\delta(F, l)$ be the degree of the node l of a spanning tree $F \in \mathcal{F}_d$. Consider a prior distribution on F which satisfies

$$\pi(F) \propto \prod_{\ell \in V_F} \delta(F, \ell)^{-\alpha}, \tag{8}$$

for some $\alpha > 1$. This prior distribution favors the spanning trees whose degrees have a power law distribution, and thus reflects our prior beliefs. Plugging this in (7), we obtain

$$\tilde{F}_\pi = \arg\max_{F \in \mathcal{F}_d} \left\{ \sum_{(i,j)\in E_F} \widehat{I}(X_i; X_j) - \lambda \sum_{\ell \in V_F} \log(\delta(F, \ell)) \right\},$$

where $\lambda = \alpha/n$ can be now viewed as a tuning parameter. To solve this optimization problem, we first rewrite the objective function as

$$f(F) = \sum_{i<j} w_{ij} F_{ij} - \lambda \sum_{i=1}^{d} \log\left(\sum_{j=1}^{d} F_{ij}\right), \tag{9}$$

where $w_{ij} = \widehat{I}(X_i; X_j)$. Here we also abuse our notation by writing F as the adjacency matrix of F, that

is, $F_{ij} = 1$ if and only if $(i,j) \in E_F$. Note that we have the additional constraint that the graph F is a spanning tree. Given a current estimate $\check{F}$, we first lower bound $f(F)$ by linearizing it at $\check{F}$:

$$\begin{aligned} f(F) &\geq \sum_{i<j} w_{ij} F_{ij} - \lambda \sum_{i=1}^{d} \left(\log \left(\sum_{j=1}^{d} \check{F}_{ij} \right) \right. \\ &\qquad \left. + \frac{\sum_{j=1}^{d} F_{ij} - \sum_{j=1}^{d} \check{F}_{ij}}{\sum_{j=1}^{d} \check{F}_{ij}} \right) \\ &= \sum_{i<j} \left(w_{ij} - \frac{\lambda}{\sum_{\ell=1}^{d} \check{F}_{i\ell}} - \frac{\lambda}{\sum_{\ell=1}^{d} \check{F}_{j\ell}} \right) F_{ij} + C, \end{aligned}$$

where C is a constant which doesn't depend on F. We can maximize this lower bound by applying Kruskal's algorithm to the graph with edge weights

$$\check{w}_{ij} = w_{ij} - \frac{\lambda}{\sum_{\ell=1}^{d} \check{F}_{i\ell}} - \frac{\lambda}{\sum_{\ell=1}^{d} \check{F}_{j\ell}}. \tag{10}$$

We see that the weights are updated at each iteration based on the current estimate of the graph. Each edge weight is penalized by two quantities that are inversely proportional to the degrees of the two endpoints of the edge. An edge weight is thus penalized less if its endpoints are already highly connected and vice versa. With such a "rich gets richer" procedure, the algorithm encourages some vertices to have high connectivity and hence the overall degree distribution to have a heavy tail. We iterate through such minorization and maximization steps until convergence. Since the objective function is always increasing, the algorithm is guaranteed to converge to a local maximum.

3.3 Multiple graphs with similar structure

In this part, we illustrate how the framework can be modified to facilitate the case where we have multiple graphs that are believed to have similar but not necessarily identical structures. Instead of one single graph, suppose that we now have K graphical models with underlying forests $F^{(1)}, \ldots, F^{(K)}$, and for the kth one, we observe data $X^{(k)} = (X^{(k)}_{1,1:d}, \ldots, X^{(k)}_{n_k,1:d})$. Given a joint prior distribution π on $(F^{(1)}, \ldots, F^{(K)})$, we combine the likelihood for the K models and update the posterior distribution (5) to be

$$\begin{aligned} &p(F^{(1:K)} | X^{(1:K)}) \\ &\propto \prod_{k=1}^{K} \prod_{t=1}^{n} \left(\prod_{(i,j) \in E_{F^{(k)}}} \frac{p^{(k)}_{ij}(X^{(k)}_{ti}, X^{(k)}_{tj})}{p^{(k)}_{i}(X^{(k)}_{ti}) p^{(k)}_{j}(X^{(k)}_{tj})} \right. \\ &\qquad \left. \cdot \prod_{\ell \in V_{F^{(k)}}} p^{(k)}_{\ell}(X^{(k)}_{t\ell}) \right) \cdot \pi(F^{(1:K)}). \end{aligned} \tag{11}$$

Next, we design a prior distribution on the set of K spanning trees which reflects our belief that the structures across the K of them share some similarity. Again we use $F^{(k)}$ to denote the adjacency matrix of the corresponding graph, that is, $F^{(k)}_{ij} = 1$ if and only if $(i,j) \in E_{F^{(k)}}$. We consider the following hierarchical model:

$$\begin{aligned} \tau_{ij} &\sim \text{Beta}(\alpha, \beta) \text{ for all } i < j, \\ F^{(k)}_{ij} \,|\, \tau_{ij} &\sim \text{Bernoulli}(\tau_{ij}) \text{ for all } k \text{ and } i < j. \end{aligned}$$

According to this model, the same edge across multiple graphs is governed by the same parameter τ_{ij}, and hence encourage similarity across them. This essentially gives a prior distribution on $F^{(1:K)}$:

$$\begin{aligned} &\pi(F^{(1:K)}) \\ &\propto \prod_{i<j} \int_{\tau_{ij}} p(F_{ij} \,|\, \tau_{ij}) p(\tau_{ij}) d\tau_{ij} \cdot \mathbb{1}\{F^{(k)} \in \mathcal{F}_d \text{ for all } k\} \\ &\propto \prod_{i<j} B(\alpha + \|F_{ij}\|_1, \beta + K - \|F_{ij}\|_1) \\ &\quad \cdot \mathbb{1}\{F^{(k)} \in \mathcal{F}_d \text{ for all } k\}, \end{aligned}$$

where F_{ij} is the vector containing the (i,j)th entries of $F^{(k)}$ for $k = 1, \ldots, K$, $\|\cdot\|_1$ denotes the ℓ_1 norm, and $B(\cdot,\cdot)$ denotes the Beta function. Now combining this with (11) and following the reasoning in Subsection 3.1, we obtain our estimator in this case

$$\begin{aligned} \tilde{F}^{(1:K)}_{\pi} = \underset{F^{(k)} \in \mathcal{F}_d, \forall k}{\arg\max} \left\{ \sum_{k=1}^{K} \sum_{(i,j) \in E_{F^{(k)}}} \widehat{I}(X^{(k)}_i; X^{(k)}_j) \right. \\ \left. + \lambda \sum_{i<j} \log B(\alpha + \|F_{ij}\|_1, \beta + K - \|F_{ij}\|_1) \right\}. \end{aligned} \tag{12}$$

Note that we include an extra tuning parameter λ in front of the penalty term to give us a bit more flexibility in controlling its magnitude. The function $k \mapsto \log B(\alpha + k, \beta + K - k)$ is convex and takes larger values when k is close to 0 or K compared to those in between. Using it as a penalty thus favors the set of graphs which share common edges.

To solve (12), we again adopt a minorize-maximization procedure. Specifically, write the objective function as

$$\begin{aligned} f(F^{(1:K)}) = &\sum_{k=1}^{K} \sum_{i<j} w^{(k)}_{ij} F^{(k)}_{ij} \\ &+ \lambda \sum_{i<j} \log B(\alpha + \|F_{ij}\|_1, \beta + K - \|F_{ij}\|_1), \end{aligned}$$

where $w^{(k)}_{ij} = \widehat{I}(X^{(k)}_i; X^{(k)}_j)$. Given a current solution

$\check{F}^{(k)}$, we linearize $f(F^{(1:K)})$ at $\check{F}^{(k)}$ and get

$$f(F^{(1:K)}) \geq \sum_{k=1}^{K} \sum_{i<j} \Big(w_{ij}^{(k)} + \lambda \left(\psi(\alpha + \|\check{F}_{ij}\|_1) - \psi(\beta + K - \|\check{F}_{ij}\|_1)\right) \Big) F_{ij}^{(k)} + C,$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function. This gives the following weights updating rule:

$$\check{w}_{ij}^{(k)} = w_{ij}^{(k)} + \lambda \left(\psi(\alpha + \|\check{F}_{ij}\|_1) - \psi(\beta + K - \|\check{F}_{ij}\|_1)\right).$$

Note that $k \mapsto \psi(\alpha + k) - \psi(\beta + K - k)$ is an increasing function. Therefore, this updating rule borrows strength across the K graphs – it increases an edge's weight when $\|F_{ij}\|_1$ is large, i.e., when other graphs also have edge (i, j) present.

3.4 Algorithms

As a short conclusion, we summarize the two procedures, which share a lot of similarity but work for different applications, here in Algorithm 2 and 3. After getting the output of the algorithm, we will prune the resulting spanning tree to obtain a forest estimate (to avoid overfitting in high dimensions). We do this by going through the last iteration of the algorithm and stop at the step where the likelihood is maximized on a held-out dataset.

Algorithm 2 Scale-free graph estimation

input Weight matrix $W = (w_{ij})_{d\times d}$, tuning parameter λ
$F \leftarrow$ output of Algorithm 1 on W
do
$\check{w}_{ij} \leftarrow w_{ij} - \frac{\lambda}{\sum_{\ell=1}^{d} F_{i\ell}} - \frac{\lambda}{\sum_{\ell=1}^{d} F_{j\ell}}$
$F \leftarrow$ output of Algorithm 1 on $\check{W} = (\check{w}_{ij})_{d\times d}$
while F has not converged
output F

Algorithm 3 Joint estimation for multiple graphs

input Weight matrices $W^{(k)} = (w_{ij}^{(k)})_{d\times d}$ for $k = 1, \ldots, K$, tuning parameters λ, α, β
$F^{(k)} \leftarrow$ output of Algorithm 1 on $(w_{ij}^{(k)})_{d\times d}$ for $k = 1, \ldots, K$
do
$\check{w}_{ij}^{(k)} \leftarrow w_{ij}^{(k)} + \lambda \left(\psi(\alpha + \|F_{ij}\|_1) - \psi(\beta + K - \|F_{ij}\|_1)\right)$
$F^{(k)} \leftarrow$ output of Algorithm 1 on $\check{W}^{(k)} = (\check{w}_{ij})_{d\times d}$ for $k = 1, \ldots, K$
while $F^{(1:K)}$ have not converged
output $F^{(1:K)}$

4 Statistical Properties

In this section, we present a theoretical result on structure selection consistency of the scale-free graph estimation procedure. We follow a similar argument in Liu et al. (2011). The consistency result for joint estimation of multiple graphs is similar.

Instead of assuming the true density is Markov to a forest, we focus on the comparison of the estimated forest with the *oracle* forest, which minimize the risk. Specifically, let F_d^* be the optimal spanning tree within $\mathcal{F}_d$ that minimizes the negative log-likelihood loss. Let $\widehat{F}_{d,\lambda}^{\mathrm{SF}}$ be the scale-free spanning tree which is obtained from Algorithm 2.

To prove selection consistency, we need some assumptions on the true density function and the kernel functions. We give the detailed assumptions in the supplementary material. Essentially, Assumption 1 ensures that the univariate and bivariate densities are smooth with order β and can be lower and upper bounded. Assumption 2 assumes that the kernel function is well-behaved and β-valid (Tsybakov, 2008). In addition, we define the *crucial set* $\mathcal{T}$ be a set of pairs of edges $((i, j), (i', j'))$ such that $I(X_i; X_j) \neq I(X_{i'}; X_{j'})$ and with positive probability, flipping the relative order of $I(X_i; X_j)$ and $I(X_{i'}; X_{j'})$ changes the learned forest structure in the population Chow-Liu algorithm. We obtain the following result on selection consistency.

Theorem 1. *Suppose Assumption 1 and Assumption 2 in the supplementary material hold with β being the smooth parameter. Suppose further that*

$$\min_{((i,j),(i',j'))\in\mathcal{T}} |I(X_i; X_j) - I(X_{i'}; X_{j'})| \geq 6L_n,$$

where $L_n = \Omega\left(\sqrt{\frac{\log(n)+\log(d)}{n^{\beta/(1+\beta)}}}\right)$. If the tuning parameter $\lambda < L_n$, then we have as $n \to \infty$

$$\mathbb{P}(\widehat{F}_{d,\lambda}^{\mathrm{SF}} = F_d^*) \to 1.$$

We give the proof in the supplementary material. This theorem implies that the scale-free forest density estimation method satisfies the structure selection consistency if $d = o(\exp(n^{\beta/(1+\beta)}))$.

5 Related work

Before proceeding to present the performance of the proposed nonparametric methods on simulated and real datasets, we pause to review some of the existing approaches on estimation of scale-free graphical models and joint estimation of multiple graphical models.

Most existing methods for estimating graphical models with prior information assume that the data follow

	General	With prior information	
		Scale-free graph	Multiple graphs
Parametric	`Glasso` (Friedman et al., 2008)	`SFGlasso`* (Liu and Ihler, 2011) `HubGlasso`† (Tan et al., 2014)	`GuoGlasso`* (Guo et al., 2011) `JointGlasso`† (Danaher et al., 2014)
Nonparametric	`FDE` (Liu et al., 2011)	`SF-FDE`‡	`J-FDE`‡

$*$: non-convex method $\dagger$: convex method $\ddagger$: this paper

Table 1: Summary and comparison between different methods in graphical modeling.

multivariate Gaussian distributions. To encourage a scale-free graph, Liu and Ihler (2011) propose to replace the ℓ_1 penalty in the formulation of the graphical lasso by a non-convex power law regularization term. Along the same line, Defazio and Caetano (2012) impose a convex penalty by using submodular functions and their Lovász extension. Essentially, both methods try to penalize the log degree of each node, but end up using a continuous/convex surrogate to avoid the combinatorial problems involving the degrees. Tan et al. (2014) propose a general framework to accommodate networks with hub nodes, using a convex formulation that involves a row-column overlap norm penalty.

Methods for inferring Gaussian graphical models on multiple units have also been proposed in recent years. Guo et al. (2011) propose a method for joint estimation of Gaussian graphical models by penalizing the graphical lasso objective function by the square root of ℓ_1 norms of the edge vector across all graphs, which results in a non-convex problem. A convex joint graphical lasso approach is developed in Danaher et al. (2014), which is based on employing generalized fused lasso or group lasso penalties. Peterson et al. (2015) and Zhu and Barber (2015) propose Bayesian approaches for inference on multiple Gaussian graphical models.

We summarize in Table 1 the aforementioned methods, which will be implemented and compared next. Methods proposed in this paper can be viewed as nonparametric counterparts to the parametric methods.

6 Experiments

6.1 Synthetic data

In this subsection, we evaluate the performance of the proposed methods and other existing methods on synthetic data.

Graph structures We consider the following types of graph structures with $d = 100$ vertices.

- **Scale-free graph**: We use a preferential attachment process to generate a scale-free graph (Albert and Barabási, 2002). We start with a chain of 4 nodes (i.e., with edges 1–2, 2–3, and 3–4). New nodes are added one at a time, and each new node is connected to one existing node with probability $p_i \propto \delta_i^\alpha$, where δ_i is the current degree of the ith node, and α is a parameter, which we set to be 1.5 in our experiments. A typical realization of such networks is shown in Figure 1 (left).
- **Stars**: The graph has 5 stars of size 20; each star is a tree with one root and 19 leaves. An illustration is shown in Figure 1 (right).
- **Multiple graphs**: We follow the above two mechanisms to generate multiple graphs with similar structures. In particular, we generate a set of $K = 3$ scale-free graphs, which share 80 common edges (this is done by applying the above generative model to grow a common tree of size 80 to be shared across the 3 units; each unit then continues this growing process independently until obtaining a tree of 100 vertices), and another set of $K = 3$ stars graphs, which have 4 common stars and one individual star with distinct roots.

We also consider scenarios where the true graph is not forest. The results are included in the supplementary material.

Probability distributions Given a particular graph, we generate 200 samples according to two types of probability distributions that are Markov to the

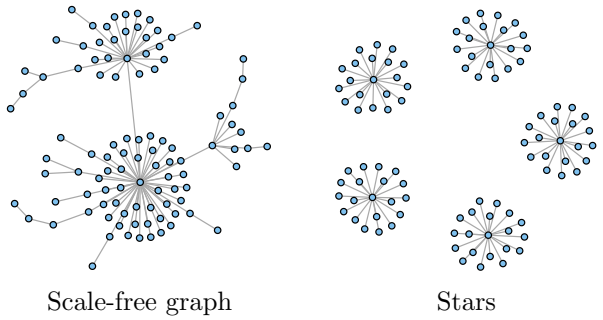


Figure 1: An illustration of simulated graph patterns.

Graphs with hubs

Graph × Dist.	FDE	SF-FDE	Glasso		SFGlasso		HubGlasso	
			held-out	oracle	held-out	oracle	held-out	oracle
Scale-free × $\mathcal{N}$	0.79	0.92	0.86	0.91	0.88	0.92	0.84	0.88
Stars × $\mathcal{N}$	0.82	0.96	0.90	0.93	0.96	0.98	0.97	0.99
Scale-free × t	0.89	0.98	0.05	0.43	0.28	0.53	0.53	0.55
Stars × t	0.93	0.98	0.52	0.56	0.65	0.67	0.79	0.79

Multiple graphs

Graph × Dist.	FDE	J-FDE	Glasso		GuoGlasso		JointGlasso	
			held-out	oracle	held-out	oracle	held-out	oracle
Scale-free × $\mathcal{N}$	0.78	0.90	0.85	0.92	0.97	0.97	0.95	0.97
Stars × $\mathcal{N}$	0.80	0.92	0.89	0.92	0.94	0.95	0.89	0.96
Scale-free × t	0.91	0.98	0.03	0.44	0.54	0.64	0.65	0.66
Stars × t	0.92	0.98	0.47	0.53	0.67	0.70	0.70	0.71

Table 2: Averaged F_1 scores for methods applied on the simulated data.

graph: Gaussian copulas and t copulas (Demarta and McNeil, 2005). The Gaussian copula (resp., the t copula) can be thought of as representing the dependence structure implicit in a multivariate Gaussian (multivariate t) distribution, while each variable follows a uniform distribution on $[0, 1]$ marginally. Since the graph structures we consider are trees or forests, we generate the data sequentially, first sampling for an arbitrary node in a tree, and then drawing samples for the neighboring nodes according to the conditional distribution given by the copula until going through all nodes in the tree. In our simulations, the degree of freedom of the t copula is set to be 1, and the correlation coefficients are chosen to be 0.4 and 0.25 for the Gaussian and the t copula.

Methods We implement methods that are summarized in Table 1. For the forest-based methods, we use a held-out set of size 100 to select tuning parameter and prune the estimated spanning trees. To implement the Gaussian-based methods, we first transform the data marginally to be approximately Gaussian. We choose the tuning parameters by searching through a fine grid, and selecting those that maximize the likelihood on the held-out set. We refer to this as *held-out tuning*. The results obtained by the held-out tuning reflect the performance of the methods in a fully data-driven way. To prevent over-selection of edges, we use the *refit* method, which is a two-step procedure – in the first step, a sparse precision matrix is obtained; in the second step, a Gaussian model is refitted without regularization, but enforcing the sparsity pattern obtained in the first step. In addition, we also consider what we call *oracle tuning*, where the tuning parameters are chosen to maximize the F_1 score of the estimated graph. An F_1 score is the harmonic mean of a method's precision and recall and hence a measure of its accuracy. It's a number between 0 and 1; a higher score means better accuracy and 1 means perfect labelling. This tuning method requires the knowledge of the true graph, and hence it's not obvious that there would exist a data-driven way to achieve this. We include the oracle tuning since it reflects the optimal performance that can be possibly achieved by the methods.

Results For both scale-free graphs and multiple graphs, we carry out four sets of experiments, with data generated from the two types of graphs and the two types of distributions. For each set of experiments, we repeat the simulations 10 times and record the F_1 scores of the estimated graphs for each method. The average F_1 scores are shown in Table 2. From the table, we see that SF-FDE and J-FDE always outperform FDE on these particular situations. Also, SF-FDE and J-FDE perform as well as or better than the other three methods. In particular, when the true copula is Gaussian, the graphical lasso-based methods all have very high scores; they fail to deliver good performance when the true copula is no longer Gaussian. On the other hand, the forest-based methods are not affected too much by the true distribution.

6.2 Real data

Stock price data We test our methods on the daily closing prices for $d = 417$ stocks that are constantly in the S&P 500 index from Yahoo! Finance. The log returns of each stock are replaced by their respective

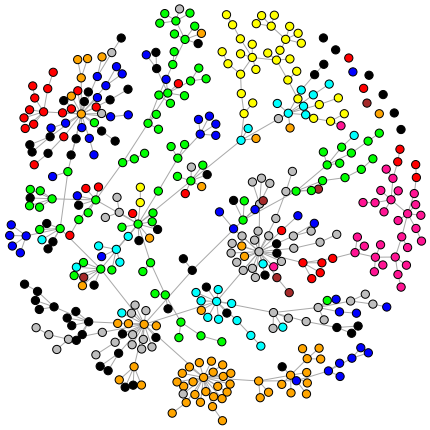


Figure 2: Estimated graph for SF-FDE applied on the stock price data. The stocks are colored according to their Global Industry Classification Standard categories.

normal scores, subject to a Winsorized truncation.

In the application of learning scale-free forests, we use the data from the first 9 months of 2014 as the training data and the data from the last 3 months of 2014 as the held-out data. The result turns out that SF-FDE yields a larger held-out log-likelihood than FDE (64.5 compared to 62.6), implying that a scale-free approximation is helpful in predicting the relationships. The estimated graph by SF-FDE is shown in Figure 2. We see that the resulting clusters tend to be consistent with the Global Industry Classification Standard categories, which are indicated by different colors in the graph. To complete the comparison, we include in the supplementary material results for fitting the stock price data using Gaussian-based methods, which, however, do not provide as interpretable results as the tree-based methods.

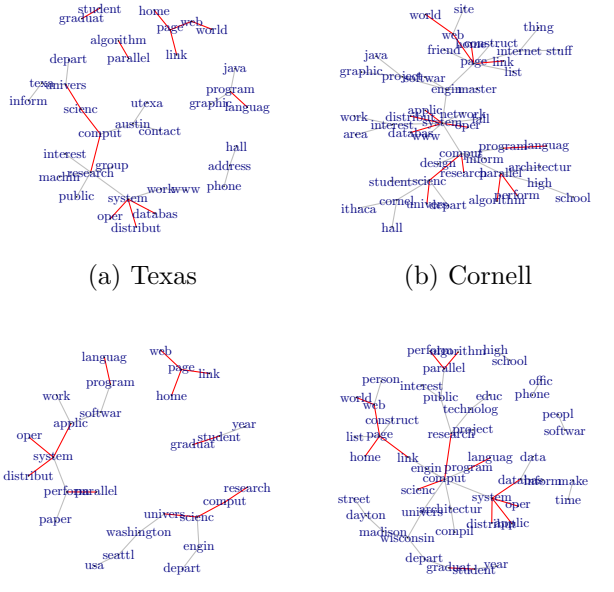
We also consider the application of learning multiple forests by dividing the data into 4 periods from 2009 to 2012, one for a year, and model the 4-unit data using our proposed method. The aggregated held-out log-likelihood over the 4 units are 193.4 for J-FDE and 185.5 for FDE. The numbers of common edges across the 4 graphs are 111 for J-FDE and 24 for FDE, respectively. The plots are included in the supplementary material.

University webpage data As a second example, we apply our methods to the university webpage data from the "World Wide Knowledge Base" project at Carnegie Mellon University, which consists of the occurrences of various terms on student webpages from 4 computer science departments at Texas, Cornell, Washington, and Wisconsin. We choose a subset of 100 terms with the largest entropy. In the analysis, we compute the empirical distributions instead of kernel density estimates since the data is discrete.

To understand the relationships among the terms, we first wish to identify terms that are hubs. Results show that SF-FDE detects 4 highly connected nodes of degree greater than 10: *comput*, *system*, *page*, and *interest*. Then we model the 4-unit data, one for a university. Figure 3 shows the estimated graphs by J-FDE (isolated nodes are not displayed in each graph). These results provides an intuitive explanation of the relationships among the terms across the 4 universities.

7 Conclusion

In this paper, we introduce a nonparametric framework for incorporating prior knowledge to assist estimation of graphical models. Instead of Gaussianity assumptions, it assumes the density is Markov to a forest, thus allowing arbitrary distribution. A key ingredient is to design a prior distribution on graphs that favors those consistent with the prior belief. We illustrate the idea by proposing such prior distributions, which lead to two algorithms, for the problems of estimating scale-free networks and multiple graphs with similar structures. An interesting future direction is to apply this idea to more applications and different types of prior information.



(a) Texas (b) Cornell

(c) Washington (d) Wisconsin

Figure 3: Estimated graphs for J-FDE applied on the university webpage data. Edges shared by at least 3 units are colored in red.

References

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

C Chow and C Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

Aaron Defazio and Tiberio S Caetano. A convex formulation for learning scale-free networks via submodular relaxation. In *Advances in Neural Information Processing Systems*, pages 1250–1258, 2012.

Stefano Demarta and Alexander J McNeil. The t copula and related copulas. *International Statistical Review*, 73(1):111–129, 2005.

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

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1): 30–37, 2004.

Mikko Koivisto and Kismat Sood. Exact bayesian structure discovery in bayesian networks. *Journal of Machine Learning Research*, 5(May):549–573, 2004.

Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1): 48–50, 1956.

Han Liu, Min Xu, Haijie Gu, Anupam Gupta, John Lafferty, and Larry Wasserman. Forest density estimation. *The Journal of Machine Learning Research*, 12:907–951, 2011.

Qiang Liu and Alexander T Ihler. Learning scale free networks by reweighted ℓ_1 regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 40–48, 2011.

Christine Peterson, Francesco C. Stingo, and Marina Vannucci. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

Kean Ming Tan, Palma London, Karthik Mohan, Su-In Lee, Maryam Fazel, and Daniela Witten. Learning graphical models with hubs. *The Journal of Machine Learning Research*, 15(1):3297–3331, 2014.

Qingming Tang, Siqi Sun, and Jinbo Xu. Learning scale-free networks by dynamic node specific degree prior. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2247–2255, 2015.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics, 1st edition, 2008.

Yang Zhou. Structure learning of probabilistic graphical models: a comprehensive survey. *arXiv preprint arXiv:1111.6925*, 2011.

Yuancheng Zhu and Rina Foygel Barber. The log-shift penalty for adaptive estimation of multiple Gaussian graphical models. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 1153–1161, 2015.