Learning Nonparametric Forest Graphical Models with Prior Information: Supplementary Material

1 Proof of Proposition 1

The following proposition shows the convexity of the function $\log B(\alpha + x, \beta + K - x)$.

Proposition 1. *For any $\alpha > 0, \beta > 0$, $-\log B(\alpha + x, \beta + K - x)$ is concave for $0 \leq x \leq K$.*

Proof. Note that $\log B(\alpha + x, \beta + K - x) = \log \Gamma(\alpha + x) + \log \Gamma(\beta + K - x) - \log \Gamma(\alpha + \beta + K)$. Using the fact that the logarithm of Gamma function is convex on positive real numbers, for any $0 \leq x_1, x_2 \leq K$ and $t \in [0, 1]$, we have

$$\begin{aligned}
&\log \Gamma(\alpha + tx_1 + (1-t)x_2) + \log \Gamma(\beta + K - tx_1 - (1-t)x_2) \\
&= \log \Gamma(t(\alpha + x_1) + (1-t)(\alpha + x_2)) + \log \Gamma(t(\beta + K - x_1) + (1-t)(\beta + K - x_2)) \\
&\leq t \log \Gamma(\alpha + x_1) + (1-t) \log \Gamma(\alpha + x_2) + t \log \Gamma(\beta + K - x_1) + (1-t) \log \Gamma(\beta + K - x_2) \\
&= t(\log \Gamma(\alpha + x_1) + \log \Gamma(\beta + K - x_1)) + (1-t)(\log \Gamma(\alpha + x_2) + \log \Gamma(\beta + K - x_2))).
\end{aligned}$$

Thus the proposition follows. □

2 Assumptions and Proof for Theorem 1

2.1 Assumptions

We follow Liu et al. (2011) to introduce the assumptions on the density and on the kernel function.

Let p^* be the true density and $\mathcal{X}_i$ be the domain of p_i^*. Fix $\beta > 0$, let $\Sigma(\beta, L, r, x_0)$ be the locally Hölder class of functions with degree β, Hölder constant L, radius r and center x_0. Given a set A, define $\Sigma(\beta, L, r, A) = \cap_{x_0 \in A} \Sigma(\beta, L, r, x_0)$. We need the following two assumptions on the true density.

Assumption 1 (Assumption on the density). *For any $1 \leq i < j \leq d$, we assume*

1. *there exists $L_1 > 0$ and $L_2 > 0$ such that for any $c > 0$ the true bivariate and univariate densities satisfy*

$$p^*(x_i, x_j) \in \Sigma(\beta, L_2, c(\log n/n)^{\frac{1}{2\beta+2}}, \mathcal{X}_i \times \mathcal{X}_j)$$

and

$$p^*(x_i) \in \Sigma(\beta, L_1, c(\log n/n)^{\frac{1}{2\beta+2}}, \mathcal{X}_i);$$

2. *there exist two constants* c_1 *and* c_2 *such that*

$$c_1 \leq \inf_{x_1 \times x_2 \in \mathcal{X}_i \times \mathcal{X}_j} p^*(x_i, x_j) \leq \sup_{x_1 \times x_2 \in \mathcal{X}_i \times \mathcal{X}_j} p^*(x_i, x_j) \leq c_2$$

almost surely.

Next, we state the assumptions on the kernel functions.

Assumption 2 (Assumptions on the kernel)**.** *We assume the kernel* K *satisfies*

1. $\int K(u)du = 1$, $\int K^2(u)du \leq \infty$ *and* $\sup_u K(u) \leq c$ *for some constant* c.
2. K *is a finite linear combination of functions* g *whose epigraphs* $\mathrm{epi}(g) = \{(s, u) : g(s) \geq u\}$, *can be represented as a finite number of Boolean operations (union and intersection) among sets of the form* $\{(s, u) : Q(s, u) \geq \phi(u)\}$, *where* Q *is a polynomial on* $\mathbb{R} \times \mathbb{R}$ *and* ϕ *is an arbitrary real function.*
3. K *has a compact support and for any* $\ell \geq 1$ *and* $1 \leq \ell' \leq \lfloor \beta \rfloor$

$$\int |t|^\beta |K(t)| dt < \infty, \ \textit{and} \int |K(t)|^\ell dt < \infty, \ \int t^{\ell'} K(t) dt = 0.$$

The assumptions on the kernel are mild. For example, the boxcar kernel satisfies the three assumptions.

2.2 Proof

Here we give the proof of Theorem 1.

Proof. Denote $\widehat{F}_{d,\lambda}^{\mathtt{SF},(t)}$ as the spanning tree structure learned from the tth iteration of Algorithm 2, with the edge weight being $\breve{w}_{ij}^t = \widehat{I}(X_i; X_j) - \frac{\lambda}{\sum_{l=1}^d \Theta_{il}^{(t-1)}} - \frac{\lambda}{\sum_{l=1}^d \Theta_{jl}^{(t-1)}}$. Here, $\Theta^{(t-1)}$ is the adjacency matrix for $\widehat{F}_{d,\lambda}^{\mathtt{SF},(t-1)}$. Following a similar argument in Liu et al. (2011), note that the event $\widehat{F}_{d,\lambda}^{\mathtt{SF},(t)} \neq F_d^*$ implies that there must exist some pair of edges (i, j) and (i', j') such that

$$(\breve{w}_{ij}^t - \breve{w}_{i'j'}^t) \cdot (I(X_i; X_j) - I(X_{i'}; X_{j'})) \leq 0.$$

Recall that $\mathcal{T}$ is 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. Apply a union bound to obtain

$$\begin{aligned}
&\mathbb{P}((\breve{w}_{ij}^t - \breve{w}_{i'j'}^t) \cdot (I(X_i; X_j) - I(X_{i'}; X_{j'})) \leq 0, \text{ for some } (i, j), (i', j')) \\
&\quad \leq d^4 \max_{((i,j),(i',j')) \in \mathcal{T}} \mathbb{P}((\breve{w}_{ij}^t - \breve{w}_{i'j'}^t) \cdot (I(X_i; X_j) - I(X_{i'}; X_{j'})) \leq 0).
\end{aligned}$$

Under the assumption that

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

we obtain

$$\max_{((i,j),(i',j'))\in\mathcal{T}} \mathbb{P}((\check{w}_{ij}^t - \check{w}_{i'j'}^t)\cdot(I(X_i;X_j) - I(X_{i'};X_{j'})) \le 0) \le \max_{i\neq j} \mathbb{P}(|I(X_i;X_j) - \check{w}_{ij}^t| > 3L_n).$$

Further observe that

$$\begin{aligned}|I(X_i;X_j) - \check{w}_{ij}^t| &\le |I(X_i;X_j) - \widehat{I}(X_i;X_j)| + |\widehat{I}(X_i;X_j) - \check{w}_{ij}^t| \\ &\le |I(X_i;X_j) - \widehat{I}(X_i;X_j)| + 2\lambda.\end{aligned}$$

The second inequality follows from the fact that $\sum_{l=1}^{d}\Theta_{ul}^{(t-1)} \ge 1$ for $u = i$ or j. Under the assumption that $\lambda < L_n$, we have

$$\max_{i\neq j}\mathbb{P}(|I(X_i;X_j) - \check{w}_{ij}^t| > 4L_n) \le \max_{i\neq j}\mathbb{P}(|I(X_i;X_j) - \widehat{I}(X_i;X_j)| > L_n).$$

Putting together the above arguments to obtain

$$\begin{aligned}\mathbb{P}(\widehat{F}_{d,\lambda}^{\mathtt{SF},(t)} \neq F_d^*) &\le d^4\cdot\max_{i\neq j}\mathbb{P}(|I(X_i;X_j) - \widehat{I}(X_i;X_j)| > L_n) \\ &\le o(\exp(4\log d - c\cdot(\log n)^{1/(1+\beta)}\log d)) \\ &= o(1),\end{aligned}$$

where the last inequality follows from Lemma 23 in Liu et al. (2011) and the definition of L_n. □

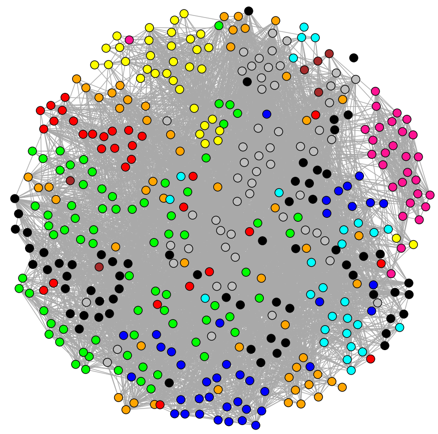
3 More Results on Synthetic Data Analysis

When the true graphs are not trees, the forest-based method can be shown to yield optimal (in the sense of KL divergence) forest approximation of the true graph. For joint learning of multiple graphs, we conduct additional analysis on synthetic Gaussian data where the true graph structure is “random”. Given the adjacency matrix Θ, the graph patterns are generated as below: each pair of off-diagonal elements are randomly set $\Theta_{ij} = \Theta_{ji} = 1$ for $i \neq j$ with probability 0.02, and 0 otherwise. This leads to about $0.01\cdot d(d-1)$ edges in the graph. The rest of simulation setup is similar to that in the main manuscript. In particular, we generate a set of $K = 3$ random graphs each with $d = 100$ nodes, which share common subgraph with 80 nodes. The result turns out that the average F_1 score for `FDE` is 0.83 and the average F_1 score for `J-FDE` is 0.92.

4 More Results on Real Data Analysis

When we fit forest graphical models to real datasets, we do not necessarily assume or expect the true structures to be forest. What we do believe though is that a forest provides a close approximation to the truth. For the stock and webpage data, we think our fitted graphs serve as a concise and interpretable “skeleton” for the true graphs.

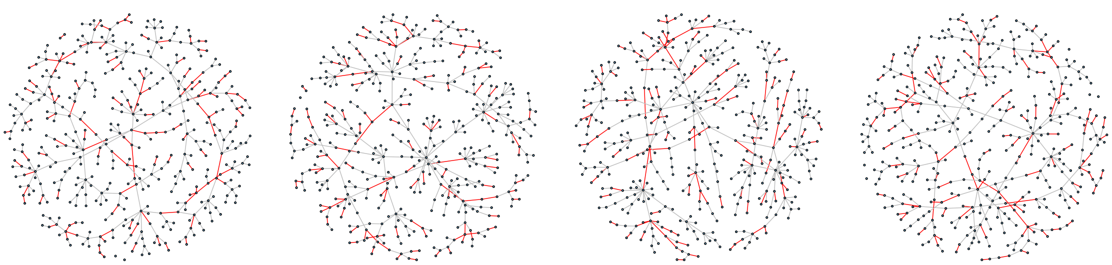
Stock price data Figure 1 shows the estimated graphs by `Glasso` and `SFGlasso` on stock price data, where we use the refit method to determine the tuning parameters, as in the simulation studies. The resulting estimated graphs are much less interpretable comparing to that obtained by



`Glasso` `SFGlasso`

Figure 1: Estimated graphs for `Glasso` and `SFGlasso` applied on the stock price data. The stocks are colored according to their Global Industry Classification Standard categories.

`SF-FDE`. The forest-based method usually provides a "skeleton" for the true graph, which helps to understand the structure and to identify possible hubs and clusters. Figure 2 displays the estimated graphs by `J-FDE` on stock price data. Common edges across the 4 graphs are colored in red.



(a) 2009 (b) 2010 (c) 2011 (d) 2012

Figure 2: Estimated graphs by `J-FDE` on stock price data. Common edges across the 4 graphs are colored in red.

University webpage data Figure 3 shows the estimated graph by `SF-FDE` on webpage data. Given the discrete data in this analysis, it is not appropriate to apply Gaussian-based methods. This further highlights the advantages of our nonparametric forest-based method.

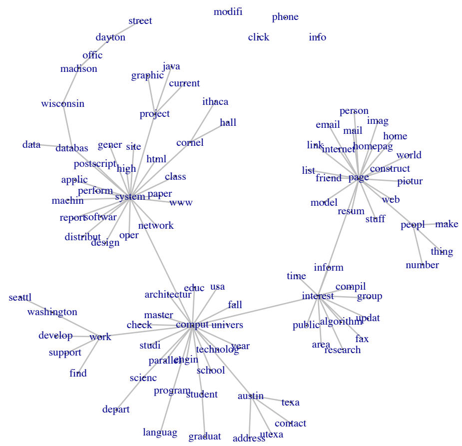


Figure 3: Estimated graph by `SF-FDE` on webpage data.

5 R Package for Scale-Free Graphical Model Estimation

An implementation of the proposed scale-free graphical model estimation method is available as an R package at `http://github.com/zhejosephliu/scalefreeForest`.

References

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.