Supplementary Material: Learning High-dimensional Gaussian Graphical Models under Total Positivity without Adjustment of Tuning Parameters

A Additional discussion of Condition 3.1

In this section, we explain why a sufficient condition for "$\lambda_{\min}(\Sigma_S) \geq \sigma_{\min}$" is that all diagonal entries of Θ scale as constants:

When all diagonal entries of Θ scale as constants, standard results on the Schur complement yield that all diagonal entries in $(\Sigma_S)^{-1}$ also scale as constants. Hence, $\lambda_{\max}((\Sigma_S)^{-1}) \leq \mathrm{trace}((\Sigma_S)^{-1}) = \sum_{i \in S}[(\Sigma_S)^{-1}]_{ii}$ is also upper bounded by a constant (since $|S| \leq d+4$). By combining this with the fact that $\lambda_{\min}(\Sigma_S) = \lambda_{\max}^{-1}(\Sigma_S^{-1})$, we can conclude that $\lambda_{\min}(\Sigma_S) \geq \sigma_{\min}$ for a positive constant $\sigma_{\min}$.

B Proof of Lemma 3.7

B.1 Characterization of maximal overlaps

Our proof of Lemma 3.7 relies on the following lemma that characterizes the size of maximal overlaps between any two batches.

Lemma B.1 (Tail-bounds on maximum overlap of subsets)**.** *Consider a set of data $B := \{x^{(i)}\}_{i=1}^{N}$ with size N. Let $B_1, \cdots, B_K \subseteq B$ denote K subsets where each B_k is created by uniformly drawing M samples from the set B, then $\forall \epsilon > 0$,*

$$\Pr\left(\max_{i,j} |B_i \cup B_j| < \frac{M^2}{N} + \epsilon N\right) \geq 1 - \exp(-2\epsilon^2 N + 2\log K).$$

Proof. By union bound, we have for any $T > 0$,

$$\Pr(\max_{i,j} |B_i \cap B_j| > T) \leq \binom{K}{2} \Pr(|B_i \cap B_j| > T). \tag{1}$$

For any $i \neq j$, let the random variable $y_\ell := \mathbf{1}\{x^{(\ell)} \in B_i\} \cdot \mathbf{1}\{x^{(\ell)} \in B_j\}$, it follows that $|B_i \cap B_j| = \sum_{l=1}^{N} y_\ell$ and thus

$$\Pr\left(|B_i \cap B_j| > T\right) = \Pr\left(\sum_{\ell=1}^{N} y_\ell > T\right). \tag{2}$$

In addition, y_ℓ is a binary variable satisfying $\Pr(y_\ell = 1) = \left(\frac{M}{N}\right)^2$.

In this case, it suffices to provide an upper bound on the probability $\Pr\left(\sum_{\ell=1}^{N} y_\ell > T\right)$. Using basic results in combinatorics, one can rewrite the conditional probability $\Pr(y_\ell = 1 | y_{\ell'} = 1)$ as follows:

$$\Pr(y_\ell = 1 | y_{\ell'} = 1) = \frac{|\{B_i : x^{(\ell')}, x^{(\ell)} \in B_i\}| \cdot |\{B_j : x^{(\ell')}, x^{(\ell)} \in B_j\}|}{|\{B_i : x^{(\ell')} \in B_i\}| \cdot |\{B_j : x^{(\ell')} \in B_j\}|} = \frac{\binom{N-2}{M-2}^2}{\binom{N-1}{M-1}^2}.$$

It follows that

$$\Pr(y_\ell = 1|y_{\ell'} = 1) = \frac{\binom{N-2}{M-2}^2}{\binom{N-1}{M-1}^2} = \left(\frac{M-1}{N-1}\right)^2 \leq \left(\frac{M}{N}\right)^2 = \Pr(y_\ell = 1),$$

which means for any $\ell \neq \ell'$, the random variables y_ℓ and $y_{\ell'}$ are negatively correlated. By applying Chernoff-Hoeffding bounds on sum of negatively associated random variables (see e.g. (Dubhashi et al., 1996, Theorem 14)), we obtain

$$\Pr\left(\sum_{\ell=1}^{N}(y_\ell - \mathbb{E}(y_\ell)) > \epsilon N\right) \leq \exp(-2\epsilon^2 N). \tag{3}$$

Combining (1), (2) and (3) and that $\mathbb{E}\left(y_\ell\right) = \frac{M^2}{N^2}$, we obtain the statement in the lemma. □

B.2 Proof of Lemma 3.7

Notations and proof ideas for Lemma 3.7. To simplify notation, we denote each $\hat{\rho}_{ij|S_k}$ as $\hat{\rho}_k$ and denote the subset of data points used to estimate $\hat{\rho}_k$ as B_k. Let $\hat{\Sigma}_k \in \mathbb{R}^{|S_k|+2\times|S_k|+2}$ denote the sample covariance matrix of the nodes $S_k \cup \{i, j\}$. Note that here $\hat{\Sigma}_k$ is estimated from the data in B_k. Let $\hat{\sigma}_k$ denote the vectorized form of $\hat{\Sigma}_k$ and let σ_k denote the expectation of $\hat{\sigma}_k$. Standard results in calculating partial correlation coefficients show that $\hat{\rho}_k$ can be taken as a function of $\hat{\sigma}_k$, which we denote as

$$\hat{\rho}_k = g_k(\hat{\sigma}_k).$$

Moreover, since the derivatives of all orders of $g_k(\cdot)$ at the point $\hat{\sigma}_k$ can be expressed as polynomials of $\hat{\sigma}_k$ and its inverse (see e.g. Eq. 36 in Wasserman et al. (2014) and the two equations after that), $g_k(\cdot)$ is infinitely differentiable whenever the inputs are non-singular matrices. Let ℓ_k denote the first order derivative of g_k at the point σ_k. It follows that $\ell_k(\hat{\sigma}_k - \sigma_k)$ is the first order approximation of $g_k(\hat{\sigma}_k)$. Let the residual

$$r_k := g_k(\hat{\sigma}_k) - \ell_k(\hat{\sigma}_k - \sigma_k). \tag{4}$$

Let $\|\hat{\sigma}_k - \sigma_k\|_\infty$ denote the ℓ_∞ norm of the vector $\hat{\sigma}_k - \sigma_k$. Standard results in Taylor expansion show that when $\|\hat{\sigma}_k - \sigma_k\|_\infty$ is negligible, one can rewrite the residual as

$$r_k = \frac{1}{2}(\hat{\sigma}_k - \sigma_k)^T H_k(\tilde{\sigma}_k)(\hat{\sigma}_k - \sigma_k),$$

where $H_k(\cdot)$ is the Hessian matrix of g_k and $\tilde{\sigma}_k$ is some point in the middle between $\hat{\sigma}_k$ and σ_k. Let $\boldsymbol{\rho} := (\hat{\rho}_1, \cdots, \hat{\rho}_K)^T$, $\mathbf{L} := (\ell_1(\hat{\sigma}_1 - \sigma_1), \cdots, \ell_K(\hat{\sigma}_K - \sigma_K))^T$ and $\mathbf{R} := (r_1, \cdots, r_K)^T$. Since each $\hat{\sigma}_k$ is estimated using a subset of data with batch size M, there may be overlaps between the set of data used to calculate different $\hat{\sigma}_k$'s. Let $\hat{\sigma}_k^{(1)}$ denote the sample covariance matrix estimated from the data in $B_k \backslash \big(\underset{k' \neq k}{\cup} B_{k'}\big)$ and let $\hat{\sigma}_k^{(2)}$ denote the sample covariance matrix estimated from the data in the overlaps, i.e., the data in $B_k \cap \big(\underset{k' \neq k}{\cup} B_{k'}\big)$. Then one can decompose $\hat{\sigma}_k$ as $\hat{\sigma}_k = \frac{M-T_k}{M}\hat{\sigma}_k^{(1)} + \frac{T_k}{M}\hat{\sigma}_k^{(2)}$, where T_k is the size of data in the overlaps. It is obvious that the $\hat{\sigma}_k^{(1)}$'s are independent from each other. Based on the above decomposition, we denote $\mathbf{L} = \mathbf{L}^{(1)} + \mathbf{L}^{(2)}$, where

$$\mathbf{L}^{(1)} := \Big(\frac{M-T_k}{M}\ell_1(\hat{\sigma}_1^{(1)} - \sigma_1), \cdots, \frac{M-T_k}{M}\ell_K(\hat{\sigma}_K^{(1)} - \sigma_K)\Big)^T$$

and

$$\mathbf{L}^{(2)} := \Big(\frac{T_k}{M}\ell_1(\hat{\sigma}_1^{(2)} - \sigma_1), \cdots, \frac{T_k}{M}\ell_K(\hat{\sigma}_K^{(2)} - \sigma_K)\Big)^T.$$

In addition, for any vector $\mathbf{a}$, we write $\mathbf{a} \geq 0$ whenever all elements of the vector $\mathbf{a}$ are greater than or equal to zero.

Let the random event

$$\mathcal{B} := \Big\{(B_1, \cdots, B_K) : \max_{k,k' \in [K]} |B_k \cap B_{k'}| \leq 2\frac{M^2}{N}\Big\}.$$

By applying Lemma B.1, it follows that there exists some positive constant C that depends on γ such that $\Pr(\mathcal{B}) \geq 1 - \exp(-CN^{4\gamma-3})$. By combining this with the decomposition, we have

$$\begin{aligned}\Pr(\boldsymbol{\rho} \geq 0) &= \Pr(\boldsymbol{\rho} \geq 0, \mathcal{B}) + \Pr(\boldsymbol{\rho} \geq 0, \neg\mathcal{B}) \leq \Pr(\boldsymbol{\rho} \geq 0 \mid \mathcal{B}) \Pr(\mathcal{B}) + \Pr(\neg\mathcal{B}) \\ &\leq \Pr(\boldsymbol{\rho} \geq 0 \mid \mathcal{B}) + \Pr(\neg\mathcal{B}),\end{aligned}$$

where $\neg\mathcal{B}$ denotes the complement of the random event $\mathcal{B}$. It is sufficient to prove Lemma 3.7 by proving

$$\Pr(\boldsymbol{\rho} \geq 0 \mid \mathcal{B}) \leq \exp(-CN^{\frac{1-\gamma}{2}}) \tag{5}$$

for some positive constant C that depends on $\sigma_{\max}$, $\sigma_{\min}$ and d. In other words, it remains to prove that $\Pr(\boldsymbol{\rho} \geq 0) \leq \exp(-CN^{\frac{1-\gamma}{2}})$ when we are under a *particular* subsampling assignment $(B_1, \cdots, B_K)$ that is in the random event $\mathcal{B}$.

Preliminary lemmas for Lemma 3.7.

Since the only remaining task is to deal with Eq. (5), for the remainder of the proof of Lemma 3.7 we can assume that we are under a *particular* subsampling assignment $(B_1, \cdots, B_K)$ in $\mathcal{B}$. To simplify notation we omit "$\mid \mathcal{B}$" in the remainder of the proof.

Lemma B.2. *For all $\epsilon > 0$, there exists some positive constant C that depends on d, $\sigma_{\max}$ and $\sigma_{\min}$ such that the following inequality holds:*

$$\Pr(\|\hat{\sigma}_k - \sigma_k\|_\infty > \epsilon) \leq 2(d+2)^2 e^{-CM\epsilon^2}.$$

Proof. This is a direct consequence of (Wasserman et al., 2014, Lemma 7) and the Gaussianity of the underlying distribution. □

Lemma B.3. *For all $\epsilon > 0$, there exist positive constants C_1 and C_2 that depend on $\sigma_{\max}$, $\sigma_{\min}$ and d such that*

$$\Pr(\|\mathbf{R}\|_\infty \leq \epsilon) \geq 1 - 2(d+2)^2 e^{\frac{1-\gamma}{2}\log N - C_1 M\epsilon} - 2(d+2)^2 e^{\frac{1-\gamma}{2}\log N - C_2\sqrt{M}}.$$

Proof. For each r_k, let $C_1 - C_3$ denote a positive constant that depends on $\sigma_{\min}$, $\sigma_{\max}$ and d and may vary from line to line. We have that

$$\begin{aligned}\Pr(|r_k| > \epsilon) &= \Pr(|r_k| > \epsilon, \|\hat{\sigma}_k - \sigma_k\|_\infty \leq M^{-1/4}) + \Pr(|r_k| > \epsilon, \|\hat{\sigma}_k - \sigma_k\|_\infty \geq M^{-1/4}) \\ &\leq \Pr((|r_k| > \epsilon, \|\hat{\sigma}_k - \sigma_k\|_\infty \leq M^{-1/4}) + \Pr(\|\hat{\sigma}_k - \sigma_k\|_\infty \geq M^{-1/4}).\end{aligned} \tag{6}$$

Under the random event where $\|\hat{\sigma}_k - \sigma_k\|_\infty \leq M^{-1/4}$, standard results in Taylor expansion show that r_k can be expressed in the form $r_k = (\hat{\sigma}_k - \sigma_k)^T H_k(\tilde{\sigma}_k)(\hat{\sigma}_k - \sigma_k)$. Thus one can rewrite (6) as

$$\begin{aligned}\Pr(|r_k| > \epsilon) \leq \Pr\Big(\Big|\frac{1}{2}(\hat{\sigma}_k - \sigma_k)^T H_k(\tilde{\sigma}_k)(\hat{\sigma}_k - \sigma_k)\Big| > \epsilon, \|\hat{\sigma}_k - \sigma_k\|_\infty \leq M^{-1/4}\Big) \\ + \Pr(\|\hat{\sigma}_k - \sigma_k\|_\infty \geq M^{-1/4}).\end{aligned}$$

Under the random event $\|\hat{\sigma}_k - \sigma_k\|_\infty \leq M^{-1/4}$, $\tilde{\sigma}_k$ is in the middle of $\hat{\sigma}_k$ and σ_k. It follows that $\|\tilde{\sigma}_k - \sigma_k\|_\infty \leq M^{-1/4}$. By combining this with the fact that the Hessian function $H_k(\cdot)$ is infinitely differentiable at the point σ_k, there exists some positive constant C_1 such that $\|H_k(\tilde{\sigma}_k) - H_k(\sigma_k)\|_\infty \leq C_1$. Using that $\|H_k(\sigma_k)\|_\infty$ is also bounded by a positive constant (since it is a function of σ_k, see e.g. (Wasserman et al., 2014, Section 6.5) and (Magnus and Neudecker, 1988, Page 185) for the explicit form), we further obtain that $\|H_k(\tilde{\sigma}_k)\|_\infty \le C_1$. As a consequence, one can further rewrite (6) as

$$
\begin{aligned}
\Pr(|r_k| > \epsilon) \le \Pr(&d\sqrt{C_1}\|\hat{\sigma}_k - \sigma_k\|_\infty > \sqrt{\epsilon}, \|\hat{\sigma}_k - \sigma_k\|_\infty \le M^{-1/4}) \\
&+ \Pr(\|\hat{\sigma}_k - \sigma_k\|_\infty \ge M^{-1/4}) \\
&\le \Pr(d\sqrt{C_1}\|\hat{\sigma}_k - \sigma_k\|_\infty > \sqrt{\epsilon}) + \Pr(\|\hat{\sigma}_k - \sigma_k\|_\infty \ge M^{-1/4}).
\end{aligned}
$$

By applying Lemma B.2, we conclude that $\Pr(|r_k| > \epsilon) \le 2(d+2)^2 e^{-C_2 M\epsilon} + 2(d+2)^2 e^{-C_3\sqrt{M}}$. By taking the union bound over all $k \in [K]$, we obtain the desired statement in the lemma. □

Lemma B.4. *Let $T := \max_k T_k$. For all $\epsilon > 0$, there exists some positive constant C that depends on $\sigma_{\max}$, $\sigma_{\min}$ and d such that*

$$\Pr(\|\mathbf{L}^{(2)}\|_\infty \le \epsilon) \ge 1 - 2(d+2)^2 e^{\frac{1-\gamma}{2}\log N - C\frac{M^2}{T}\epsilon^2}.$$

Proof. For each $\hat{\sigma}_k^{(2)}$, it follows from Lemma B.2 that for all $\epsilon > 0$, there exists some positive constant C that depends on $\sigma_{\max}, \sigma_{\min}$ as well as d such that

$$\Pr(|\ell_k(\hat{\sigma}_k^{(2)} - \sigma_k)| > \epsilon) \le \Pr(\|\ell_k\|_1\|\hat{\sigma}_k^{(2)} - \sigma_k\|_\infty > \epsilon) \le 2(d+2)^2 e^{-CT_k\epsilon^2},$$

where the term $\|\ell_k\|_1$ is absorbed into the positive constant C since $\|\ell_k\|_1$ is a constant that depends on $\sigma_{\max}, \sigma_{\min}$ and d. By taking the union bound and using that $T_k \le T$, we obtain

$$
\begin{aligned}
\Pr(\|\mathbf{L}^{(2)}\|_\infty > \epsilon) \le \sum_{k=1}^{K} \Pr\big(\frac{T_k}{M}|\ell_k(\hat{\sigma}_k^{(2)} - \sigma_k)| > \epsilon\big) &\le 2(d+2)^2 N^{\frac{1-\gamma}{2}} e^{-C\frac{M^2}{T_k}\epsilon^2} \\
&\le 2(d+2)^2 N^{\frac{1-\gamma}{2}} e^{-C\frac{M^2}{T}\epsilon^2},
\end{aligned}
$$

which completes the proof. □

With these preparations we can now prove Lemma 3.7.

Proof of Lemma 3.7. Let $C_1 - C_6$ denote positive constants that depend on $\sigma_{\min}$, $\sigma_{\max}$ and d and may vary from line to line. For any $\epsilon > 0$, standard results in probability yield that

$$
\begin{aligned}
\Pr(\boldsymbol{\rho} \ge 0) &= \Pr(\boldsymbol{\rho} \ge 0, \|\mathbf{R}\|_\infty \le \epsilon) + \Pr(\boldsymbol{\rho} \ge 0, \|\mathbf{R}\|_\infty \ge \epsilon) \\
&\le \Pr(\mathbf{L} + \mathbf{R} \ge 0, \|\mathbf{R}\|_\infty \le \epsilon) + \Pr(\|\mathbf{R}\|_\infty \ge \epsilon) \\
&\le \Pr(\mathbf{L} \ge -\epsilon, \|\mathbf{R}\|_\infty \le \epsilon) + \Pr(\|\mathbf{R}\|_\infty \ge \epsilon) \le \Pr(\mathbf{L} \ge -\epsilon) + \Pr(\|\mathbf{R}\|_\infty \ge \epsilon).
\end{aligned}
$$

Then using the decomposition that $\mathbf{L} = \mathbf{L}^{(1)} + \mathbf{L}^{(2)}$, it follows from the same derivation as the above inequality that for any $\epsilon > 0$,

$$
\begin{aligned}
\Pr(\boldsymbol{\rho} \ge 0) &\le \Pr(\mathbf{L} \ge -\epsilon) + \Pr(\|\mathbf{R}\|_\infty \ge \epsilon) \\
&\le \Pr(\mathbf{L}^{(1)} \ge -2\epsilon) + \Pr(\|\mathbf{L}^{(2)}\|_\infty \ge \epsilon) + \Pr(\|\mathbf{R}\|_\infty \ge \epsilon).
\end{aligned}
$$

Then by choosing $\epsilon = \frac{1}{2\sqrt{M}}$, it follows directly from Lemmas B.3 and B.4 that there exist positive constants C_1, C_2 and C_3 such that

$$
\begin{aligned}
\Pr(\boldsymbol{\rho} \ge 0) \le &\Pr(\mathbf{L}^{(1)} \ge -\frac{1}{\sqrt{M}}) + 2(d+2)^2 e^{\frac{1-\gamma}{2}\log N - C_1\sqrt{M}} \\
&+ 2(d+2)^2 e^{\frac{1-\gamma}{2}\log N - C_2\sqrt{M}} + 2(d+2)^2 e^{\frac{1-\gamma}{2}\log N - C_3\frac{M}{T}}.
\end{aligned}
\tag{7}
$$

Using that the subsampling assignment is from the random event $\mathcal{B}$, it follows that $T \leq \frac{2M^2}{N} \cdot N^{\frac{1-\gamma}{2}}$. By combining this with (7) and the fact that $M = N^\gamma$, we obtain

$$\Pr(\boldsymbol{\rho} \geq 0) \leq \Pr(\mathbf{L}^{(1)} \geq -\frac{1}{\sqrt{M}}) + e^{\log(2(d+2)^2)+\frac{1-\gamma}{2}\log N - C_1 N^{\gamma/2}} \tag{8}$$
$$+ e^{\log(2(d+2)^2)+\frac{1-\gamma}{2}\log N - C_2 N^{\gamma/2}} + e^{\log(2(d+2)^2)+\frac{1-\gamma}{2}\log N - C_3 N^{\frac{1-\gamma}{2}}}.$$

Then using $\log N = o(N^{\gamma/2 \wedge \frac{1-\gamma}{2}})$ and $\log(2(d+2)^2) = o(N^{\gamma/2\wedge\frac{1-\gamma}{2}})$, we can absorb the terms $\log(2(d+2)^2)$ and $\frac{1-\gamma}{2}\log N$ into $N^{\gamma/2}$ and $N^{\frac{1-\gamma}{2}}$ respectively and obtain

$$\Pr(\boldsymbol{\rho} \geq 0) \leq \Pr(\mathbf{L}^{(1)} \geq -\frac{1}{\sqrt{M}}) + e^{-C_1 N^{\gamma/2}} + e^{-C_2 N^{\frac{1-\gamma}{2}}}.$$

It remains to bound the term $\Pr(\mathbf{L}^{(1)} \geq -\frac{1}{\sqrt{M}})$. Since all the $\hat{\sigma}_k^{(1)}$'s are independent random vectors, we have

$$\begin{aligned}\Pr(\mathbf{L}^{(1)} \geq -\frac{1}{\sqrt{M}}) &= \prod_{k=1}^{K} \Pr(\frac{M-T_k}{M}\ell_k(\hat{\sigma}_k^{(1)} - \sigma_k) \geq -\frac{1}{\sqrt{M}}) \\ &\leq \prod_{k=1}^{K} \Pr(\ell_k(\hat{\sigma}_k^{(1)} - \sigma_k) \geq -\frac{2}{\sqrt{M}}),\end{aligned}$$

where the last inequality is based on the fact that $T_k \ll M$ on the event $\mathcal{B}$ and therefore $\frac{M-T_k}{M} \geq \frac{1}{2}$. Let $\nu_k := M \cdot \mathrm{var}(\ell_k(\hat{\sigma}_k^{(1)} - \sigma_k))$. By further applying the standard Berry-Essen theorem, we obtain

$$|\Pr(\ell_k(\hat{\sigma}_k^{(1)} - \sigma_k) \geq -\frac{2}{\sqrt{M}}) - \Pr(Z \geq -2/\sqrt{\nu_k})| \leq C_5/\sqrt{M},$$

where Z represents a standard Gaussian random variable. Using that $\ell_k(\hat{\sigma}_k^{(1)} - \sigma_k)$ can be expressed as the mean of $M - T_k$ independent random variables and that $T_k \ll M$, we obtain that there exists some positive constant C_4 such that for all $k \in [K]$, $\nu_k \geq C_4$. Hence, $\Pr(Z \geq -2/\sqrt{\nu_k}) \leq \Pr(Z \geq -2/\sqrt{C_4})$ and

$$\Pr(\ell_k(\hat{\sigma}_k^{(1)} - \sigma_k) \geq -\frac{2}{\sqrt{M}}) \leq \Pr(Z \geq -2/\sqrt{C_4}) + C_5/\sqrt{M} \leq C_6$$

for some positive constant $C_6 < 1$. Hence, one can rewrite (8) as

$$\Pr(\boldsymbol{\rho} \geq 0) \leq (C_6)^K + e^{-C_1 N^{\gamma/2}} + e^{-C_2 N^{\frac{1-\gamma}{2}}},$$

which finally yields

$$\Pr(\boldsymbol{\rho} \geq 0) \leq e^{-(\log\frac{1}{C_6})\cdot N^{\frac{1-\gamma}{2}}} + e^{-C_1 N^{\gamma/2}} + e^{-C_2 N^{\frac{1-\gamma}{2}}}$$

under the random event $\mathcal{B}$, which completes the proof. □

C Proof of Theorem 3.6

Proof of Theorem 3.6. For any $i \neq j$, without loss of generality, we assume that $|\mathrm{adj}_i(G)| \leq |\mathrm{adj}_j(G)|$. Also, let $S_{ij} := \mathrm{adj}_i(G) \setminus \{j\}$. We denote the random event $\mathcal{A}$ by:

$$\mathcal{A} := \Big\{\text{for any } (i,j) \notin G, \exists t \in [p] \setminus S_{ij} \cup \{i,j\} \text{ such that } \hat{\rho}_{i,j|S_{ij}\cup\{t\}} \leq 0\Big\}.$$

Similarly, for each $(i,j) \notin G$, we let

$$\mathcal{A}_{ij} := \Big\{ \exists t \in [p] \setminus S_{ij} \cup \{i,j\} \text{ such that } \hat{\rho}_{i,j|S_{ij}\cup\{t\}} \leq 0 \Big\}.$$

Let $t_1, \cdots, t_K \in [p] \setminus S_{ij} \cup \{i,j\}$ denote a list of nodes with size $K = N^{\frac{1-\gamma}{2}}$ (this is a valid choice since Condition 3.3 gives us that $p \geq N^{\frac{1-\gamma}{2}} + d + 2$ for any $\gamma \in (\frac{3}{4}, 1)$). It is straightforward to show that $\rho_{ij|S_{ij}\cup\{t_k\}} = 0$ for all $k \in [K]$. Then by setting each S_k in Lemma 3.7 as $S_k := S_{ij} \cup \{t_k\}$, it follows from Lemma 3.7 that with probability at least $1 - \exp(-CN^{\frac{1-\gamma}{2}\wedge 4\gamma-3})$, there exists some t_k such that $\hat{\rho}_{i,j|S_{ij}\cup\{t_k\}} \leq 0$, which yields $\Pr(\mathcal{A}_{ij}) \geq 1 - \exp(-CN^{\frac{1-\gamma}{2}\wedge 4\gamma-3})$. By taking the union bound over all the edges $(i,j) \notin G$, we obtain that $\Pr(\mathcal{A}) \geq 1 - p^2 e^{-C\frac{1-\gamma}{2}\wedge 4\gamma-3}$.

Thus, to complete the proof of the theorem, it remains to prove that under the random event $\mathcal{A}$, all edges $(i,j) \notin G$ are deleted by Algorithm 1 when the algorithm is at iteration $\ell = d+1$. We prove this by contradiction. Suppose there exists an edge $(i,j) \notin G$ that is not deleted by the algorithm at $\ell = d+1$. By applying Theorem 3.5, we obtain that the estimated graph $\hat{G}$ in the iteration $\ell = |\text{adj}_i(G)|$ satisfies $\text{adj}_i(G) \subseteq \text{adj}_i(\hat{G})$ and as a consequence the edge (i,j) will be selected at Step 5 of Algorithm 1 at iteration $\ell = |\text{adj}_i(G)|$. Then by choosing the S at Step 7 to be S_{ij} and using that we are on the event $\mathcal{A}$, we obtain that there exists a node k such that $\hat{\rho}_{ij|S\cup\{k\}} \leq 0$. As a consequence, the edge (i,j) will be deleted at Step 8. This contradicts the fact that the edge (i,j) exists in the final output, which completes the proof. □

D Proof of Theorem 3.5

Lemma D.1. *Consider a Gaussian random vector $X = (X_1, \cdots, X_p)^T$ that follows an MTP$_2$ distribution. Then for any $i,j \in [p]$ and any $S \subseteq [p] \setminus \{i,j\}$, it holds that $\rho_{ij|S} \geq \rho_{ij|[p]\setminus\{i,j\}}$.*

Proof. For $\rho_{ij|S}$, if we let $M = S_{i,j}$, we have

$$\rho_{ij|S} = -\frac{((\Sigma_M)^{-1})_{i_M,j_M}}{\sqrt{((\Sigma_M)^{-1})_{i_M,i_M}((\Sigma_M)^{-1})_{j_M,j_M}}}.$$

Using that the precision matrix Θ is an M-matrix, it follows from basic calculations using Schur complements that $((\Sigma_M)^{-1})_{i_M,i_M} \leq \Theta_{ii}$, $((\Sigma_M)^{-1})_{j_M,j_M} \leq \Theta_{jj}$ and $((\Sigma_M)^{-1})_{i_M,j_M} \leq \Theta_{ij} \leq 0$. By combining this with the fact that $\rho_{ij|[p]\setminus\{i,j\}} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}$, we obtain the lemma. □

With this, we can now provide the proof of Theorem 3.5.

Proof. For any edge $(i,j) \in G$ and any conditioning set $S \subseteq [p] \setminus \{i,j\}$ with $|S| \leq d+2$, by using the same decomposition as in (8), we can decompose the random variable $\hat{\rho}_{ij|S}$ as

$$\hat{\rho}_{ij|S} = \rho_{ij|S} + \ell_{ij|S} + r_{ij|S},$$

where the random variable $\ell_{ij|S}$ is the first order approximation of $\hat{\rho}_{ij|S} - \rho_{ij|S}$ and $r_{ij|S}$ is the residual. It follows from Lemma B.2 and the proof of Lemma B.3 that there exists some positive constant τ such that with probability at least $1 - p^{-(\tau+d+4)}$,

$$|\hat{\rho}_{ij|S} - \rho_{ij|S}| \leq C_1 \sqrt{(\tau + d + 4)\frac{\log p}{N^\gamma}},$$

where C_1 is some positive constant that depends on $\sigma_{\min}, \sigma_{\max}$ and d. By further taking union bound over all $(i,j) \in G$ and all $S \subseteq [p] \setminus \{i,j\}$ with $|S| \leq d+2$, it follows that

$$\Pr\left\{ \forall (i,j) \in G,\ \forall S \subseteq [p] \setminus \{i,j\} \text{ with } |S| \leq d+2, |\hat{\rho}_{ij|S} - \rho_{ij|S}| \leq C_1 \sqrt{(\tau + d + 4)\frac{\log p}{N^\gamma}} \right\} \\ \geq 1 - p^{-\tau}.$$

As a consequence, by assuming that c_ρ in Condition 3.2 is sufficiently large such that $c_\rho > C_1\sqrt{d+4}$ and choosing τ such that $\tau < \left(\frac{c_\rho}{C_1}\right)^2 - d - 4$, it follows from Lemma D.1 that with probability at least $1 - p^{-\tau}$, $\hat{\rho}_{ij|S} > 0$ for all the (i, j, S)'s where $(i, j) \in G$ and $|S| \leq d + 2$. Hence, we obtain that the edges $(i, j) \in G$ will not be deleted by Algorithm 1, which completes the proof. □

E Additional comments on empirical evaluation

E.1 Stability selection

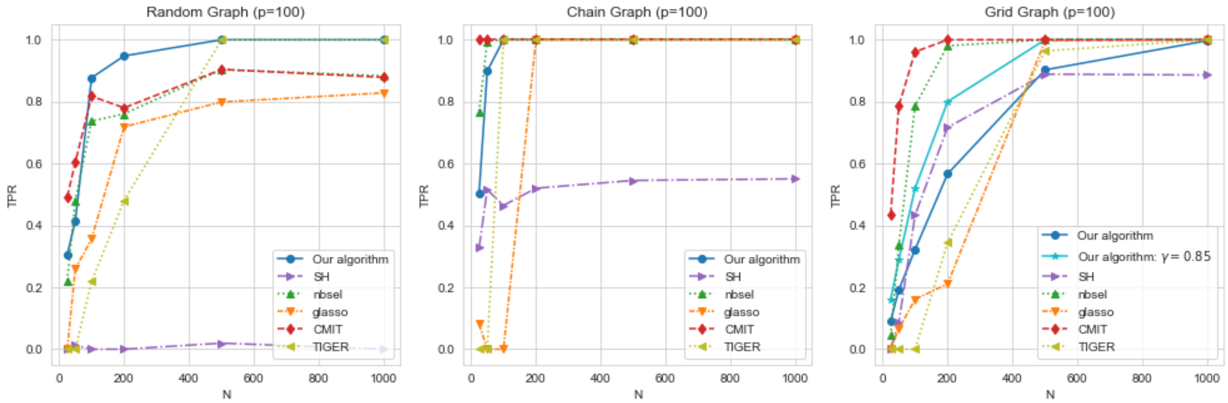
Overview of stability selection: Stability selection (Meinshausen and Bühlmann, 2006) is a well-known technique for enhancing existing variable selection algorithms with tuning parameters. Stability selection works by taking an existing algorithm with a tuning parameter and running it multiple times on different subsamples of the data with various reasonable values for the tuning parameter. A variable is selected if there exists a tuning parameter for which it is selected often enough (in our case we use the threshold $\pi = 0.8$, meaning a variable must be present in at least 80% of trials for a given tuning parameter). Because for each tuning parameter, the algorithm is run many times on different subsamples of the data, stability selection is very computationally expensive. It is important to note that stability selection is *better* than simply choosing the best tuning parameter for a given algorithm, as it is able to combine information across various tuning parameters where appropriate and adapt to different settings.

The advantages of stability selection: As can be seen from Figure 1(c), the purple line corresponds to the *SH* algorithm with stability selection and the pink line corresponds to the *SH* algorithm where the *best* tuning parameter is chosen for each different N (i.e. the y-axis contains the *best* MCC across *all* tuning parameters). Note that the pink line is not a realistic scenario, as in a real-world application we would not have access to the evaluation metric on the test dataset as we do in this simulated example. However this example is instructive in showing that *even when* a particular algorithm is evaluated with the best possible tuning parameter, stability selection is able to outperform it, showing that stability selection truly offers a tremendous advantage for the performance of algorithms with tuning parameters. Thus it is remarkable that our algorithm with theoretically optimal γ is able to compete with other algorithms using stability selection.

Variation of γ and our algorithm with stability selection: It is also worth noting that although our algorithm doesn't have a "tuning parameter" in a traditional sense (i.e. our consistency guarantees are valid for all $\gamma \in (0.75, 1)$), it is still possible to perform stability selection with our algorithm by using various choices of γ in the valid range. In particular, we see from Figure 1(c) that our algorithm with $\gamma = 0.85$ out-performs the theoretically "optimal" value of $\gamma = 7/9$. Thus in practice, because different values of γ lead to different performance (and in some cases better performance than the theoretically optimal value), our algorithm would likely be improved by performing stability selection. This would likely offer an improvement in performance for our algorithm at the expense of higher computational costs. Although it is worth noting that in our experiments our algorithm *without* stability selection performed quite competitively.

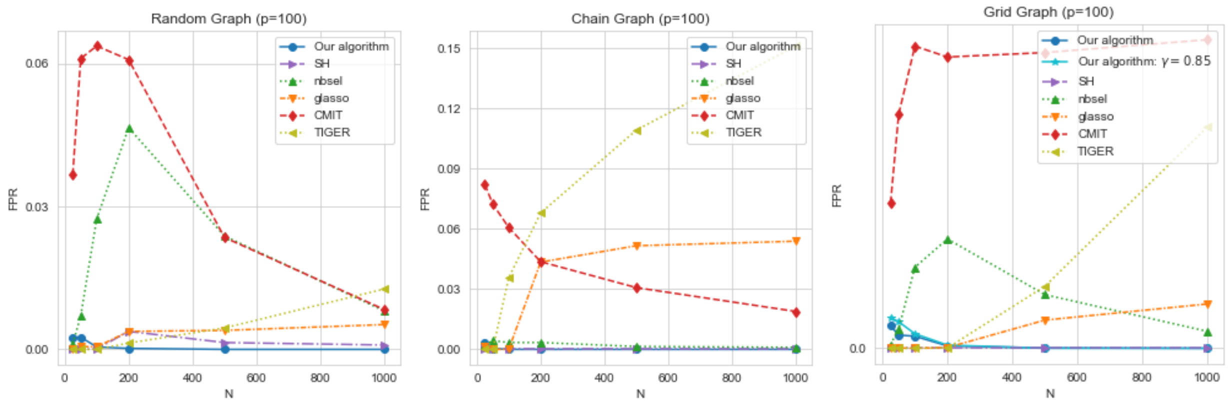
E.2 FPR and TPR

In Figures E.1 and E.2 we report performance of various methods based on the false positive rate (FPR) and true positive rate (TPR) respectively. From these figures we can get similar conclusion as using the MCC measure. In particular, it is important to note that although the TPR of CMIT is higher than our algorithm across all simulation set ups, its FPR is also high, which makes the overall performance less compelling than our algorithm. The performance of TIGER is worse than our method in terms of both TPR and FPR.



(a) Random graphs (b) Chain graphs (c) Grid graphs

Figure E.1: Comparison of different algorithms evaluated on TPR.



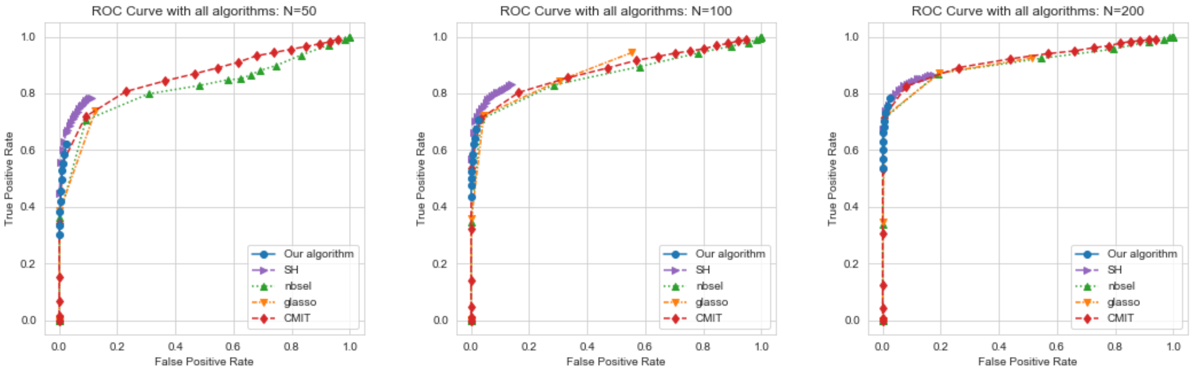
(a) Random graphs (b) Chain graphs (c) Grid graphs

Figure E.2: Comparison of different algorithms evaluated on FPR.

E.3 ROC Curves

To generate the ROC curve for each setting of N, we sample 30 different random graphs (*random* as defined in Section 4)) and then draw N samples from a multivariate normal with the resulting precision matrix. For each of the 30 trials, we get an ROC curve for each algorithm based on the range of tuning parameters tried. To get a mean ROC curve for each algorithm, we average together the 30 trials. The averaged ROC curves are shown Figure 2(a) as well as Figure E.3. The range of tuning parameters tried for each algorithm is listed below:

- *SH:* 20 equally spaced points for $q \in [0.00, 1.0]$.
- *glasso, nbsel:* 20 equally spaced points in log space for $\lambda \in [10^{-6}, 10^{1.2}]$.
- *CMIT:* For computational reasons, we always set $\eta = 1$. However the tuning threshold λ is varied as 20 equally spaced points in log space between $\lambda \in [10^{-4}, 10^{1.2}]$.
- *Our algorithm:* We varied $\gamma \in [0.75, 0.95]$ for 10 equally spaced points in this interval.



(a) ROC Curve $N = 50$

(b) ROC Curve $N = 100$

(c) ROC Curve $N = 200$

Figure E.3: ROC curves for $N = 50, 100, 200$ respectively averaged across 30 trials of a random graph with $p = 100$.

E.4 Normalization of Tuning Parameters

For each algorithm there is a *reasonable* range of tuning parameters that one might consider while attempting to perform structure recovery for Gaussian graphical models with the particular algorithm *in practice*. For *glasso* and *nbsel* it is well known that $\lambda = O\left(\sqrt{\frac{\log p}{N}}\right)$ is theoretically optimal (Friedman et al., 2008; Meinshausen and Bühlmann, 2006). For all of the experiments shown in Figure 2, we have that $p = 100$ and $N = 500$, giving $\sqrt{\frac{\log p}{N}} \approx 0.1$. To test the sensitivity of these algorithms' performance to choice of λ close to this optimal quantity, we let the minimum and maximum λ for both of these algorithms be a factor of 5 within 0.1. Thus, $\lambda_{\min(\text{glasso})} = \lambda_{\min(\text{nbsel})} = 0.02$ and $\lambda_{\max(\text{glasso})} = \lambda_{\max(\text{nbsel})} = 0.5$. We ran both algorithms with a variety of tuning parameters in this range and mapped the tuning parameters linearly to $[0, 1]$ so that 0.02 is mapped to 0 and 0.5 is mapped to 1 in the normalized tuning parameter x-axis in Figures 2(b) and (c). For *CMIT*, the threshold is also optimal for $O\left(\sqrt{\frac{\log p}{N}}\right)$, so we chose $\eta = 1$ for computational reasons and let the threshold vary similarly as *glasso* and *nbsel* and be mapped to $[0, 1]$ similarly for normalization.

For *SH*, we let the threshold $q \in [0.7, 1.]$ as that is the range of threshold quantiles that the authors used in their paper (Slawski and Hein, 2015). Once again, we performed a linear transformation such that the interval of tuning parameters gets mapped to the unit interval.

For our algorithm, we let $\gamma \in [0.75, 0.95]$ and also mapped this interval to $[0, 1]$ for normalizing the γ "tuning parameter". We decided this was an appropriate range for γ since the Algorithm is consistent for $\gamma \in (0.75, 1)$. We make a minor note that in our mapping, we let smaller values of γ correspond to higher values of the normalized tuning parameter (still a linear mapping, simply a reflection of the x-axis) since as γ decreases, it performs similarly to providing more regularization since more edges are removed. In general, an increase in the normalized tuning parameter corresponds to more regularization.

Throughout, we wanted to use a reasonable range of tuning parameters for all algorithms to map onto the unit interval after normalization, so that we could have a fair comparison of the sensitivity of different algorithms' performance to their respective choice of tuning parameters.

F Real data analysis

In this analysis, we consider the following metric that evaluates the community structure of a graph.

Modularity. Given an estimated graph $G := ([p], E)$ with vertex set $[p]$ and edge set E, let A denote the adjacency matrix of G. For each stock j let c_j denote the sector to which stock j belongs and let k_j denote the number of neighbors of stock j in G. Then the *modularity coefficient* Q is given by

$$Q = \frac{1}{2|E|} \sum_{i,j \in [p]} \left(A_{ij} - \frac{k_i k_j}{2|E|} \right) \delta(c_i, c_j),$$

where $\delta(\cdot, \cdot)$ denotes the δ-function with $\delta(i, j) = 1$ if $i = j$ and 0 otherwise.

The modularity coefficient measures the difference between the fraction of edges in the estimated graph that are within a sector as compared to the fraction that would be expected from a random graph. A high coefficient Q means that stocks from the same sector are more likely to be grouped together in the estimated graph, while a low Q means that the community structure of the estimated graph does not deviate significantly from that of a random graph. Table 1 in the main paper shows the modularity scores of the graphs estimated from the various methods; our method using fixed $\gamma = 7/9$ outperforms all the other methods.

References

D. P. Dubhashi, V. Priebe, and D. Ranjan. Negative dependence through the FKG inequality. 1996.

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

X. Magnus and H. Neudecker. *Matrix differential calculus.* New York, 1988.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

M. Slawski and M. Hein. Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, 473:145–179, 2015.

L. Wasserman, M. Kolar, and A. Rinaldo. Berry-Esseen bounds for estimating undirected graphs. *Electronic Journal of Statistics*, 8(1):1188–1224, 2014.