High Dimensional Non-Gaussian DAG Selection

April 13, 2017

1 Notation

For $j \in C \subseteq V$, let $\beta_{ij.C}$ be regression coefficient of j when i is regressed on C

$$\beta_{ij.C} = \left[\left(\mathbb{E}\left(Y_C Y_C^T \right) \right)^{-1} \mathbb{E}(Y_C^T Y_i) \right]_j .$$

Denote the residuals when i is regressed on C as

$$Y_{i.C} = Y_i - \sum_{j \in C} \beta_{ij.C} Y_j .$$

For some set of nodes $D \subseteq V$ and $a \in \mathbb{N}^{|D|}$, let

$$m_{D(a)} = \mathbb{E}\left(\prod_{d \in D} (Y_d)^{a_d} \right).$$

We use the same notation if instead of node i we include $i.C$ such that

$$m_{\{i.C,j,k\}(\alpha)} = \mathbb{E}\left(Y_{i.C}^{\alpha_1} Y_j^{\alpha_2} Y_k^{\alpha_3} \right).$$

2 Parameter and Test Statistic

The statement could probably be strengthened in 2 ways. Explicit characterization of amount of Gaussianity. Also, fix B to be faithful, then generically wrt to error moments

Theorem 1. *Consider the SEM associated with DAG $\mathcal{G} = \{V, E\}$.*

(1) Suppose $j \notin \mathrm{pa}(i)$. There exists some set C such that $C \bigcap \{\mathrm{de}(i) \bigcup \mathrm{de}(j)\} = \emptyset$ such that

$$\tau_{i.C \to j} = m_{\{i.C,j\}(K-1,1)} m_{\{i.C,j\}(2,0)} - m_{\{i.C,j\}(K,0)} m_{\{i.C,j\}(1,1)} = 0$$

(2) Suppose $j \in \mathrm{pa}(i)$, then generically with respect to the error moments and edgeweights, for any C such that $j \notin C$ and $C \bigcap \mathrm{de}(i) = \emptyset$

$$\tau_{i.C \to j} = m_{\{i.C,j\}(K-1,1)} m_{i.C(2)} - m_{j.C(K)} m_{\{i.C,j\}(1,1)} \neq 0$$

Proof. The statements are shown via direct calculation.

Statement (1): Suppose $j \notin \text{pa}(i)$. There exists some set C such that $C \bigcap \{\text{de}(i) \bigcup \text{de}(j)\} = \emptyset$ such that

$$\tau_{i.C \to j} = m_{\{i.C,j\}(K-1,1)} m_{\{i.C,j\}(2,0)} - m_{\{i.C,j\}(K,0)} m_{\{i.C,j\}(1,1)} = 0$$

Consider the set $C = \text{pa}(i)$, then

$$\begin{aligned} Y_{i.C} &= Y_i - \sum_{k \in \text{pa}(i)} \beta_{ik.C} Y_k - \sum_{k \in C \setminus \text{pa}(i)} \beta_{ik.C} Y_k \\ &= Y_i - \sum_{k \in \text{pa}(i)} \beta_{ik.C} Y_k \\ &= \epsilon_i \end{aligned} \tag{1}$$

Let π_{jz} be the sum of all directed path weights from z to j, then

$$\begin{aligned} \tau_{i.C \to j} &= m_{\{i.C,j\}(K-1,1)} m_{\{i.C,j\}(2,0)} - m_{\{i.C,j\}(K,0)} m_{\{i.C,j\}(1,1)} \\ &= \mathbb{E}(Y_{i.C}^{K-1} Y_j) \mathbb{E}(Y_{i.C}^2) - \mathbb{E}(Y_{i.C}^K) \mathbb{E}(Y_{i.C} Y_j) \\ &= \mathbb{E}\left(\epsilon_i^{K-1} \left[\epsilon_j + \pi_{ji} \epsilon_i + \sum_{z \in \text{an}(j)} \pi_{jz} \epsilon_z \right] \right) \mathbb{E}\left(\epsilon_i^2\right) \\ &\quad - \mathbb{E}\left(\epsilon_i^K\right) \mathbb{E}\left(\epsilon_i \left[\epsilon_j + \pi_{ji} \epsilon_i + \sum_{z \in \text{an}(j)} \pi_{jz} \epsilon_z \right] \right) \\ &= \pi_{ji} \mathbb{E}\left(\epsilon_i^K\right) \mathbb{E}\left(\epsilon_i^2\right) - \pi_{ji} \mathbb{E}\left(\epsilon_i^K\right) \mathbb{E}\left(\epsilon_i^2\right) \\ &= 0 \end{aligned} \tag{2}$$

Where the penultimate equality follows from the DAG assumption. Note that if there is no directed path from i to j, then $\pi_{ij} = 0$ and the statement still holds.

Statement (2): Suppose $j \in \text{pa}(i)$, then generically with respect to the error moments and edgeweights, for any C such that $j \notin C$ and $C \bigcap \text{de}(i) = \emptyset$:

$$\tau_{i.C \to j} = m_{\{i.C,j\}(K-1,1)} m_{i.C(2)} - m_{j.C(K)} m_{\{i.C,j\}(1,1)} \neq 0$$

Since each Y_k is a linear combination of the error terms, the moments of Y_k is a polynomial of the error moments and the edge weights. Since $\tau_{i.C \to j}$ is a rational function of the raw moments (not simply a polynomial since the $\beta_{ij.C}$ coefficients involve a ratio of the moments), then selecting a single point where the quantity $\tau_{i.C \to j}$ is non-zero is sufficient for showing that the quantity is generically non-zero.

Consider the point where $\beta_{iq} = 0$ for all $q \in \text{pa}(i) \setminus \{j\}$, $\beta_{jq} = 0$ for all $q \in \text{pa}(j)$, $\beta_{qj} = 0$ for all $q \in \text{ch}(j)$, and $\beta_{ij} \neq 0$. Under this construction, Y_i is correlated with Y_j but is uncorrelated with any $q \notin \text{de}(i)$. Also, let moments of degree $1, \dots K-1$ of ϵ_i and ϵ_j be consistent with a Gaussian distribution but $\mathbb{E}(\epsilon_i^K)$ and $\mathbb{E}(\epsilon_j^K)$ not be consistent with the Gaussian distribution implied by the $1, \dots K-1$ moments.

Then, for any set C, where $C \bigcap \text{de}(i) = \emptyset$-

$$\begin{aligned} Y_{i.C} &= Y_i - \sum_{k \in C} \beta_{ik.C} Y_k \\ &= Y_i = \epsilon_i + \beta_{ij} \epsilon_j \end{aligned} \tag{3}$$

and

$$Y_j = \epsilon_j \tag{4}$$

Thus,

$$\begin{aligned}
\tau_{i.C\to j} &= \mathbb{E}(Y_{i.C}^{K-1}Y_j)\mathbb{E}(Y_{i.C}^2) - \mathbb{E}(Y_{i.C}^K)\mathbb{E}(Y_{i.C}Y_j) \\
&= \mathbb{E}\left((\epsilon_i + \beta_{ij}\epsilon_j)^{K-1}\epsilon_j\right)\mathbb{E}\left((\epsilon_i + \beta_{ij}\epsilon_j)^2\right) \\
&\quad - \mathbb{E}\left((\epsilon_i + \beta_{ij}\epsilon_j)^K\right)\mathbb{E}\left((\epsilon_i + \beta_{ij}\epsilon_j)(\epsilon_j)\right) \\
&= \left(\sum_{a=0}^{K-1}\binom{K-1}{a}\beta_{ij}^{K-1-a}\mathbb{E}\left(\epsilon_i^a\right)\mathbb{E}\left(\epsilon_j^{K-a}\right)\right)\left(\mathbb{E}(\epsilon_i^2) + \beta_{ij}^2\mathbb{E}(\epsilon_j^2)\right) \\
&\quad - \left(\sum_{a=0}^{K}\binom{K}{a}\mathbb{E}\left(\epsilon_i^a\right)\beta_{ij}^{K-a}\mathbb{E}\left(\epsilon_j^{K-a}\right)\right)\left(\beta_{ij}\mathbb{E}\left(\epsilon_j^2\right)\right) \\
&= \left(\sum_{a=1}^{K-1}\binom{K-1}{a}\beta_{ij}^{K-1-a}\mathbb{E}\left(\epsilon_i^a\right)\mathbb{E}\left(\epsilon_j^{K-a}\right)\right)\left(\mathbb{E}(\epsilon_i^2) + \beta_{ij}^2\mathbb{E}(\epsilon_j^2)\right) \\
&\quad - \left(\sum_{a=1}^{K-1}\binom{K}{a}\mathbb{E}\left(\epsilon_i^a\right)\beta_{ij}^{K-a}\mathbb{E}\left(\epsilon_j^{K-a}\right)\right)\left(\beta_{ij}\mathbb{E}\left(\epsilon_j^2\right)\right) \\
&\quad + \beta_{ij}^{K-1}\mathbb{E}(\epsilon_j^K)\left(\mathbb{E}(\epsilon_i^2) + \beta_{ij}^2\mathbb{E}(\epsilon_j^2)\right) - \left(\beta_{ij}^K\mathbb{E}(\epsilon_j^K) + \mathbb{E}(\epsilon_i^K)\right)\left(\beta_{ij}\mathbb{E}(\epsilon_j^2)\right) \\
&= \left(\sum_{a=1}^{K-1}\binom{K-1}{a}\beta_{ij}^{K-1-a}\mathbb{E}\left(\epsilon_i^a\right)\mathbb{E}\left(\epsilon_j^{K-a}\right)\right)\left(\mathbb{E}(\epsilon_i^2) + \beta_{ij}^2\mathbb{E}(\epsilon_j^2)\right) \\
&\quad - \left(\sum_{a=1}^{K-1}\binom{K}{a}\mathbb{E}\left(\epsilon_i^a\right)\beta_{ij}^{K-a}\mathbb{E}\left(\epsilon_j^{K-a}\right)\right)\left(\beta_{ij}\mathbb{E}\left(\epsilon_j^2\right)\right) \\
&\quad + \beta_{ij}^{K-1}\mathbb{E}(\epsilon_j^K)\mathbb{E}(\epsilon_i^2) - \beta_{ij}\mathbb{E}(\epsilon_i^K)\mathbb{E}(\epsilon_j^2)
\end{aligned} \tag{5}$$

If the errors have moments consistent with a Gaussian up to degree $K-1$, and K is odd, then $\mathbb{E}(\epsilon_i^a)\mathbb{E}(\epsilon_j^{K-a}) = 0$ for all $a = 1, \dots K-1$, so we are left with

$$\tau_{i.C\to j} = \beta_{ij}^{K-1}\mathbb{E}(\epsilon_j^K)\mathbb{E}(\epsilon_i^2) - \beta_{ij}\mathbb{E}(\epsilon_i^K)\mathbb{E}(\epsilon_j^2) \tag{6}$$

By construction, $\mathbb{E}(\epsilon_i^K) \neq 0$, $\mathbb{E}(\epsilon_i^K) \neq 0$. So fixing β_{ij} to any non-zero value, and letting

$$\mathbb{E}(\epsilon_i^K) = \frac{\beta_{ij}^{K-2}\mathbb{E}(\epsilon_i^2)\mathbb{E}(\epsilon_j^K)}{\mathbb{E}(\epsilon_j^2)} + \eta$$

implies that $\tau_{i.C\to j} = \eta$.

If K is even, then $\mathbb{E}(\epsilon_i^a)\mathbb{E}(Y_j^{K-a}) = 0$ when a is odd, so we are left with-

$$\begin{aligned}
\tau_{i.C\to j} &= \left(\sum_{a=2,4,\dots K-2}\binom{K-1}{a}\beta_{ij}^{K-1-a}\mathbb{E}\left(\epsilon_i^a\right)\mathbb{E}\left(\epsilon_j^{K-a}\right)\right)\left(\mathbb{E}(\epsilon_i^2) + \beta_{ij}^2\mathbb{E}(\epsilon_j^2)\right) \\
&\quad - \left(\sum_{a=2,4,\dots K-2}\binom{K}{a}\mathbb{E}\left(\epsilon_i^a\right)\beta_{ij}^{K-a}\mathbb{E}\left(\epsilon_j^{K-a}\right)\right)\left(\beta_{ij}\mathbb{E}\left(\epsilon_j^2\right)\right) \\
&\quad + \beta_{ij}^{K-1}\mathbb{E}(\epsilon_j^K)\mathbb{E}(\epsilon_i^2) - \beta_{ij}\mathbb{E}(\epsilon_i^K)\mathbb{E}(\epsilon_j^2)
\end{aligned} \tag{7}$$

By construction, the lower order moments are consistent with some Gaussian distribution, so denoting $\mathbb{E}(\epsilon_i^2)$ and $\mathbb{E}(\epsilon_j^2)$ by σ_i^2 and σ_j^2,

$$\mathbb{E}(\epsilon_i^a) = a!!\sigma_i^a$$

for $a < K$. Further evaluating $\tau_{i.C\to j}$ then yields

$$\begin{aligned}
&= \left(\sum_{a=2,\dots K-2} \binom{K-1}{a} \beta_{ij}^{K-1-a} a!!\sigma_i^a (K-a)!!\sigma_j^{K-a}\right)\left(\sigma_i^2 + \beta_{ij}^2\sigma_j^2\right) \\
&\quad - \left(\sum_{a=2,\dots K-2} \binom{K}{a} a!!\sigma_i^a \beta_{ij}^{K-a}(K-a)!!\sigma_j^{K-a}\right)\left(\beta_{ij}\sigma_j^2\right) \\
&\quad + \beta_{ij}^{K-1}\mathbb{E}(\epsilon_j^K)\sigma_i^2 - \beta_{ij}\mathbb{E}(\epsilon_i^K)\sigma_j^2 \\
&= \left(\sum_{a=2,\dots K-2} \binom{K-1}{a} \beta_{ij}^{K-1-a} a!!\sigma_i^{a+2} (K-a)!!\sigma_j^{K-a}\right) \\
&\quad + \beta_{ij}\sigma_j^2\left(\sum_{a=2,\dots K-2} \binom{K-1}{a} \beta_{ij}^{K-1-a} a!!\sigma_i^a (K-a)!!\sigma_j^{K-a}\right) \\
&\quad - \left(\sum_{a=2,\dots K-2} \binom{K-1}{a} \frac{K}{K-a} a!!\sigma_{\epsilon_i}^a \beta_{ij}^{K-a}(K-a)!!\sigma_j^{K-a}\right)\left(\beta_{ij}\sigma_{Y_j}^2\right) \\
&\quad + \beta_{ij}^{K-1}\mathbb{E}(\epsilon_j^K)\sigma_i^2 - \beta_{ij}\mathbb{E}(\epsilon_i^K)\sigma_j^2 \qquad (8) \\
&= \left(\sum_{a=2,\dots K-2} \binom{K-1}{a} \beta_{ij}^{K-1-a} a!!\sigma_i^{a+2} (K-a)!!\sigma_j^{K-a}\right) \\
&\quad + \left(\sum_{a=2,\dots K-2} \binom{K-1}{a} \left(1 - \frac{K}{K-a}\right) a!!\sigma_{\epsilon_i}^a \beta_{ij}^{K-a}(K-a)!!\sigma_j^{K-a}\right)\left(\beta_{ij}\sigma_{Y_j}^2\right) \\
&\quad + \beta_{ij}^{K-1}\mathbb{E}(\epsilon_j^K)\sigma_i^2 - \beta_{ij}\mathbb{E}(\epsilon_i^K)\sigma_j^2 \\
&= \left(\sum_{a=2,\dots K-4} \binom{K-1}{a} \beta_{ij}^{K-1-a} a!!\sigma_i^{a+2} (K-a)!!\sigma_j^{K-a}\right) \\
&\quad - \left(\sum_{a=4,\dots K-2} \binom{K-1}{a} \frac{a}{K-a} a!!\sigma_{\epsilon_i}^a \beta_{ij}^{K-a}(K-a)!!\sigma_j^{K-a}\right)\left(\beta_{ij}\sigma_{Y_j}^2\right) \\
&\quad + \beta_{ij}K!!\sigma_i^K\sigma_j^2 - \beta_{ij}^{K-1}K!!\sigma_j^K\sigma_i^2 \\
&\quad + \beta_{ij}^{K-1}\mathbb{E}(\epsilon_j^K)\sigma_i^2 - \beta_{ij}\mathbb{E}(\epsilon_i^K)\sigma_j^2
\end{aligned}$$

Rewriting terms and a change of variables show that the first two lines cancel leaving

$$
\begin{aligned}
&= \beta_{ij}\sigma_j^2 \left[\left(\sum_{a=2,\dots K-4} \binom{K-1}{a+2} \frac{(a+1)(a+2)}{(K-(a+1))(K-(a+2))} \beta_{ij}^{K-(a+2)} a!!\sigma_i^{a+2} (K-a)!!\sigma_j^{K-(a+2)} \right) \right. \\
&\quad \left. - \left(\sum_{a=4,\dots K-2} \binom{K-1}{a} \left(\frac{a}{K-a} \right) a!!\sigma_{\epsilon_i}^a \beta_{ij}^{K-a} (K-a)!!\sigma_j^{K-a} \right) \right] \\
&\quad + \beta_{ij} K!!\sigma_i^K \sigma_j^2 - \beta_{ij}^{K-1} K!!\sigma_j^K \sigma_i^2 \\
&\quad + \beta_{ij}^{K-1} \mathbb{E}(\epsilon_j^K)\sigma_i^2 - \beta_{ij}\mathbb{E}(\epsilon_i^K)\sigma_j^2 \\
&= \beta_{ij}\sigma_j^2 \left[\left(\sum_{a=2,\dots K-4} \binom{K-1}{a+2} \frac{a+2}{K-(a+2)} \beta_{ij}^{K-(a+2)} (a+2)!!\sigma_i^{a+2} (K-(a+2))!!\sigma_j^{K-(a+2)} \right) \right. \\
&\quad \left. - \left(\sum_{a=4,\dots K-2} \binom{K-1}{a} \left(\frac{a}{K-a} \right) a!!\sigma_{\epsilon_i}^a \beta_{ij}^{K-a} (K-a)!!\sigma_j^{K-a} \right) \right] \\
&\quad + \beta_{ij} K!!\sigma_i^K \sigma_j^2 - \beta_{ij}^{K-1} K!!\sigma_j^K \sigma_i^2 \\
&\quad + \beta_{ij}^{K-1} \mathbb{E}(\epsilon_j^K)\sigma_i^2 - \beta_{ij}\mathbb{E}(\epsilon_i^K)\sigma_j^2 \\
&= \beta_{ij} K!!\sigma_i^K \sigma_j^2 - \beta_{ij}^{K-1} K!!\sigma_j^K \sigma_i^2 + \beta_{ij}^{K-1} \mathbb{E}(\epsilon_j^K)\sigma_i^2 - \beta_{ij}\mathbb{E}(\epsilon_i^K)\sigma_j^2
\end{aligned}
\tag{9}
$$

Note that if the Kth moments of ϵ were consistent with a Gaussian distribution such that $\mathbb{E}(\epsilon_i^K) = K!!\sigma_i^K$ and $\mathbb{E}(\epsilon_j^K) = K!!\sigma_j^K$, then $\tau_{i.C\to j} = 0$. For fixed $\eta_1 \neq 0$, let

$$
\begin{aligned}
\mathbb{E}(\epsilon_i^K) &= K!!\sigma_i^K + \eta_1 \\
\mathbb{E}(\epsilon_j^K) &= K!!\sigma_j^K + \eta_2 \\
\eta_2 &= \frac{\eta_1 \sigma_j^2}{\beta_{ij}^{K-2}\sigma_i^2} + 1
\end{aligned}
\tag{10}
$$

so that

$$
\begin{aligned}
&\beta_{ij} K!!\sigma_i^K \sigma_j^2 - \beta_{ij}^{K-1} K!!\sigma_j^K \sigma_i^2 + \beta_{ij}^{K-1} \mathbb{E}(\epsilon_j^K)\sigma_i^2 - \beta_{ij}\mathbb{E}(\epsilon_i^K)\sigma_j^2 \\
&= \beta_{ij}^{K-1}\eta_2\sigma_i^2 - \beta_{ij}\eta_1\sigma_j^2 \\
&= \beta_{ij}^{K-1}\sigma_i^2
\end{aligned}
\tag{11}
$$

□

3 Algorithm

Consider the following algorithm for discovering the topological ordering. The method iteratively tests whether a node has a parent node in the remaining unordered nodes. If a node does not have any parents in the remaining unordered nodes, it is assumed to be the root in the remaining sub-graph and added to the discovered ordering.

Theorem 2. *For DAG, $\mathcal{G} = \{V, E\}$, assume*

1. *The maximum in-degree is J*
2. *$|m_{V(\alpha)}| < M_1 - 1 \quad \forall |\alpha| < J$ (the raw moments of Y are bounded)*

Algorithm 1 Naive Topological Ordering

$\Omega = \emptyset$; $\Psi = V$; $z = 1$
Set $\hat{\tau}_i^{(0)}$ to some suitably large constant
while $|\Psi| > 1$ **do**
  $\mathcal{C} = \{C \subseteq \Omega : \Omega(z-1) \in C \text{ and } |C| = \min(J, |\Omega|)$
  for $i \in \Psi$ **do**
    Compute $\hat{\tau}_i^{(z)} = \min\left(\tau_i^{(z-1)}, \min_{C \in \mathcal{C}} \max_{j \in \Psi \setminus i} \tau_{i \to j.C}\right)$
  end for
  $r = \arg\min_{i \in \Psi} \hat{\tau}_i$
  $\Omega(z) = r$
  $\Psi = \Psi \setminus r$
  $z = z + 1$
end while
Return Ω as topological ordering of V

3. $\Sigma = \mathbb{E}(Y^T Y)$ *has minimum eigenvalue* λ_{min} *(min eigenvalue is needed for matrix inversion in estimation of regression coefficients)*
4. $|\hat{m}_{V(\alpha)} - m_{V(\alpha)}| < \delta < \lambda_{min}/J$ *for all* $|\alpha| < J$ *and* $\delta < 1$ *(estimated raw moments are close to true raw moments)*
5. *If* $j \in \text{pa}(i)$*, then* $\tau_{j \to i.C} > \gamma > 0$

Then the output of Algorithm 1 will be a correct topological ordering of V*.*

Algorithmic speedups

- Prune nodes which are ancestors but not parents
- For fixed i Test $\tau_{i \to j}$ in order from largest to smallest to see if you can decrease the max statistic
- Allow for multi-roots. Once $tau_{i \to j}$ is below an adaptive cut-off (rising max stat of selected root), then don't re-test

4 Regression coefficients

Lemma 1. *Assume*

1. $|m_{V(\alpha)}| < M_1 - 1 \quad \forall |\alpha| < K$
2. $\Sigma = \mathbb{E}(Y^T Y)$ *has minimum eigenvalue* λ_{min}
3. $||\hat{\Sigma} - \Sigma||_\infty < \delta_1 < \lambda_{min}/(2J)$

Then

$$||\hat{\beta}_{iC.C} - \beta_{iC.C}||_\infty < \frac{2J\delta_1 \left(\lambda_{min} + JM_1\right)}{\lambda_{min}} = \delta_2$$

for any $C \subseteq V$ *such that* $|C| < J$ *and*

Proof. Via Theorem 4.3.15 (Horn and Johnson) for any $C \subseteq V$ with $|C| \leq J$, $\lambda(\Sigma_{CC})_{min} \geq \lambda_{min}$, so that

$$\delta_1 \leq \lambda_{min}/J \leq \lambda_{min}/|C| \leq \lambda(\Sigma_{CC})_{min}/|C|$$

Let $\omega_{ij} = (\Sigma_{CC})^{-1}_{ij}$. Via Harris and Drton (Lemma 5) for any $C \subseteq V$, we have

$$\max_{ij} |\hat{\omega}_{ij} - \omega_{ij}| \leq \frac{|C|\delta_1/\lambda(\Sigma_{CC})^2_{min}}{1 - |C|\delta_1/\lambda(\Sigma_{CC})_{min}} \tag{12}$$

Further note that $\eta_c = (C\delta/\lambda^2)/(1 - |C|\delta/\lambda)$ is decreasing in λ for $\lambda > J\delta_1$ and increasing in $|C| \leq J$ since

$$\frac{\partial \eta_c}{\partial \lambda} = \frac{1}{\lambda^2} - \frac{1}{(\lambda - |C|\delta_1)^2} < 0 \tag{13}$$

$$\frac{\partial \eta_c}{\partial |C|} = \frac{\delta_1}{(|C|\delta_1 - \lambda)^2} > 0 \tag{14}$$

Which yields the global bound for any C such that $|C| \leq J$

$$\max_{ij} |\hat{\omega}_{ij} - \omega_{ij}| \leq \frac{J\delta_1/\lambda^2_{min}}{1 - J\delta_1/\lambda_{min}} = \eta \tag{15}$$

Furthermore, $||(\Sigma_{CC})^{-1}||_\infty \leq \frac{1}{\lambda(\Sigma_{CC})_{min}} \leq \frac{1}{\lambda_{min}}$.

Then for $C \subseteq V$ and $i \in V \setminus C$

$$\begin{aligned}
||\hat{\beta}_{iC.C} - \beta_{iC.C}||_\infty &= \max_{c \in C} |\hat{\beta}_{ic.C} - \beta_{ic.C}| \\
&= \max_{c \in C} |\sum_{s \in C} (\hat{\Sigma}_{C,C})^{-1}_{cs} \hat{m}_{\{i,s\}(1,1)} - \sum_{s \in C} (\Sigma_{CC})^{-1}_{cs} m_{\{i,s(1,1)\}}| \\
&= \max_{c \in C} |\sum_{s \in C} (\hat{\omega}_{cs} \hat{m}_{\{i,s\}(1,1)} - \sum_{s \in C} (\omega_{cs} m_{\{i,s(1,1)\}})| \\
&= \max_{c \in C} |\sum_{s \in C} ((\omega_{cs} + \eta_{cs})(m_{\{i,s\}(1,1)} + \xi_{is}) - \sum_{s \in C} (\omega_{cs} m_{\{i,s(1,1)\}})| \\
&= \max_{c \in C} |\sum_{s \in C} (\eta_{cs} m_{\{i,s\}(1,1)} + \omega_{cs}\xi_{is} + \xi_{is}\eta_{cs})| \\
&\leq |C|\eta M_1 + |C|\frac{1}{\lambda_{min}}\delta_1 + |C|\delta_1\eta \\
&= |C|\eta(M_1 + \delta_1) + |C|\frac{1}{\lambda_{min}}\delta_1 \\
&\leq J\frac{J\delta_1}{\lambda_{min}(\lambda_{min} - J\delta_1)}(M_1 + \delta_1) + \frac{J\delta_1}{\lambda_{min}} \\
&= \left(\frac{J\delta_1}{\lambda_{min}}\right)\left(1 + \frac{J(M_1 + \delta_1)}{\lambda_{min} - J\delta_1}\right) \\
&= \left(\frac{J\delta_1}{\lambda_{min}}\right)\left(\frac{\lambda_{min} + JM_1}{\lambda_{min} - J\delta_1}\right)
\end{aligned} \tag{16}$$

Since $2J\delta_1 < \lambda_{min}$ by assumption,

$$\lambda_{min} - J\delta_1 > \frac{\lambda_{min}}{2}$$

so that

$$\left(\frac{J\delta_1}{\lambda_{min} - J\delta_1}\right)\left(\frac{\lambda_{min} + JM_1}{\lambda_{min}}\right) < \frac{2J\delta_1\left(\lambda_{min} + JM_1\right)}{\lambda_{min}} = \delta_2 \tag{17}$$

□

5 First and fourth terms of τ

Lemma 2. *Assume*

1. $|m_{V(\alpha)}| < M_1 - \delta_1$ *for all* $|\alpha| < K$
2. $|\beta_{ic.C}| < M_2 - \delta_1$
3. $|\hat{m}_{V(\alpha)} - m_{V(\alpha)}| < \delta_1 < 1$
4. $|\hat{\beta}_{ic.C} - \beta_{ic.C}| < \delta_2 < 1 \quad \forall c \in C$

Then

$$|\hat{m}_{\{i,j\}(s,1).C} - m_{\{i,j\}(s,1).C}| < (J+s)^{s+.5} s! M_1 M_2^s \sqrt{(J+s)^s \delta_1^2 + J\delta_2^2}$$

Proof. Let $f_{\{i,j\}(s,1).C}$ be the map from the raw cross-moments of Y and $\beta_{iC.C}$ to $m_{\{i,j\}(s,1).C}$. Thus,

$$\begin{aligned} f_{\{i,j\}(s,1).C}\left(m_{V(a)}, \beta_{iC.C}, \beta_{jC.C}\right) &= \mathbb{E}\left[(Y_i - \beta_{iC}Y_C)^s (Y_j)\right] \\ &= \mathbb{E}\left[\left(\sum_{|a|=s} \binom{s}{a} \prod_{c \in C} (-\beta_{ic.C}^{a_c} Y_c)^{a_c} Y_i^{\alpha_{C+1}}\right)(Y_j)\right] \\ &= \sum_{|\alpha|=s} \binom{s}{\alpha} m_{\{C,i,j\}(\alpha,1)} \prod_{c \in C} (-\beta_{ic.C})^{\alpha_c} \end{aligned} \tag{18}$$

This is a polynomial, so we derive a Lipschitz constant over that will hold over the bounded domain.

For each of the moments $m_{\{\cdot\}}$

$$\left|\frac{\partial f_1}{\partial m_{\{\cdot\}}}\right| \leq s! M_2^s \tag{19}$$

and for any of the regression coefficients $\beta_{iz.C}$

$$
\begin{aligned}
\left|\frac{\partial f_1}{\partial \beta_{iz.C}}\right| &= \left|\sum_{\substack{|\alpha|=s\\ \alpha_z>0}} \binom{s}{\alpha} m_{\{C,i,j\}(\alpha,1)}(-\alpha_z)(-\beta_{iz.C})^{\alpha_z-1} \prod_{c\in C\setminus z}(-\beta_{ic.C})^{\alpha_c}\right| \\
&\leq \sum_{\substack{|\alpha|=s\\ \alpha_z>0}} \left|\binom{s}{\alpha} m_{\{C,i,j\}(\alpha,1)}(-\alpha_z)(-\beta_{iz.C})^{\alpha_z-1} \prod_{c\in C\setminus z}(-\beta_{ic.C})^{\alpha_c}\right| \\
&\leq \sum_{\substack{|\alpha|=s\\ \alpha_z>0}} \binom{s}{\alpha} M_1 s M_2^{s-1} \\
&\leq \sum_{\substack{|\alpha|=s\\ \alpha_z>0}} \binom{s}{\alpha} \left[M_1 s M_2^{s-1}\right] \\
&\leq (C+1)^s \left[M_1 s M_2^{s-1}\right]
\end{aligned} \tag{20}
$$

So that over the domain $(-M_1, M_1)^{\binom{C+s}{C}} \times (-M_2, M_2)^C$, the function is Lipschitz continuous with constant

$$
\begin{aligned}
&\sqrt{\binom{C+s}{C}(s!M_2^s)^2 + C\left((C+1)^s\left[M_1 s M_2^{s-1}\right]\right)^2} \\
&\leq \sqrt{(C+s)^{2s+1}(s!M_1M_2^s)^2} \\
&= (C+s)^{s+.5} s! M_1 M_2^s
\end{aligned} \tag{21}
$$

So that

$$
\begin{aligned}
|\hat{m}_{\{i,j\}(s,1).C} - m_{\{i,j\}(s,1).C}| &\leq (C+s)^{s+.5} s! M_1 M_2^{s-1} \sqrt{\binom{C+s}{C}\delta_1^2 + C\delta_2^2} \\
&\leq (C+s)^{s+.5} s! M_1 M_2^s \sqrt{(C+s)^s \delta_1^2 + C\delta_2^2} \\
&\leq (J+s)^{s+.5} s! M_1 M_2^s \sqrt{(J+s)^s \delta_1^2 + J\delta_2^2} = \delta_3
\end{aligned} \tag{22}
$$

For τ, we need $s = 1, K-1$, so the error in both cases is bounded by

$$
(J+(K-1))^{K-1+.5}(K-1)! M_1 M_2^{(K-1)} \sqrt{(J+(K-1))^{(K-1)}\delta_1^2 + J\delta_2^2}
$$

□

6 Second and third term

Lemma 3. *Assume*

1. $|m_{V(\alpha)}| < M_1 - \delta_1$ *for all* $|\alpha| < K$
2. $|\beta_{ic.C}| < M_2 - \delta_1$

3. $|\hat{m}_{V(\alpha)} - m_{V(\alpha)}| < \delta_1$

4. $|\hat{\beta}_{ic.C} - \beta_{ic.C}| < \delta_2 \quad \forall c \in C$

Then

$$|\hat{m}_{\{i\}(s).C} - m_{\{i\}(s).C}| < bound$$

Proof. Similarly, for

$$\begin{aligned} f_{\{i\}(s).C} &= \mathbb{E}\left((Y_i - \beta_{iC}Y_C)^s\right) \\ &= \mathbb{E}\left(\sum_{|\alpha|=s}\binom{s}{\alpha}\prod_{c\in C}(-\beta_{ic.C}^{\alpha_c}Y_c)^{\alpha_c}Y_i^{\alpha_{C+1}}\right) \\ &= \sum_{|\alpha|=s}\binom{s}{\alpha}m_{\{C,i\}(\alpha)}\prod_{c\in C}(-\beta_{ic.C})^{\alpha_c} \end{aligned} \tag{23}$$

$$\left|\frac{\partial f_{i(s).C}}{\partial m_{\{C,i\}(\alpha)}}\right| < s!M_2^s$$

and

$$\left|\frac{\partial f_{i(s).C}}{\beta_{ic.C}}\right| < s(C+1)^s M_1 M_2^{s-1}$$

So that over the domain $(-M_1, M_1)^{\binom{C+s}{C}} \times (-M_2, M_2)^C$, the function is Lipschitz continuous with constant

$$\begin{aligned} \sqrt{\binom{C+s}{C}\left(s!M_2^s\right)^2 + C\left(s(C+1)^s M_1 M_2^{s-1}\right)^2} &< \sqrt{(C+s)^s\left(s!M_2^s\right)^2 + C\left(s(C+1)^s M_1 M_2^{s-1}\right)^2} \\ &\leq \sqrt{(C+s)^{2s+1}s!^2 M_1^2 M_2^{2s}} \\ &= (C+s)^{s+.5}s!M_1M_2^s \end{aligned} \tag{24}$$

So that

$$\begin{aligned} |\hat{m}_{\{i\}(s).C} - m_{\{i\}(s).C}| &\leq (C+s)^{s+.5}s!M_1M_2^s\sqrt{\binom{C+s}{C}\delta_1^2 + C\delta_2^2} \\ &\leq (C+s)^{s+.5}s!M_1M_2^s\sqrt{(C+s)^s\delta_1^2 + C\delta_2^2} \\ &\leq (J+s)^{s+.5}s!M_1M_2^s\sqrt{(J+s)^s\delta_1^2 + J\delta_2^2} \end{aligned} \tag{25}$$

For τ we need $s = 2, K$ so the error in both cases is bounded by

$$(J+K)^{K+.5}K!M_1M_2^K\sqrt{(J+K)^K\delta_1^2 + J\delta_2^2} = \delta_3$$

□

Note that this value is strictly larger than the bound on the first and fourth terms, so δ_3 is a global bound on all 4 terms required for τ

7 Pairwise decision

Lemma 4. *1.* $|m_{V(\alpha)}| < M_1$ *for all* $|\alpha| < K$

2. For $i, j \in V$ *and* $C \subseteq V \setminus \{i, j\}$, $|\hat{m}_{\{i,j\}(s,r).C} - m_{\{i,j\}(s,r).C}| < \delta_3$ *for* $(s = K, r = 0)$, $(s = 2, r = 0)$, $(s = 1, r = 1)$ *and* $(s = K - 1, r = 1)$

Then

$$|\hat{\tau}_{i\to j.C} - \tau_{i\to j.C}| < 4M_1\delta_3 + 2\delta_3^2$$

Proof.

$$\begin{aligned}|\hat{\tau}_{i\to j} - \tau_{i\to j}| &= |\hat{m}_{\{i,j\}(K-1,1).C}\hat{m}_{i(2).C} - \hat{m}_{i(K)}\hat{m}_{\{i,j\}(1,1)} - (m_{\{i,j\}(K-1,1).C}m_{i(2).C} - m_{i(K)}m_{\{i,j\}(1,1)})| \\ &\leq |\hat{m}_{\{i,j\}(K-1,1).C}\hat{m}_{i(2).C} - m_{\{i,j\}(K-1,1).C}m_{i(2).C}| + |\hat{m}_{i(K)}\hat{m}_{\{i,j\}(1,1)} - m_{i(K)}m_{\{i,j\}(1,1)})|\end{aligned} \tag{26}$$

Consider each of the two terms separately. For some $|\eta_1| < \delta_3$ and $|\eta_2| < \delta_3$

$$\begin{aligned}|\hat{m}_{\{i,j\}(K-1,1).C}\hat{m}_{i(2).C} - m_{\{i,j\}(K-1,1).C}m_{i(2).C}| &= |(m_{\{i,j\}(K-1,1).C} + \eta_1)(m_{i(2).C} + \eta_2) - m_{\{i,j\}(K-1,1).C}m_{i(2).C}| \\ &= |(m_{\{i,j\}(K-1,1).C}\eta_2 + m_{i(2).C}\eta_1) + \eta_1\eta_2| \\ &\leq M_1\eta_2 + M_1\eta_1 + \eta_1\eta_2 \\ &= 2M_1\delta_3 + \delta_3^2\end{aligned} \tag{27}$$

Using the analogous argument for the second term, we can bound the entire term such that

$$|\hat{\tau}_{i\to j} - \tau_{i\to j}| < 4M_1\delta_3 + 2\delta_3^2$$

□

8 Algorithm Correctness

Theorem 3. *For DAG,* $\mathcal{G} = \{V, E\}$*, assume*

1. *The maximum in-degree is* J
2. $|m_{V(\alpha)}| < M_1 - \delta_1 \quad \forall |\alpha| < J$ *(the raw moments of* Y *are bounded)*
3. $\Sigma = \mathbb{E}(Y^T Y)$ *has minimum eigenvalue* λ_{min} *(min eigenvalue is needed for matrix inversion in estimation of regression coefficients)*
4. $|\hat{m}_{V(\alpha)} - m_{V(\alpha)}| < \delta_1 < \lambda_{min}/J$ *for all* $|\alpha| < K$ *(estimated raw moments are close to true raw moments)*
5. *If* $j \in \text{pa}(i)$*, then* $\tau_{j\to i.C} > \gamma \geq 2\left(4M_1\delta_3 + 2\delta_3^2\right) > 0$ *for all* $C \subseteq V$ *with* $|C| < J$

Then the output of Algorithm 1 will be a correct topological ordering of V*.*

Proof. Since

$$j \in pa(i) \Rightarrow \tau_{i\to j.C} > \gamma \tag{28}$$

For any step z, we will correctly identify any root node (relative to the nodes in Ψ) since every non-root node (relative to Ψ) i will have a parent $j \in \Psi$ so that $\tau_i = \max_j \min_C \tau_{i\to j.C} > \gamma$, but for any root node r, there will exist $C = \text{pa}(r)$ such that $\hat{\tau}_i = \max_j \tau_{r\to j.C} < \gamma$. If there are multiple roots then we can pick one at random and proceed. □

9 High Dimensional Consistency

There are $\binom{V}{K}$ moments $m_{V(\alpha)}$ such $|\alpha| = K$, thus via union bound, the probability that all moments will be within δ_1 of the expectation is bounded by

$$V^K \max_{|\alpha|=K} P\left(|m_{V(\alpha)} - m_{V(\alpha)}| > \delta_1(\gamma, M_1, \lambda_{min}, C)\right)$$

Note that $M_2 = \frac{JM_1}{\lambda_{min}}$, since

$$\beta_{ij.C} = \left[(\Sigma_{CC})^{-1}\mathbb{E}(Y_C^T Y_i)\right]_j \quad = \sum_{c\in C} \omega_{jc} m_{\{c,i\}(1,1)} \leq C\frac{1}{\lambda_{min}} M_1 \quad \leq \frac{JM_1}{\lambda_{min}} \tag{29}$$

Note that

$$\begin{aligned}
\delta_3 &= (J+K)^{K+.5} K! M_1 M_2^K \sqrt{(J+K)^K \delta_1^2 + J\delta_2^2} \\
&= (J+K)^{K+.5} K! M_1 \left(\frac{JM_1}{\lambda_{min}}\right)^K \sqrt{(J+K)^K \delta_1^2 + J\left(\frac{2J\delta_1\left(\lambda_{min} + JM_1\right)}{\lambda_{min}}\right)^2} \\
&= (J+K)^{K+.5} K! M_1 \left(\frac{JM_1}{\lambda_{min}}\right)^K \sqrt{\delta_1^2\left((J+K)^K + J\left(\frac{2J\left(\lambda_{min} + JM_1\right)}{\lambda_{min}}\right)^2\right)} \\
&= (J+K)^{K+.5} K! M_1 \left(\frac{JM_1}{\lambda_{min}}\right)^K \delta_1 \sqrt{(J+K)^K + J\left(\frac{2J\left(\lambda_{min} + JM_1\right)}{\lambda_{min}}\right)^2}
\end{aligned} \tag{30}$$

So for fixed γ,

$$\gamma \geq 2\left(4M_1\delta_3 + 2\delta_3^2\right)$$

implies that

$$\delta_1 \leq -M_1\xi + \frac{\sqrt{4M_1^2\xi^2 + \gamma}}{2} \tag{31}$$

where

$$\xi = (J+K)^{K+.5} K! M_1 \left(\frac{JM_1}{\lambda_{min}}\right)^K \sqrt{(J+K)^K + J\left(\frac{2J\left(\lambda_{min} + JM_1\right)}{\lambda_{min}}\right)^2}$$

Via Lemma B.3 (Lin et al 2016)

Lemma 5. *Consider a degree z polynomial $f(X) = f(X_1, \ldots, X_m)$ where $X_1, \ldots, X_m$ are rv with log-concave joint distributions on $\mathbb{R}^m$. Let $L > 0$ be the constant from B.2. Then, for all δ such that*

$$K := \frac{2}{L}\left(\frac{\delta}{e\sqrt{Var[f(X)]}}\right)^{1/z} \geq 2$$

we have,

$$P(|f(X) - \mathbb{E}[f(X)]| > \delta) \leq \exp\left(\frac{-2}{L}\left(\frac{\delta}{e\sqrt{Var[f(X)]}}\right)^{1/z}\right)$$

10 High Dimensional Consistency

For fixed M_1, γ, K and J, we have

$$P(\hat{\Omega} \neq \Omega) \leq V^K \exp\left(\frac{-2}{L}\left(\frac{\sqrt{n}\delta}{e\sqrt{M_1}}\right)^{1/K}\right) \tag{32}$$

So the estimate of the topological ordering will be consistent as long as

$$K\log(V) - \frac{-2n^{1/(2K)}}{L}\left(\frac{\delta}{e\sqrt{M_1}}\right)^{1/K} \to -\infty \tag{33}$$

11 Algorithm refinements

There are a few ideas for modifying the algorithm that may make the estimation more robust or run faster-

- Use sum rather than max: We could choose the next root node via summing over $j \neq i$ rather than simply taking the max. That is

  $$\text{root} = \arg\min_i \min_C \sum_j |\tau_{i\to j.C}|$$

  rather than

  $$\text{root} = \arg\min_i \min_C \max_j |\tau_{i\to j}|$$

  This doesn't quite fit the theory we've written up because then the "non-Gaussian-ness" assumption about γ now depends on the number of variables we sum over. However, in practice this seems to improve estimation significantly.

- Pruning ancestors: In general, we can use any sort variable selection technique to remove ancestors which aren't parents. However, using something else may require making additional assumptions. If we want to stay with the assumptions we currently have, for $k \in \text{an}(i)\setminus\text{pa}(i)$, then there will eventually be a $C \in \Omega$ such that C blocks all paths from k to i and so $\mathbb{E}(\tau_{k\to i.C}) = 0$. Worst case scenario, the only C that satisfies this is $\text{pa}(i)$ so we may not be able to eliminate any ancestor nodes early, but in practice, we can use this to prune out ancestors which are not parents by only consider nodes k such that $\min_C \tau_{k\to j.C} > \zeta$. We can set ζ before hand to be some small constant, or we can adaptively choose ζ to be driven from the of the statistics from the nodes selected so far.

  $$\arg\min_i \min_C \sum_j |\tau_{i\to j.C}|$$

- Ordering of min-max (or min-sum): For $C = \text{pa}(i)$, $\tau_{i\to j.C} = 0$ for all $j \notin C$. However, for a general set C, $\tau_{i\to j.C}$ could be 0 for some j when $C \not\subseteq \text{pa}(i)$. Thus, we could also pick a root via $\arg\max_j \min_C \tau_{i\to j.C}$. Since there should be some dependency across $\tau_{i\to j.C}$ for fixed C, simulations show that the min-max statistic seems to identify the causal ordering better. However, using the max-min can lead to a considerable speed up. For instance, for each i, if we store $\hat{\tau}_{i\to j}^{(t-1)} = \min_C \tau_{i\to j.C}$ for each j, When we remove node r from the set of unordered nodes and place it into Ω, we can update

  $$\hat{\tau}_{i\to j} = \min\left(\hat{\tau}_{i\to j}^{(t-1)}, \min_{C: r\in C} \tau_{i\to j.C}\right)$$

where we do not have to recalculate everything and only consider the new sets C which contain the latest identified root r. If we use the min-max statistic, we then have to recalculate everything, unless we have stored each individual $\tau_{i \to j.C}$ for each j and C. This is not such a big speed-up when you prune ancestors aggressively, but if you don't prune ancestors, this can help a lot by reducing the effective size of the max in-degree by 1.

In terms of speed, right now, using 6 cores on my desktop and using a "moderate" amount of pruning, we can topologically order a 100 node graph with max in-degree of 3 in roughly 340 seconds