I. Causal Graphical Model with Soft Interventions

To represent causal effects, consider a DAG with structure $(\mathcal{V}, \mathcal{B})$, where $\mathcal{V} = [N] \doteq \{1, \dots, N\}$ is the set of N nodes and $\mathcal{B}$ is the set of directed edges. The observational (without intervention) edge-weight matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ captures the strength of causal effects, where the (i,j)-th entry represents the weight of the edge $i \to j$.

To model causal effects under intervention, consider node-wise intervention, defined as

$$\boldsymbol{a} = (a_1, \dots, a_N)^\top \in \{0,1\}^N, \tag{1}$$

where a_i represents whether node i is intervened (1) or not (0). Specifically, instead of hard interventions, we consider soft interventions, which do not necessarily cut off causal relationships between the intervened node and its parents, but change the incoming edges to the node. We denote the set of parents of node i by $\mathcal{P}_i(a_i)$, the estimated set of parents by $\hat{\mathcal{P}}_i(a_i)$. The set difference of the estimated and true parent sets is denoted by $\hat{\mathcal{P}}_i \backslash \mathcal{P}_i(a_i)$.

Further, we denote the interventional edge-weight matrix by $\boldsymbol{B'} \in \mathbb{R}^{N \times N}$, such that the post-intervention weight matrix $\boldsymbol{B_a}$ can be constructed as

$$[\boldsymbol{B_a}]_i = \mathbb{I}(a_i = 1)\boldsymbol{B}'_i + \mathbb{I}(a_i = 0)\boldsymbol{B}_i, \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $[\cdot]_i$ represents the i-th column of a matrix. The i-th column of the post-intervention weight matrix determines the set of parents of node i and how these parents causally influence node i.

As a result of the intervention, the vector of stochastic values associated with the nodes is represented by $\boldsymbol{x} \in \mathbb{R}^N$. The causal relationship among nodes is described by a linear structural equation model (LinSEM),

$$\boldsymbol{x} = (\boldsymbol{B_a})^\top \boldsymbol{x} + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{e}$ is a vector of exogenous/noise variables. We assume that $\boldsymbol{e}$ contains independent elements, with known means and unknown variances represented by $\boldsymbol{\nu}$ and $\boldsymbol{\epsilon}$. The causal relationship described in (3) can be further manipulated, resulting in

$$\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{B_a})^{-\top} \boldsymbol{e} \doteq (\boldsymbol{C_a})^\top \boldsymbol{e}. \tag{4}$$

We define $\boldsymbol{C_a}$ as the post-intervention flow-weight matrix, whose (i,j)-th entry represents the weight of the net flow from node i to j. In this way, each random variable x_i can be considered as a linear combination of exogenous variables in $\boldsymbol{e}$, weighted by the corresponding flow strength. Thus, under a specific intervention $\boldsymbol{a}$, $\boldsymbol{x}$ follows a multivariate distribution with mean and covariance defined as

$$\boldsymbol{\mu}(\boldsymbol{a}) \doteq \mathbb{E}[\boldsymbol{x}|\boldsymbol{a}] = (\boldsymbol{C_a})^\top \boldsymbol{\nu}, \tag{5}$$

$$\boldsymbol{\Sigma}(\boldsymbol{a}) \doteq \mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{a}))(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{a}))^\top \middle| \boldsymbol{a}\right] \tag{6}$$

$$= (\boldsymbol{C_a})^\top \mathbf{diag}(\boldsymbol{\epsilon})\ \boldsymbol{C_a}. \tag{7}$$

Further, given an intervention selection strategy π, the post-intervention matrix and the second moment matrix can be defined as

$$\boldsymbol{C}^\pi \doteq \mathbb{E}_\pi[\boldsymbol{C_a}] = \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}) \boldsymbol{C_a}, \tag{8}$$

$$\boldsymbol{M}^\pi \doteq \mathbb{E}_\pi[\boldsymbol{x}\boldsymbol{x}^\top] = \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}) \big(\boldsymbol{\Sigma}(\boldsymbol{a}) + \boldsymbol{\mu}(\boldsymbol{a})\boldsymbol{\mu}(\boldsymbol{a})^\top\big). \tag{9}$$

II. Proofs

With the whole set of estimated parents, the estimated weights and residuals are given by MMSE estimation as

$$\hat{\boldsymbol{B}}_{j,\hat{\mathcal{P}}_j} = \big(\boldsymbol{X}_{\hat{\mathcal{P}}_j}^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j}\big)^{-1} \boldsymbol{X}_{\hat{\mathcal{P}}_j}^\top (\boldsymbol{X}_j - \mathbf{1}\nu_j), \tag{10}$$

$$\boldsymbol{r}_j(\hat{\mathcal{P}}_j) = \big[\boldsymbol{I} - \boldsymbol{X}_{\hat{\mathcal{P}}_j} \big(\boldsymbol{X}_{\hat{\mathcal{P}}_j}^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j}\big)^{-1} \boldsymbol{X}_{\hat{\mathcal{P}}_j}^\top\big](\boldsymbol{X}_j - \mathbf{1}\nu_j). \tag{11}$$

Note that $\boldsymbol{X}_{\mathcal{P}}$ represents the sub-matrix consisting of columns corresponding to the nodes in the set $\mathcal{P}$. Denote the projection matrices onto the column and left null space of $\boldsymbol{X}_{\mathcal{P}}$ by

$$\Phi(\boldsymbol{X}_{\mathcal{P}}) \doteq \boldsymbol{X}_{\mathcal{P}} \left(\boldsymbol{X}_{\mathcal{P}}^\top \boldsymbol{X}_{\mathcal{P}}\right)^{-1} \boldsymbol{X}_{\mathcal{P}}^\top, \tag{12}$$

$$\Phi^{\mathcal{C}}(\boldsymbol{X}_{\mathcal{P}}) \doteq \boldsymbol{I} - \boldsymbol{X}_{\mathcal{P}} \left(\boldsymbol{X}_{\mathcal{P}}^\top \boldsymbol{X}_{\mathcal{P}}\right)^{-1} \boldsymbol{X}_{\mathcal{P}}^\top. \tag{13}$$

which allow us to rewrite the residual vector as

$$\boldsymbol{r}_j(\hat{\mathcal{P}}_j) = \big[\boldsymbol{I} - \Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j}\big)\big](\boldsymbol{X}_j - \mathbf{1}\nu_j) \tag{14}$$

$$= \big[\boldsymbol{I} - \Phi\big(\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i\big)\big]\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)(\boldsymbol{X}_j - \mathbf{1}\nu_j) \tag{15}$$

$$= \Phi^{\mathcal{C}}\big(\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i\big) \cdot \boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i). \tag{16}$$

Essentially, the residual vector is successively projected onto orthogonal subspaces [1]. Lastly, we define the *normalized difference in squared residual norms* as

$$\Delta_{ij} \doteq \left(\big\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i)\big\|_2^2 - \big\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j)\big\|_2^2\right)/t_j, \tag{17}$$

where t_j denotes the number of time slots, or equivalently, number of samples of interest.

Lemma 1. If $\mathcal{P}_j \subseteq \hat{\mathcal{P}}_j$ and $i \in \hat{\mathcal{P}}_j \backslash \mathcal{P}_j$, Δ_{ij} converges almost surely to a limit for sufficiently large t_j,

$$\Delta_{ij} \xrightarrow{\text{a.s.}} \frac{\left[C_{ji}^\pi \epsilon_j^2 - \epsilon_j^2\ \boldsymbol{C}_{j,\hat{\mathcal{P}}_j\backslash i}^\pi \big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i}^\pi\big)^{-1} \boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i,i}^\pi\right]^2}{M_{ii}^\pi - \boldsymbol{M}_{i,\hat{\mathcal{P}}_j\backslash i}^\pi \big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i}^\pi\big)^{-1} \boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i,i}^\pi}$$

$$\doteq \Delta_{ij}^*(\text{F parent}). \tag{18}$$

Proof: In both cases, the difference in squared norms can be reformulated as

$$\begin{aligned}
&\left\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i)\right\|_2^2-\left\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j)\right\|_2^2 \\
&=\left[\boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i)+\boldsymbol{r}_j(\hat{\mathcal{P}}_j)\right]^\top\left[\boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i)-\boldsymbol{r}_j(\hat{\mathcal{P}}_j)\right] \qquad (19)\\
&\overset{(a)}{=}\boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i)^\top\left[2\boldsymbol{I}-\Phi\big(\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i\big)\right] \\
&\qquad\cdot\Phi\big(\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i\big)\ \boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i) \qquad (20)\\
&\overset{(b)}{=}\boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i)^\top\Phi\big(\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i\big)\ \boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i), \qquad (21)
\end{aligned}$$

where (a) comes from the substitution of (16) and (b) from the property of the projection matrix Φ.

Given the fact that $i\in\hat{\mathcal{P}}_j\backslash\mathcal{P}_j$, we can rewrite the residual vector as

$$\begin{aligned}
\boldsymbol{r}_j(\hat{\mathcal{P}}_j\backslash i)&=\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)(\boldsymbol{X}_j-\mathbf{1}\nu_j) \qquad (22)\\
&\overset{(c)}{=}\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\left[\boldsymbol{X}_{\mathcal{P}_j}\boldsymbol{B}_{j,\mathcal{P}_j}+\boldsymbol{e}_j-\mathbf{1}\nu_j\right] \qquad (23)\\
&\overset{(d)}{=}\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\cdot\tilde{\boldsymbol{e}}_j, \qquad (24)
\end{aligned}$$

where $\tilde{\boldsymbol{e}}_j\doteq\boldsymbol{e}_j-\mathbf{1}\nu_j$ denotes the vector of centered exogenous variable. Equation (c) comes from the substitution of (3) while (d) holds because value vectors of parent nodes exist exactly in the projection space. Based on the new formulation, the normalized difference can be rewritten as

$$\begin{aligned}
\Delta_{ij}&=\frac{1}{t_j}\cdot\frac{\tilde{\boldsymbol{e}}_j^\top\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i\boldsymbol{X}_i^\top\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\tilde{\boldsymbol{e}}_j}{\boldsymbol{X}_i^\top\Phi^{\mathcal{C}}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i} \qquad (25)\\
&=\frac{\big(\tilde{\boldsymbol{e}}_j^\top\left[\boldsymbol{I}-\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\right]\boldsymbol{X}_i/t_j\big)^2}{\boldsymbol{X}_i^\top\left[\boldsymbol{I}-\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\right]\boldsymbol{X}_i/t_j} \qquad (26)\\
&=\frac{\big(\tilde{\boldsymbol{e}}_j^\top\boldsymbol{X}_i/t_j-\tilde{\boldsymbol{e}}_j^\top\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i/t_j\big)^2}{\boldsymbol{X}_i^\top\boldsymbol{X}_i/t_j-\boldsymbol{X}_i^\top\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i/t_j}. \qquad (27)
\end{aligned}$$

As the sample size t_j approaches infinity, the strong law of large numbers guarantees that

$$\begin{aligned}
\frac{\boldsymbol{X}_i^\top\boldsymbol{X}_i}{t_j}&=\frac{\sum_{\tau=1}^{t_j}(x_i^\tau)^2}{t_j}\xrightarrow{\text{a.s.}}\mathbb{E}[x_i^2]=M_{ii}^\pi, \qquad (28)\\
\frac{\tilde{\boldsymbol{e}}_j^\top\boldsymbol{X}_i}{t_j}&=\frac{\sum_{\tau=1}^{t_j}\tilde{e}_j^\tau\sum_k\left[C_{\boldsymbol{a}^\tau}\right]_{ki}e_k^\tau}{t_j} \qquad (29)\\
&\xrightarrow{\text{a.s.}}\mathbb{E}\Big[\tilde{e}_j\sum_k\left[C_{\boldsymbol{a}}\right]_{ki}e_k\Big]\overset{(e)}{=}\mathbb{E}\Big[\tilde{e}_j\left[C_{\boldsymbol{a}}\right]_{ji}e_j\Big] \qquad (30)\\
&=C_{ji}^\pi\epsilon_j^2, \qquad (31)
\end{aligned}$$

where $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence and (e) is a direct consequence of the independence among exogenous variables. As for the weighted inner products, we have

$$\begin{aligned}
\frac{\boldsymbol{X}_i^\top\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i}{t_j}&=\frac{\boldsymbol{X}_i^\top\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)^{-1}\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_i}{t_j}\\
&=\left(\frac{\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_i}{t_j}\right)^\top\left(\frac{\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}}{t_j}\right)^{-1}\frac{\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_i}{t_j} \qquad (32)\\
&\xrightarrow{\text{a.s.}}\boldsymbol{M}_{i,\hat{\mathcal{P}}_j\backslash i}^\pi\big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i}^\pi\big)^{-1}\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i,i}^\pi, \qquad (33)
\end{aligned}$$

where the convergence relies on the strong law of large numbers and the continuous mapping theorem [2]. Similarly, we also have

$$\begin{aligned}
\frac{\tilde{\boldsymbol{e}}_j^\top\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i}{t_j}&=\frac{\tilde{\boldsymbol{e}}_j^\top\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)^{-1}\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_i}{t_j}\\
&=\left(\frac{\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\tilde{\boldsymbol{e}}_j}{t_j}\right)^\top\left(\frac{\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}}{t_j}\right)^{-1}\frac{\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}^\top\boldsymbol{X}_i}{t_j} \qquad (34)\\
&\xrightarrow{\text{a.s.}}\epsilon_j^2\,\boldsymbol{C}_{j,\hat{\mathcal{P}}_j\backslash i}^\pi\big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i}^\pi\big)^{-1}\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i,i}^\pi. \qquad (35)
\end{aligned}$$

Lastly, since addition and multiplication preserve almost sure convergence [3], a combination of the aforementioned results completes the proof. □

Lemma 2. If $\mathcal{P}_j\subseteq\hat{\mathcal{P}}_j$ and $i\in\mathcal{P}_j$, Δ_{ij} converges almost surely to a limit for sufficiently large t_j,

$$\Delta_{ij}\xrightarrow{\text{a.s.}}\Delta_{ij}^*(\text{T parent})\doteq \frac{\left[B_{ij}M_{ii}^\pi-\big(B_{ij}\boldsymbol{M}_{i,\hat{\mathcal{P}}_j\backslash i}^\pi+\epsilon_j^2\boldsymbol{C}_{j,\hat{\mathcal{P}}_j\backslash i}^\pi\big)\big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i}^\pi\big)^{-1}\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i,i}^\pi\right]^2}{M_{ii}^\pi-\boldsymbol{M}_{i,\hat{\mathcal{P}}_j\backslash i}^\pi\big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i}^\pi\big)^{-1}\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i,i}^\pi} \qquad (36)$$

Proof: Since $i\in\mathcal{P}_j$, we can rewrite the normalized difference based on (21) as

$$\begin{aligned}
\Delta_{ij}&=\frac{1}{t_j}\cdot\frac{\big((\boldsymbol{X}_j-\mathbf{1}\nu_j)^\top\left[\boldsymbol{I}-\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\right]\boldsymbol{X}_i\big)^2}{\boldsymbol{X}_i^\top\left[\boldsymbol{I}-\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\right]\boldsymbol{X}_i} \qquad (37)\\
&\overset{(a)}{=}\frac{\left[\big(B_{ij}\boldsymbol{X}_i+\tilde{\boldsymbol{e}}_j\big)^\top\left[\boldsymbol{I}-\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\right]\boldsymbol{X}_i/t_j\right]^2}{\boldsymbol{X}_i^\top\left[\boldsymbol{I}-\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\right]\boldsymbol{X}_i/t_j}, \qquad (38)
\end{aligned}$$

where (a) relies on the fact that vectors of all parents except for node i lie within the projection space. Again, to understand the asymptotic behavior of Δ_{ij}, we study the limits of its building components.

Based on (28) and (31), we have

$$\frac{\big(B_{ij}\boldsymbol{X}_i+\tilde{\boldsymbol{e}}_j\big)^\top\boldsymbol{X}_i}{t_j}\xrightarrow{\text{a.s.}}B_{ij}M_{ii}^\pi+C_{ji}^\pi\epsilon_j^2=B_{ij}M_{ii}^\pi, \qquad (39)$$

where $C_{ji}^\pi=0$ due to the fact that $i\in\mathcal{P}_j$. Similar to (33), the weighted inner product converges as

$$\begin{aligned}
&\frac{\big(B_{ij}\boldsymbol{X}_i+\tilde{\boldsymbol{e}}_j\big)^\top\Phi\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j\backslash i}\big)\boldsymbol{X}_i}{t_j}\xrightarrow{\text{a.s.}}\\
&\qquad\big(B_{ij}\boldsymbol{M}_{i,\hat{\mathcal{P}}_j\backslash i}^\pi+\epsilon_j^2\boldsymbol{C}_{j,\hat{\mathcal{P}}_j\backslash i}^\pi\big)\big(\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i}^\pi\big)^{-1}\boldsymbol{M}_{\hat{\mathcal{P}}_j\backslash i,i}^\pi. \qquad (40)
\end{aligned}$$

The convergence relies on the strong law of large numbers and the continuous mapping theorem [2]. Also, note that the asymptotic limit of the denominator comes directly from (28) and (33). Eventually, a combination of the convergence results completes the proof. □

References

[1] G. Strang, *Introduction to linear algebra*. SIAM, 2022.

[2] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge University Press, 2000, vol. 3.

[3] G. Casella and R. Berger, *Statistical inference*. CRC Press, 2024.