I. Causal Graphical Model with Soft Interventions

To represent causal effects, consider a DAG with structure $(\mathcal{V}, \mathcal{B})$, where $\mathcal{V} = [N] \doteq \{1, \ldots, N\}$ 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

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

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 cut off causal relationships between the intervened node and its parents, but change the incoming edges to the node. We denote the sets 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 structure equation model (LinSEM),

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

where $\boldsymbol{e}$ is a vector of exogenous variables, independent of $\boldsymbol{x}$. 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

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

where $\boldsymbol{C_a}$ denotes the post-intervention flow-weight matrix, whose (i, j)-th entry represents the net flow weight from node i to j. In this way, each variable x_i can be considered as a linear combination of exogenous variables in $\boldsymbol{e}$, weighted by the strength of causal flow. 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 mean and variance can be defined as

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

$$\boldsymbol{\Sigma}^\pi \doteq \mathbb{E}_\pi\left[(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{a}))(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{a}))^\top\right] = \sum_{\boldsymbol{a}} \pi(\boldsymbol{a})\boldsymbol{\Sigma}(\boldsymbol{a}), \tag{9}$$

where $\pi(\boldsymbol{a})$ stands for the intervention selection probability.

II. Proofs

With the whole set of potential parents, the estimated weights, $\hat{\boldsymbol{B}}_{j,\hat{\mathcal{P}}_j}$ and residuals, $\boldsymbol{r}_j(\hat{\mathcal{P}}_j)$ are given by linear minimum mean square error (MMSE) estimation as

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

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

Denote the projection matrix onto the column space of $\boldsymbol{X}_{\mathcal{P}}$ by

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

which allows us to rewrite the residual vector as

$$\boldsymbol{r}_j(\hat{\mathcal{P}}_j) = \left[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j)\right](\boldsymbol{X}_j - \nu_j \mathbf{1}) \tag{13}$$

$$= \left[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(i)\right]\left[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\right](\boldsymbol{X}_j - \nu_j \mathbf{1}) \tag{14}$$

$$= \left[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(i)\right]\boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i). \tag{15}$$

We define the *normalized difference in squared residual norms* as

$$\Delta_{ij} \doteq \left(\left\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i)\right\|_2^2 - \left\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j)\right\|_2^2\right) / t_j(a_j), \tag{16}$$

where $t_j(a_j)$ denotes the number of time slots when node j is under intervention a_j.

Lemma 1. If $\mathcal{P}_j \subseteq \hat{\mathcal{P}}_j$ and $i \in \hat{\mathcal{P}}_j \backslash \mathcal{P}_j$, the following upper bound on Δ_{ij} holds almost surely for sufficiently large t_j.

$$\Delta_{ij} \leq \overline{\Delta}_{ij}(\text{F parent}) \doteq \frac{\left(\epsilon_j^2 \sum_{\boldsymbol{a}} \pi(\boldsymbol{a})\,[C_{\boldsymbol{a}}]_{ji}\right)^2}{\Sigma_{ii}^\pi + (\mu_i^\pi)^2}.$$

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

$$\left\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i)\right\|_2^2 - \left\|\boldsymbol{r}_j(\hat{\mathcal{P}}_j)\right\|_2^2$$

$$= \left[\boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i) + \boldsymbol{r}_j(\hat{\mathcal{P}}_j)\right]^\top \left[\boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i) - \boldsymbol{r}_j(\hat{\mathcal{P}}_j)\right] \tag{17}$$

$$\overset{(a)}{=} \boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i)^\top \left[2\boldsymbol{I} - \Phi_{\boldsymbol{X}}(i)\right] \Phi_{\boldsymbol{X}}(i)\ \boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i) \tag{18}$$

$$\overset{(b)}{=} \boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i)^\top \Phi_{\boldsymbol{X}}(i)\ \boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i), \tag{19}$$

where (a) comes from the substitution of (15) and (b) from the fact that $\Phi_{\boldsymbol{X}}(i)$ is a projection matrix.

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

$$\boldsymbol{r}_j(\hat{\mathcal{P}}_j \backslash i) = \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big](\boldsymbol{X}_j - \nu_j \mathbf{1}) \tag{20}$$

$$\overset{(a)}{=} \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big]\big[\boldsymbol{X}_{\mathcal{P}_j} \boldsymbol{B}_{j,\mathcal{P}_j} + \boldsymbol{e}_j - \nu_j \mathbf{1}\big] \tag{21}$$

$$\overset{(b)}{=} \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big]\ \tilde{\boldsymbol{e}}_j, \tag{22}$$

where $\tilde{\boldsymbol{e}}_j \doteq \boldsymbol{e}_j - \nu_j \mathbf{1}$ denotes the vector of centered exogenous variable. Equation (a) is a direct subsequence of (3) while (b) holds because value vectors of parent nodes $\boldsymbol{X}_{\mathcal{P}_j}$ exist exactly in the projection space. Based on the new formulation, the normalized difference can be bounded from above as

$$\Delta_{ij} = \frac{\tilde{\boldsymbol{e}}_j^\top \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big] \boldsymbol{X}_i \boldsymbol{X}_i^\top \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big] \tilde{\boldsymbol{e}}_j}{t_j\ \boldsymbol{X}_i^\top \boldsymbol{X}_i} \tag{23}$$

$$= \frac{\Big(\tilde{\boldsymbol{e}}_j^\top \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big] \boldsymbol{X}_i / t_j\Big)^2}{\boldsymbol{X}_i^\top \boldsymbol{X}_i / t_j} \tag{24}$$

$$\overset{(a)}{\leq} \frac{\Big[\sum_{\tau=1}^{t_j} (e_j^\tau - \nu_j) \left(\sum_k \left[C_{\boldsymbol{a}^\tau}\right]_{ki} e_k^\tau\right) / t_j\Big]^2}{\sum_{\tau=1}^{t_j} \left(x_i^\tau\right)^2 / t_j} \tag{25}$$

$$\doteq \tilde{\Delta}_{ij}. \tag{26}$$

Inequality (a) holds because magnitude of the inner product is smaller in the projection space. As the number of samples t_j approaches infinity, the strong law of large numbers guarantees the following convergence result:

$$\frac{\sum_{\tau=1}^{t_j(0)} (e_j^\tau - \nu_j) \sum_k \left[C_{\boldsymbol{a}^\tau}\right]_{ki} e_k^\tau}{t_j} \overset{\text{a.s.}}{\to} \mathbb{E}\Big[\tilde{e}_j \sum_k \left[C_{\boldsymbol{a}}\right]_{ki} e_k\Big] \tag{27}$$

$$\overset{(a)}{=} \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}) \mathbb{E}\left[\tilde{e}_j \left[C_{\boldsymbol{a}}\right]_{ji} e_j \middle| \boldsymbol{a}\right] \tag{28}$$

$$= \epsilon_j^2 \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}) \left[C_{\boldsymbol{a}}\right]_{ji}, \tag{29}$$

where $\overset{\text{a.s.}}{\to}$ denotes almost sure convergence and equality (a) comes from the independence among exogenous variables. Similarly, the denominator of (25) converges as

$$\frac{\sum_{\tau=1}^{t_j} \left(x_i^\tau\right)^2}{t_j} \overset{\text{a.s.}}{\to} \mathbb{E}[x_i^2] = \Sigma_{ii}^\pi + (\mu_i^\pi)^2. \tag{30}$$

Lastly, according to the continuous mapping theorem [1] and the fact that multiplication preserves almost sure convergence [2], we have

$$\tilde{\Delta}_{ij} \overset{\text{a.s.}}{\to} \frac{\Big(\epsilon_j^2 \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}) \left[C_{\boldsymbol{a}}\right]_{ji}\Big)^2}{\Sigma_{ii}^\pi + (\mu_i^\pi)^2}, \text{ as } t_j \to \infty, \tag{31}$$

which completes the proof. □

Lemma 2. If $\mathcal{P}_j \subseteq \hat{\mathcal{P}}_j$ and $i \in \mathcal{P}_j$, the following lower bound on Δ_{ij} holds almost surely for sufficiently large t_j.

$$\Delta_{ij} \geq \underline{\Delta}_{ij}(\text{T parent}) \doteq \frac{B_{ij}^2}{\Sigma_{ii}^\pi + (\mu_i^\pi)^2} \times \max\left[0, \Sigma_{ii}^\pi + (\mu_i^\pi)^2 - \frac{\sum_{k \in \hat{\mathcal{P}}_j \backslash i} \left(\Sigma_{ik}^\pi + \mu_i^\pi \mu_k^\pi\right)^2}{\lambda_{\min}\big(\boldsymbol{\Sigma}_{\hat{\mathcal{P}}_j \backslash i}^\pi + \boldsymbol{\mu}_{\hat{\mathcal{P}}_j \backslash i}^\pi \big(\boldsymbol{\mu}_{\hat{\mathcal{P}}_j \backslash i}^\pi\big)^\top\big)}\right]^2, \tag{32}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue of a matrix.

Proof: Since $i \in \mathcal{P}_j$, we can rewrite the normalized difference based on (19) as

$$\Delta_{ij} = \frac{\big((\boldsymbol{X}_j - \nu_j \mathbf{1})^\top \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big] \boldsymbol{X}_i\big)^2}{t_j\ \boldsymbol{X}_i^\top \boldsymbol{X}_i} \tag{33}$$

$$\overset{(a)}{=} \frac{\big[\big(B_{ij} \boldsymbol{X}_i + \tilde{\boldsymbol{e}}_j\big)^\top \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big] \boldsymbol{X}_i / t_j\big]^2}{\boldsymbol{X}_i^\top \boldsymbol{X}_i / t_j} \tag{34}$$

$$= \frac{1}{\boldsymbol{X}_i^\top \boldsymbol{X}_i / t_j} \Bigg(\frac{\tilde{\boldsymbol{e}}_j^\top \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big] \boldsymbol{X}_i}{t_j} + B_{ij} \frac{\boldsymbol{X}_i^\top \boldsymbol{X}_i - \boldsymbol{X}_i^\top \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i) \boldsymbol{X}_i}{t_j} \Bigg)^2. \tag{35}$$

Equation (a) holds because the value vectors of all parents except for node i lie within the projection space. To understand the asymptotic behavior of Δ_{ij}, we need to study its building components.

First, the convergence result of $\boldsymbol{X}_i^\top \boldsymbol{X}_i / t_j$ is provided in (30). The third component can be bounded as

$$\frac{\boldsymbol{X}_i^\top \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i) \boldsymbol{X}_i}{t_j} = \frac{\boldsymbol{X}_i^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i} \big(\boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i}^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i}\big)^{-1} \boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i}^\top \boldsymbol{X}_i}{t_j}$$

$$\overset{(a)}{\leq} \frac{\lambda_{\max}\big(\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i}^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i}\big)^{-1}\big) \big\|\boldsymbol{X}_i^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i}\big\|_2^2}{t_j} \tag{36}$$

$$= \frac{\sum_{k \in \hat{\mathcal{P}}_j \backslash i} \left(\boldsymbol{X}_i^\top \boldsymbol{X}_k / t_j\right)^2}{\lambda_{\min}\big(\boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i}^\top \boldsymbol{X}_{\hat{\mathcal{P}}_j \backslash i} / t_j\big)} \tag{37}$$

$$\overset{\text{a.s.}}{\to} \frac{\sum_{k \in \hat{\mathcal{P}}_j \backslash i} \left(\Sigma_{ik}^\pi + \mu_i^\pi \mu_k^\pi\right)^2}{\lambda_{\min}\big(\boldsymbol{\Sigma}_{\hat{\mathcal{P}}_j \backslash i}^\pi + \boldsymbol{\mu}_{\hat{\mathcal{P}}_j \backslash i}^\pi \big(\boldsymbol{\mu}_{\hat{\mathcal{P}}_j \backslash i}^\pi\big)^\top\big)}. \tag{38}$$

Inequality (a) follows directly from the definition of maximum eigenvalue. The convergence relies on the strong law of large numbers, the fact that eigenvalues are continuous functions of the matrix [3] and the continuous mapping theorem [1]. Since magnitude of the last component is bounded from above by zero asymptotically,

$$\left|\frac{\tilde{\boldsymbol{e}}_j^\top \big[\boldsymbol{I} - \Phi_{\boldsymbol{X}}(\hat{\mathcal{P}}_j \backslash i)\big] \boldsymbol{X}_i}{t_j}\right| \leq \left|\frac{\tilde{\boldsymbol{e}}_j^\top \boldsymbol{X}_i}{t_j}\right| \overset{\text{a.s.}}{\to} |\mathbb{E}[\tilde{e}_j x_i]| = 0, \tag{39}$$

the first component converges to zero almost surely.

Eventually, since addition and multiplication preserve almost sure convergence [2], combination of the aforementioned results completes the proof. □

REFERENCES

[1] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[2] G. Casella and R. Berger, *Statistical inference*. CRC press, 2024.

[3] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.