Appendix I: The proposed SICE algorithm

This section details our approach for estimating sparse inverse covariance matrix from data, which can be achieved through solving for the optimization problem in (1). Our approach is based on the block coordinate descent (BCD) algorithm, but with an extended capacity of allowing for prior domain knowledge to be incorporated into the problem solving process.

The basic idea of the BCD algorithm is to update each column (or row) of $\mathbf{\Theta}$ iteratively while fixing all other columns (or rows), until convergence. Because the BCD algorithm works by iterations, we will only illustrate the steps in one iteration and other iterations work in a similar way. At a certain iteration, we first need to partition the current $\mathbf{\Theta}$ as follows. Let $\mathbf{\Theta}_{\backslash j\backslash j}$ be the matrix produced by removing row j and column j from $\mathbf{\Theta}$, θ_{jj} be the element at row j and column j of $\mathbf{\Theta}$, and $\mathbf{\Theta}_j$ be the column j of $\mathbf{\Theta}$ with θ_{jj} removed. Then, $\mathbf{\Theta}$ can be partitioned as $\mathbf{\Theta} = \begin{bmatrix} \mathbf{\Theta}_{\backslash j\backslash j} & \mathbf{\Theta}_j \\ \mathbf{\Theta}_j^T & \theta_{jj} \end{bmatrix}$, and correspondingly $\mathbf{S}$ can be partitioned as $\mathbf{S} = \begin{bmatrix} \mathbf{S}_{\backslash j\backslash j} & \mathbf{S}_j \\ \mathbf{S}_j^T & s_{jj} \end{bmatrix}$. Next, we want to update $\mathbf{\Theta}_j$ and θ_{jj} while holding other elements in $\mathbf{\Theta}$ constant. To do this, let f represent the objective function in (1), i.e., $f = log|\mathbf{\Theta}| - tr(\mathbf{S\Theta}) - \lambda\|\mathbf{\Theta}\|_1$; take the partial derivatives of f with respect to $\mathbf{\Theta}_j$ and θ_{jj}, respectively; and then make the partial derivatives to be zero, i.e.,

$$\frac{\partial f}{\partial \mathbf{\Theta}_j} = -\frac{2}{\theta_{jj} - \mathbf{\Theta}_j^T \mathbf{\Theta}_{\backslash j\backslash j}^{-1} \mathbf{\Theta}_j} \mathbf{\Theta}_{\backslash j\backslash j}^{-1} \mathbf{\Theta}_j - \mathbf{S}_j - \lambda\, SGN(\mathbf{\Theta}_j) = 0, \text{ and} \tag{A-1}$$

$$\frac{\partial f}{\partial \theta_{jj}} = \frac{1}{\theta_{jj} - \mathbf{\Theta}_j^T \mathbf{\Theta}_{\backslash j\backslash j}^{-1} \mathbf{\Theta}_j} - s_{jj} - \lambda = 0, \tag{A-2}$$

where $SGN(\mathbf{\Theta}_j)$ denotes the partial derivative of $\|\mathbf{\Theta}\|_1$ with respect to $\mathbf{\Theta}_j$. It is difficult to solve for $\mathbf{\Theta}_j$ and θ_{jj} from (A-1) and (A-2) directly. Therefore, we adopt the following strategies.

Letting $\mathbf{a} = -\frac{\mathbf{\Theta}_j}{\theta_{jj} - \mathbf{\Theta}_j^T \mathbf{\Theta}_{\backslash j\backslash j}^{-1} \mathbf{\Theta}_j}$, then (A-1) and (A-2) become

$$2\mathbf{\Theta}_{\backslash j\backslash j}^{-1}\mathbf{a} - \mathbf{S}_j + \lambda\, SGN(\mathbf{a}) = 0 \tag{A-3}$$

$$\mathbf{a} = -(s_{jj} + \lambda)\mathbf{\Theta}_j. \tag{A-4}$$

It is clear that (A-3) is also the result of making the partial derivative of g with respect to $\mathbf{a}$ to be zero in the following optimization problem:

$$\min_{\mathbf{a}} g = \mathbf{a}^T \mathbf{\Theta}_{\backslash j\backslash j}^{-1} \mathbf{a} - \mathbf{S}_j^T \mathbf{a} + \lambda\|\mathbf{a}\|_1, \tag{A-5}$$

which is equivalent to the following min-max problem:

$$\max_{\boldsymbol{\kappa}} \min_{\mathbf{a}} g = 2\left(-\frac{1}{2}\boldsymbol{\kappa}^T \Theta_{\backslash j\backslash j} \boldsymbol{\kappa} + \boldsymbol{\kappa}^{\mathrm{T}}\mathbf{a}\right) - \mathbf{S}_j^T \mathbf{a} + \lambda\|\mathbf{a}\|_1. \tag{A-6}$$

This min-max problem can be solved by the prox method.

After $\mathbf{a}$ and $\boldsymbol{\kappa}$ are obtained, (A-4) can be used to find $\mathbf{\Theta}_j$, i.e., $\mathbf{\Theta}_j = -\frac{\mathbf{a}}{s_{jj}+\lambda}$. Furthermore, based on (A-2), θ_{jj} can be obtained, i.e., $\theta_{jj} = \frac{(-\mathbf{a}^T\boldsymbol{\kappa}+\mathbf{1})}{s_{jj}+\lambda}$.

Furthermore, suppose that some prior domain knowledge is available, e.g., nodes X_i and X_j are disconnected in the IC model, which means that $\theta_{ij} = 0$ in $\mathbf{\Theta}$. Then, we can forces the corresponding entry in $\mathbf{a}$ to be zero in each iteration. As a result, we can re-formulate (A-5) as follows:

$$\min_{\mathbf{a}} g = \mathbf{a}^T \mathbf{\Theta}_{\backslash j\backslash j}^{-1} \mathbf{a} - \mathbf{S}_j^T \mathbf{a} + \lambda\|\mathbf{a}\|_1$$

$$s.t.\, \mathbf{a}_i = \mathbf{0} \quad if \;\; i \in \mathbf{V}$$

where $\mathbf{V}$ is the set of indices (based on prior domain knowledge) corresponding to zero entries in $\mathbf{a}$. Note that this problem is also strictly convex and can be solved efficiently. □

Appendix II: Proof of the monotone property of the SICE algorithm

A sufficient and necessary condition of the monotone property is as follow:

Theorem 1: Let $\{\mathbf{C}_1^{\lambda_1}, \dots, \mathbf{C}_{L_1}^{\lambda_1}\}$ and $\{\mathbf{C}_1^{\lambda_2}, \dots, \mathbf{C}_{L_2}^{\lambda_2}\}$ denote the clusters of nodes in the SICE-based graphical models, with λ equal to λ_1 and λ_2 ($\lambda_1 < \lambda_2$), respectively. Then, for any $C_i^{\lambda_2}$, $i \in \{1,2,\dots,L_2\}$, there must exist a $C_j^{\lambda_1}$, $j \in \{1,2,\dots,L_1\}$ such that $C_i^{\lambda_2} \subseteq C_j^{\lambda_1}$.

This section proves the monotone property by proving that Theorem 1 is true.

(1) can be equivalently written as

$$\widehat{\boldsymbol{\Sigma}} \;=\; \text{argmin}\; log\, det(\boldsymbol{\Sigma}) + tr(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + \lambda\|\boldsymbol{\Sigma}^{-1}\|_1 . \tag{B-1}$$

It is known from [17] that the solution, $\widehat{\boldsymbol{\Sigma}}$, is unique with a fixed positive λ, and $\widehat{\boldsymbol{\Sigma}}$ must satisfy the equations in (B-2):

$$\begin{aligned} &(\mathbf{S})_{kl} - (\boldsymbol{\Sigma})_{kl} = -\lambda, \quad \text{for } \left(\boldsymbol{\Sigma}^{-1}\right)_{kl} > 0; \\ &(\mathbf{S})_{kl} - (\boldsymbol{\Sigma})_{kl} = \lambda, \quad \text{for } \left(\boldsymbol{\Sigma}^{-1}\right)_{kl} < 0; \\ &|(\mathbf{S})_{kl} - (\boldsymbol{\Sigma})_{kl}| \le \lambda, \quad \text{for } \left(\boldsymbol{\Sigma}^{-1}\right)_{kl} = 0; \end{aligned} \tag{B-2}$$

where $(\cdot)_{kl}$ denotes the element at the k-th row, l-th column of a matrix.

When $\lambda = \lambda_1$, denote the solution to (B-1) by $\widehat{\boldsymbol{\Sigma}}^{\lambda_1}$. Furthermore, we can rearrange the rows and columns of $\widehat{\boldsymbol{\Sigma}}^{\lambda_1}$, such that $\widehat{\boldsymbol{\Sigma}}^{\lambda_1}$ becomes a block diagonal matrix and each sub-matrix along the main diagonal of the rearranged $\widehat{\boldsymbol{\Sigma}}^{\lambda_1}$ correspond to a cluster of nodes in the SICE-based graphical model. Denote the sub-matrices by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_1}$, $j = 1, \dots, L_1$. Recall that $\mathbf{C}_j^{\lambda_1}$is the j-th cluster of nodes in the graphical model. As a result, $\widehat{\boldsymbol{\Sigma}}^{\lambda_1}$ can be written as:

$$\widehat{\boldsymbol{\Sigma}}^{\lambda_1} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{\mathbf{C}_1^{\lambda_1}}^{\lambda_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{\Sigma}}_{\mathbf{C}_2^{\lambda_1}}^{\lambda_1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\boldsymbol{\Sigma}}_{\mathbf{C}_{L_1}^{\lambda_1}}^{\lambda_1} \end{bmatrix}. \tag{B-3}$$

A sufficient condition for Theorem 1 being true is that the solution to (B-1) when $\lambda = \lambda_2$, denoted by $\widehat{\boldsymbol{\Sigma}}^{\lambda_2}$, must share the same structure as (B-3), i.e., $\widehat{\boldsymbol{\Sigma}}^{\lambda_2}$ can be written as:

$$\widehat{\boldsymbol{\Sigma}}^{\lambda_2} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{\mathbf{C}_1^{\lambda_1}}^{\lambda_2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{\Sigma}}_{\mathbf{C}_2^{\lambda_1}}^{\lambda_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\boldsymbol{\Sigma}}_{\mathbf{C}_{L_1}^{\lambda_1}}^{\lambda_2} \end{bmatrix}. \tag{B-4}$$

To prove this sufficient condition, our strategy will include two steps: step one aims to find a matrix having the same structure as $\widehat{\boldsymbol{\Sigma}}^{\lambda_1}$; step two aims to prove that this matrix is a solution to (B-2) with $\lambda = \lambda_2$.

Step One:

The rows and columns of the sample covariance matrix, $\mathbf{S}$, can be rearranged in the same way as $\widehat{\boldsymbol{\Sigma}}^{\lambda_1}$, i.e.,

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{\mathbf{C}_1^{\lambda_1}} & \cdots & \cdots & \cdots \\ \cdots & \mathbf{S}_{\mathbf{C}_2^{\lambda_1}} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \mathbf{S}_{\mathbf{C}_{L_1}^{\lambda_1}} \end{bmatrix}. \tag{B-5}$$

Next, one optimization problem can be formulated corresponding to one sub-matrix $\mathbf{S}_{\mathbf{C}_j^{\lambda_1}}$, $j = 1, \dots, L_1$, i.e.,

$$\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2} = \operatorname{argmin} log\ det\left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right) + tr\left(\mathbf{S}_{\mathbf{C}_j^{\lambda_1}}\left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)^{-1}\right) + \lambda_2\left\|\left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)^{-1}\right\|_1, \tag{B-6}$$

Furthermore, the solutions to (B-6), i.e., $\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}$, $j = 1, \dots, L_1$, can be put together and form a big matrix $\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}$, i.e.,

$$\widehat{\boldsymbol{\Upsilon}}^{\lambda_2} = \begin{bmatrix} \widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_1^{\lambda_1}}^{\lambda_2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_2^{\lambda_1}}^{\lambda_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_{L_1}^{\lambda_1}}^{\lambda_2} \end{bmatrix}. \tag{B-7}$$

It is obvious that $\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}$ has the same structure as $\widehat{\boldsymbol{\Sigma}}^{\lambda_1}$.

Step Two:

This step aims to prove that the $\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}$ in (B-7) satisfies (B-2) with $\lambda = \lambda_2$. To prove this, we need to prove that (i) the elements in $\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}$, $j = 1, \dots, L_1$, satisfy (B-2), and that (ii) the elements not in $\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}$, all of which are equal to zero, also satisfy (B-2).

(i) Suppose that $\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)_{kl}$ is an element in $\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}$, $j \in \{1, \dots, L_1\}$; more specifically, suppose that $\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)_{kl}$ is the element at the h-th row, s-th column of $\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}$, i.e.,

$$\left(\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)_{hs} = \left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)_{kl}. \tag{B-8}$$

Because $\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}$ is the solution to the optimization in (B-6), it must satisfy (B-9):

$$\begin{aligned} &\left(\mathbf{S}_{\mathbf{C}_j^{\lambda_1}}\right)_{hs} - \left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)_{hs} = -\lambda_2, \quad \text{for } \left(\left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)^{-1}\right)_{hs} > 0; \\ &\left(\mathbf{S}_{\mathbf{C}_j^{\lambda_1}}\right)_{hs} - \left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)_{hs} = \lambda_2, \quad \text{for } \left(\left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)^{-1}\right)_{hs} < 0; \\ &\left|\left(\mathbf{S}_{\mathbf{C}_j^{\lambda_1}}\right)_{hs} - \left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)_{hs}\right| \le \lambda_2, \quad \text{for } \left(\left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)^{-1}\right)_{hs} = 0; \end{aligned} \tag{B-9}$$

It is easy to know that $\left(\mathbf{S}_{\mathbf{C}_j^{\lambda_1}}\right)_{hs}$ is in fact the element at the k-th row, l-th column of $\mathbf{S}$, i.e.,

$$\left(\mathbf{S}_{\mathbf{C}_j^{\lambda_1}}\right)_{hs} = (\mathbf{S})_{kl}; \tag{B-10}$$

and $\left(\left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)^{-1}\right)_{hs}$ is the element at the k-th row, l-th column of $\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)^{-1}$, i.e.,

$$\left(\left(\boldsymbol{\Upsilon}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}\right)^{-1}\right)_{hs} = \left(\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)^{-1}\right)_{kl}. \tag{B-11}$$

Inserting (B-8), (B-10), and (B-11) into (B-9) results in (B-2) with $\lambda = \lambda_2$.

(ii) Suppose that $\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)_{kl}$ is an element not in $\widehat{\boldsymbol{\Upsilon}}_{\mathbf{C}_j^{\lambda_1}}^{\lambda_2}$, $j = 1, \dots, L_1$, i.e., $\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)_{kl} = 0$. Furthermore, it can be known that $\left(\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)^{-1}\right)_{kl} = 0$, because $\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}$ is a block diagonal matrix. Since $\left(\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)^{-1}\right)_{kl} = 0$, to prove that $\left(\widehat{\boldsymbol{\Upsilon}}^{\lambda_2}\right)_{kl}$

satisfies (B-2) with $\lambda = \lambda_2$ is to prove that $\left|(\mathbf{S})_{kl} - \left(\widehat{\mathbf{Y}}^{\lambda_2}\right)_{kl}\right| \leq \lambda_2$. It can be derive that $\left|(\mathbf{S})_{kl} - \left(\widehat{\mathbf{Y}}^{\lambda_2}\right)_{kl}\right| = |(\mathbf{S})_{kl}| = \left|(\mathbf{S})_{kl} - \left(\widehat{\mathbf{\Sigma}}^{\lambda_1}\right)_{kl}\right| \leq \lambda_1$, where the second equality holds because $\left(\widehat{\mathbf{\Sigma}}^{\lambda_1}\right)_{kl} = 0$, and the "$\leq$" holds due to the last equation in (B-1) with $\lambda = \lambda_1$. Also, it has been known that $\lambda_1 \leq \lambda_2$. Therefore, $\left|(\mathbf{S})_{kl} - \left(\widehat{\mathbf{Y}}^{\lambda_2}\right)_{kl}\right| \leq \lambda_1 \leq \lambda_2$. □