Supplement 3. Mathematical Notation of Input Representation, Edge Construction, and Loss Function in the Graph Autoencoder

(a)

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix}$$

(b)

$$\mathrm{A}_{ij} = \begin{cases} 1 & \text{if } |\rho_{ij}| \geq 0.4 \\ 0 & \text{otherwise} \end{cases}$$

(c)

$$\hat{A} = \sigma(ZZ^{\top})$$

(d)

$$L = -\sum_{i,j} [a_{ij} \log \hat{a}_{ij} + (1 - a_{ij}) \log(1 - \hat{a}_{ij})]$$

This figure summarizes the core mathematical components applied in the data preprocessing and Graph Autoencoder modeling process.

(a) Node Feature Matrix:
Represents the input data as a matrix where each row corresponds to a node (either an air pollutant or a disease variable), and each column encodes monthly or regional features.

(b) Adjacency Matrix Construction:
Defines whether an edge exists between pairs of nodes based on the absolute Pearson correlation coefficient. Edges are created only when the correlation exceeds a predefined threshold (0.4).

(c) Adjacency Reconstruction:
The GAE decodes structural similarity among nodes by computing the sigmoid-transformed inner product of learned embeddings, yielding a matrix of predicted connection strengths.

(d) Loss Function:
Binary cross-entropy is used to quantify the difference between the original adjacency matrix and the reconstructed adjacency matrix, guiding the learning process to capture latent structural patterns among variables.

This formulation illustrates the process of transforming raw input data into graph representations suitable for learning and visualizing complex structural relationships.