Appendix A: Proofs of the simple properties

Here, we provide the proofs for the three lemmas in Section 2.3 in the paper.

Lemma 1. *The joint distribution of $X_1, \ldots, X_p$ and the latent variables $L_1, \ldots, L_k$ is Gaussian: $(\mathbf{X}, \mathbf{L}) \sim N(\mathbf{0}, \mathbf{\Sigma_{XL}})$, where $\mathbf{\Sigma_{XL}}$ is a $(p+k) \times (p+k)$ covariance matrix.*

Proof.

$$\begin{aligned} P(\mathbf{X}, \mathbf{L}|\mathbf{\Sigma_L}) &= P(\mathbf{X}|\mathbf{L})P(\mathbf{L}|\mathbf{\Sigma_L}) \\ &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2_{Z_i}}} \exp\left\{-\frac{(X_i - L_{Z_i})^2}{2\sigma^2_{Z_i}}\right\} \cdot \frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma_L}|}} \exp\left\{-\frac{1}{2}\mathbf{L}^\intercal {\mathbf{\Sigma_L}}^{-1}\mathbf{L}\right\} \\ &= \frac{1}{\sqrt{(2\pi)^{p+k}(\sigma^2_{Z_i})^p|\mathbf{\Sigma_L}|}} \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{p} \frac{(X_i - L_{Z_i})^2}{\sigma^2_{Z_i}} + \mathbf{L}^\intercal {\mathbf{\Sigma_L}}^{-1}\mathbf{L}\right)\right\}. \quad (1) \end{aligned}$$

The above distribution can be rewritten as:

$$P(\mathbf{X}, \mathbf{L}|\mathbf{\Sigma_L}) = \frac{1}{\sqrt{(2\pi)^{p+k}|\mathbf{\Sigma_{XL}}|}} \exp\left\{-\frac{1}{2}\left[\mathbf{X}^\intercal \mathbf{L}^\intercal\right] {\mathbf{\Sigma_{XL}}}^{-1} \left[\mathbf{X}^\intercal \mathbf{L}^\intercal\right]^\intercal\right\}, \quad (2)$$

where

$$\mathbf{\Sigma_{XL}} = \begin{bmatrix} \mathbf{A} & \mathbf{C}^\intercal \\ \mathbf{C} & \mathbf{B} \end{bmatrix}^{-1},$$

where $\mathbf{A}$ is a $p \times p$ matrix whose element $A_{ij} = 1/\sigma^2_{Z_i}$ if $i = j$ and 0 otherwise; $\mathbf{C}$ is a $k \times p$ matrix whose element $C_{ij} = -(1/\sigma^2_i)$ if $X_j \in \mathcal{M}_i$ and 0 otherwise; and

$$\mathbf{B} = {\mathbf{\Sigma_L}}^{-1} + \begin{bmatrix} |\mathcal{M}_1|/\sigma^2_1 & \ldots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \ldots & |\mathcal{M}_k|/\sigma^2_k \end{bmatrix}.$$

□

Lemma 2. *The marginal probability distribution of the observed variables $X_1, \ldots, X_p$ is Gaussian: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma_X})$, where $\mathbf{\Sigma_X}$ is a $p \times p$ covariance matrix.*

Proof. If joint distribution of $\mathbf{X} = \{X_1, \ldots, X_p\}$ and the latent variables $\mathbf{L} = \{L_1, \ldots, L_k\}$ is a multivariate Gaussian, then the marginal distribution over a subset of variables is also a multivariate Gaussian distribution. □

Lemma 3. *Let $\mathbf{\Sigma_L}$ be a $k \times k$ covariance matrix of $\mathbf{L}$. The relationship between $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_L}$ is as*

$$\mathbf{\Sigma_X} = \left\{\mathbf{A} - \mathbf{C}^\intercal \mathbf{B}^{-1} \mathbf{C}\right\}^{-1}, \quad (3)$$

where $\mathbf{A}$ is a $p \times p$ matrix whose element $A_{ij} = 1/\sigma^2_{Z_i}$ if $i = j$ and 0 otherwise; $\mathbf{C}$ is a $k \times p$ matrix whose element $C_{ij} = -(1/\sigma^2_i)$ if $X_j \in \mathcal{M}_i$ and 0 otherwise; and

$$\mathbf{B} = {\mathbf{\Sigma_L}}^{-1} + \begin{bmatrix} |\mathcal{M}_1|/\sigma^2_1 & \ldots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \ldots & |\mathcal{M}_k|/\sigma^2_k \end{bmatrix}$$

$|\mathcal{M}_k|$ *meaning the number of* X *variables in the module* k.

Proof. From *Lemma* 1, the distribution described in (1) is equivalent to $N(\mathbf{0}, \mathbf{\Sigma_{XL}})$, which leads to the following:

$$\mathbf{\Sigma_{XL}}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{C}^\intercal \\ \mathbf{C} & \mathbf{B} \end{bmatrix},$$

where $\mathbf{B}$ and $\mathbf{C}$ are defined in (3). Making use of the Schur complement, we obtain the following:

$$\mathbf{\Sigma_{XL}} = \begin{bmatrix} (\mathbf{A} - \mathbf{C}^\intercal\mathbf{B}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{C}^\intercal\mathbf{B}^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{B}^{-1} \\ -\mathbf{B}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{C}^\intercal\mathbf{B}^{-1}\mathbf{C})^{-1} & \mathbf{B}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{C}^\intercal\mathbf{B}^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{B}^{-1} + \mathbf{B}^{-1} \end{bmatrix}.$$

Therefore, the covariance matrix $\mathbf{\Sigma_X}$ is the following:

$$\mathbf{\Sigma_X} = \left\{\mathbf{A} - \mathbf{C}^\intercal\mathbf{B}^{-1}\mathbf{C}\right\}^{-1}.$$

□

Appendix B: Derivation of MGL algorithm

Here, we present our learning algorithm that optimizes the likelihood function based on the joint distribution in the following form:

$$\begin{aligned} P(\mathbf{X}, \mathbf{L}, \mathbf{Z}, \mathbf{\Sigma_L}) &= \prod_{i=1}^{p} P(X_i|L_{Z_i})P(\mathbf{L}|\mathbf{\Sigma_L})P(\mathbf{\Sigma_L}^{-1})P(\mathbf{Z}) \\ &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{Z_i}^2}} \exp\left\{-\frac{(X_i - L_{Z_i})^2}{2\sigma_{Z_i}^2}\right\} \cdot \prod_{j\neq j'} \frac{\lambda}{2} \exp\left\{-\lambda|(\mathbf{\Sigma_L}^{-1})_{jj'}|\right\} \\ &\qquad \cdot \frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma_L}|}} \exp\left\{-\frac{1}{2}\mathbf{L}^\intercal\mathbf{\Sigma_L}^{-1}\mathbf{L}\right\} P(\mathbf{Z}). \end{aligned} \tag{4}$$

We begin by summarizing our model's variables:

- $\mathbf{L} = \{L_1, \ldots, L_k\}$: A set of latent variables $\mathbf{L} \sim N(\mathbf{0}, \mathbf{\Sigma_L})$.
- $\mathbf{\Sigma_L}$: A $k \times k$ covariance matrix of $\mathbf{L}$. The non-zero pattern of $\mathbf{\Sigma_L}^{-1}$ corresponds to the graph structure.
- $\mathbf{Z} = \{Z_1, \ldots, Z_p\}$: Each Z_i indicates the module which X_i is assigned to. $(1 \leq Z_i \leq k)$
- $\mathbf{X} = \{X_1, \ldots, X_p\}$: A set of observed variables each of which is modeled as $X_i \sim N(L_{Z_i}, \sigma_{Z_i}^2)$.
- $\mathbf{\Sigma_X}$: A $p \times p$ covariance matrix of $\mathbf{X}$. We can compute $\mathbf{\Sigma_X}$ based on $\mathbf{\Sigma_L}$ and $\mathbf{Z}$ (see *Lemma* 3).

Estimation of L

In order to estimate L given Z and Θ_L based on the joint log-likelihood function (4), we solve the following optimization problem:

$$\underset{L_1,\ldots,L_k}{\text{maximize}} \left\{ -\text{trace}\left(LL^\intercal \Theta_L\right) - \sum_{i=1}^{p} \frac{\parallel X_i - L_{Z_i} \parallel_2^2}{\sigma_{Z_i}^2} \right\}. \tag{5}$$

where X_i denotes the ith row of the matrix X, L_i denotes the ith row of the matrix L, and the trace term can be rewritten as:

$$\begin{aligned} \text{trace}(LL^\intercal \Theta_L) &= L_1 L_1{}^\intercal (\Theta_L)_{11} + L_1 L_2{}^\intercal (\Theta_L)_{12} + \ldots + L_1 L_k{}^\intercal (\Theta_L)_{1k} \\ &+ L_2 L_1{}^\intercal (\Theta_L)_{21} + L_2 L_2{}^\intercal (\Theta_L)_{22} + \ldots + L_2 L_k{}^\intercal (\Theta_L)_{2k} \\ &+ \ldots \\ &+ L_k L_1{}^\intercal (\Theta_L)_{k1} + L_k L_2{}^\intercal (\Theta_L)_{k2} + \ldots + L_k L_k{}^\intercal (\Theta_L)_{kk}. \end{aligned}$$

Setting the derivative of the objective function in Eq (5) to zero with respect to L_m, we obtain:

$$((\Theta_L)_{1m} L_1 + (\Theta_L)_{2m} L_2 + \ldots + (\Theta_L)_{km} L_k) - \frac{1}{\sigma_m^2} \sum_{X_i \in \mathcal{M}_m} (X_i - L_m) = 0. \tag{6}$$

Here, $\mathcal{M}_m$ means a set of X_i that belongs to the mth module: $\mathcal{M}_m = \{X_i | Z_i = m\}$, and $|\mathcal{M}_m|$ means the number of variables that belong to $\mathcal{M}_m$. Solving the Eq (6) with respect to L_m leads to:

$$L_m = \frac{\sum_{X_i \in \mathcal{M}_m} X_i - \sigma_m^2 \sum_{i \neq m} (\Theta_L)_{im} L_i}{|\mathcal{M}_m| + \sigma_m^2 (\Theta_L)_{mm}}. \tag{7}$$

We update L_m for each m ($1 \leq m \leq k$), based on the current values on the other latent variables $L_1, \ldots, L_{(m-1)}, L_{(m+1)}, \ldots, L_k$.

If all entries in Θ_L equal to zero, L_m would be updated to be the average of X_i's that belong to the mth module. In other words, L_m would be set to be the centroid of the mth module. This leads to a nice interpretation of the MGL learning algorithm with respect to the k-means clustering algorithm.

The k-means clustering algorithm is the special case of the MGL when the latent variables are assumed to be independent; this means that the elements of Θ_L are forced to be zero.

Estimation of Z

In order to estimate Z given L and Θ_L, we solve the following optimization problem:

$$\underset{Z_1,\ldots,Z_p}{\text{maximize}} \left\{ - \sum_{i=1}^{p} \frac{\parallel X_i - L_{Z_i} \parallel_2^2}{\sigma_{Z_i}^2} \right\}, \tag{8}$$

where X_i denotes the ith row of the matrix X and L_i denotes the ith row of the matrix L. When $\sigma_1, \ldots, \sigma_k = 1$, this amounts to finding the assignment for each X_i that minimizes the Euclidean distance with the latent variable.

$$Z_i = \arg \min_{Z_i \in \{1,\ldots,k\}} \left\{ \| X_i - L_{Z_i} \|_2^2 \right\}, \tag{9}$$

Estimation of Θ_L

To estimate Θ_L given L and Z, we solve the following optimization problem:

$$\underset{\Theta_L \succ 0}{\text{maximize}} \left\{ \log\det \Theta_L - \text{trace}\left(S_L \Theta_L\right) - \lambda \sum_{j \neq j'} |(\Theta_L)_{jj'}| \right\}, \tag{10}$$

where S_L is the empirical estimate of the covariance matrix of L: $S_L = \frac{1}{n} L L^\intercal$. Since L is given, the optimization problem (10) can be solved by the standard graphical lasso algorithm applied to L.