

Graphic Lasso: Possible Accuracy for Multi-Layer Model

Jiixin Hu

January 6, 2021

1 Discussion about Identifiability

Suppose we have a dataset with p variables and K categories. In multi-layer model, we assume the rank of decomposition r is known, and the precision matrices are of form

$$\Omega^k = \Theta_0 + \sum_{l=1}^r u_{lk} \Theta_l, \quad \text{for } k = 1, \dots, K. \quad (1)$$

The identifiability problem for $\{\Theta_0, \Theta_1, \dots, \Theta_r, \mathbf{u}_1, \dots, \mathbf{u}_r\}$ is actually an identifiability problem for tensor decomposition.

Let $\mathcal{Y} \in \mathbb{R}^{p \times p \times K}$ denote the collection of K networks, where $\mathcal{Y}[:, k] = \Omega^k, k \in [K]$. Let $\mathcal{C} \in \mathbb{R}^{p \times p \times (r+1)}$ denote the collection of “core” networks, where $\mathcal{C}[:, 1] = \sqrt{K} \Theta_0, \mathcal{C}[:, l] = \Theta_{l-1}, l = 2, \dots, (r+1)$. Let $\mathbf{U} \in \mathbb{R}^{K \times (r+1)} = (\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_r)$ denote the factor matrix, where $\mathbf{u}_0 = \mathbf{1}_K / \sqrt{K}$. Rewrite the model (1) in tensor form.

$$\mathcal{Y} = \mathcal{C} \times_3 \mathbf{U}. \quad (2)$$

Therefore, the identifiability problem for $\{\Theta_l, \mathbf{u}_l\}$ becomes the identifiability problem for $\{\mathcal{C}, \mathbf{U}\}$. Before we discuss the identifiable condition case by case, we first assume \mathcal{C} is full rank on mode 3.

1. No sparsity constrain on \mathbf{U} .

Proposition 1. *The decomposition \mathcal{C} and \mathbf{U} are identifiable if \mathbf{U} is an orthonormal matrix, i.e., $\mathbf{U}^T \mathbf{U} = \mathbf{I}_{r+1}$.*

Proof. Let $\text{Unfold}(\cdot)$ denote the unfold representation of a tensor on mode 3. The model (2) is equal to

$$\text{Unfold}(\mathcal{Y}) = \mathbf{U} \text{Unfold}(\mathcal{C}).$$

By matrix SVD, we have $\text{Unfold}(\mathcal{Y}) = \tilde{\mathbf{U}} \Sigma \mathbf{V}^T$, where $\tilde{\mathbf{U}}$ is an orthonormal matrix. The SVD decomposition is unique up to orthogonal rotation (ignore row permutation).

Note that $\mathbf{u}_0 = \mathbf{1}_K / \sqrt{K}$. There always has a unique orthonormal matrix \mathbf{R} such that the first column of $\tilde{\mathbf{U}} \mathbf{R}$ is equal to $\mathbf{1}_K / \sqrt{K}$. Let $\mathbf{U} = \tilde{\mathbf{U}} \mathbf{R}$ and $\text{Unfold}(\mathcal{C}) = \mathbf{R}^T \Sigma \mathbf{V}$. Then, \mathbf{U} and \mathcal{C} are identifiable. \square

2. Membership constrain on \mathbf{U} . (Without intercept Θ_0)

If \mathbf{U} is a membership matrix, we are clustering K categories into r groups. Then, the model (1) becomes

$$\Omega^k = \Theta_{i_k}, \quad \text{for } k = 1, \dots, K,$$

where $i_k \in [r]$ is the group for the k -th category. Then, let $\mathcal{C} \in \mathbb{R}^{p \times p \times r}$, where $\mathcal{C}[:, l] = \Theta_l, l = 1, \dots, r$, and $\mathbf{U} \in \mathbb{R}^{K \times r} = (\mathbf{u}_1, \dots, \mathbf{u}_r)$.

Proposition 2. *The decomposition \mathcal{C} and \mathbf{U} are identifiable up to permutation if \mathbf{U} is a membership matrix, i.e., in each row of \mathbf{U} there is only 1 copy of 1 and massive 0.*

Proof. If \mathbf{U} is a membership matrix, the model (2) is a special case of tensor block model. By Proposition 1 in Wang, the matrix \mathbf{U} is identifiable if \mathcal{C} is irreducible on mode 3. In our case, we assume \mathcal{C} is full rank on mode 3, and thus $\{\mathbf{U}, \mathcal{C}\}$ are identifiable. \square

Remark 1. The sparsity of Θ_l won't affect the identifiability in these two cases under the assumption that \mathcal{C} is full rank on mode 3. In no sparsity constrain case, we only need the full rankness of $\text{Unfold}(\mathcal{C})$, and the sparsity on the first and second mode of \mathcal{C} does not affect the rank of $\text{Unfold}(\mathcal{C})$. In membership constrain, we only need the mode 3 irreducibility of \mathcal{C} .

Remark 2. The two cases above are two extreme cases. Intermediate cases include the fuzzy clustering, where $\sum_{l=1}^r u_{lk} = 1, k \in [K]$, and the sparsity constrain for the column, where $|\mathbf{u}_l|_0 < a, l \in [r]$.

2 A simple extension

bad notation. The function $Q(\cdot)$ varies depending on k , because \mathbf{S} depends on k .

Let $Q(\Omega) = \text{tr}(S\Omega) - \log |\Omega|$. Assume the rank of decomposition r is known. Consider the constrained optimization problem

$$\begin{aligned} \min_{\mathcal{C}} \quad & \sum_{k=1}^K [Q(\Omega^k)] \\ \text{s.t.} \quad & \Omega^k = \Theta_0 + \sum_{l=1}^r u_{lk} \Theta_l, \quad \text{for } k = 1, \dots, K, \\ & \|\Theta_l\|_0 \leq b, \quad \text{for } l = 1, \dots, r, \\ & \|\Theta_0\|_0 \leq b_0, \\ & \mathbf{u}_l^T \mathbf{u}_l = 1, \quad \text{for } l = 1, \dots, r, \\ & \mathbf{u}_k^T \mathbf{u}_l = 0, \quad \text{for } k \neq l. \end{aligned}$$

where a, b, b_0 are fixed positive constants, $|\cdot|_0$ refers to the vector L_0 norm, and $\|\cdot\|_0$ refers to the matrix L_0 norm. For simplicity, let $\hat{\mathcal{C}} = \{\hat{\Theta}_0, \hat{\Theta}_1, \dots, \hat{\Theta}_r, \hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_r\}$ denote the estimation, and $\hat{\Omega}^k = \hat{\Theta}_0 + \sum_{l=1}^r \hat{u}_{lk} \hat{\Theta}_l$ for $k = 1, \dots, K$.

For true precision matrices Ω^k , let $T^k = \{(j, j') | \omega_{j,j'}^k \neq 0\}$ and $q^k = |T^k|$. Let $T = T^1 \cup \dots \cup T^K$ and $q = |T|$.

Theorem 2.1. *Suppose two assumptions hold. Let $\{\Omega^k\}$ denote the true precision matrices. For the estimation $\hat{\mathcal{C}}$ such that $\sum_{k=1}^K [Q(\hat{\Omega}^k)] \leq \sum_{k=1}^K [Q(\Omega^k)]$ and satisfies the constrains, the following accuracy bound holds with probability tending to 1.*

What is n ?

How does the convergence depend on K ?

Write down the closed form.

$$\sum_{k=1}^K \left\| \hat{\Omega}^k - \Omega^k \right\|_F = \mathcal{O}_p \left[\left\{ \frac{(p+q) \log p}{n} \right\}^{1/2} \right].$$

Proof. Let Ω^k denote the true precision matrices for $k = 1, \dots, K$. Consider the estimation $\hat{\mathcal{C}}$ such that $\sum_{k=1}^K [Q(\hat{\Omega}^k)] \leq \sum_{k=1}^K [Q(\Omega^k)]$. Let $\Delta^k = \hat{\Omega}^k - \Omega^k$. Define the function

$$G(\{\Delta^k\}) = \sum_{k=1}^K \text{tr}(S(\Omega^k + \Delta^k)) - \text{tr}(\Omega^k) - \log |\Omega^k + \Delta^k| + \log |\Omega^k| = I_1 + I_2,$$

where

$$I_1 = \sum_{k=1}^K \text{tr}((S^k - \Sigma^k)\Delta^k), \quad I_2 = \sum_{k=1}^K (\tilde{\Delta}^k)^T \int_0^1 (1-v)(\Omega^k + v\Delta^k)^{-1} \otimes (\Omega^k + v\Delta^k)^{-1} dv \tilde{\Delta}^k.$$

explain the notation

With probability tending to 1, we have

$$I_1 \leq C_1 \left(\frac{\log p}{n} \right)^{1/2} \sum_{k=1}^K (|\Delta_{T^k}^k|_1 + |\Delta_{T^{k,c}}^k|_1) + C_2 \left(\frac{p \log p}{n} \right)^{1/2} \sum_{k=1}^K \|\Delta^k\|_F, \quad I_2 \geq \frac{1}{4\tau_2^2} \sum_{k=1}^K \|\Delta^k\|_F^2.$$

Note that $|\Delta_{T^k}^k|_1 \leq q^{1/2} \|\Delta^k\|_F$. Then, we only need to deal with $|\Delta_{T^{k,c}}^k|_1$. Rewrite the term, we have

$$|\Delta_{T^{k,c}}^k|_1 = |\hat{\Theta}_{0,T^{k,c}} + \hat{u}_{1k}\hat{\Theta}_{1,T^{k,c}} + \dots + \hat{u}_{rk}\hat{\Theta}_{r,T^{k,c}}|_1 \leq (b_0 + rb) \|\Delta^k\|_{\max} \leq (b_0 + rb) \|\Delta^k\|_F.$$

why not q_k?

Then, by Guo et al., we have

How? elaborate.

Pay special attention to factors involving Δ^k

$$\sum_{k=1}^K \|\Delta^k\|_F = \sum_{k=1}^K \|\hat{\Omega}^k - \Omega^k\|_F = \mathcal{O}_p \left[\left\{ \frac{(p+q) \log p}{n} \right\}^{1/2} \right]. \quad (3)$$

□

Remark 3. Note that q can be replaced by $\max_k q^k$, where $q^k \leq (b_0 + rb)$ for all $k = 1, \dots, K$. Also, the accuracy (3) holds when q^k are fixed. Otherwise, the accuracy is of order $\mathcal{O}_p \left[q \left\{ \frac{\log p}{n} \right\}^{1/2} \right]$.

Remark 4. This proof does not utilize the special structure of Ω^k . We can go through the proof with the constrain $|\Omega^k| < s$.

3 Next

- Think about the identifiability of the intermediate cases (sparse matrix factorization).
- Think about the proof which utilizes the special structure of the Ω^k .