# Graphic Lasso: Possible Accuracy for Multi-Layer Model

Jiaxin 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

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

The identifiability problem for $\{\Theta_0, \Theta_1, ..., \Theta_r, \mathbf{u}_1, ..., \mathbf{u}_r\}$ 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 $\mathcal{C}[,,1] = \sqrt{K}\Theta_0$, $\mathcal{C}[,,l] = \Theta_{l-1}, l = 2, ..., (r+1)$. Let $\boldsymbol{U} \in \mathbb{R}^{K\times(r+1)} = (\mathbf{u}_0, \mathbf{u}_1, ..., \mathbf{u}_r)$ denote the factor matrix, where $\mathbf{u}_0 = \mathbf{1}_K/\sqrt{K}$. Rewrite the model (1) in tensor form.

$$\mathcal{Y} = \mathcal{C} \times_3 \boldsymbol{U}. \tag{2}$$

Therefore, the identifiability problem for $\{\Theta_l, \mathbf{u}_l\}$ becomes the identifiability problem for $\{\mathcal{C}, \boldsymbol{U}\}$. 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 $\boldsymbol{U}$.**

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

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

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

   By matrix SVD, we have Unfold$(\mathcal{Y}) = \tilde{\boldsymbol{U}}\Sigma\boldsymbol{V}^T$, where $\tilde{\boldsymbol{U}}$ 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 $\boldsymbol{R}$ such that the first column of $\tilde{\boldsymbol{U}}\boldsymbol{R}$ is equal to $\mathbf{1}_K/\sqrt{K}$. Let $\boldsymbol{U} = \tilde{\boldsymbol{U}}\boldsymbol{R}$ and Unfold$(\mathcal{C}) = \boldsymbol{R}^T\Sigma\boldsymbol{V}$. Then, $\boldsymbol{U}$ and $\mathcal{C}$ are identifiable. □

2. **Membership constrain on $\boldsymbol{U}$. (Without intercept $\Theta_0$)**

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

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

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 $\mathcal{C}[,,l] = \Theta_l, l = 1, ..., r$, and $\boldsymbol{U} \in \mathbb{R}^{K\times r} = (\mathbf{u}_1, ..., \mathbf{u}_r)$.

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

*Proof.* If $\boldsymbol{U}$ is a membership matrix, the model (2) is a special case of tensor block model. By Proposition 1 in Wang, the matrix $\boldsymbol{U}$ 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 $\{\boldsymbol{U}, \mathcal{C}\}$ are identifiable. □

**Remark 1.** The sparsity of $\Theta_l$ 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 Unfold($\mathcal{C}$), and the sparsity on the first and second mode of $\mathcal{C}$ does not affect the rank of 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 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

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

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 $\hat{\mathcal{C}} = \{\hat{\Theta}_0, \hat{\Theta}_1, ..., \hat{\Theta}_r, \hat{\mathbf{u}}_1, ...., \hat{\mathbf{u}}_r\}$ denote the estimation, and $\hat{\Omega}^k = \hat{\Theta}_0 + \sum_{l=1}^r \hat{u}_{lk}\hat{\Theta}_l$ for $k = 1, ..., K$.

For true precision matrices $\Omega^k$, let $T^k = \{(j, j')|\omega^k_{j,j'} \ne 0\}$ and $q^k = |T^k|$. Let $T = T^1 \cup \cdots \cup T^k$ 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 $\sum_{k=1}^K \left[Q(\hat{\Omega}^k)\right] \le \sum_{k=1}^K \left[Q(\Omega^k)\right]$ 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 $\Omega^k$ denote the true precision matrices for $k = 1, ..., K$. Consider the estimation $\hat{\mathcal{C}}$ such that $\sum_{k=1}^K \left[Q(\hat{\Omega}^k)\right] \le \sum_{k=1}^K \left[Q(\Omega^k)\right]$. Let $\Delta^k = \hat{\Omega}^k - \Omega^k$. Define the function

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

where

explain the notation

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

With probability tending to 1, we have

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

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

why not q_k?

$$|\Delta^k_{T^{k,c}}|_1 = |\hat{\Theta}_{0,T^{k,c}} + \hat{u}_{1k}\hat{\Theta}_{1,T^{k,c}} + \cdots + \hat{u}_{rk}\hat{\Theta}_{r,T^{k,c}}|_1 \le (b_0 + rb)\left\|\Delta^k\right\|_{\max} \le (b_0 + rb)\left\|\Delta^k\right\|_F.$$

Then, by Guo et al, we have

How? elaborate.
Pay special attention to factors involving K

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

□

**Remark 3.** Note that $q$ can be replaced by $\max_k q^k$, where $q^k \le (b_0 + rb)$ for all $k = 1, ..., K$. 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 $\Omega^k$. We can go through the proof with the constrain $|\Omega^k| < s$.

add a section: How does the accuracy result differs between our constrained estimator vs. penalized estimator.
Any reason?

# 3 Next

- Think about the identifiability of the intermediate cases (spare matrix factorization).
- Think about the proof which utilizes the special structure of the $\Omega^k$.