

# Graphic Lasso: Possible Accuracy for Multi-Layer Model

Jiixin Hu

January 11, 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 $\Theta_0$ )

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  $\Theta_l$  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  $\text{Unfold}(\mathcal{C})$ , 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

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

$$\begin{aligned} \min_{\mathcal{C}} \quad & \sum_{k=1}^K [Q^k(\Omega^k)] \\ \text{s.t.} \quad & \Omega^k = \Theta_0 + \sum_{l=1}^r u_{lk} \Theta_l, \quad \text{for } k = 1, \dots, K, \\ & \|\Theta_l\|_0 \leq b, \quad \text{for } l = 1, \dots, r, \\ & \|\Theta_0\|_0 \leq b_0, \\ & \text{with more identifiability conditions,} \end{aligned}$$

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  $\Omega^k$ , let  $T^k = \{(j, j') | \omega_{j, j'}^k \neq 0\}$  and  $q^k = |T^k|$ . Assume  $0 < \tau_1 < \phi_{\min}(\Omega^k) \leq \phi_{\max}(\Omega^k) < \tau_2, k = 1, \dots, K$ , for some positive constant  $\tau_1, \tau_2$ . **This condition can be transferred as some conditions for  $\{\Theta_l, \mathbf{u}_l\}$ . I will specify the conditions later.**

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

$$\sum_{k=1}^K \left\| \hat{\Omega}^k - \Omega^k \right\|_F \leq CK \left( C_1(b_0 + rb) \left( \frac{\log p}{n} \right)^{1/2} + C_2 \left( \frac{p \log p}{n} \right)^{1/2} \right),$$

where  $n$  is the sample size for each category, and  $C, C_1, C_2$  are positive constants independent with  $p, n$ .

*Proof.* Let  $\Omega^k$  denote the true precision matrices for  $k = 1, \dots, K$ . Consider the estimation  $\hat{C}$  such that  $\sum_{k=1}^K [Q^k(\hat{\Omega}^k)] \leq \sum_{k=1}^K [Q^k(\Omega^k)]$ . Let  $\Delta^k = \hat{\Omega}^k - \Omega^k$ . For a matrix  $M$ , let  $M_T$  denote the matrix  $M$  with all elements outside the index set  $T$  replaced by 0, and  $\tilde{M} = \text{vec}(M)$  be the vectorization of  $M$ . Define the function

$$G(\{\Delta^k\}) = \sum_{k=1}^K \text{tr}(S(\Omega^k + \Delta^k)) - \text{tr}(\Omega^k) - \log |\Omega^k + \Delta^k| + \log |\Omega^k| = I_1 + I_2, \quad (3)$$

where

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

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

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

To let the equation (3) smaller than 0, we have

$$I_2 \leq -I_1 \leq |I_1|. \quad (4)$$

Plugging the upper bound of  $|I_1|$  and the lower bound of  $I_2$  into the inequality (4), we have

$$\frac{1}{4\tau_2^2} \sum_{k=1}^K \|\Delta^k\|_F^2 \leq C_1 \left( \frac{\log p}{n} \right)^{1/2} \sum_{k=1}^K (\sqrt{q^k} \|\Delta^k\|_F + (b_0 + rb) \|\Delta^k\|_F) + C_2 \left( \frac{p \log p}{n} \right)^{1/2} \sum_{k=1}^K \|\Delta^k\|_F.$$

By Cauchy Schwartz inequality, we know that  $\sum_{k=1}^K \|\Delta^k\|_F^2 \geq \frac{1}{K} (\sum_{k=1}^K \|\Delta^k\|_F)^2$ . Also, note that  $q^k \leq (b_0 + rb), k = 1, \dots, K$ . Dividing by  $\sum_{k=1}^K \|\Delta^k\|_F$  on both sides of the inequality, we obtain the accuracy rate

$$\sum_{k=1}^K \|\Delta^k\|_F = \sum_{k=1}^K \|\hat{\Omega}^k - \Omega^k\|_F \leq 4\tau_2^2 K \left( C_1(b_0 + rb) \left( \frac{\log p}{n} \right)^{1/2} + C_2 \left( \frac{p \log p}{n} \right)^{1/2} \right). \quad (5)$$

□

**Remark 3.** The accuracy (5) 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  $|\Omega^k| < s$ .

**Remark 5.** Both accuracy results of our constrained estimator and penalized estimator are of order  $F(p, q) \left( \frac{\log p}{n} \right)^{1/2}$ , where  $F(p, q) = (p + q)^{1/2}$  for penalized estimator and  $F(p, q) = (p + q^2)^{1/2}$  with  $q = b_0 + rb$  in our estimator. In case of growing  $(p, n)$  and fixed  $q$ , the two estimators share the same accuracy rate.

### 3 Others

#### Can the factor $K$ be improved?

First, consider the case  $r = 0, K > 1$ . Then, we have  $\Theta_0 = \Omega^k, k = 1, \dots, K$ . Let  $\hat{\Theta}_0$  be the estimator of  $\Omega^k, k = 1, \dots, K$ , and thus  $\Delta^k = \hat{\Omega}^k - \Omega^k = \Delta, k = 1, \dots, K$ . Define the function

$$G(\Delta) = \frac{1}{K} \sum_{k=1}^K \text{tr}(S^k(\Theta_0 + \Delta)) - \text{tr}(S^k \Theta_0) - \log |\Theta_0 + \Delta| + \log |\Theta_0| = I_1 + I_2,$$

where

$$I_1 = \text{tr}\left(\left(\frac{1}{K} \sum_{k=1}^K S^k - \Sigma\right)\Delta\right), \quad I_2 = (\tilde{\Delta})^T \int_0^1 (1-v)(\Theta_0 + v\Delta)^{-1} \otimes (\Theta_0 + v\Delta)^{-1} dv \tilde{\Delta}.$$

Note that  $\frac{1}{K} \sum_{k=1}^K S^k$  can be considered as the sample covariance matrix with sample size  $nK$ . Then, the upper bound for  $|I_1|$  is

$$|I_1| \leq C_1 \left(\frac{\log p}{nK}\right)^{1/2} (|\Delta|_1) + C_2 \left(\frac{p \log p}{nK}\right)^{1/2} \|\Delta\|_F.$$

Since  $I_2 \geq \frac{1}{4\tau_2^2} \|\Delta\|_F^2$ ,  $|\Delta|_1 \leq \sqrt{q} \|\Delta\|_F + (b_0 + rb) \|\Delta\|_F$ , and we need  $I_2 \leq |I_1|$ , we obtain the error bound

$$\|\Delta\|_F^2 \leq \frac{4\tau_2^2}{K^{1/2}} F(p, q, n) \|\Delta\|_F,$$

and thus  $\|\Delta\|_F = \|\hat{\Theta}_0 - \Theta_0\|_F$  decreases in  $K$  of order  $\mathcal{O}(K^{-1/2})$ .

This result agrees with the intuition. As  $K$  growing, the sample size for estimating the  $\Theta_0$  becomes larger. Then, the error for the estimation goes smaller.

#### My thoughts.

Consider the problem for scalar. Let  $Y_{ij} \sim_{i.i.d.} N(\mu, \sigma^2), i = 1, \dots, n, j = 1, \dots, K$ . Then, we have

$$\sum_{j=1}^K \sum_{i=1}^n (Y_{ij} - \bar{Y})^2 = \sum_{j=1}^K \sum_{i=1}^n (Y_{ij} - Y_{.j})^2 + \sum_{j=1}^K n(Y_{.j} - \bar{Y})^2,$$

where  $\bar{Y} = \frac{1}{nK} \sum_{i,j} Y_{ij}$  and  $Y_{.j} = \frac{1}{n} \sum_i Y_{ij}$ . Note that  $Y_{.j} \sim_{i.i.d.} N(\mu, \frac{\sigma^2}{n})$ . We have

$$\frac{1}{K} \sum_{j=1}^K (Y_{.j} - \bar{Y})^2 \rightarrow_{a.s.} \frac{\sigma^2}{n},$$

as  $K \rightarrow \infty$ . For all  $\epsilon > 0$ , we have  $n, K$  large enough such that

$$\frac{1}{nK} \sum_{j=1}^K \sum_{i=1}^n (Y_{ij} - \bar{Y})^2 = \frac{1}{nK} \sum_{j=1}^K \sum_{i=1}^n (Y_{ij} - Y_{.j})^2 + \epsilon.$$

The term  $\frac{1}{nK} \sum_{j=1}^K \sum_{i=1}^n (Y_{ij} - \bar{Y})^2$  is the sample variance with sample size  $nK$ , and  $\frac{1}{nK} \sum_{j=1}^K \sum_{i=1}^n (Y_{ij} - Y_{.j})^2$  is the average of the sample variance for each group.

In multi-layer model, let  $S$  denote the sample covariance matrix with sample size  $nK$  and  $S^k$  be the sample covariance matrix for each group. Similarly as the scalar example, we may have  $S$  and  $\frac{1}{K} \sum_k S^k$  close enough when  $n, K$  are large, and then we can go through the above proof.

For the log-determinant term,

$$\sum_{k=1}^K \log |\Omega^k| - \log |\Omega^k + \Delta^k| \quad \text{is replaced by} \quad K \log |\Theta_0| - K \log |\Theta_0 + \Delta|.$$

Consider the function  $f(t) = \log |\Omega' + t\Delta'|$ . By Taylor expansion for  $t = 1$  around  $t = 0$ , we have

$$f(1) - f(0) = \log |\Omega'| - \log |\Omega' + \Delta'| = \text{tr}(\Sigma' \Delta') + (\tilde{\Delta}')^T \int_0^1 (1-v)(\Omega' + v\Delta')^{-1} \otimes (\Omega' + v\Delta')^{-1} dv \tilde{\Delta}'.$$

The Taylor expansion takes derivatives of  $t$ . It seems impossible to let  $K \log |\Theta_0| - K \log |\Theta_0 + \Delta|$  unrelated with  $K$ ?

I do not know. I do believe increasing K should increases the accuracy. Here, the optimal Delta

**Comparison Table** in red depends on K, so hard to tell whether the entire thing is related to K or not.

Try searching related literature and see whether others' work can give you some inspiration.

## 4 Next

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

	Penalized	
	$L_0$	$L_1$
Ground Truth	For $k \in [K]$ , $ \Omega^k _0 < s$ , where $ \cdot _0$ denote the number of nonzero elements in the matrix, and $s > 1$ is a positive constant.	For $k \in [K]$ , $ \Omega^k _1 < c$ , where norm $ \Omega _1 = \sum_{(i,j)}  \omega_{ij} $ .
Fitting techniques	The estimator is the solution to the optimization problem $\min_{\{\Omega^k\}} Q^k(\Omega^k) + \lambda \sum_{k=1}^K  \Omega^k _0.$	The estimator is the solution to the optimization problem $\min_{\{\Omega^k\}} Q^k(\Omega^k) + \lambda \sum_{k=1}^K  \Omega^k _1.$
Accuracy	For $\lambda \geq 0$ , $\sum_k \ \Delta_k\ _F^2 \leq 4\tau_2^2 \left( F(p, q, n) \sum_k \ \Delta^k\ _F + K\lambda q \right) \sum_k \ \Delta_k\ _F^2 \leq 4\tau_2^2 (\lambda\sqrt{q} + F(p, q, n)) \sum_k \ \Delta^k\ _F,$ where $F(p, q, n) = C_1 p \left( \frac{\log p}{n} \right)^{1/2} + C_2 \left( \frac{p \log p}{n} \right)^{1/2}$ . Then, the error $\sum_k \ \Delta_k\ _F = \mathcal{O}\left(\frac{K\sqrt{\lambda}}{n^{1/4}}\right).$ If $\lambda \geq \Lambda_1 \left( \frac{\log p}{n} \right)^{1/2}$ , the error becomes of order $\mathcal{O}(K/n^{1/2})$ .	For $\lambda \geq \Lambda_1 \left( \frac{\log p}{n} \right)^{1/2}$ , we have $\sum_k \ \Delta_k\ _F^2 \leq 4\tau_2^2 (\lambda\sqrt{q} + F(p, q, n)) \sum_k \ \Delta^k\ _F,$ where $F(p, q, n) = C_1 \left( \frac{q \log p}{n} \right)^{1/2} + C_2 \left( \frac{p \log p}{n} \right)^{1/2}$ . Then, we have $\sum_k \ \Delta^k\ _F \leq 4\tau_2^2 K (\lambda\sqrt{q} + F(p, q, n)).$

	Constrained	
	$L_0$	$L_1$
Ground Truth	<p>For <math>k \in [K]</math>,</p> $ \Omega^k _0 < s$ <p>, where <math> \cdot _0</math> denote the number of nonzero elements in the matrix, and <math>s &gt; 1</math> is a positive constant.</p>	<p>For <math>k \in [K]</math>,</p> $ \Omega^k _1 < c,$ <p>where norm <math> \Omega _1 = \sum_{(i,j)}  \omega_{ij} </math>.</p>
Fitting techniques	<p>The estimator is the solution to the optimization problem</p> $\min_{\{\Omega^k\}} Q^k(\Omega^k)$ <p>s.t. <math> \Omega^k _0 &lt; s, \quad k \in [K]</math></p>	<p>The estimator is the solution to the optimization problem</p> $\min_{\{\Omega^k\}} Q^k(\Omega^k)$ <p>s.t. <math> \Omega^k _1 &lt; c, \quad k \in [K]</math></p>
Accuracy	<p>We have</p> $\sum_k \ \Delta_k\ _F^2 \leq 4\tau_2^2 F(p, s, n) \sum_k \ \Delta^k\ _F,$ <p>where <math>F(p, s, n) = C_1 s \left(\frac{\log p}{n}\right)^{1/2} + C_2 \left(\frac{p \log p}{n}\right)^{1/2}</math>. Then</p> $\sum_k \ \Delta_k\ _F \leq 4\tau_2^2 K F(p, s, n).$	<p>We have</p> $\sum_k \ \Delta_k\ _F^2 \leq 4\tau_2^2 F(p, q, n) \sum_k \ \Delta^k\ _F,$ <p>where <math>F(p, q, n) = C_1 p \left(\frac{\log p}{n}\right)^{1/2} + C_2 \left(\frac{p \log p}{n}\right)^{1/2}</math>. Then,</p> $\sum_k \ \Delta_k\ _F \leq 4\tau_2^2 K F(p, q, n).$