Graphic Lasso: Numerical Implementation and Statistical Property

Jiaxin Hu

December 27, 2020

- Numerical evidence for R implementation.
- Review of *“Joint estimation of multiple graphical models”*

1 Numerical evidence for R implementation

First, the tuning parameter ρ has the same definition in Matlab and R: the regularization parameter in the marginal lasso.

Next, we consider the random network case ($k = 5$). Figure 1 plots the ground truth, optimal Matlab output, and optimal R output, which implies optimal R output may recover better than Matlab. Note that “optimal” refers to the case where the output sparsity close to the ground truth sparsity.

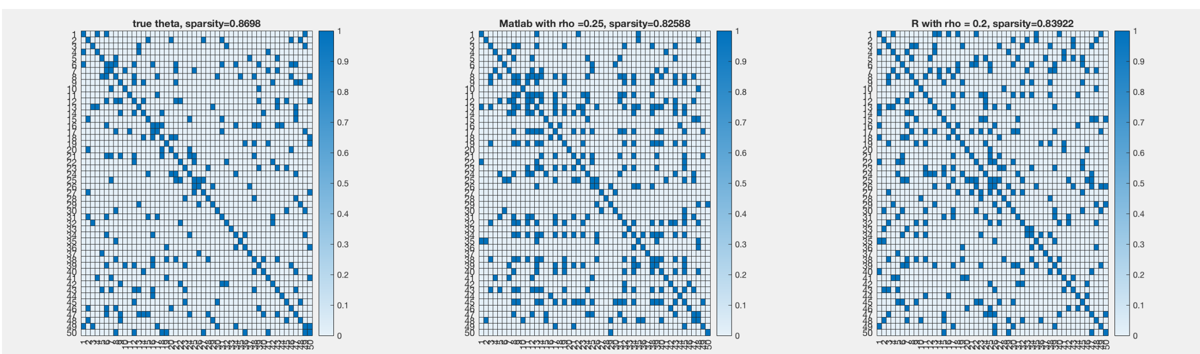

Figure 1: Ground truth, optimal Matlab output, and optimal R output. Here $k = 5$.

Tables 1 and 2 show the false positivity rate (FPR) and false negativity rate (FNR) of Matlab and R implementation with different ρ.

FP	$\rho = 0.15$	$\rho = 0.2$	$\rho = 0.25$	$\rho = 0.3$
Matlab	0.822	0.468	**0.123**	0.043
R	0.193	**0.098**	0.039	0.016

Table 1: False positivity rate of Matlab and R implementation with different ρ. The bold numbers are results under the optimal cases.

Under the optimal cases, the R implementation outperforms Matlab in both FPR and FNR. Therefore, I believe R implementation is better.

FN	$\rho = 0.15$	$\rho = 0.2$	$\rho = 0.25$	$\rho = 0.3$
Matlab	0.095	0.336	**0.689**	0.853
R	0.448	**0.603**	0.759	0.871

Table 2: False negativity rate of Matlab and R implementation with different ρ. The bold numbers are results under the optimal cases.

2 Review of Guo's paper

2.1 Model

Consider a dataset with p variables and K categories. The k-th category contains n_k observations $(x_1^{(k)}, ..., x_{n_k}^{(k)})$, where $x_i^{(k)} = (x_{i,1}^{(k)}, ..., x_{i,p}^{(k)})$ is a p-dimensional vector for $i = 1, ..., n_k$. Assume $x_1^{(k)}, ..., x_{n_k}^{(k)}$ i.i.d. follows a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma^{(k)})$, for $k = 1, ..., K$. Let $\Omega^{(k)} = (\Sigma^{(k)})^{-1} = [\![\omega_{ij}^{(k)}]\!] \in \mathbb{R}^{p\times p}$. Our goal is to estimate $\{\Omega^{(k)}\}_{k=1}^K$.

Since different categories may share some common structure, we further assume $\omega_{j,j'} = \theta_{j,j'}\gamma_{j,j'}^{(k)}$, where $\theta_{j,j'} \geq 0$ controls the link of nodes j and j' across all categories, and $\gamma_{j,j'}$ reflects the differences between categories. Let $\Theta = [\![\theta_{j,j'}]\!] \in \mathbb{R}^{p\times p}$ and $\Gamma^{(k)} = [\![\gamma_{j,j'}]\!] \in \mathbb{R}^{p\times p}$ for $k = 1, ..., K$. The proposed joint estimator is the solution to the following minimization problem.

$$\min_{\Theta, \{\Gamma^{(k)}\}_{k=1}^K} \sum_{k=1}^K \left[\mathrm{tr}(S^{(k)}\Omega^{(k)}) - \log\det(\Omega^{(k)})\right] + \eta_1 \sum_{j\neq j'} \theta_{j,j'} + \eta_2 \sum_{j\neq j'}\sum_{k=1}^K |\gamma_{j,j'}^{(k)}|, \tag{1}$$

where $S^{(k)}$ is the sample covariance matrix of k-th categories, η_1 and η_2 are two tuning parameter, $\Omega^{(k)} = \Theta \cdot \Gamma^{(k)}$ and $\cdot$ refers to the Schur-Hadamard product (element-wise product).

Reformulate the minimization problem (1). We obtain the single tuning parameter problem (2).

$$\min_{\Theta, \{\Gamma^{(k)}\}_{k=1}^K} \sum_{k=1}^K \left[\mathrm{tr}(S^{(k)}\Omega^{(k)}) - \log\det(\Omega^{(k)})\right] + \sum_{j\neq j'} \theta_{j,j'} + \eta \sum_{j\neq j'}\sum_{k=1}^K |\gamma_{j,j'}^{(k)}|, \tag{2}$$

where $\eta = \eta_1\eta_2$.

Further, we reformulate the minimization problem to the problem (3), which is a minimization problem of $\{\Omega^{(k)}\}$ for computational purposes.

$$\min_{\{\Omega^{(k)}\}_{k=1}^K} \sum_{k=1}^K \left[\mathrm{tr}(S^{(k)}\Omega^{(k)}) - \log\det(\Omega^{(k)})\right] + \lambda \sum_{j\neq j'} \left(\sum_{k=1}^K |\omega_{j,j'}^{(k)}|\right)^{1/2}, \tag{3}$$

where $\lambda = 2\eta^{1/2} = 2(\eta_1\eta_2)^{1/2}$.

To compare, the equation (4) is the separate multiple network estimation problem.

$$\min_{\{\Omega^{(k)}\}_{k=1}^K} \sum_{k=1}^K \left[\mathrm{tr}(S^{(k)}\Omega^{(k)}) - \log\det(\Omega^{(k)})\right] + \lambda \sum_{j\neq j'}\sum_{k=1}^K |\omega_{j,j'}^{(k)}|, \tag{4}$$

where λ is a tuning parameter. Note that equation (4) can be decomposed to K marginal minimization problem while the joint estimation (3) can not be separated because of the $\left(\sum_{k=1}^{K} |\omega_{j,j'}^{(k)}|\right)^{1/2}$ term.

2.2 Asymptotic properties

Let $(\Omega^{(1)}, ..., \Omega^{(K)})$ denote the true precision matrices and $(\hat{\Omega}^{(1)}, ..., \hat{\Omega}^{(K)})$ denote the local minimizer of (3). Let $T_k = \{(j, j') : j \neq j', \omega_{j,j'}^{(k)} \neq 0\}$, $T = T_1 \cup \cdots \cup T_K$, and $q_k = |T_k|$ and $q = |T|$.

Assumptions

1. There exist two constants τ_1, τ_2 such that $0 < \tau_1 < \phi_{\min}(\Omega^{(k)}) \leq \phi_{\max}(\Omega^{(k)}) < \tau_2 < \infty$, for all $p \geq 1, k = 1, ..., K$, where $\phi_{\min}(\cdot), \phi_{\max(\cdot)}$ denote the minimal and maximal eigenvalues, respectively.

2. There exists a constant $\tau_3 > 0$ such that $\min_{k=1,...,K,(j,j')\in T_k} |\omega_{j,j'}^{(k)}| \geq \tau_3$.

The first assumption ensures the existence of the inverse matrix, and the second assumption ensures that nonzero elements are bounded away from 0.

Theorems

Theorem 2.1 (Consistency). *Suppose the above assumptions hold,* $\frac{(p+q)\log p}{n} = o(1)$, *and there exist two positive constants* Λ_1, Λ_2 *such that* $\Lambda_1 \left\{\frac{\log p}{n}\right\}^{1/2} \leq \lambda \leq \Lambda_2 \left\{\frac{(1+p/q)\log p}{n}\right\}^{1/2}$. *There exists a local minimizer of* (3) *such that*

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

Theorem 2.2 (Sparsity). *Suppose the above assumptions hold,* $\frac{(p+q)\log p}{n} = o(1)$, *and there exist two positive constants* Λ_1, Λ_2 *such that* $\Lambda_1 \left\{\frac{\log p}{n}\right\}^{1/2} \leq \lambda \leq \Lambda_2 \left\{\frac{(1+p/q)\log p}{n}\right\}^{1/2}$. *Further assume*

$$\sum_{k=1}^{K} \left\|\hat{\Omega}^{(k)} - \Omega^{(k)}\right\|^2 = O_p(\eta_n), \quad \text{where} \quad \eta_n \to 0, \quad \left\{\frac{\log p}{n}\right\}^{1/2} + \eta_n^{1/2} = O(\lambda),$$

where $\|\cdot\|$ *denotes the matrix 2-norm. Then, with probability tending to 1, the local minimizer of* (3) *satisfies that* $\hat{\omega}_{j,j'}^{(k)} = 0$ *for all* $(j, j') \in T_k^c, k = 1, ..., K$.

Note that consistency requires the upper and lower bound of λ while the sparsity requires an additional lower bound of λ. Therefore, it is harder to have sparsity than consistency. To guarantee the consistency and sparsity at the same time, we must have

$$\left\{\frac{\log p}{n}\right\}^{1/2} + \eta_n^{1/2} = O\left[\left\{\frac{(1+p/q)\log p}{n}\right\}^{1/2}\right]. \tag{5}$$

Otherwise, the lower bound required by sparsity will exceed the upper bound of λ in consistency.

By the equivalence of norm, we have $\|A\|_F^2/p \leq \|A\|^2 \leq \|A\|_F$, for a matrix $A \in \mathbb{R}^{p\times p}$. This leads to two extreme cases.

1. If $\sum_k \left\|\hat{\Omega}^{(k)} - \Omega^{(k)}\right\|^2$ has the same rate as $\sum_k \left\|\hat{\Omega}^{(k)} - \Omega^{(k)}\right\|_F$, it requires $\eta_n = O\left[\frac{(p+q)\log p}{n}\right]$. To satisfy the compatible condition (5), we need $O(p+q) = O(p/q)$ and thus $q = O(1)$.

2. If $\sum_k \left\|\hat{\Omega}^{(k)} - \Omega^{(k)}\right\|^2$ has the same rate as $\sum_k \left\|\hat{\Omega}^{(k)} - \Omega^{(k)}\right\|_F / p^{1/2}$, it requires $\eta_n = O\left[\frac{(1+q/p)\log p}{n}\right]$. To satisfy the compatible condition (5), we need $O(q/p) = O(p/q)$ and thus $q = O(p)$.

3 Our model

Let r denote the rank of decomposition. Our joint estimation problem is following.

$$\min_{\{\Theta_i\}_{i=0}^r, \{u_i\}_{i=1}^r} \sum_{k=1}^K \left[\operatorname{tr}(S^{(k)}\Omega^{(k)}) - \log\det(\Omega^{(k)})\right] + \sum_{l=1}^r \lambda_{1l} \|u_l\|_0 + \sum_{l=1}^r \lambda_{2l} \|\Theta_l\|_1,$$

where $u_i^T u_i = 1$, for all $i = 1, ..., r$ and $\Omega^{(k)} = \Theta_0 + \sum_{l=1}^r u_{lk}\Theta_l$.
Compared with Guo's method in which $\Omega^{(k)} = \Theta_0 \cdot \Gamma^{(k)}$, our method reduces the number of parameter if r is small.