Graphic Lasso: Review for
Simultaneous Clustering and Estimation of Heterogeneous Graphical Models

Jiaxin Hu

February 4, 2021

1 Model

Suppose n sample vectors $\{X_i\}_{i=1}^n \in \mathbb{R}^p$ is able to be clustered into K groups. Assume the sample vectors follow the multivariate normal distribution, i.e.,

$$X_i \sim \mathcal{N}_p(\mu_k, \Sigma_k), \quad \text{if} \quad i \in \mathcal{A}_k,$$

where $\mathcal{A}_k$ refers to the index set of the observations in the group k. Since the true cluster label may not be available, we introduce the probability $\{\pi_k\}_{k=1}^K$ for an observation belongs to the k-th group. Therefore, the density of each observations is

$$f(X_i, \Theta) = \sum_{k=1}^K \pi_k f_k(X_i, \Theta_k),$$

where

$$f_k(X_i, \Theta_k) = (2\pi)^{-p/2} \det(\Sigma_k)^{-1/2} \exp\left\{ -\frac{1}{2}(X_i - \mu_k)^T \Sigma_k^{-1} (X_i - \mu_k) \right\},$$

,Θ is the vectorized parameters $\Theta = (\Theta_1, ..., \Theta_K)^T \in \mathbb{R}^{K(p^2+p)}$ with $\Theta_k = \text{vec}(\mu_k, \Omega_k)$, and $\Omega_k = \Sigma_k^{-1}$. The ultimate goal is to estimate the parameters Θ and $\{\pi_k\}$.

2 Optimization

Note that each observation X_i only belongs to 1 group. We introduce the cluster assignment matrix $L = [\![L_{ik}]\!] \in \mathbb{R}^{n \times K}$, where $L_{ik} = \boldsymbol{I}\{X_i \in \mathcal{A}_k\}$. Therefore, the objective function of the optimization problem is

$$F(\Theta|X, L) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K L_{ik} \left[\log \pi_k + \log f_k(X_i; \Theta_k)\right] - \mathcal{P}(\Theta),$$

where

$$\mathcal{P}(\Theta) = \lambda_1 \sum_{k=1}^K \sum_{j=1}^p |\mu_{kj}| + \lambda_2 \sum_{k=1}^K \sum_{i \neq j} |\omega_{kij}| + \lambda_3 \sum_{i \neq j} \left(\sum_{k=1}^K \omega_{kij}^2 \right)^{1/2}.$$

In each iteration, the algorithm maximize the expectation conditional on the parameters from the last step. That is

$$(\pi_k^{(t)}, \Theta^{(t)}) = \arg\max_{\pi_k, \Theta} \mathbb{E}_{L|X,\Theta^{(t-1)}}\left[F(\Theta|X, L)\right] = \arg\max_{\pi_k, \Theta} Q_n(\Theta|\Theta^{(t-1)}) - P(\Theta),$$

where

$$Q_n(\Theta|\Theta^{(t-1)}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} L_{\Theta^{(t-1)},k}(X_i)\left[\log \pi_k + \log f_k(X_i; \Theta_k)\right],$$

and

$$L_{\Theta^{(t-1)},k}(X_i) = \frac{\pi_k^{(t-1)} f_k(X_i, \Theta_k^{(t-1)})}{\sum_{k=1}^{K} \pi_k^{(t-1)} f_k(X_i, \Theta_k^{(t-1)})}.$$

3 Statistical Guarantees

The statistical guarantees are established for the estimate given by the optimization algorithm in last section, which is not necessarily a MLE or a local maximizer. Let Θ^* denote hte true parameters and $\mathcal{B}(\Theta^*) = \{\Theta : \|\Theta - \Theta^*\|_2 \le \alpha\}$. The error upper bound requires three conditions:

1. (Sufficiently Separable Condition) Suppose $L_{\Theta,k}(X)L_{\Theta,j}(X)$ is close to 0, for $k \neq j$ and $\Theta \in \mathcal{B}(\Theta^*)$. See Condition 6 in the paper for the detailed condition.

2. (Bounded singular values) Suppose there exist positive constants β_1, β_2, such that $0 < \beta_1 < \min_k \sigma_{\min}(\Omega_k^*) \le \max_k \sigma_{\max}(\Omega_k^*) < \beta_2$.

3. (Bounded difference between population-based and sample-based conditional maximization) Let

   $$Q(\Theta|\Theta') = \mathbb{E}\left[\sum_{k=1}^{K} L_{\Theta^{(t-1)},k}(X)\left[\log \pi_k + \log f_k(X; \Theta_k)\right]\right],$$

   with respect to X. Then, for all $\Theta \in \mathcal{B}(\Theta^*)$, with high probability, we have

   $$\|\nabla Q_n(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta)\|_{\mathcal{P}^*} \le \epsilon_1,$$

   and

   $$\|[\nabla Q_n(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta)]_G\|_2 \le \epsilon_2,$$

   where $\|\cdot\|_{\mathcal{P}^*}$ is the dual norm of $\mathcal{P}$, and G is the index set corresponding to the diagonal elements in Ω_k.

Define the following parameters:

1. τ: Gradient Stability parameter, which satisfies

   $$\|\nabla Q(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta^*)\|_2 \le \tau \|\Theta - \Theta^*\|_2,$$

   for $\Theta \in \mathcal{B}(\Theta^*)$, under the first condition.

2. γ: Restricted strong concavity parameter, which satisfies

   $$Q_n(\Theta'|\Theta) - Q_n(\Theta^*|\Theta) - \langle \nabla Q_n(\Theta^*|\Theta), \Theta' - \Theta^* \rangle \le -\frac{\gamma}{2}\|\Theta' - \Theta^*\|_2^2,$$

   where $\gamma = c \min \beta_1, 0.5(\beta_2 + 2\alpha)^{-2}$ for some c, under the second condition.

3. $\nu(\mathcal{M}) = \sup_{\Theta \in \mathcal{M}} \frac{\mathcal{P}(\Theta)}{\|\Theta\|_2}$, where $\mathcal{M}$ is the support space for Θ.

Then, we have the theorem.

Theorem 3.1. *Suppose the three conditions are hold. Let* $\kappa = \frac{6\tau}{\gamma}$ *and the initialization* $\Theta^{(0)} \in \mathcal{B}(\Theta^*)$*. Assume the tuning parameter*

$$\lambda_n^{(t)} = \epsilon + \kappa \frac{\gamma}{\nu(\mathcal{M})} \left\| \Theta^{(t-1)} - \Theta^* \right\|_2 .$$

If the sample size is large enough such that $\epsilon \leq (1-\kappa)\frac{\gamma\alpha}{6\nu(\mathcal{M})}$*, then the estimate* $\Theta^{(t)}$ *satisfies with probability* $1 - t\delta'$*,*

$$\left\| \Theta^{(t)} - \Theta^* \right\|_2 \leq \frac{6\nu(\mathcal{M})}{(1-\kappa)\gamma}\epsilon + \kappa^t \left\| \Theta^{(0)} - \Theta^* \right\|_2 ,$$

where δ' *is small positive constant, and* $\epsilon = \epsilon_1 + \epsilon_2/\nu(\mathcal{M})$*.*

4 Comparison

Here are few points need to be noticed:

1. Though the algorithm gives an output of matrix L, the estimate of L is not a membership matrix. This implies the model does not address a "hard" clustering, but a "soft" clustering problem by estimate
$$\pi_k = \mathbb{P}(X \in \mathcal{A}_k).$$
Also, by Lemma 2 and 3 in the paper, the update of μ_k and Ω_k does not depend on π_k with given the matrix L, and the update of π_k is a function of μ_k and Ω_k. Therefore, the paper only discuss the accuracy of Θ_k.

2. The accuracy theorem is an error upper bound for the algorithm outputs, rather than the maximizer of the objective function. If we assume the model is a hard clustering model with $f(X_i, \Theta) = \sum_{k=1}^{K} \boldsymbol{I}\{X_i \in \mathcal{A}_k\} f_k$ and $\mu_k = 0$, we may get some inspirations about the penalized likelihood estimation of Ω_k when letting $t \to \infty$.