Graphic Lasso: What we learn from Cheng's paper

Jiaxin Hu

February 7, 2021

1 Proof Sketch of the Key Theorem

1.1 Model

Consider the model

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

where f_k is the density of multivariate normal distribution with parameters Θ_k. The estimate we consider in the paper satisfies the following update criterion

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

where

$$Q_n(\Theta|\Theta^{(t-1)}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} L_{\Theta^{(t-1)},k}(X_i) \left[\log \pi_k + \log f_k(X_i; \Theta_k) \right], \tag{2}$$

and

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

Define the population version of Q_n as following

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

where the expectation takes with respect to X.

1.2 Assumptions

Consider the following assumptions:

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 $0 < \beta_1 < \min_k \sigma_{\min}(\Omega_k^*) \leq \max_k \sigma_{\max}(\Omega_k^*) < \beta_2$.

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
$$\|\nabla Q_n(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta)\|_{\mathcal{P}^*} \leq \epsilon_1,$$
and
$$\|[\nabla Q_n(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta)]_G\|_2 \leq \epsilon_2,$$
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 coefficients:

1. τ: Gradient Stability parameter, which satisfies
$$\|\nabla Q(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta^*)\|_2 \leq \tau \|\Theta - \Theta^*\|_2,$$
for $\Theta \in \mathcal{B}(\Theta^*)$, under the first condition.

2. γ: Restricted strong concavity parameter, which satisfies
$$Q_n(\Theta'|\Theta) - Q_n(\Theta^*|\Theta) - \langle \nabla Q_n(\Theta^*|\Theta), \Theta' - \Theta^* \rangle \leq -\frac{\gamma}{2}\left\|\Theta' - \Theta^*\right\|_2^2,$$
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 (with the same nonzero index set as true parameters) for Θ.

1.3 Theorem

Theorem 1.1. *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 $\epsilon = \epsilon_1 + \epsilon_2/\nu(\mathcal{M})$.

1.4 Proof

Lemma 1 (Key Lemma)**.** *Suppose* $\Theta^{(t-1)} \in \mathcal{B}_\alpha(\Theta^*)$ *with choice* $\lambda_n^{(t)} = \epsilon + \tau \left\|\Theta^{(t-1)} - \Theta^*\right\|_2 / \nu(\mathcal{M})$. *The estimate from* (1) *satisfies*

$$\left\|\Theta^{(t)} - \Theta^*\right\|_2 \le \frac{6\nu(\mathcal{M})\lambda_n^{(t)}}{\gamma},$$

with high probability.

Note that $\lambda_n^{(t)}$ includes a term of $\left\|\Theta^{(t-1)} - \Theta^*\right\|_2$. We can use math induction to obtain the error with term $\left\|\Theta^{(0)} - \Theta^*\right\|_2$. Therefore, Lemma 1 is the theorem of our main interest if we would like to modify the techniques for precision matrix model.

Proof for Lemma 1. Consider the function

$$f(\Delta) = Q_n(\Theta^* + \Delta|\Theta^{(t-1)}) - Q_n(\Theta^*|\Theta^{(t-1)}) - \lambda_n^{(t)}(\mathcal{P}(\Theta^* + \Delta) - P(\Theta^*)).$$

Note that $f(0) = 0$ and $f(\hat{\Delta}) \ge 0$, where $\hat{\Delta} = \Theta^{(t)} - \Theta^*$. The original proof follows idea below:

1. Show that $f(\Delta) < 0$ if $\|\Delta\|_2 = \xi$, where $\xi = \frac{6\nu(\mathcal{M})\lambda_n^{(t)}}{\gamma}$.

2. Show that $\hat{\Delta}$ is inside of the set $C(\xi) = \{\Delta : \|\Delta\|_2 \le \xi\}$.

Step 2 is proved by contradiction using the convexity of $\mathcal{P}(\Theta)$, while the proof for step 1 is more interesting. Therefore, we only summarize the proof for step 1 here.

1. By the **restricted strong concavity property (whereγ comes from)**, we have

   $$Q_n(\Theta^* + \Delta|\Theta^{(t-1)}) - Q_n(\Theta^*|\Theta^{(t-1)}) \le \langle \nabla Q_n(\Theta^*|\Theta^{(t-1)}), \Delta\rangle - \frac{\gamma}{2}\|\Delta\|_2^2.$$

   Adding the term $\lambda_n^{(t)}(\mathcal{P}(\Theta^* + \Delta) - \mathcal{P}(\Theta))$ on the both sides, we have an upper bound for $f(\Delta)$. That is

   $$f(\Delta) \le I_1 - \lambda_n^{(t)} I_2 - \frac{\gamma}{2}\|\Delta\|_2^2,$$

   where

   $$I_1 = \langle \nabla Q_n(\Theta^*|\Theta^{(t-1)}), \Delta\rangle, \quad I_2 = \mathcal{P}(\Theta^* + \Delta) - \mathcal{P}(\Theta).$$

2. For part I_1, we the upper bound

   $$I_1 \le |I_1| \le SE + OE,$$

   where

   $$SE = |\langle \nabla Q_n(\Theta^*|\Theta^{(t-1)}) - \nabla Q(\Theta^*|\Theta^{(t-1)}), \Delta\rangle|,$$

   and

   $$OE = |\langle \nabla Q(\Theta^*|\Theta^{(t-1)}) - \nabla Q(\Theta^*|\Theta^*), \Delta\rangle|.$$

   The upper bound follows by the self-consistency property of Θ^*, i.e.,

   $$\Theta^* = \arg\max_{\Theta} Q(\Theta|\Theta^*), \quad \text{and thus} \quad \nabla Q(\Theta^*|\Theta^*) = 0.$$

- For SE, by **generalized Cauchy-Schwartz inequality**, we have

$$SE \le \left\| h(\Theta^*|\Theta^{(t-1)})_{G^c} \right\|_{\mathcal{P}^*} \mathcal{P}(\Delta) + \left\| h(\Theta^*|\Theta^{(t-1)})_G \right\|_2 \|\Delta\|_2 ,$$

where $\|\cdot\|_{\mathcal{P}^*}$ is the dual norm of $\mathcal{P}$, and G is the set for the diagonal elements, and $h(\Theta^*|\Theta^{(t-1)}) = \nabla Q_n(\Theta^*|\Theta^{(t-1)}) - \nabla Q(\Theta^*|\Theta^{(t-1)})$.

By **Condition 3**, with high probability, we have

$$SE \le \epsilon_1 \mathcal{P}(\Delta) + \epsilon_2 \|\Delta\|_2 . \tag{3}$$

- For OE, by **Gradient stability (where τ comes from)**, we have

$$OE \le \left\| \nabla Q(\Theta^*|\Theta^{(t-1)}) - \nabla Q(\Theta^*|\Theta^*) \right\|_2 \|\Delta\|_2 \le \tau \left\| \Theta^{(t-1)} - \Theta^* \right\|_2 \|\Delta\|_2 . \tag{4}$$

3. For part I_2, by **triangle inequality** and the decomposition of penalty, we have

$$I_2 = \mathcal{P}(\Theta^* + \Delta_{\mathcal{M}}) + \mathcal{P}(\Delta_{\mathcal{M}^\perp}) - \mathcal{P}(\Theta^*) \ge \mathcal{P}(\Delta_{\mathcal{M}^\perp}) - \mathcal{P}(\Delta_{\mathcal{M}}) \tag{5}$$

4. Combining the inequalities (3), (4), (5), and following basic inequalities

$$\mathcal{P}(\Delta) \le \mathcal{P}(\Delta_{\mathcal{M}}) + \mathcal{P}(\Delta_{\mathcal{M}^\perp}), \quad \mathcal{P}(\Delta_{\mathcal{M}}) \le \nu(\mathcal{M}) \|\Delta\|_2 ,$$

we finally have

$$f(\Delta) \le -\frac{\gamma}{2} \|\Delta\|_2^2 + 2\lambda_n^{(t)} \nu(\mathcal{M}) \|\Delta\|_2 , \tag{6}$$

which is negative after plugging $\|\Delta\|_2 = \xi$.

Remark 1 (Proof Idea)**.** Note that we may simply use the fact $f(\hat{\Delta}) \ge 0$ to get the result. To obtain the inequality (6), we only requires $\Theta^{(t-1)} \in \mathcal{B}(\Theta^*)$. Then, the inequality (6) will directly leads to

$$\|\Delta\|_2 \le \frac{4\lambda_n^{(t)} \nu(\mathcal{M})}{\gamma}.$$

□

2 Possible extension to precision matrix model

Remark 2. (The error bound for Θ includes the clustering accuracy) Let

$$\mathcal{L}(\Theta|\Theta^{(t-1)}) = Q_n(\Theta|\Theta^{(t-1)}) - P(\Theta)$$

denote the objective function for each iteration step. Check the definition of (2). The cluster assignment parameter L is the only term depend on $\Theta^{(t-1)}$. This implies the membership estimate in this model is a function of $\Theta^{(t-1)}$. Recall the Key Lemma 1. With a more explicit form, we have

$$\left\| \Theta^{(t)} - \Theta^* \right\|_2 \le \frac{6\nu(\mathcal{M})\epsilon}{\gamma} + \frac{6\tau}{\gamma} \left\| \Theta^{(t-1)} - \Theta^* \right\| .$$

The second term in some sense includes the error from misclassification. What's more, in the proof for I_1, the SE can be considered as deviation for the estimation of $\hat{\Theta}$ with given membership, and OE can be considered as the error for the mismatching of the clusters. Therefore, there is no wonder why the clustering accuracy is contained in the error bound for $\Theta^{(t)}$.

Notice that "condition on $\Theta^{(t-1)}$" is equal to "condition on the given membership from last step". In the context of other models whose membership estimation is not expressed as a function of $\hat{\Theta}$, let U denote the membership matrix (maybe mixed membership). The above proof utilize the fact that

$$\mathcal{L}(\hat{\Theta}|\hat{U}) - \mathcal{L}(\Theta^*|\hat{U}) \geq 0,$$

which follows by the fact that $\hat{\Theta}$ is the maximizer of the objective function. Correspondingly, the terms OE which represents the misclassification may be upper bounded by a function of Misclassification Rate (MCR), or other terms imply the misclassification like $\left\|\hat{U} - U\right\|_2$.

Therefore, the above proof may be helpful after we have some conclusions about the cluster accuracy. We may try the above techniques for precision matrix through replacing $\Theta^{(t-1)}$ by true membership U^* firstly.

Remark 3 (Penalized optimization)**.**