Simultaneous Clustering and Estimation of Multiple Sparse Networks

Botao Hao

November 16, 2015

Joint work with Wei Sun, Guang Cheng, and Yufeng Liu

Outline

Outline

Background for clustering

- Consider each p-dimensional observation $x_i, i = 1, \cdots, n$ is drawn from a Gaussian mixture model with

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \mu_k, \Sigma_k) \tag{1.1}$$

  where π_k is the mixture weight and $f_k(x; \mu_k, \Sigma_k)$ is a multivariate normal distribution.
- For classical clustering problem, we assume $\Sigma_k = \Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_p^2)$
- Output: cluster assignments and cluster means

Outline

Regularized Model-based Clustering

- Sun et al.(2012) proposed a high-dimensional cluster analysis via regularized model-based clustering.
- Regularized log-likelihood function for the observed data:

$$\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \sigma^2 I_p)\right) - \lambda \sum_{j=1}^{p} \|\boldsymbol{\mu}_{(j)}\|_2 \qquad (1.2)$$

  with $\boldsymbol{\mu}_{(j)}$ the j-th column of the center matrix $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)^T$.
- An EM algorithm can be employed to maximize (1.2), where the cluster assignment L_{ik} is treated as missing data.

Outline

Single Gaussian Graphical model

- Suppose that the observation $x_1, \cdots, x_n \in \mathbb{R}^p$ are independent and identically distributed $N(\mu, \Sigma)$, where $\mu \in \mathbb{R}^p$ and Σ is a positive definite $p \times p$ matrix.
- Goal: Estimate the *precision* matrix $\Omega = \Sigma^{-1}$
- Gaussian log-likelihood takes the form(up to a constant)

$$l(\Omega) = \frac{n}{2}[\log\{\det(\Sigma^{-1})\} - \mathrm{tr}(S\Sigma^{-1})] \tag{1.3}$$

where S denotes the empirical covariance matrix.

- Method for estimating Σ^{-1} in high-dimensional setting

$$\max_{\Omega}[\log\{\det(\Omega)\} - \mathrm{tr}(S\Omega) - \lambda\|\Omega\|_1] \tag{1.4}$$

Outline

Multiple Gaussian Graphical Model without Similarity

- Key assumption: The cluster structure is known in advance.
- Suppose that we are given K data set $X^{(1)}, \cdots, X^{(K)}$, with K fixed. $X^{(k)}$ is an $n_k \times p$ matrix, $x_1^{(k)}, \cdots, x_{n_k}^k \sim N(\mu_k, \Sigma_k)$.
- $\Omega_k = \Sigma_k^{-1}$, $\Theta = (\Omega_1, \cdots, \Omega_K)$. We let $S_k = (1/n_k)(X^{(k)})^T X^{(k)}$
- Similarly, the log-likelihood for the data takes the form(up to a constant)

$$l(\Theta) = \frac{1}{2} \sum_{k=1}^{K} n_k [\log\{\det(\Omega_k)\} - \mathrm{tr}(S_k \Omega_k)] \qquad (1.5)$$

- It is the same as to estimate each precision matrix separately.

Multiple Gaussian Graphical Model with Similarity

- Multiple graphical models share certain characteristics, such as the locations or weights of non-zero edges.
- In high-dimensional setting, joint estimation of graphical models solves

$$\max_{\Omega} \sum_{k=1}^{K} n_k [\log\{\det(\Omega_k)\} - \mathrm{tr}(S_k \Omega_k)] - \mathcal{P}(\Theta) \qquad (1.6)$$

- The choice of $\mathcal{P}(\Theta)$ could encourage estimators to share similar characteristics across classes and simultaneously encourage sparsity.

Multiple Gaussian Graphical Model with Similarity

- Guo et al. (2011) employed a non-convex penalty

$$P(\Theta) = \lambda \sum_{i \neq j} \Big(\sum_{k=1}^{K} |\omega_{ij}^{(k)}| \Big)^2, \tag{1.7}$$

where ω_{ij} is the entry of precision matrix Ω.

- Danaher et al. (2014) applied two convex penalties: fused graphical lasso

$$P(\Theta) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{ij}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i,j} |\omega_{ij}^{(k)} - \omega_{ij}^{(k')}|. \tag{1.8}$$

group graphical lasso

$$P(\Theta) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{ij}^{(k)}| + \lambda_2 \sum_{i \neq j} \Big(\sum_{k=1}^{K} {\omega_{ij}^{(k)}}^2 \Big)^{1/2}. \tag{1.9}$$

Multiple Gaussian Graphical Model with Similarity

- Lee and Liu(2015) extend the joint estimation method to non-Gaussian cases.

All the literature above treat the cluster structure as known!

Outline

Motivation

- From a clustering point of view, we consider the dependencies among variables within clusters.
- From a perspective of joint estimation for precision matrix, we do not assume that the clustering structure is given in advance.

Outline

Some notations

- Let $x_1, \cdots, x_n \sim f(x) = \sum_{k=1}^{K} \pi_k f_k(x, \mu_k, \Sigma_k)$
- The set of parameters $\Theta := \{(\mu_k, \Omega_k), k = 1, \cdots, K\}$, where $\Omega_k = \Sigma_k^{-1}$
- Denote the K clusters as $\mathcal{A}_1, \cdots, \mathcal{A}_K$ and L as the cluster assignment matrix with $L_{ik} = \mathcal{I}(x_i \in \mathcal{A}_k)$. So $\sum_{k=1}^{K} L_{ik} = 1$.
- In EM algorithm, L can be treated as a latent variable.

Joint Clustering and Graphical Lasso

- Our optimization problem is formulated as

$$\max_{\pi_k,\mu_k,\Omega_k} \sum_{i=1}^{n} \log(\sum_{k=1}^{K} \pi_k f_k(x_i;\mu_k,(\Omega_k)^{-1})) - \mathcal{P}(\Theta) \tag{2.1}$$

where

$$\mathcal{P}(\Theta) = \lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p} |\mu_{kj}| + \lambda_2 \sum_{k=1}^{K}\sum_{i\neq j} |\omega_{kij}| + \lambda_3 \sum_{i\neq j}(\sum_{k=1}^{K} \omega_{kij}^2)^{1/2} \tag{2.2}$$

- If L_{ik} is available, the regularized log-likelihood function for the compete data can be calculated as

$$\log L_c(\Theta) := \sum_{i=1}^{n}\sum_{k=1}^{K} L_{ik}[\log \pi_k + \log f_k(x_i;\Theta_k)] - P(\Theta)$$

Outline

Expectation Step

- Given the current estimator $\widehat{\Theta}^{(t)}$, the conditional expectation of (18) is computed as

$$Q_m(\Theta|\widehat{\Theta}^{(t)}) - P(\Theta) \tag{2.3}$$

where

$$Q_m(\Theta|\widehat{\Theta}^{(t)}) := \sum_{i=1}^{n}\sum_{k=1}^{K}\widehat{L}_{ik}^{(t)}[\log \pi_k + \log f_k(x_i;\Theta_k)] \tag{2.4}$$

$$\widehat{L}_{ik}^{(t)} = \frac{\widehat{\pi}_k^{(t)} f_k(x_i;\widehat{\Theta}_k^{(t)})}{\sum_{k=1}^{K}\widehat{\pi}_k^{(t)} f_k(x_i;\widehat{\Theta}_k^{(t)})} \tag{2.5}$$

Maximization Step

- Update of π_k:

$$\widehat{\pi}_k^{(t+1)} = \sum_{i=1}^{n} \frac{\widehat{L}_{ik}^{(t)}}{n} \tag{2.6}$$

- Update of μ_{kj} (*follows KKT condition*):

$$\text{If } \left| \sum_{i=1}^{n} \widehat{L}_{ik}^{(t)} \Big(\sum_{l=1, l\neq j}^{p} (x_{il} - \widehat{\mu}_{kl}^{(t)}) \widehat{\omega}_{klj}^{(t)} + x_{ij} \widehat{\omega}_{kjj}^{(t)} \Big) \right| \leq \lambda_1, \text{ then } \widehat{\mu}_{kj}^{(t+1)} = 0; \tag{2.7}$$

$$\text{Else } \widehat{\mu}_{kj}^{(t+1)} = \Big(\widehat{\omega}_{kjj}^{(t)} \sum_{i=1}^{n} \widehat{L}_{ik}^{(t)} \Big)^{-1} \Big\{ \sum_{i=1}^{n} \widehat{L}_{ik}^{(t)} \big(\sum_{l=1}^{p} x_{il} \widehat{\omega}_{klj}^{(t)} \big)$$

$$- \big(\sum_{i=1}^{n} \widehat{L}_{ik}^{(t)} \big) \Big(\sum_{l=1}^{p} \widehat{\mu}_{kl}^{(t)} \widehat{\omega}_{klj}^{(t)} - \widehat{\mu}_{kj}^{(t)} \widehat{\omega}_{kjj}^{(t)} \Big) - \lambda_1 \text{sign}(\widehat{\mu}_{kj}^{(t)}) \Big\} \tag{2.8}$$

Maximization Step

- Update of Ω_k: Note that maximize (2.3) with respect to Ω_k is equivalent to solve the following maximization problem.

$$\max_{\Omega_k} \sum_{k=1}^{K} n_k [\log\det(\Omega_k) - \mathrm{trace}(\widetilde{S}_k \Omega_k)] - P(\Omega), \tag{2.9}$$

where

$$\widetilde{S}_k := \frac{\sum_{i=1}^{n} \widehat{L}_{ik}^{(t)} (x_i - \boldsymbol{\mu}_k)^T (x_i - \boldsymbol{\mu}_k)}{\sum_{i=1}^{n} \widehat{L}_{ik}^{(t)}} \tag{2.10}$$

- This optimization problem can be solved efficiently via the ADMM algorithm in Danaher et al. (2014).

EM algorithm

Table: Outline of Our Algorithm

Input: Training data $x_1, \ldots, x_n$ and Number of clusters K.
Output: Cluster assignment L_{ik} and graph Ω_k.

Step 1: Initialize cluster centers by the k-means++ clustering and set $\pi_k^{(0)} = \frac{1}{K}$.
Step 2: Until the termination condition is met, for $t = 1, 2, \ldots$
(a) E-step. Find the cluster assignment $L_{ik}^{(t)}$ as in (2.5).
(b) M-step. Given $L_{ik}^{(t)}$, update $\pi_k^{(t+1)}$ as in (2.6), $\boldsymbol{\mu}_k^{(t+1)}$ as in (2.7), and the precision matrix $\Omega_k^{(t+1)}$ in (2.9) via JGL algorithm.

Outline

Some previous work

- Balakrishnan et al.(2014) first considered both the statistical error and optimization error for the guarantee of EM algorithm.
- Wang et al.(2014) proposed high-dimensional EM algorithm based on truncation step.
- Yi et al.(2015) used regularization method for high-dimensional EM.

Decomposable Regularizer

Definition (Decomposability)

Regularizer $\mathcal{P} : \mathbb{R}^p \to \mathbb{R}^+$ is decomposable with respect to $(\mathcal{S}, \bar{\mathcal{S}})$ if

$$\mathcal{P}(u+v) = \mathcal{P}(u) + \mathcal{P}(v), for\ any\ u \in \mathcal{S}, v \in \bar{\mathcal{S}}^{\perp} \tag{3.1}$$

where $\mathcal{S} \subseteq \bar{\mathcal{S}}$.

- Example

$$\|\theta + \gamma\|_1 = \|(\theta_{\mathcal{S}}, 0) + (0, \gamma_{\mathcal{S}^c})\|_1 = \|\theta\|_1 + \|\gamma\|_1$$

where $\mathcal{S}$ is the support of θ.

Subspace compatibility constant

Definition (Subspace compatibility constant)

For any subspace $\mathcal{S}$ of $\mathbb{R}^p$, the subspace compatibility constant with respect to the pair $(\mathcal{P}, \|.\|)$ is given by

$$\psi(\mathcal{S}) := \sup_{u \in \mathcal{S} \setminus \{0\}} \frac{\mathcal{P}(u)}{\|u\|} \tag{3.2}$$

- Example: If $\mathcal{S}$ is a s-dimensional coordinate subspace, with regularizer $\mathcal{P}(u) = \|u\|_1$ and error norm $\|u\| = \|u\|_2$, then we have $\Psi(\mathcal{S}) = \sqrt{s}$

Outline

Population version vs. Sample version

- Sample version Q function

$$Q_n(\Theta'|\Theta) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} L_{ik}(x_i)[\log \pi'_k + \log f_k(x_i;\Theta'_k)] \tag{3.3}$$

- Population version of Q function

$$Q(\Theta'|\Theta) := E\Big[\sum_{k=1}^{K} L_{ik}(X)[\log \pi'_k + \log f_k(X;\Theta'_k)]\Big] \tag{3.4}$$

Condition for Population version

Condition (Strong Concavity and Smoothness)

$$Q(\Theta_2|\Theta^*) - Q(\Theta_1|\Theta^*) - \langle \nabla Q(\Theta_1|\Theta^*), \Theta_2 - \Theta_1 \rangle \leq -\frac{\gamma}{2}\|\Theta_2 - \Theta_1\|_2^2,$$

$$Q(\Theta_2|\Theta) - Q(\Theta_1|\Theta) - \langle \nabla Q(\Theta_1|\Theta), \Theta_2 - \Theta_1 \rangle \geq -\frac{\mu}{2}\|\Theta_2 - \Theta_1\|_2^2, \quad (3.5)$$

for any $\Theta_1, \Theta_2 \in \mathcal{B}_\alpha(\Theta^*)$.

- We will focus on the Euclidean ball of radius $\alpha > 0$ for the basin of attraction space. That is,

$$\mathcal{B}_\alpha(\Theta^*) := \{\Theta \in \mathbb{R}^{K(p^2+p)} : \|\Theta - \Theta^*\|_2 \leq \alpha\}. \quad (3.6)$$

Condition for Population version

Condition (τ-Lipschitz-Gradient)

The function $\nabla_{\Theta'} Q(\bar{\boldsymbol{\mu}}, \bar{\Omega}|\cdot)$ satisfies,

$$\left\|\nabla_{\Theta'} Q(\bar{\boldsymbol{\mu}}, \bar{\Omega}|\Theta) - \nabla_{\Theta'} Q(\bar{\boldsymbol{\mu}}, \bar{\Omega}|\Theta^*)\right\|_2 \leq \tau \cdot \|\Theta - \Theta^*\|_2, \quad (3.7)$$

for any $\Theta \in \mathcal{B}_\alpha(\Theta^*)$.

- Recall that $(\bar{\boldsymbol{\mu}}_k, \bar{\Omega}_k), k = 1, \ldots, K$ is the true maximizer of the population objective function $Q(\Theta^{'}|\Theta)$.

Condition for Sample version

Condition (Restricted Strong Concavity)

For any fixed $\Theta \in \mathcal{B}(\alpha; \Theta^*)$, with probability at least 1-δ, we have that for all $\Theta' - \Theta^* \in \Omega \bigcap \mathcal{C}(\mathcal{S}, \bar{\mathcal{S}}; \mathcal{P})$,

$$Q_n(\Theta'|\Theta) - Q_n(\Theta^*|\Theta) - \langle \nabla Q_n(\Theta^*|\Theta), \Theta' - \Theta^* \rangle \leq -\frac{\gamma_n}{2} ||\Theta' - \Theta^*||^2 \tag{3.8}$$

where $\mathcal{C}(\mathcal{S}, \bar{\mathcal{S}}; \mathcal{P})$ is a particular set.

Condition for Sample version

Condition (Statistical Error($\epsilon_n, \alpha, \delta$))

For any fixed $\Theta \in \mathcal{B}_\alpha(\Theta^*)$, with probability at least $1-\delta$, we have

$$\left\|\nabla Q_n(\Theta^*|\Theta) - \nabla Q(\Theta^*|\Theta)\right\|_{\mathcal{P}^*} \leq \epsilon_n \tag{3.9}$$

- $\|.\|_{\mathcal{P}^*}$ is the dual norm of $\mathcal{P}$
- There is *no sparsity* here for statistical error.

Outline

1. Introduction
   - Review of Clustering
   - Clustering for high-dimensional data
   - Review of Estimation for Graphical Model
   - Multiple Gaussian Graphical Model
2. Simultaneous Clustering and Estimation
   - Motivation
   - Joint Clustering and Graphical Lasso
   - EM algorithm
3. Theoretical Guarantee
   - Preliminary
   - Several Conditions
   - Main Result
4. Conclusion

Main Result

Theorem

Assume the model parameter $\Theta^* \in \mathcal{S}$ *and regularizer* $\mathcal{P}$ *is decomposable with respect to* $(\mathcal{S}, \bar{\mathcal{S}})$. *Given n samples and T iterations, we let* $m := n/T$. *Further, assume* $Q(\cdot|\cdot)$ *and* $Q_m(\cdot|\cdot)$ *satisfy the conditions above. If* $\Theta^{(t-1)} \in \mathcal{B}_\alpha(\Theta^*)$ *and*

$$\lambda_m^{(t)} \geq 3\epsilon_m + \frac{\alpha\mu\tau}{\gamma}\left\|\Theta^{(t-1)} - \Theta^*\right\|\frac{1}{\Psi(S)} \tag{3.10}$$

then we have

$$\left\|\Theta^{(t)} - \Theta^*\right\| \leq 5\Psi(S)\frac{\lambda_m^{(t)}}{\gamma_m} \tag{3.11}$$

Main Result

Remark

Here we focus on the analysis of solution $\Theta^{(t)}$ directly from the algorithm instead of the optimal solution

$$\widehat{\Theta} \in \arg\max_{\Theta} \sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} \pi_k f_k(x_i; \mu_k, (\Omega_k)^{-1})\Big) - \mathcal{P}(\Theta)$$

Main Result

Corollary

Under certain condition and careful choice of λ_m, we have

$$\left\|\Theta^{(t+1)} - \Theta^*\right\|_2 \lesssim \underbrace{\frac{1-\kappa^t}{\gamma_m(1-\kappa)}(\sqrt{d+Ks}+\sqrt{Kp})\sqrt{\frac{\log(Kp^2-Kp)}{n}}}_{statistical\ error} + \underbrace{\kappa^t\left\|\Theta^{(0)} - \Theta^*\right\|_2}_{optimization\ error}$$

with high probability.

- Denote the number of non-zero entries in $\boldsymbol{\mu}_k^*$ as $d_k := \|\boldsymbol{\mu}_k^*\|_0$, and let $d = \sum_k d_k$, s is the sparsity of precision matrix Ω_k
- K is the number of cluster and $1/2 \leq \kappa \leq 3/4$

Comparison with Another Result

- Similar result is derived by Martin J. Wainwright(2014). If we assume Θ^* is exactly group-sparse, say, supported on a group subset $S_{\mathcal{G}} \subseteq \{1, 2, \cdots, N_{\mathcal{G}}\}$ of cardinality $s_{\mathcal{G}}$, then

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim \sqrt{s_{\mathcal{G}}(\frac{m}{n} + \frac{\log |\mathcal{G}|}{n})} \tag{3.12}$$

where $m = \max_{t=1,\cdots,N_{\mathcal{G}}} |G_t|$.

- In this paper, we propose a regularized Gaussian mixture model with a joint graphical lasso penalty to borrow strength across various clusters in estimating multiple graphical models which share some common structures.
- For theory, we consider the potential gap between statistical and computational guarantees in application of our algorithm.
- The statistical rate is nearly mini-max rate.

Reference

- Balakrishnan, S., Wainwright, M., and Yu, B.(2014), Statistical guarantees for the EM algorithm: From population to sample-based analysis, *arXiv*.
- Danaher, P., Wang, P., and Witten, D.(2014), The joint graphical lasso for inverse covariance estimation across multiple classes, *Journal of Royal Statistical Society, Series B* 76, 373-397.
- Guo, J., Levina, E., Michailidis, G., and Zhu, J.(2011), Joint estimation of multiple graphical models, *Biometrika*, 98, 1-15.
- Lee, W. and Liu, Y.(2015), Joint estimation of multiple precision matrices with common structures, *Manuscript*.

Reference

- Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B.(2012), A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, *Statistical Science*, 27, 538-557.
- Sun, W., Wang, J., and Fang, Y.(2012), Regularized k-means clustering of high-dimensional data and its asymptotic consistency, *Electronic Journal of Statistics*, 6, 148-167.
- Yi, X., Caramanis, C.(2015), Regularized EM Algorithms: A Unified Framework and Statistical Guarantees. *NIPS*