Simultaneous Clustering And EstimatioN of Multiple Graphical Models

Botao Hao
Department of Statistics
Purdue University

Statistical Learning and Data Science Conference
July 25, 2017

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

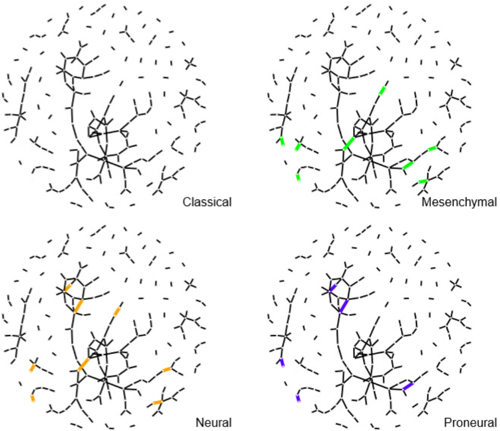
Outline

1. Motivation
   - Real Data Motivation
   - Literature Review
2. Simultaneous Clustering and Estimation
   - Methodology
   - Theoretical Result
   - Simulation Results

Outline

1. Motivation
   - Real Data Motivation
   - Literature Review
2. Simultaneous Clustering and Estimation
   - Methodology
   - Theoretical Result
   - Simulation Results

Joint Estimation Motivation: Glioblastoma Cancer Data



- **Homogeneity**: All Glioblastoma cancer data.
- **Heterogeneity**: Four different subtypes.

Big Data Motivation: Online Advertising

- Goal: both **user clustering** and the knowledge of **conditional dependence** among user attributes could benefit online advertising.

Outline

1. Motivation
   - Real Data Motivation
   - Literature Review
2. Simultaneous Clustering and Estimation
   - Methodology
   - Theoretical Result
   - Simulation Results

Literature Review & Open Problems

- Existing literature focuses on joint estimation from **labeled** data set. (Guo et al., 2011; Danaher et al., 2014; Cai et al., 2016)
- Big Data $\Longrightarrow$ huge **unknown** clusters. Joint estimation approach will dramatically fail when is given incorrect cluster labels.
- Known clustering label $\Longrightarrow$ **convex** optimization. Unknown clustering label $\Longrightarrow$ **non-convex** optimization.

Literature Review & Open Problems

- Existing literature focuses on joint estimation from **labeled** data set. (Guo et al., 2011; Danaher et al., 2014; Cai et al., 2016)
- Big Data $\Longrightarrow$ huge **unknown** clusters. Joint estimation approach will dramatically fail when is given incorrect cluster labels.
- Known clustering label $\Longrightarrow$ **convex** optimization.
  Unknown clustering label $\Longrightarrow$ **non-convex** optimization.

Literature Review & Open Problems

- Existing literature focuses on joint estimation from **labeled** data set. (Guo et al., 2011; Danaher et al., 2014; Cai et al., 2016)
- Big Data $\Longrightarrow$ huge **unknown** clusters. Joint estimation approach will dramatically fail when is given incorrect cluster labels.
- Known clustering label $\Longrightarrow$ **convex** optimization. Unknown clustering label $\Longrightarrow$ **non-convex** optimization.

Our Contribution

- In methodology, we propose a high dimensional EM algorithm to clustering and estimation simultaneously. Within one iteration, E-step stands the role of clustering, while M-step conducts the high-dimensional joint estimation.
- In theory, we analyze the estimator arising in each iteration step, which leads to an interesting *trade-off* between **statistical error** and **optimization error**.
- Our analysis is non-asymptotic.

Outline

Multiple Graphical Models

- Consider K clusters $\mathcal{A}_1, \ldots, \mathcal{A}_K$ with cluster assignment matrix $\boldsymbol{L} \in \mathbb{R}^{n \times K}$. Observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ are drawn from Gaussian mixture model, which can be specified as mixed Gaussian density by the form

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, \ldots, K,$$

  where π_k is k-th mixture weight, $\boldsymbol{\mu}_k$ is k-th cluster mean and $\boldsymbol{\Sigma}_k$ is k-the covariance matrix.
- Let $\boldsymbol{\Theta} := \{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K\}$, where $\boldsymbol{\Theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$.
- Goal: **Estimation K precision matrices $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K$.**

If the cluster assignment matrix is known...

- The log-likelihood function is referred to *complete data*:

$$\log \mathcal{L}(\mathbf{\Theta}|\boldsymbol{X}, \boldsymbol{L}) := \sum_{i=1}^{n}\sum_{k=1}^{K} L_{ik}[\log \pi_k + \log f_k(\boldsymbol{x}_i; \mathbf{\Theta}_k)].$$

- Joint estimation is summarized as the following optimization problem:

$$\underset{\mathbf{\Omega}_1,\dots,\mathbf{\Omega}_K \succ 0}{\text{argmax}} \ \log \mathcal{L}(\mathbf{\Theta}|\boldsymbol{X}, \boldsymbol{L}) - \mathcal{P}(\mathbf{\Theta}).$$

- $\mathcal{P}(\mathbf{\Theta})$ encourages common structure across different clusters.
- If we use convex penalty, the whole problem is a convex optimization problem.

If the cluster assignment matrix is unknown...

- **S**imultaneous **C**lustering **A**nd Estimatio**N** (SCAN).
- The log-likelihood function for the *observed data* can be specified by

$$\log \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{X}) := \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k\Big(\boldsymbol{x}_i; \boldsymbol{\mu}_k, (\boldsymbol{\Omega}_k)^{-1}\Big)\right).$$

- The **non-convex** optimization problem is formulated as

$$\max_{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k} \log \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{X}) - \mathcal{P}(\boldsymbol{\Theta}).$$

- SCAN Penalty

$$\mathcal{P}(\boldsymbol{\Theta}) = \underbrace{\lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p} |\mu_{kj}|}_{\text{feature selection}} + \underbrace{\lambda_2 \sum_{k=1}^{K}\sum_{i\neq j} |\omega_{kij}|}_{\text{sparse estimate}} + \underbrace{\lambda_3 \sum_{i\neq j}(\sum_{k=1}^{K} \omega_{kij}^2)^{1/2}}_{\text{common characteristic}}.$$

Expectation-Maximization Algorithm

- As SCAN is based on Gaussian mixture model, overall it's an non-convex problem.
- E-step is a clustering step, and M-step is a joint estimation step. The interaction between E-step and M-step makes the cluster structure more and more refined.

Outline of Our Algorithm

Input: Training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, number of clusters K, tuning parameter $\lambda_1, \lambda_2, \lambda_3$.
Output: Cluster assignment L_{ik}, cluster means μ_k and graph Ω_k.

Step 1: Randomly initialize cluster centers $\boldsymbol{\mu}_k^{(0)}$, precision matrices $\boldsymbol{\Omega}_k^{(0)}$ 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 $L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)$
(b) M-step. Given $L_{\boldsymbol{\Theta}^{(t-1)},k}(\boldsymbol{x}_i)$, update mixture weight $\pi_k^{(t)}$, cluster mean $\boldsymbol{\mu}_k^{(t)}$, and the precision matrix $\boldsymbol{\Omega}_k^{(t)}$ accordingly.

Outline

Limitation of Classical Result

- Classical convergence result of EM algorithm by Wu (1983):
  - **Unimodal** $\rightarrow$ Global optimum
  - **Muti-modal** $\rightarrow$ Local optimum (Non-convexity of GMM)
- Statisticians are interested in *Maximum Likelihood Estimate* (MLE), which shows good statistical performance.
- There is a significant gap when we move from **practical** use of EM algorithm to its **theoretical** understanding.

$$\underbrace{\text{Iterative Estimator}}_{\text{practice use}} \overset{\text{gap}}{\Longleftrightarrow} \underbrace{\text{Maximum Likelihood Estimator}}_{\text{theoretical use}}$$

- Our results show that if we are given an **appropriate initialization**, the EM update will converge to a local optima or stationary point but within a **statistical precision** of a global optima.

Limitation of Classical Result

- Classical convergence result of EM algorithm by Wu (1983):
  - **Unimodal** → Global optimum
  - **Muti-modal** → Local optimum (Non-convexity of GMM)
- Statisticians are interested in *Maximum Likelihood Estimate* (MLE), which shows good statistical performance.
- There is a significant gap when we move from **practical** use of EM algorithm to its **theoretical** understanding.

$$\underbrace{\text{Iterative Estimator}}_{\text{practice use}} \overset{\text{gap}}{\Longleftrightarrow} \underbrace{\text{Maximum Likelihood Estimator}}_{\text{theoretical use}}$$

- Our results show that if we are given an **appropriate initialization**, the EM update will converge to a local optima or stationary point but within a **statistical precision** of a global optima.

Limitation of Classical Result

- Classical convergence result of EM algorithm by Wu (1983):
  - **Unimodal** → Global optimum
  - **Muti-modal** → Local optimum (Non-convexity of GMM)
- Statisticians are interested in *Maximum Likelihood Estimate* (MLE), which shows good statistical performance.
- There is a significant gap when we move from **practical** use of EM algorithm to its **theoretical** understanding.

$$\underbrace{\text{Iterative Estimator}}_{\text{practice use}} \overset{\text{gap}}{\Longleftrightarrow} \underbrace{\text{Maximum Likelihood Estimator}}_{\text{theoretical use}}$$

- Our results show that if we are given an **appropriate initialization**, the EM update will converge to a local optima or stationary point but within a **statistical precision** of a global optima.

Limitation of Classical Result

- Classical convergence result of EM algorithm by Wu (1983):
  - **Unimodal** $\rightarrow$ Global optimum
  - **Muti-modal** $\rightarrow$ Local optimum (Non-convexity of GMM)
- Statisticians are interested in *Maximum Likelihood Estimate* (MLE), which shows good statistical performance.
- There is a significant gap when we move from **practical** use of EM algorithm to its **theoretical** understanding.

$$\underbrace{\text{Iterative Estimator}}_{\text{practice use}} \overset{\text{gap}}{\Longleftrightarrow} \underbrace{\text{Maximum Likelihood Estimator}}_{\text{theoretical use}}$$

- Our results show that if we are given an **appropriate initialization**, the EM update will converge to a local optima or stationary point but within a **statistical precision** of a global optima.

Limitation of Classical Result

- Classical convergence result of EM algorithm by Wu (1983):
  - **Unimodal** $\rightarrow$ Global optimum
  - **Muti-modal** $\rightarrow$ Local optimum (Non-convexity of GMM)
- Statisticians are interested in *Maximum Likelihood Estimate* (MLE), which shows good statistical performance.
- There is a significant gap when we move from **practical** use of EM algorithm to its **theoretical** understanding.

$$\underbrace{\text{Iterative Estimator}}_{\text{practice use}} \overset{\text{gap}}{\Longleftrightarrow} \underbrace{\text{Maximum Likelihood Estimator}}_{\text{theoretical use}}$$

- Our results show that if we are given an **appropriate initialization**, the EM update will converge to a local optima or stationary point but within a **statistical precision** of a global optima.

Population Q-function & Finite-sample Q-function

Definition (Finite-sample Q-function)

$$Q_n(\boldsymbol{\Theta}'|\boldsymbol{\Theta}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} L_{\boldsymbol{\Theta},k}(\boldsymbol{x}_i)[\log \pi_k + \log f_k(\boldsymbol{x}_i; \boldsymbol{\Theta}'_k)].$$

It corresponds to *statistical error*.

Definition (Population Q-function)

$$Q(\boldsymbol{\Theta}'|\boldsymbol{\Theta}) := \mathbb{E}\left[\sum_{k=1}^{K} L_{\boldsymbol{\Theta},k}(\boldsymbol{X})[\log \pi_k' + \log f_k(\boldsymbol{X}; \boldsymbol{\Theta}_k')]\right].$$

It corresponds to *optimization error*.

Final Estimation Error

Theorem

Consider the SCAN procedure with initialization $\mathbf{\Theta}^{(0)} \in \mathcal{B}_\alpha(\mathbf{\Theta}^*)$ *for some constant radius* α. *Let* $\kappa < 1$ *be a contractive parameter. If the sample size* n *is large enough, the iterative estimator* $\mathbf{\Theta}^{(t)}$ *satisfies*

$$\left\|\mathbf{\Theta}^{(t)} - \mathbf{\Theta}^*\right\|_2 \lesssim \underbrace{\varepsilon\left(n, p, K, \Psi(\mathcal{M})\right)}_{\textbf{Statistical Error(SE)}} + \underbrace{\kappa^t \left\|\mathbf{\Theta}^{(0)} - \mathbf{\Theta}^*\right\|_2}_{\textbf{Optimiation Error(OE)}}, \quad (2.1)$$

with high probability. $\Psi(\mathcal{M})$ *measures the sparsity of cluster means and precision matrices. Here* $\mathbf{\Theta}^* \in \mathbb{R}^{K(p^2+p)}$ *is the true parameter.*

Statistical Error

Corollary

When the iteration step t is large enough such that

$$t \geq T = \log_{1/\kappa} \frac{\left\| \boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^* \right\|_2}{\varphi(n, p, K)},$$

the optimization error is dominated by the statistical error, namely

$$\left\| \boldsymbol{\Theta}^{(T)} - \boldsymbol{\Theta}^* \right\|_2 = O_P \left(\underbrace{\sqrt{\frac{K^5 d \log p}{n}}}_{\text{Cluster means error}} + \underbrace{\sqrt{\frac{K^3 (Ks + p) \log p}{n}}}_{\text{Precision matrices error}} \right),$$

with probability converging to 1. Here d and s represent the sparsity of cluster mean and precision matrix for a single cluster.

Some Remarks

Remark

- This corollary tells us if we are given an **appropriate initialization**, the EM update will converge to a local optima or stationary point within a **statistical precision** of a global optima.
- In our framework, we allow the number of cluster K diverging with n and p, which fits our big data set up.
- When K is fixed, if the cluster label is given in advance, the rate $O_P(\sqrt{(s+p)\log p/n})$ is the optimal rate for precision matrix estimation. If the covariance matrix is an identity matrix, $O_P(\sqrt{d\log p/n})$ is the optimal rate for cluster mean estimation.

Outline

Simulation Set Up

- $K = 3, n = 300, p = 100$. Label Y_i is uniformly generated from $\{1, 2, 3\}$.
- $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}(Y_i), \Omega(Y_i))$.
- The first 10 variables of $\boldsymbol{\mu}(Y_i)$ are

$$\begin{cases} (\mu\mathbf{1}_5^T, -\mu\mathbf{1}_5^T)^T & \text{if} \quad Y_i = 1 \\ \mu\mathbf{1}_{10} & \text{if} \quad Y_i = 2 \\ (-\mu\mathbf{1}_5^T, -\mu\mathbf{1}_5^T)^T & \text{if} \quad Y_i = 3 \end{cases},$$

  and the last 90 variables of $\boldsymbol{\mu}(Y_i)$ are zeros.
- Ω_i are tridiagonal matrix. The off-diagonal term of $\Omega_1, \Omega_2, \Omega_3$ is $\eta, 0.9 * \eta, 1.1 * \eta$, respectively.
- **Model 1**: $\mu = 1, \eta = 0.4$; **Model 2**: $\mu = 0.8, \eta = 0.3$; **Model 3**: $\mu = 0.8, \eta = 0.4$.

Simulation Results

The clustering errors(CE), mean estimation error (MEE), precision estimation errors (PEE), true positive rate (TPR1) and false positive rate (FPR1) of mean estimation, true positive rate (TPR2) and false positive rate (FPR2) of precision matrix estimation of various clustering algorithms in three simulations.

Models	Methods	CE	MEE	PEE	TPR1/FPR1	TPR2 /FPR2
	K-means	0.219	3.65	NA	1/1	NA / NA
Model 1	Zhou et al. (2009)	0.134	2.33	11.62	0.99 /0.24	0.996 /0.154
$\mu = 1$	K-means + JGL	0.226	3.68	10.53	1 /1	0.991 /0.04
$\eta = 0.4$	SCAN	**0.070**	**1.97**	**8.57**	0.97 /0.23	0.998 /0.04
	K-means	0.186	2.58	NA	1 /1	NA / NA
Model 2	Zhou et al. (2009)	0.104	1.64	10.93	0.97 /0.16	0.96 /0.1
$\mu = 0.8$	K-means + JGL	0.112	1.79	8.05	1 /1	0.99 /0.007
$\eta = 0.3$	SCAN	**0.039**	**1.30**	**7.52**	1 /0.15	1 /0.006
	K-means	0.015	1.30	NA	1 /1	NA / NA
Model 3	Zhou et al. (2009)	0.024	1.38	10.42	0.99 /0	0.97 /0.099
$\mu = 0.8$	K-means + JGL	0.015	1.30	7.55	1 /1	0.999 /0.006
$\eta = 0.4$	SCAN	**0.007**	**1.29**	**7.50**	1 /0	0.999 /0.006

References I

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2016). Statistical guarantees for the em algorithm: From population to sample-based analysis. *AoS*.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *JRSS-B*.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2014). High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint*.

Yi, X. and Caramanis, C. (2015). Regularized em algorithms: A unified framework and statistical guarantees. *arXiv preprint*.

Zhou, H., Pan, W., and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Statist.*

Thank you!
Botao Hao
Purdue University