Statistical and Computational Tradeoffs in Biconvex Optimization

Wei Sun

Department of Statistics
Purdue University

Joint work with Guang Cheng (Purdue), Yufeng Liu (UNC), Zhaoran Wang (Princeton), and Han Liu (Princeton)

Outline

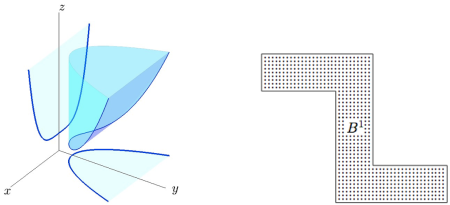
- Motivations
- General theory of biconvex optimization
- Application 1: bigraphical model
- Application 2: joint clustering and network estimation
- Future work

Background: Biconvex Optimization

- A function $g(x, y) : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is biconvex if $g(x, y)$ is convex in x for fixed $y \in \mathcal{B}$, and convex in y for fixed $x \in \mathcal{A}$.
- Biconvex optimization:

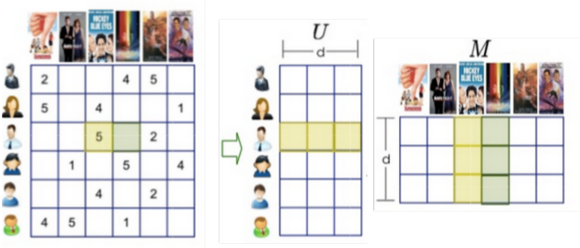
$$\min \; g(x, y)$$
$$s.t. \; x \in \mathcal{A}, y \in \mathcal{B}$$

Figure : Biconvex function and biconvex set



Resource: https://inst.eecs.berkeley.edu/ ee127a

Motivation: Non-negative matrix factorization



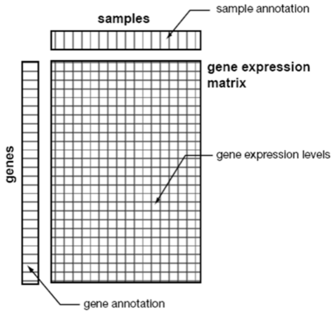
Source: A. Karatzoglou, ESSIR 2013 Recommender Systems tutorial

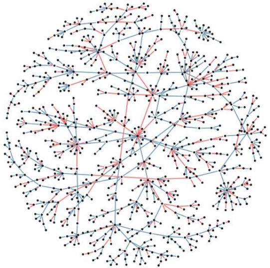
- Non-negative matrix factorization solves

$$\begin{aligned} \min_{U,M} \quad & \frac{1}{2}\|X - UM\|_F^2 \\ s.t. \quad & U_{ij} \geq 0, M_{ij} \geq 0. \end{aligned}$$

Motivation: Bigraphical model

Traditional Graphical Model:





https://galton.uchicago.edu/ lafferty/research.html

Motivation: Bigraphical model

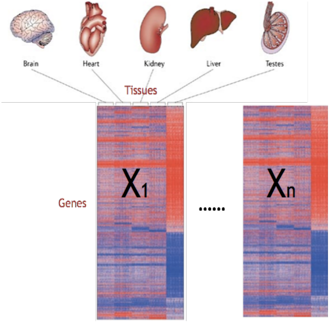


Figure : Matrix-variate data

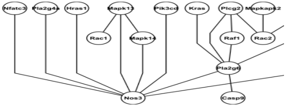


Figure : Gene network

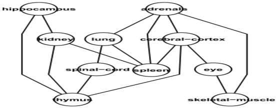


Figure : Tissue network

Source: Yin and Li (2012)

Biconvex Optimization

- Population objective function: $g(\mathbf{a}, \mathbf{b})$

$$
\begin{aligned}
(\mathbf{a}^*, \mathbf{b}^*) &= \arg\min \; g(\mathbf{a}, \mathbf{b}) \\
& \quad s.t. \; \mathbf{a} \in \mathcal{A}, \mathbf{b} \in \mathcal{B}
\end{aligned}
$$

- Sample objective function: $g_n(\mathbf{a}, \mathbf{b})$
- Goal: Find a minimizer $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}})$ via $g_n(\mathbf{a}, \mathbf{b})$ s.t. $\|\widehat{\mathbf{a}} - \mathbf{a}^*\|_2$ and $\|\widehat{\mathbf{b}} - \mathbf{b}^*\|_2$ are small given limited computational resources.

Alternative Update Algorithm

Input: function $g_n(\mathbf{a}, \mathbf{b})$, maximal number of iterations T.

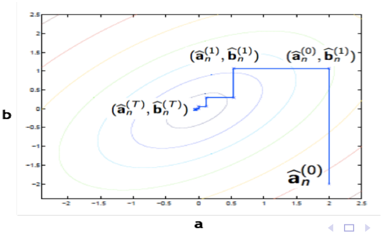
Initialize: $\widehat{\mathbf{a}}_n^{(0)}$.

For $t = 1$ **to** T:

- Fix $\widehat{\mathbf{a}}_n^{(t-1)}$, update $\widehat{\mathbf{b}}_n^{(t)} = \arg\min_{\mathbf{b}} g_n(\widehat{\mathbf{a}}_n^{(t-1)}, \mathbf{b})$;
- Fix $\widehat{\mathbf{b}}_n^{(t)}$, update $\widehat{\mathbf{a}}_n^{(t)} = \arg\min_{\mathbf{a}} g_n(\mathbf{a}, \widehat{\mathbf{b}}_n^{(t)})$;

End For

Output: $\widehat{\mathbf{a}} = \widehat{\mathbf{a}}_n^{(T)}$ and $\widehat{\mathbf{b}} = \widehat{\mathbf{b}}_n^{(T)}$.



Theory: Outline

$$\begin{tikzcd} \boxed{g_n(\mathbf{a}, \mathbf{b})} \arrow[d] & \boxed{g(\mathbf{a}, \mathbf{b})} \arrow[d] \\ (\widehat{\mathbf{a}}_n^{(t)}, \widehat{\mathbf{b}}_n^{(t)}) \arrow[d] & (\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \arrow[d] \\ {} & {} \end{tikzcd}$$

$$\left\|\widehat{\mathbf{a}}_n^{(t)} - \mathbf{a}^*\right\|_2 \leq \underbrace{\left\|\mathbf{a}^{(t)} - \mathbf{a}^*\right\|_2}_{\textit{Population}} + \underbrace{\left\|\widehat{\mathbf{a}}_n^{(t)} - \mathbf{a}^{(t)}\right\|_2}_{\textit{Sample}}$$

Theory: Population Version

- We focus on the Euclidean ball of radius $\alpha > 0$ for $\mathcal{A}$ and $\mathcal{B}$.

$$\begin{aligned}\mathcal{A} = \mathcal{B}(\alpha; \mathbf{a}^*) &:= \{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a} - \mathbf{a}^*\|_2 \leq \alpha\} \\ \mathcal{B} = \mathcal{B}(\alpha; \mathbf{b}^*) &:= \{\mathbf{b} \in \mathbb{R}^q : \|\mathbf{b} - \mathbf{b}^*\|_2 \leq \alpha\}\end{aligned}$$

- Denote $\nabla_1 g(\mathbf{a}, \mathbf{b})$ be the gradient w.r.t. $\mathbf{a}$ and $\nabla_2 g(\mathbf{a}, \mathbf{b})$ be the gradient w.r.t. $\mathbf{b}$.

Condition ((λ, λ')-Strong-Convexity)

The function $g(\mathbf{a}^*, \cdot)$ is λ-strongly convex, and $g(\cdot, \mathbf{b}^*)$ is λ'-strongly convex. That is, for any $\mathbf{b}_1, \mathbf{b}_2 \in \mathcal{B}$ and $\mathbf{a}_1, \mathbf{a}_2 \in \mathcal{A}$,

$$\begin{aligned}g(\mathbf{a}^*, \mathbf{b}_1) - g(\mathbf{a}^*, \mathbf{b}_2) - \langle \nabla_2 g(\mathbf{a}^*, \mathbf{b}_2), \mathbf{b}_1 - \mathbf{b}_2 \rangle &\geq \frac{\lambda}{2} \cdot \|\mathbf{b}_1 - \mathbf{b}_2\|_2^2 \\ g(\mathbf{a}_1, \mathbf{b}^*) - g(\mathbf{a}_2, \mathbf{b}^*) - \langle \nabla_1 g(\mathbf{a}_2, \mathbf{b}^*), \mathbf{a}_1 - \mathbf{a}_2 \rangle &\geq \frac{\lambda'}{2} \cdot \|\mathbf{a}_1 - \mathbf{a}_2\|_2^2\end{aligned}$$

Theory: Population Version

- Denote population minimization functions:

$$M_1(\mathbf{a}) = \arg\min_{\mathbf{b}} g(\mathbf{a}, \mathbf{b}); \quad M_2(\mathbf{b}) = \arg\min_{\mathbf{a}} g(\mathbf{a}, \mathbf{b}).$$

Condition ((γ, γ')-Lipschitz-Gradient)

The function $\nabla_2 g(\mathbf{a}, \cdot)$ satisfies γ-Lipschitz gradient condition, and the function $\nabla_1 g(\cdot, \mathbf{b})$ satisfies γ'-Lipschitz gradient condition. That is, for any $\mathbf{a} \in \mathcal{A}$ and any $\mathbf{b} \in \mathcal{B}$,

$$\left\| \nabla_2 g(\mathbf{a}^*, M_1(\mathbf{a})) - \nabla_2 g(\mathbf{a}, M_1(\mathbf{a})) \right\|_2 \leq \gamma \cdot \|\mathbf{a}^* - \mathbf{a}\|_2$$

$$\left\| \nabla_1 g(M_2(\mathbf{b}), \mathbf{b}^*) - \nabla_1 g(M_2(\mathbf{b}), \mathbf{b}) \right\|_2 \leq \gamma' \cdot \|\mathbf{b}^* - \mathbf{b}\|_2.$$

Theory: Population Version

Theorem

Under $(\lambda, \lambda^{'})$-Strong-Convexity and $(\gamma, \gamma^{'})$-Lipschitz-Gradient conditions, we have

$$\|M_1(\mathbf{a}) - \mathbf{b}^*\|_2 \leq (\gamma/\lambda) \cdot \|\mathbf{a} - \mathbf{a}^*\|_2 \quad \textit{for any } \mathbf{a} \in \mathcal{A},$$
$$\|M_2(\mathbf{b}) - \mathbf{a}^*\|_2 \leq (\gamma^{'}/\lambda^{'}) \cdot \|\mathbf{b} - \mathbf{b}^*\|_2 \quad \textit{for any } \mathbf{b} \in \mathcal{B}.$$

Moreover, for any initialization $\mathbf{a}^{(0)}$, the solutions from the population alternative updates converge linearly,

$$\|\mathbf{b}^{(t)} - \mathbf{b}^*\|_2 \leq (\frac{\gamma}{\lambda})^t (\frac{\gamma^{'}}{\lambda^{'}})^{t-1} \cdot \|\mathbf{a}^{(0)} - \mathbf{a}^*\|_2$$
$$\|\mathbf{a}^{(t)} - \mathbf{a}^*\|_2 \leq (\frac{\gamma}{\lambda})^t (\frac{\gamma^{'}}{\lambda^{'}})^t \cdot \|\mathbf{a}^{(0)} - \mathbf{a}^*\|_2.$$

Theory: Sample Version

- Denote sample minimization functions:

$$M_{1n}(\mathbf{a}) = \arg\min_{\mathbf{b}} g_n(\mathbf{a}, \mathbf{b}); \quad M_{2n}(\mathbf{b}) = \arg\min_{\mathbf{a}} g_n(\mathbf{a}, \mathbf{b}).$$

Condition (Statistical-Error(ϵ_s, δ, n))

Uniformly over all $\mathbf{a} \in \mathcal{A}$ with $\mathbf{b} \in \mathcal{B}$, we have that

$$\max\left\{ \left\| M_{1n}(\mathbf{a}) - M_1(\mathbf{a}) \right\|_2, \left\| M_{2n}(\mathbf{b}) - M_2(\mathbf{b}) \right\|_2 \right\} \le \epsilon_s$$

with probability at least $1 - \delta$.

Theory: Main Result

Denote the initialization error as $\epsilon_0 := \|\widehat{\mathbf{a}}_n^{(0)} - \mathbf{a}^*\|_2$.

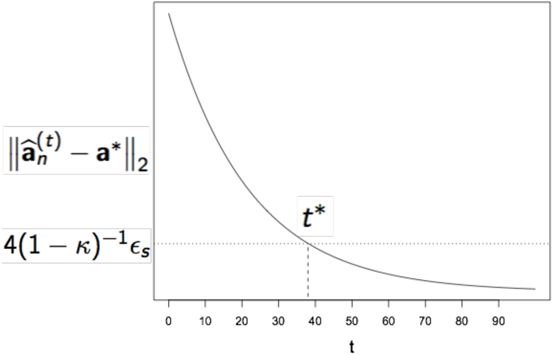
Theorem

Under above assumptions and assume $\gamma < \lambda$ and $\gamma' < \lambda'$ s.t. $\kappa = \gamma\gamma'/(\lambda\lambda') < 1$. Assume $\widehat{\mathbf{a}}_n^{(0)} \in \mathcal{B}(\alpha; \mathbf{a}^*)$, and n is sufficiently large such that $\epsilon_s \le \min\{(1-\gamma/\lambda)\alpha, (1-\gamma'/\lambda')\alpha\}$. Then

$$
\begin{aligned}
\left\|\widehat{\mathbf{b}}_n^{(t)} - \mathbf{b}^*\right\|_2 &\le 2(1-\kappa)^{-1}\epsilon_s + \kappa^{t-1}\epsilon_0, \\
\left\|\widehat{\mathbf{a}}_n^{(t)} - \mathbf{a}^*\right\|_2 &\le \underbrace{2(1-\kappa)^{-1}\epsilon_s}_{Statistical\ Error} + \underbrace{\kappa^t\epsilon_0}_{Optimization\ Error}.
\end{aligned}
$$

Theory: Main Result

$$\left\|\widehat{\mathbf{a}}_n^{(t)} - \mathbf{a}^*\right\|_2 \le \underbrace{2(1-\kappa)^{-1}\epsilon_s}_{\text{Statistical Error}} + \underbrace{\kappa^t \epsilon_0}_{\text{Optimization Error}}$$



Application 1: Bigraphical model

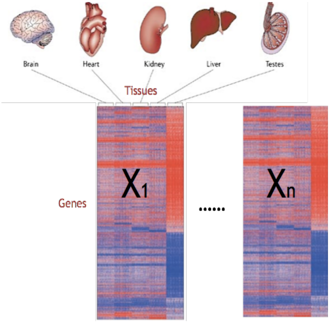


Figure : Matrix-variate data

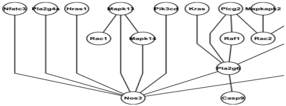


Figure : Gene network

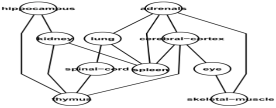


Figure : Tissue network

Application 1: Bigraphical Model

A random matrix $\mathbf{X} \in \mathbb{R}^{p \times q}$ follows a matrix-variate normal if

$$\mathrm{vec}(\mathbf{X}) \sim N_{pq}(\mathbf{0}, \Psi \otimes \Sigma).$$

- Denote the precision matrices $\Lambda = \Psi^{-1}$ and $\Omega = \Sigma^{-1}$
- Zeros in Λ (or Ω) define pairwise conditional independence of corresponding entries given all other entries.
- Goal: Estimate sparse Λ and Ω.

Bigraphical Model

- Given i.i.d. data $\mathbf{X}_1 \ldots, \mathbf{X}_n$, the lasso penalized likelihood estimator minimizes

$$g_n(\Omega, \Lambda) = \frac{1}{npq} \sum_{i=1}^{n} \mathrm{tr}(\mathbf{X}_i \Lambda \mathbf{X}_i^\top \Omega) - \frac{1}{p} \mathrm{logdet}(\Omega) - \frac{1}{q} \mathrm{logdet}(\Lambda) + \lambda_1 \|\Omega\|_1 + \lambda_2 \|\Lambda\|_1.$$

- The population version likelihood function is

$$g(\Omega, \Lambda) = \frac{1}{pq} \mathbb{E}\big[\mathrm{tr}(\mathbf{X} \Lambda \mathbf{X}^\top \Omega)\big] - \frac{1}{p} \mathrm{logdet}(\Omega) - \frac{1}{q} \mathrm{logdet}(\Lambda).$$

- The objective function $g(\Omega, \Lambda)$ is biconvex in (Ω, Λ).

Bigraphical Model: Algorithm

Input: samples $\mathbf{X}_1 \ldots, \mathbf{X}_n$, maximal number of iterations T, tuning parameters λ_1, λ_2.
Initialize $\Omega^{(0)}$.
For $t = 1$ **to** T: Alternatively update $\Omega^{(t)}, \Lambda^{(t)}$ as

- Given $\Omega^{(t-1)}$, compute $\mathbf{S}_1 = (np)^{-1} \sum_{i=1}^n \mathbf{X}_i^\top \Omega^{(t-1)} \mathbf{X}_i$ and solve the glasso problem

$$\Lambda^{(t)} = \arg\min_{\Lambda} \left\{ \frac{1}{q} \mathrm{tr}(\mathbf{S}_1 \Lambda) - \frac{1}{q} \mathrm{logdet}(\Lambda) + \lambda_\Lambda \|\Lambda\|_1 \right\}$$

- Given $\Lambda^{(t)}$, compute $\mathbf{S}_2 = (nq)^{-1} \sum_{i=1}^n \mathbf{X}_i \Lambda^{(t)} \mathbf{X}_i^\top$ and solve the glasso problem

$$\Omega^{(t)} = \arg\min_{\Omega} \left\{ \frac{1}{p} \mathrm{tr}(\mathbf{S}_2 \Omega) - \frac{1}{p} \mathrm{logdet}(\Omega) + \lambda_\Omega \|\Omega\|_1 \right\}$$

End For
Output: $\widehat{\Omega} = \Omega^{(T)}$ and $\widehat{\Lambda} = \Lambda^{(T)}$.

Bigraphical Model: Convergence Analysis

- Step 1: Verify Strongly-Convexity of $g(\Omega^*, \cdot)$ and $g(\cdot, \Lambda^*)$, and Lipschitz-Gradient conditions of $\nabla_1 g(\cdot, \Lambda)$ and $\nabla_2 g(\Omega, \cdot)$.
- Define

$$\begin{aligned} \mathcal{B}(\alpha; \Omega^*) &:= \{\Omega \in \mathbb{R}^{p \times p} : \|\Omega - \Omega^*\|_F \le \alpha\} \\ \mathcal{B}(\alpha; \Lambda^*) &:= \{\Lambda \in \mathbb{R}^{q \times q} : \|\Lambda - \Lambda^*\|_F \le \alpha\}. \end{aligned}$$

Bigraphical Model: Convergence Analysis

The alternative update algorithm via the population objective function $g(\Omega, \Lambda)$ is locally contractive.

Corollary

For the population objective function $g(\Omega, \Lambda)$, we have that $g(\Omega^*, \cdot)$ and $g(\cdot, \Lambda^*)$ are strongly convex with parameters, respectively,

$$\lambda = p^{-1}[\|\Omega^*\|_2 + 3\alpha]^{-2} \ \mathit{and} \ \lambda' = q^{-1}[\|\Lambda^*\|_2 + 3\alpha]^{-2}.$$

Both $\nabla_1 g(\cdot, \Lambda)$ and $\nabla_2 g(\Omega, \cdot)$ satisfy the Lipschitz-Gradient conditions with

$$\gamma = \gamma' = (pq)^{-1}\|\Sigma^* \otimes \Psi^*\|_F.$$

If $\|\Sigma^* \otimes \Psi^*\|_F$ is bounded and if there exist constants $C_1, C_2 > 0$ such that $C_1 \leq \|\Omega^*\|_2, \|\Lambda^*\|_2 \leq C_2$, then

$$\gamma < \lambda, \gamma' < \lambda'.$$

- Step 2: Compute the statistical error.

- Let $\mathcal{S}_1 := \{(i,j) : \Omega^*_{ij} \neq 0\}$ and $\mathcal{S}_2 := \{(i,j) : \Lambda^*_{ij} \neq 0\}$.
- Denote $s_1 = |\mathcal{S}_1| - p$ and $s_2 = |\mathcal{S}_2| - q$.
- Remind that

$$M_{1n}(\Omega) := \arg\min_{\Lambda} g_n(\Omega, \Lambda), \;\; M_{2n}(\Lambda) = \arg\min_{\Omega} g_n(\Omega, \Lambda),$$
$$M_{1}(\Omega) := \arg\min_{\Lambda} g(\Omega, \Lambda), \;\; M_{2}(\Lambda) = \arg\min_{\Omega} g(\Omega, \Lambda).$$

Condition (Bounded Eigenvalues)

There are positive constants C_1 and C_2 such that

$$0 < C_1 \leq \lambda_{\min}(\Sigma^*) \leq \lambda_{\max}(\Sigma^*) \leq 1/C_1 < \infty$$
$$0 < C_2 \leq \lambda_{\min}(\Psi^*) \leq \lambda_{\max}(\Psi^*) \leq 1/C_2 < \infty.$$

Condition (Tuning)

The tuning parameters satisfy

$$\lambda_\Omega = O\left(\sqrt{\frac{\log p}{np^2 q}}\right), \lambda_\Lambda = O\left(\sqrt{\frac{\log q}{npq^2}}\right).$$

Bigraphical Model: Convergence Analysis

Corollary

Under above two conditions, the statistical errors are

$$\sup_{\Omega \in \mathcal{B}(\alpha;\Omega^*)} \left\| M_{1n}(\Omega) - M_1(\Omega) \right\|_F = O_p\left(\sqrt{\frac{(q+s_2)\log q}{np}} \right),$$

$$\sup_{\Lambda \in \mathcal{B}(\alpha;\Lambda^*)} \left\| M_{2n}(\Lambda) - M_2(\Lambda) \right\|_F = O_p\left(\sqrt{\frac{(p+s_1)\log p}{nq}} \right).$$

- Step 1: Exploit the independence structure in $(\Omega^*)^{\frac{1}{2}}\mathbf{X}(\Lambda^*)^{\frac{1}{2}}$.
- Step 2: Use Talagrand inequality for the convergence rate of

$$(np)^{-1}\sum_{i=1}^{n} \mathbf{X}_i^\top \Omega \mathbf{X}_i - p^{-1}\mathbb{E}\left[\mathbf{X}^\top \Omega \mathbf{X}\right].$$

Bigraphical Model: Main Result

Combine above two Corollaries, we have

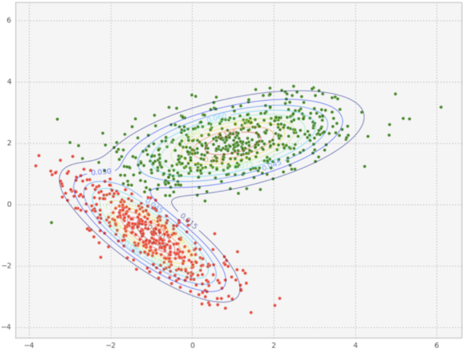
$$\|\widehat{\Lambda}^{(t)} - \Lambda^*\|_F \leq C\sqrt{\frac{(p+s_1)\log p}{nq}} + \kappa^{t-1}\epsilon_0,$$

$$\|\widehat{\Omega}^{(t)} - \Omega^*\|_F \leq C\sqrt{\frac{(q+s_2)\log q}{np}} + \kappa^{t}\epsilon_0.$$

- When $n = 1$, we can still consistently estimate Λ^* or Ω^*.
- Leng and Tang (2012) showed there existed a local minimizer which can obtain above statistical error.
- We prove that our algorithm can find such minimizer.
- The convergence rates showed in Yin and Li (2012), Tsiligkaridis et al. (2013) are slower than ours and they require at least $n > (p+q)(\log p + \log q)$.

Application 2: Joint Clustering and Network Estimation

- Gaussian mixture model (GMM) $\pi_k N(\boldsymbol{\mu}_k, \Sigma_k)$, $k = 1, \ldots, K$.



Source: http://www.nehalemlabs.net/

Application 2: Background

- Samples $\mathbf{x}_i, \ldots, \mathbf{x}_n$ follows a GMM with $\pi_k f_k(x; \boldsymbol{\mu}_k, \Sigma_k)$.
- If assume $\Sigma_k = \sigma_k \mathbb{1}_p$ (Pan and Shen, 2007, Sun et al., 2012), the clustering can be solved by minimizing

$$\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k) \right) - P(\boldsymbol{\mu}).$$

- If assume clustering assignment is given, the networks are estimated jointly (Guo et al., 2011, Danaher et al., 2014) via

$$\begin{aligned} \max_{\Omega_1,\ldots,\Omega_K} \quad & \sum_{k=1}^{K} n_k [\log\det(\Omega_k) - \mathrm{tr}(S_k \Omega_k)] - P(\Omega) \\ \text{s.t.} \quad & \Omega_1, \ldots, \Omega_K \text{ are positive definite.} \end{aligned}$$

Application 2: Joint Clustering and Network Estimation

- Denote the set of parameters as $\Theta := \{(\boldsymbol{\mu}_k, \Omega_k), k \in [K]\}$. Our optimization is formulated as

$$\max_{\pi_k, \boldsymbol{\mu}_k, \Omega_k} \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_k \Big(\mathbf{x}_i; \boldsymbol{\mu}_k, \Omega_k \Big) \right) - P(\Theta).$$

- We focus on the l_1 penalty on $\boldsymbol{\mu}_k$ and fussed graphical lasso penalty (Danaher et al., 2014) on $\Omega_k = (\omega_{kij})$,

$$P(\Theta) = \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p} |\mu_{kj}| + \lambda_2 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{kij}| + \lambda_3 \sum_{k<k'} \sum_{i,j} |\omega_{kij} - \omega_{k'ij}|.$$

Application 2: EM Algorithm

- Denote the K clusters as $\mathcal{A}_1, \cdots, \mathcal{A}_K$, denote cluster assignment matrix L with entry $L_{ik} = \mathbf{1}(X_i \in \mathcal{A}_k)$.
- The regularized complete log-likelihood function is

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

- E-step: compute the conditional expectation

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

- M-step: maximize $Q(\Theta | \widehat{\Theta}^{(t)})$ w.r.t. π_k, $\boldsymbol{\mu}_k$, Ω_k.

Application 2: EM Algorithm

M-step is a tri-convex optimization.

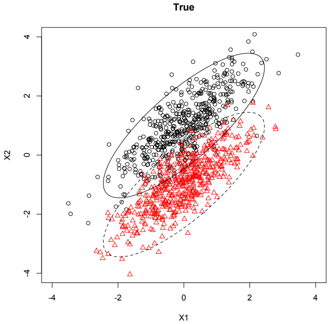
- Update π_k: $\widehat{\pi}_k^{(t+1)} = n^{-1} \sum_{i=1}^{n} \widehat{L}_{ik}^{(t)}$.
- Update $\boldsymbol{\mu}_k$: solve sparse mean via KKT condition.
- Update Ω_k: solve sparse networks via existing joint graphical lasso algorithms,

$$\max_{\Omega_1,\ldots,\Omega_K} \sum_{k=1}^{K} n_k [\log\det(\Omega_k) - \mathrm{trace}(\widetilde{S}_k \Omega_k)] - P(\Omega).$$

Application 2: Illustration

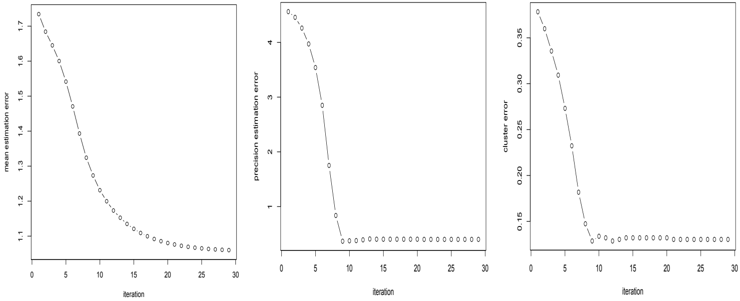
- $n = 1000$ with 500 from $N(\boldsymbol{\mu}_1, \Sigma)$ and 500 from $N(\boldsymbol{\mu}_2, \Sigma)$,

$$\boldsymbol{\mu}_1 = (0,1)^T, \boldsymbol{\mu}_2 = (0,-1)^T, \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$



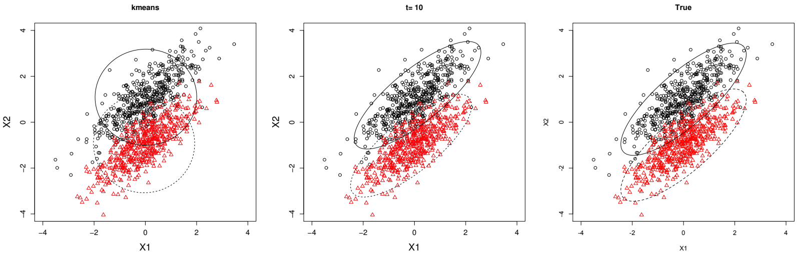
Application 2: Illustration

Figure : Mean vector estimation errors, precision matrix estimation errors, and cluster errors versus # of iterations.



Application 2: Illustration

Figure : Kmeans, Iteration $t = 10$ of our algorithm, and the truth.

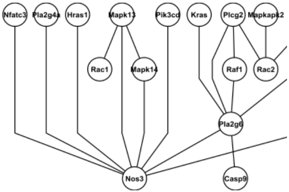


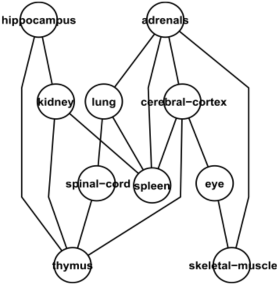
Summary

- A general convergence study of bi-convex problems.
- It reveals statistical and computational tradeoffs.
- Our theory is widely applicable to many models:
  - bigraphical model,
  - joint clustering and network estimation,
  - non-negative matrix factorization,
  - sparse tensor decomposition...

Future Work: Statistical Inference

- From parameter estimation to statistical inference.
- In the bigraphical model, test $H_0 : \omega_{ij} = 0$ v.s. $H_1 : \omega_{ij} \neq 0$
- Tools: Desparsify Lasso (van de Geer et al., 2014), De-correlated Score Test (Ning and Liu, 2014).





Wei Sun
Department of Statistics
Purdue University
sun244@purdue.edu