Alternating Proximal Gradient Method for Sparse and Low-Rank Optimization

Shiqian Ma

Institute for Mathematics and Its Applications, University of Minnesota

SPOC, Hefei, Anhui
July 1, 2012

Covariance estimation from portfolio selection

- Mean-variance model (Markowitz 1952)

$$\begin{array}{ll} \min & x^\top \Sigma x \\ \text{s.t.} & \mu^\top x = r \\ & e^\top x = 1. \end{array}$$

- Solution: $x^* = \alpha \Sigma^{-1} \mu + \beta \Sigma^{-1} e$
- Question: How to estimate a good Σ or Σ^{-1}
- Risk underestimation for mean-variance model in high dimension (El Karoui (2009))
- Data: Stock return data from history
- Decision: Estimate a good inverse covariance matrix
- Structure: sparse

Covariance estimation from portfolio selection

- Mean-variance model (Markowitz 1952)

$$\begin{array}{ll} \min & x^\top \Sigma x \\ \text{s.t.} & \mu^\top x = r \\ & e^\top x = 1. \end{array}$$

- Solution: $x^* = \alpha \Sigma^{-1} \mu + \beta \Sigma^{-1} e$
- Question: How to estimate a good Σ or Σ^{-1}
- Risk underestimation for mean-variance model in high dimension (El Karoui (2009))
- Data: Stock return data from history
- Decision: Estimate a good inverse covariance matrix
- Structure: sparse

Covariance estimation from portfolio selection

- Mean-variance model (Markowitz 1952)

$$\begin{array}{ll} \min & x^\top \Sigma x \\ \text{s.t.} & \mu^\top x = r \\ & e^\top x = 1. \end{array}$$

- Solution: $x^* = \alpha \Sigma^{-1} \mu + \beta \Sigma^{-1} e$
- Question: How to estimate a good Σ or Σ^{-1}
- Risk underestimation for mean-variance model in high dimension (El Karoui (2009))
- Data: Stock return data from history
- Decision: Estimate a good inverse covariance matrix
- Structure: sparse

Covariance estimation from portfolio selection

- Mean-variance model (Markowitz 1952)

$$\begin{array}{ll} \min & x^\top \Sigma x \\ \text{s.t.} & \mu^\top x = r \\ & e^\top x = 1. \end{array}$$

- Solution: $x^* = \alpha \Sigma^{-1} \mu + \beta \Sigma^{-1} e$
- Question: How to estimate a good Σ or Σ^{-1}
- Risk underestimation for mean-variance model in high dimension (El Karoui (2009))
- Data: Stock return data from history
- Decision: Estimate a good inverse covariance matrix
- Structure: sparse

Sparse Inverse Covariance Estimation (SICE)

- $(X_1, \ldots, X_p)$ multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$
- $(\Sigma^{-1})_{ij} = cov(X_i, X_j | rest)$. $(\Sigma^{-1})_{ij} = 0 \Leftrightarrow X_i \perp X_j \mid X_{-(i,j)}$.
- Suppose we have iid noise $\epsilon_i \sim \mathcal{N}(0, 1)$ and linear model: $x = z + \epsilon_1, \quad y = z + \epsilon_2, \quad z = \epsilon_3$.
- Covariance matrix and inverse covariance matrix:

$$\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 3 \end{pmatrix}$$

- Convex formulation (Yuan and Lin 2007; Banerjee, El Ghaoui and d'Aspremont 2007; Friedman, Hastie and Tibshirani 2008)

$$\min_S \langle \hat{\Sigma}, S \rangle - \log\det S + \rho \|S\|_1.$$

Sparse Inverse Covariance Estimation (SICE)

- $(X_1, \ldots, X_p)$ multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$
- $(\Sigma^{-1})_{ij} = cov(X_i, X_j | rest)$. $(\Sigma^{-1})_{ij} = 0 \Leftrightarrow X_i \perp X_j \mid X_{-(i,j)}$.
- Suppose we have iid noise $\epsilon_i \sim \mathcal{N}(0, 1)$ and linear model: $x = z + \epsilon_1, \quad y = z + \epsilon_2, \quad z = \epsilon_3$.
- Covariance matrix and inverse covariance matrix:

$$\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 3 \end{pmatrix}$$

- Convex formulation (Yuan and Lin 2007; Banerjee, El Ghaoui and d'Aspremont 2007; Friedman, Hastie and Tibshirani 2008)

$$\min_S \langle \hat{\Sigma}, S \rangle - \log\det S + \rho \|S\|_1.$$

Sparse Inverse Covariance Estimation (SICE)

- $(X_1, \ldots, X_p)$ multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$
- $(\Sigma^{-1})_{ij} = cov(X_i, X_j | rest)$. $(\Sigma^{-1})_{ij} = 0 \Leftrightarrow X_i \perp X_j \mid X_{-(i,j)}$.
- Suppose we have iid noise $\epsilon_i \sim \mathcal{N}(0, 1)$ and linear model: $x = z + \epsilon_1, \quad y = z + \epsilon_2, \quad z = \epsilon_3$.
- Covariance matrix and inverse covariance matrix:

$$\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 3 \end{pmatrix}$$

- Convex formulation (Yuan and Lin 2007; Banerjee, El Ghaoui and d'Aspremont 2007; Friedman, Hastie and Tibshirani 2008)

$$\min_S \langle \hat{\Sigma}, S \rangle - \log\det S + \rho \|S\|_1.$$

Latent variable covariance estimation (LVCE)

- X: observed variables; Y: latent variables (hidden factors)
- (X, Y) jointly follow a multivariate Gaussian distribution.
- Covariance matrix and inverse covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \qquad \Theta = \begin{bmatrix} \Theta_X & \Theta_{XY} \\ \Theta_{YX} & \Theta_Y \end{bmatrix}$$

- Inverse covariance matrix

$$\Sigma_X^{-1} = \Theta_X - \Theta_{XY}\Theta_Y^{-1}\Theta_{YX}$$

- Inverse covariance matrix = "sparse - low-rank"
- Convex formulation (Chandrasekaran et.al., 2010)

$$\begin{array}{ll} \min_{R,S,L} & \langle R, \hat{\Sigma}_X \rangle - \log\det(R) + \alpha\|S\|_1 + \beta \mathrm{Tr}(L) \\ \text{s.t.} & R = S - L, R \succ 0, L \succeq 0. \end{array}$$

Latent variable covariance estimation (LVCE)

- X: observed variables; Y: latent variables (hidden factors)
- (X, Y) jointly follow a multivariate Gaussian distribution.
- Covariance matrix and inverse covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \qquad \Theta = \begin{bmatrix} \Theta_X & \Theta_{XY} \\ \Theta_{YX} & \Theta_Y \end{bmatrix}$$

- Inverse covariance matrix

$$\Sigma_X^{-1} = \Theta_X - \Theta_{XY}\Theta_Y^{-1}\Theta_{YX}$$

- Inverse covariance matrix = "sparse - low-rank"
- Convex formulation (Chandrasekaran et.al., 2010)

$$\begin{array}{ll} \min_{R,S,L} & \langle R, \hat{\Sigma}_X \rangle - \log\det(R) + \alpha\|S\|_1 + \beta \mathrm{Tr}(L) \\ \text{s.t.} & R = S - L, R \succ 0, L \succeq 0. \end{array}$$

Latent variable covariance estimation (LVCE)

- X: observed variables; Y: latent variables (hidden factors)
- (X, Y) jointly follow a multivariate Gaussian distribution.
- Covariance matrix and inverse covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \qquad \Theta = \begin{bmatrix} \Theta_X & \Theta_{XY} \\ \Theta_{YX} & \Theta_Y \end{bmatrix}$$

- Inverse covariance matrix

$$\Sigma_X^{-1} = \Theta_X - \Theta_{XY}\Theta_Y^{-1}\Theta_{YX}$$

- Inverse covariance matrix = "sparse - low-rank"
- Convex formulation (Chandrasekaran et.al., 2010)

$$\begin{array}{ll} \min_{R,S,L} & \langle R, \hat{\Sigma}_X \rangle - \log\det(R) + \alpha\|S\|_1 + \beta \mathrm{Tr}(L) \\ \text{s.t.} & R = S - L, R \succ 0, L \succeq 0. \end{array}$$

Composite Optimization

- SICE:

$$\min_{S} \ \langle \hat{\Sigma}, S \rangle - \log\det S + \rho \|S\|_1$$

- Composite Optimization:

$$\min_{x} \quad f(x) + g(x)$$

- Variable splitting

$$\begin{array}{ll} \min & f(x) + g(y) \\ \text{s.t.} & x - y = 0 \end{array}$$

- Augmented Lagrangian function:

$$\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu} \|x - y\|^2$$

Composite Optimization

- SICE:

$$\min_{S} \ \langle \hat{\Sigma}, S \rangle - \log\det S + \rho \|S\|_1$$

- Composite Optimization:

$$\min_{x} \quad f(x) + g(x)$$

- Variable splitting

$$\begin{array}{ll} \min & f(x) + g(y) \\ \text{s.t.} & x - y = 0 \end{array}$$

- Augmented Lagrangian function:

$$\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu} \|x - y\|^2$$

Alternating Direction Method of Multipliers (ADMM)

- $\min_x \ F(x) \equiv f(x) + g(x)$
- $\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu}\|x - y\|^2$
- ADMM

$$\begin{cases} x^{k+1} & := & \arg\min_x \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \arg\min_y \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (x^{k+1} - y^{k+1})/\mu \end{cases}$$

- Subproblems are easy: closed-form solutions
  - SICE: $f(X) = \langle \hat{\Sigma}, X \rangle - \log\det(X)$ and $g(Y) = \|Y\|_1$
  - 1st: $\min_X \ -\log\det(X) + \frac{1}{2\tau}\|X - Z\|_F^2$
  - Solution: $X^* := U\mathrm{diag}(\gamma)U^\top$, where $Z = U\mathrm{diag}(\sigma)U^\top$ and $\gamma_i = (\sigma_i + \sqrt{\sigma_i^2 + 4\tau})/2$
  - 2nd: $\min_Y \ \|Y\|_1 + \frac{1}{2\tau}\|Y - Z\|_F^2$
  - Solution: $Y^* := Sgn(Z) \circ \max\{|Z| - \tau, 0\}$ (ℓ_1 shrinkage)
- Convergence: globally convergent.

Alternating Direction Method of Multipliers (ADMM)

- $\min_x \ F(x) \equiv f(x) + g(x)$
- $\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu}\|x - y\|^2$
- ADMM

$$\begin{cases} x^{k+1} & := & \arg\min_x \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \arg\min_y \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (x^{k+1} - y^{k+1})/\mu \end{cases}$$

- Subproblems are easy: closed-form solutions
  - SICE: $f(X) = \langle \hat{\Sigma}, X \rangle - \log\det(X)$ and $g(Y) = \|Y\|_1$
  - 1st: $\min_X \ -\log\det(X) + \frac{1}{2\tau}\|X - Z\|_F^2$
  - Solution: $X^* := U\mathrm{diag}(\gamma)U^\top$, where $Z = U\mathrm{diag}(\sigma)U^\top$ and $\gamma_i = (\sigma_i + \sqrt{\sigma_i^2 + 4\tau})/2$
  - 2nd: $\min_Y \ \|Y\|_1 + \frac{1}{2\tau}\|Y - Z\|_F^2$
  - Solution: $Y^* := Sgn(Z) \circ \max\{|Z| - \tau, 0\}$ (ℓ_1 shrinkage)
- Convergence: globally convergent.

Alternating Direction Method of Multipliers (ADMM)

- $\min_x \ F(x) \equiv f(x) + g(x)$
- $\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu}\|x - y\|^2$
- ADMM

$$\begin{cases} x^{k+1} & := & \arg\min_x \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \arg\min_y \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (x^{k+1} - y^{k+1})/\mu \end{cases}$$

- Subproblems are easy: closed-form solutions
  - SICE: $f(X) = \langle \hat{\Sigma}, X \rangle - \log\det(X)$ and $g(Y) = \|Y\|_1$
  - 1st: $\min_X \ -\log\det(X) + \frac{1}{2\tau}\|X - Z\|_F^2$
  - Solution: $X^* := U\mathrm{diag}(\gamma)U^\top$, where $Z = U\mathrm{diag}(\sigma)U^\top$ and $\gamma_i = (\sigma_i + \sqrt{\sigma_i^2 + 4\tau})/2$
  - 2nd: $\min_Y \ \|Y\|_1 + \frac{1}{2\tau}\|Y - Z\|_F^2$
  - Solution: $Y^* := Sgn(Z) \circ \max\{|Z| - \tau, 0\}$ (ℓ_1 shrinkage)
- Convergence: globally convergent.

Alternating Direction Method of Multipliers (ADMM)

- $\min_x \ F(x) \equiv f(x) + g(x)$
- $\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu}\|x - y\|^2$
- ADMM

$$\begin{cases} x^{k+1} & := & \arg\min_x \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \arg\min_y \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (x^{k+1} - y^{k+1})/\mu \end{cases}$$

- Subproblems are easy: closed-form solutions
  - SICE: $f(X) = \langle \hat{\Sigma}, X \rangle - \log\det(X)$ and $g(Y) = \|Y\|_1$
  - 1st: $\min_X \ -\log\det(X) + \frac{1}{2\tau}\|X - Z\|_F^2$
  - Solution: $X^* := U\mathrm{diag}(\gamma)U^\top$, where $Z = U\mathrm{diag}(\sigma)U^\top$ and $\gamma_i = (\sigma_i + \sqrt{\sigma_i^2 + 4\tau})/2$
  - 2nd: $\min_Y \ \|Y\|_1 + \frac{1}{2\tau}\|Y - Z\|_F^2$
  - Solution: $Y^* := Sgn(Z) \circ \max\{|Z| - \tau, 0\}$ (ℓ_1 shrinkage)
- Convergence: globally convergent.

Alternating Direction Method of Multipliers (ADMM)

- $\min_x \ F(x) \equiv f(x) + g(x)$
- $\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu}\|x - y\|^2$
- ADMM

$$\begin{cases} x^{k+1} & := & \arg\min_x \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \arg\min_y \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (x^{k+1} - y^{k+1})/\mu \end{cases}$$

- Subproblems are easy: closed-form solutions
  - SICE: $f(X) = \langle \hat{\Sigma}, X \rangle - \log\det(X)$ and $g(Y) = \|Y\|_1$
  - 1st: $\min_X \ -\log\det(X) + \frac{1}{2\tau}\|X - Z\|_F^2$
  - Solution: $X^* := U\mathrm{diag}(\gamma)U^\top$, where $Z = U\mathrm{diag}(\sigma)U^\top$ and $\gamma_i = (\sigma_i + \sqrt{\sigma_i^2 + 4\tau})/2$
  - 2nd: $\min_Y \ \|Y\|_1 + \frac{1}{2\tau}\|Y - Z\|_F^2$
  - Solution: $Y^* := Sgn(Z) \circ \max\{|Z| - \tau, 0\}$ (ℓ_1 shrinkage)
- Convergence: globally convergent.

Alternating Direction Method of Multipliers (ADMM)

- $\min_x \ F(x) \equiv f(x) + g(x)$
- $\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu}\|x - y\|^2$
- ADMM

$$\begin{cases} x^{k+1} & := & \arg\min_x \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \arg\min_y \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (x^{k+1} - y^{k+1})/\mu \end{cases}$$

- Subproblems are easy: closed-form solutions
  - SICE: $f(X) = \langle \hat{\Sigma}, X \rangle - \log\det(X)$ and $g(Y) = \|Y\|_1$
  - 1st: $\min_X \ -\log\det(X) + \frac{1}{2\tau}\|X - Z\|_F^2$
  - Solution: $X^* := U\mathrm{diag}(\gamma)U^\top$, where $Z = U\mathrm{diag}(\sigma)U^\top$ and $\gamma_i = (\sigma_i + \sqrt{\sigma_i^2 + 4\tau})/2$
  - 2nd: $\min_Y \ \|Y\|_1 + \frac{1}{2\tau}\|Y - Z\|_F^2$
  - Solution: $Y^* := Sgn(Z) \circ \max\{|Z| - \tau, 0\}$ (ℓ_1 shrinkage)
- Convergence: globally convergent.

Alternating Direction Method of Multipliers (ADMM)

- $\min_x \ F(x) \equiv f(x) + g(x)$
- $\mathcal{L}(x, y; \lambda) := f(x) + g(y) - \langle \lambda, x - y \rangle + \frac{1}{2\mu}\|x - y\|^2$
- ADMM

$$\begin{cases} x^{k+1} & := & \arg\min_x \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \arg\min_y \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (x^{k+1} - y^{k+1})/\mu \end{cases}$$

- Subproblems are easy: closed-form solutions
  - SICE: $f(X) = \langle \hat{\Sigma}, X \rangle - \log\det(X)$ and $g(Y) = \|Y\|_1$
  - 1st: $\min_X \ -\log\det(X) + \frac{1}{2\tau}\|X - Z\|_F^2$
  - Solution: $X^* := U\mathrm{diag}(\gamma)U^\top$, where $Z = U\mathrm{diag}(\sigma)U^\top$ and $\gamma_i = (\sigma_i + \sqrt{\sigma_i^2 + 4\tau})/2$
  - 2nd: $\min_Y \ \|Y\|_1 + \frac{1}{2\tau}\|Y - Z\|_F^2$
  - Solution: $Y^* := Sgn(Z) \circ \max\{|Z| - \tau, 0\}$ (ℓ_1 shrinkage)
- Convergence: globally convergent.

A natural ADMM for latent variable covariance estimation

- Rewrite LVCE as

$$\begin{array}{ll} \min_{R,S,L} & \langle \hat{\Sigma}_X, R\rangle - \log\det R + \alpha\|S\|_1 + \beta \mathrm{Tr}(L) + \mathcal{I}(L \succeq 0) \\ \text{s.t.} & R - S + L = 0. \end{array}$$

- Augmented Lagrangian function

$$\begin{aligned} &\mathcal{L}_\mu(R,S,L;\Lambda) := \langle \hat{\Sigma}_X, R\rangle - \log\det R + \alpha\|S\|_1 \\ &+\beta \mathrm{Tr}(L) + \mathcal{I}(L \succeq 0) - \langle \Lambda, R - S + L\rangle + \tfrac{1}{2\mu}\|R - S + L\|_F^2. \end{aligned}$$

- ADMM with 3 blocks

$$\begin{cases} R^{k+1} & := & \mathrm{argmin}_R\ \mathcal{L}_\mu(R, S^k, L^k; \Lambda^k) \\ S^{k+1} & := & \mathrm{argmin}_S\ \mathcal{L}_\mu(R^{k+1}, S, L^k; \Lambda^k) \\ L^{k+1} & := & \mathrm{argmin}_L\ \mathcal{L}_\mu(R^{k+1}, S^{k+1}, L; \Lambda^k) \\ \Lambda^{k+1} & := & \Lambda^k - (R^{k+1} - S^{k+1} + L^{k+1})/\mu \end{cases}$$

- Convergence is unknown!

A natural ADMM for latent variable covariance estimation

- Rewrite LVCE as

$$\begin{array}{ll} \min_{R,S,L} & \langle \hat{\Sigma}_X, R\rangle - \log\det R + \alpha\|S\|_1 + \beta \mathrm{Tr}(L) + \mathcal{I}(L \succeq 0) \\ \text{s.t.} & R - S + L = 0. \end{array}$$

- Augmented Lagrangian function

$$\begin{aligned} &\mathcal{L}_\mu(R,S,L;\Lambda) := \langle \hat{\Sigma}_X, R\rangle - \log\det R + \alpha\|S\|_1 \\ &+\beta \mathrm{Tr}(L) + \mathcal{I}(L \succeq 0) - \langle \Lambda, R - S + L\rangle + \tfrac{1}{2\mu}\|R - S + L\|_F^2. \end{aligned}$$

- ADMM with 3 blocks

$$\begin{cases} R^{k+1} & := & \mathrm{argmin}_R\ \mathcal{L}_\mu(R, S^k, L^k; \Lambda^k) \\ S^{k+1} & := & \mathrm{argmin}_S\ \mathcal{L}_\mu(R^{k+1}, S, L^k; \Lambda^k) \\ L^{k+1} & := & \mathrm{argmin}_L\ \mathcal{L}_\mu(R^{k+1}, S^{k+1}, L; \Lambda^k) \\ \Lambda^{k+1} & := & \Lambda^k - (R^{k+1} - S^{k+1} + L^{k+1})/\mu \end{cases}$$

- Convergence is unknown!

ADMM for three blocks

- ADMM for two blocks
  - $\min_{x,y}\ f(x)+g(y), \quad \text{s.t.} \quad Ax+By=b.$

$$\begin{cases} x^{k+1} & := & \operatorname{argmin}_x\ \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \operatorname{argmin}_y\ \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (Ax^{k+1}+By^{k+1}-b)/\mu. \end{cases}$$

  - Global convergence
- ADMM for three blocks
  - $\min_{x,y,z}\ f(x)+g(y)+h(z), \quad \text{s.t.} \quad Ax+By+Cz=b.$

$$\begin{cases} x^{k+1} & := & \operatorname{argmin}_x\ \mathcal{L}(x, y^k, z^k; \lambda^k) \\ y^{k+1} & := & \operatorname{argmin}_y\ \mathcal{L}(x^{k+1}, y, z^{k+1}; \lambda^k) \\ z^{k+1} & := & \operatorname{argmin}_z\ \mathcal{L}(x^{k+1}, y^{k+1}, z; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (Ax^{k+1}+By^{k+1}+Cz^{k+1}-b)/\mu. \end{cases}$$

  - Douglas and Gunn (1964): A general formulation of alternating direction methods.
  - Lions and Mercier (1979): "... However, the convergence seems difficult to prove in this general framework ..."

ADMM for three blocks

- ADMM for two blocks
  - $\min_{x,y}\ f(x)+g(y), \quad \text{s.t.} \quad Ax+By=b.$

$$\begin{cases} x^{k+1} & := & \operatorname{argmin}_x\ \mathcal{L}(x, y^k; \lambda^k) \\ y^{k+1} & := & \operatorname{argmin}_y\ \mathcal{L}(x^{k+1}, y; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (Ax^{k+1}+By^{k+1}-b)/\mu. \end{cases}$$

  - Global convergence
- ADMM for three blocks
  - $\min_{x,y,z}\ f(x)+g(y)+h(z), \quad \text{s.t.} \quad Ax+By+Cz=b.$

$$\begin{cases} x^{k+1} & := & \operatorname{argmin}_x\ \mathcal{L}(x, y^k, z^k; \lambda^k) \\ y^{k+1} & := & \operatorname{argmin}_y\ \mathcal{L}(x^{k+1}, y, z^{k+1}; \lambda^k) \\ z^{k+1} & := & \operatorname{argmin}_z\ \mathcal{L}(x^{k+1}, y^{k+1}, z; \lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (Ax^{k+1}+By^{k+1}+Cz^{k+1}-b)/\mu. \end{cases}$$

  - Douglas and Gunn (1964): A general formulation of alternating direction methods.
  - Lions and Mercier (1979): "... However, the convergence seems difficult to prove in this general framework ..."

Alternating Proximal Gradient Method

$$(P) \qquad \begin{array}{ll} \min_{x,y,z} & f(x)+g(y)+h(z) \\ \text{s.t.} & Ax+By+Cz=b. \end{array}$$

$$\mathcal{L}(x,y,z;\lambda) := f(x)+g(y)+h(z)-\langle \lambda, Ax+By+Cz-b\rangle + \frac{1}{2\mu}\|Ax+By+Cz-b\|_2^2$$

Treat $w := (y,z)$. ADMM for two blocks

$$\begin{cases} x^{k+1} & := & \operatorname{argmin}_x \mathcal{L}(x,y^k,z^k;\lambda^k) \\ w^{k+1} := \begin{pmatrix} y^{k+1} \\ z^{k+1} \end{pmatrix} & := & \operatorname{argmin}_{y,z} \mathcal{L}(x^{k+1},y,z;\lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (Ax^{k+1}+By^{k+1}+Cz^{k+1}-b)/\mu. \end{cases}$$

ADMM based on proximal gradient steps

- Chen and Teboulle, 1994
- Yang and Zhang, 2011
- Tao and Yuan, 2011
- Yin, 2012

Alternating Proximal Gradient Method

$$(P) \qquad \begin{array}{ll} \min_{x,y,z} & f(x)+g(y)+h(z) \\ \text{s.t.} & Ax+By+Cz=b. \end{array}$$

$$\mathcal{L}(x,y,z;\lambda) := f(x)+g(y)+h(z)-\langle \lambda, Ax+By+Cz-b\rangle + \frac{1}{2\mu}\|Ax+By+Cz-b\|_2^2$$

Treat $w := (y,z)$. ADMM for two blocks

$$\begin{cases} x^{k+1} & := & \operatorname{argmin}_x \mathcal{L}(x,y^k,z^k;\lambda^k) \\ w^{k+1} := \begin{pmatrix} y^{k+1} \\ z^{k+1} \end{pmatrix} & := & \operatorname{argmin}_{y,z} \mathcal{L}(x^{k+1},y,z;\lambda^k) \\ \lambda^{k+1} & := & \lambda^k - (Ax^{k+1}+By^{k+1}+Cz^{k+1}-b)/\mu. \end{cases}$$

ADMM based on proximal gradient steps

- Chen and Teboulle, 1994
- Yang and Zhang, 2011
- Tao and Yuan, 2011
- Yin, 2012

Alternating Proximal Gradient Method

$$(P) \qquad \begin{array}{ll} \min_{x,y,z} & f(x)+g(y)+h(z) \\ \text{s.t.} & Ax+By+Cz=b. \end{array}$$

$$\mathcal{L}(x,y,z;\lambda) := f(x)+g(y)+h(z)-\langle \lambda, Ax+By+Cz-b\rangle+\frac{1}{2\mu}\|Ax+By+Cz-b\|_2^2$$

APGM:

$$\begin{cases} G_x^k & := Ax^k+By^k+Cz^k-b-\mu\lambda^k \\ x^{k+1} & := \operatorname{argmin}_x \ f(x)+\frac{1}{2\mu\tau_1}\|x-(x^k-\tau_1 A^\top G_x^k)\|_2^2 \\ G^k & := Ax^{k+1}+By^k+Cz^k-b-\mu\lambda^k \\ \begin{pmatrix} y^{k+1} \\ z^{k+1}\end{pmatrix} & := \operatorname{argmin}_{y,z} \ g(y)+h(z)+\frac{1}{2\mu\tau_2}\left\|\begin{pmatrix} y \\ z\end{pmatrix}-\left(\begin{pmatrix} y^k \\ z^k\end{pmatrix}-\tau_2\begin{pmatrix} B^\top \\ C^\top\end{pmatrix}G^k\right)\right\|_2^2 \\ \lambda^{k+1} & := \lambda^k-(Ax^{k+1}+By^{k+1}+Cz^{k+1}-b)/\mu \end{cases}$$

Theorem

APGM converges to an optimal solution of (P) from any starting point as long as $\tau_1 < 1/\lambda_{\max}(A^\top A)$ and $\tau_2 < 1/\lambda_{\max}([B,C]^\top[B,C])$.

Alternating Proximal Gradient Method

$$(P) \qquad \begin{array}{ll} \min_{x,y,z} & f(x)+g(y)+h(z) \\ \text{s.t.} & Ax+By+Cz=b. \end{array}$$

$$\mathcal{L}(x,y,z;\lambda) := f(x)+g(y)+h(z)-\langle \lambda, Ax+By+Cz-b\rangle+\frac{1}{2\mu}\|Ax+By+Cz-b\|_2^2$$

APGM:

$$\begin{cases} G_x^k & := Ax^k+By^k+Cz^k-b-\mu\lambda^k \\ x^{k+1} & := \operatorname{argmin}_x \ f(x)+\frac{1}{2\mu\tau_1}\|x-(x^k-\tau_1 A^\top G_x^k)\|_2^2 \\ G^k & := Ax^{k+1}+By^k+Cz^k-b-\mu\lambda^k \\ \begin{pmatrix} y^{k+1} \\ z^{k+1}\end{pmatrix} & := \operatorname{argmin}_{y,z} \ g(y)+h(z)+\frac{1}{2\mu\tau_2}\left\|\begin{pmatrix} y \\ z\end{pmatrix}-\left(\begin{pmatrix} y^k \\ z^k\end{pmatrix}-\tau_2\begin{pmatrix} B^\top \\ C^\top\end{pmatrix}G^k\right)\right\|_2^2 \\ \lambda^{k+1} & := \lambda^k-(Ax^{k+1}+By^{k+1}+Cz^{k+1}-b)/\mu \end{cases}$$

Theorem

APGM converges to an optimal solution of (P) from any starting point as long as $\tau_1 < 1/\lambda_{\max}(A^\top A)$ and $\tau_2 < 1/\lambda_{\max}([B,C]^\top[B,C])$.

APGM for LVCE

$$\mathcal{L}_\mu(R,S,L;\Lambda) := \langle \hat{\Sigma}_X, R\rangle - \log\det R + \alpha\|S\|_1 \\ +\beta \mathrm{Tr}(L) + \mathcal{I}(L \succeq 0) - \langle \Lambda, R-S+L\rangle + \frac{1}{2\mu}\|R-S+L\|_F^2.$$

- Treat $W = (S, L)$ as one big variable.

$$\begin{cases} R^{k+1} & := & \mathrm{argmin}_R\ \mathcal{L}_\mu(R, S^k, L^k; \Lambda^k) \\ W^{k+1} = (S^{k+1}, L^{k+1}) & := & \mathrm{argmin}_{S,L}\ \mathcal{L}_\mu(R^{k+1}, S, L; \Lambda^k) \\ \Lambda^{k+1} & := & \Lambda^k - (R^{k+1} - S^{k+1} + L^{k+1})/\mu \end{cases}$$

- Solve the 2nd subprob inexactly by a proximal gradient step.
- Gradient of the quadratic term: $G^k = A^\top(R^{k+1} + AW - \mu\Lambda^k)$ with $A = [-I, I]$
-

$$\begin{pmatrix} S^{k+1} \\ L^{k+1} \end{pmatrix} := \alpha\|S\|_1 + \beta\mathrm{Tr}(L) + \mathcal{I}(L \succeq 0) + \frac{1}{2\mu\tau}\left\| \begin{pmatrix} S \\ L \end{pmatrix} - \left(\begin{pmatrix} S^k \\ L^k \end{pmatrix} - \tau G^k \right) \right\|_F^2$$

APGM for LVCE

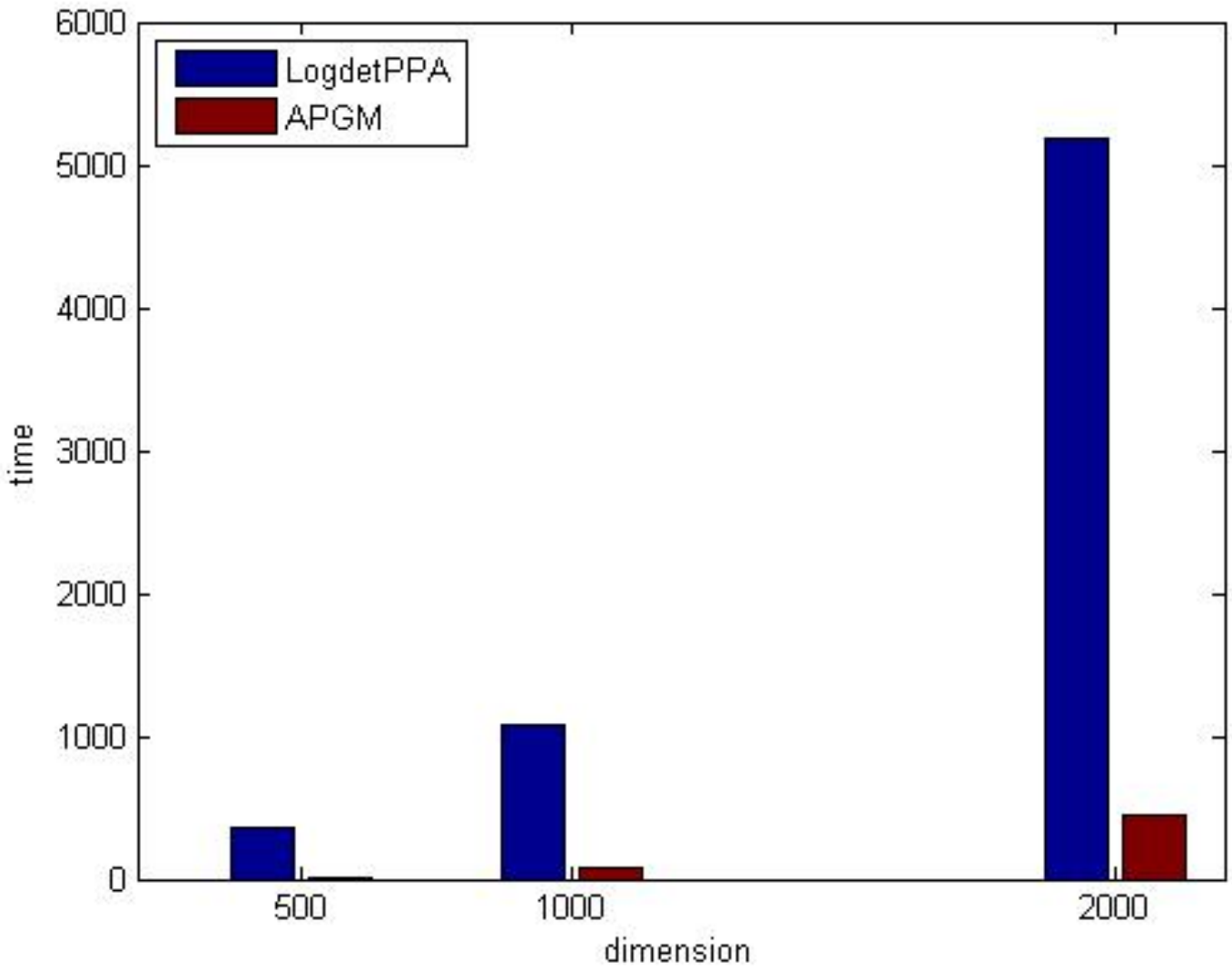
$$\begin{cases} R^{k+1} & := & \operatorname{argmin}_R \ \mathcal{L}_\mu(R, S^k, L^k; \Lambda^k) \\ \begin{pmatrix} S^{k+1} \\ L^{k+1} \end{pmatrix} & := & \alpha\|S\|_1 + \beta \mathrm{Tr}(L) + \mathcal{I}(L \succeq 0) + \frac{1}{2\mu\tau} \left\| \begin{pmatrix} S \\ L \end{pmatrix} - \left(\begin{pmatrix} S^k \\ L^k \end{pmatrix} - \tau G^k \right) \right\|_F^2 \\ \Lambda^{k+1} & := & \Lambda^k - (R^{k+1} - S^{k+1} + L^{k+1})/\mu \end{cases}$$

Theorem

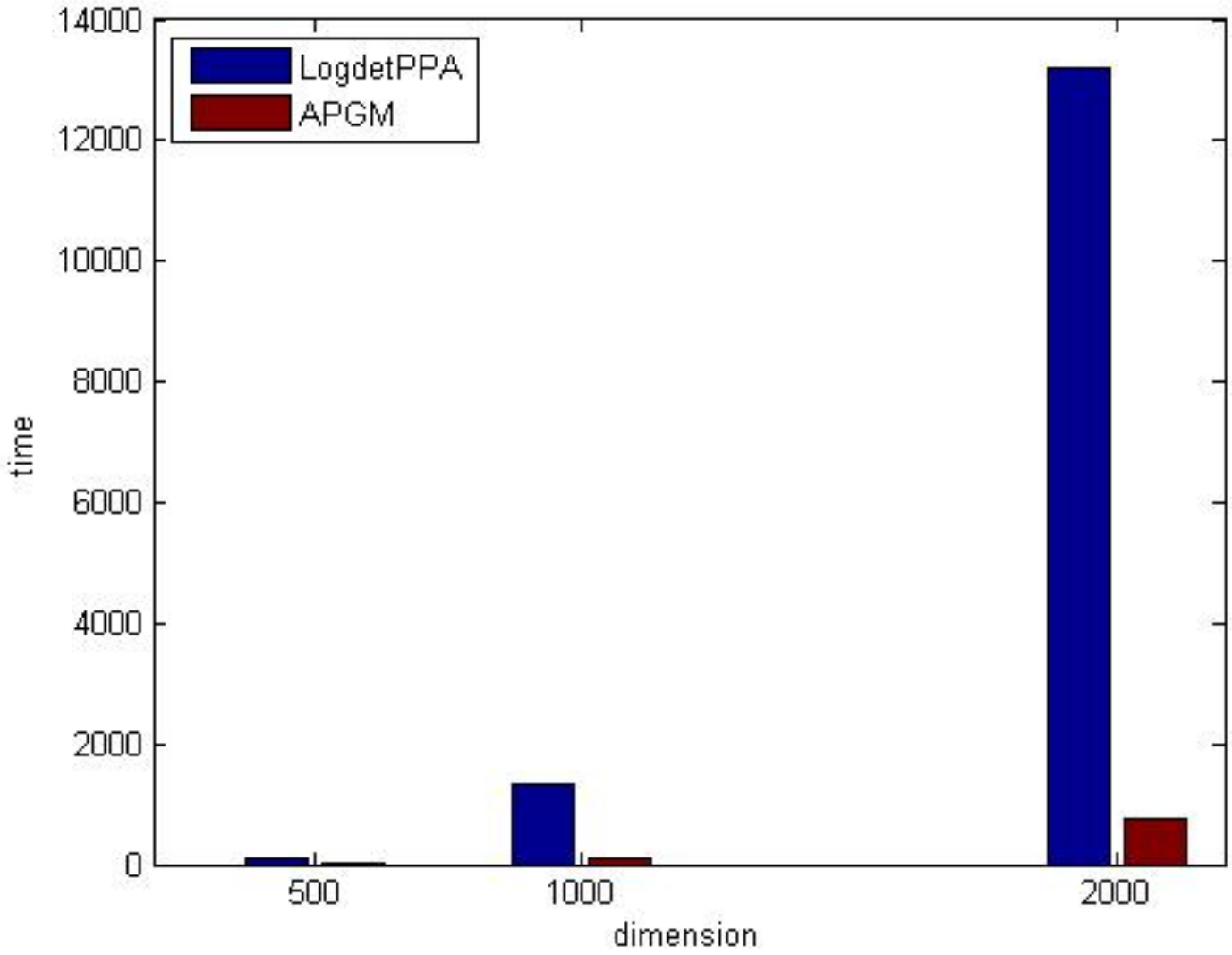
If $0 < \tau < 1/2$, then $(R^k, S^k, L^k, \Lambda^k)$ generated by APGM converges to the optimal solution from any starting point.

APGM for Latent Variable Covariance Estimation

- Compare with LogdetPPA: Wang, Sun and Toh (2009)

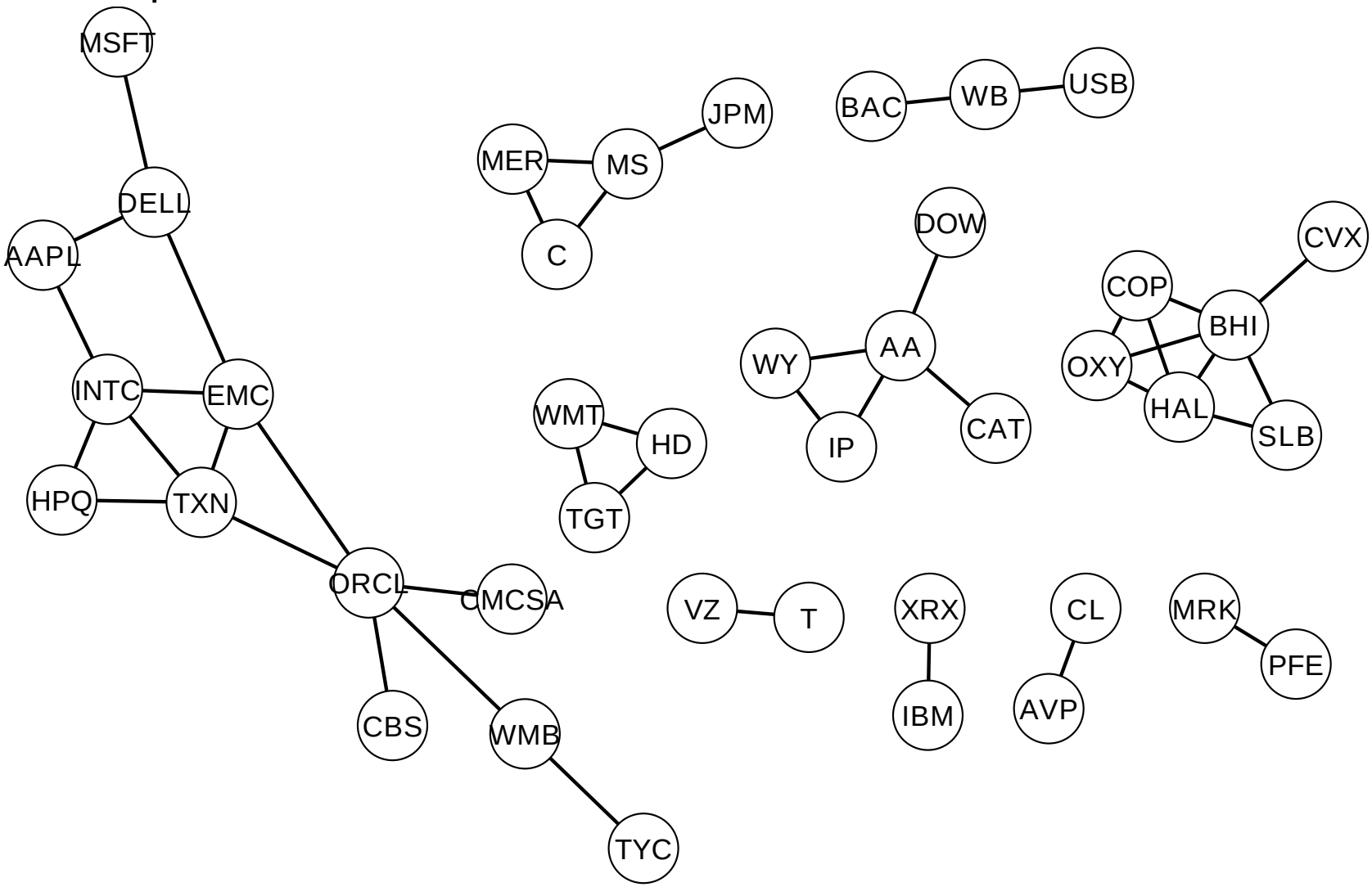


Gene Expression Data



Stock Return Data

- 84 companies from S&P 100 Index



Robust PCA: Background extraction

- Robust PCA: min $\{\|L\|_* + \rho\|S\|_1 : L + S = M\}$

- 200 frames with size 144×176, so $M \in \mathbb{R}^{25344 \times 200}$

Other Applications

- Stable PCP with Nonnegative Constraints (Candes et.al.2011)

$$\begin{array}{ll} \min_{L,S,Z} & \|L\|_* + \rho\|S\|_1 \\ \text{s.t.} & L + S + Z = M \\ & \|Z\|_F \leq \sigma \\ & L \geq 0. \end{array}$$

- General RPCA (Wright et.al.2012)

$$\min_{L,S} \quad \|L\|_* + \rho\|S\|_1, \quad \text{s.t.} \quad \mathcal{A}(L+S) = M.$$

- TILT and RASL (Ma et.al.2012)

$$\min_{x,L,S} \quad \|L\|_* + \rho\|S\|_1, \quad \text{s.t.} \quad D_i x = L + S.$$

- TV-based MRI deblurring with constraints

$$\min_{x} \quad TV(x) + \alpha\|Wx\|_1, \quad \text{s.t.} \quad \|Rx - b\|_2 \leq \sigma.$$

Other Applications

- Stable PCP with Nonnegative Constraints (Candes et.al.2011)

$$\begin{array}{ll} \min_{L,S,Z} & \|L\|_* + \rho\|S\|_1 \\ \text{s.t.} & L + S + Z = M \\ & \|Z\|_F \leq \sigma \\ & L \geq 0. \end{array}$$

- General RPCA (Wright et.al.2012)

$$\min_{L,S} \quad \|L\|_* + \rho\|S\|_1, \quad \text{s.t.} \quad \mathcal{A}(L+S) = M.$$

- TILT and RASL (Ma et.al.2012)

$$\min_{x,L,S} \quad \|L\|_* + \rho\|S\|_1, \quad \text{s.t.} \quad D_i x = L + S.$$

- TV-based MRI deblurring with constraints

$$\min_{x} \quad TV(x) + \alpha\|Wx\|_1, \quad \text{s.t.} \quad \|Rx - b\|_2 \leq \sigma.$$

Conclusion

- Problems with three or more blocks are still solvable by alternating direction type methods
- Alternating proximal gradient method
- Latent variable covariance estimation, robust PCA, TV-based MRI deblurring, ...

Thank you for your attention !