Sparse Optimization, Lecture 6

Zaiwen Wen (文再文)

Department of Mathematics and Institute of Natural Sciences
Shanghai Jiaotong University

July 20-21, 2011

Outline

- Matrix Completion
  - Simple Shrinkage based algorithm
  - Nesterov's type approach
  - Factorization model
- Sparse inverse covariance estimation
  - Block Coordinate method
  - Nesterov's smoothing technique

References

- Jianfeng Cai, Emmanuel Candes, Zuowei Shen, *Singular value thresholding algorithm for matrix completion*
- Shiqian Ma, Donald Goldfarb, Lifeng Chen, *Fixed point and Bregman iterative methods for matrix rank minimization*
- Zaiwen Wen, Wotao Yin, Yin Zhang, *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*
- Onureena Banerjee, Laurent El Ghaoui, Alexandre d'Aspremont, *Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data*
- Zhaosong Lu, *Smooth optimization approach for sparse covariance selection*

Matrix Rank Minimization

Given $X \in \mathbb{R}^{m \times n}$, $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^p$, $b \in \mathbb{R}^p$, we consider

- the matrix rank minimization problem:

$$\min \ \operatorname{rank}(X), \ \text{s.t. } \mathcal{A}(X) = b$$

- matrix completion problem:

$$\min \ \operatorname{rank}(X), \ \text{s.t. } X_{ij} = M_{ij}, (i,j) \in \Omega$$

- nuclear norm minimization:

$$\min \ \|X\|_* \ \text{ s.t. } \mathcal{A}(X) = b$$

where $\|X\|_* = \sum_i \sigma_i$ and $\sigma_i = i$th singular value of matrix X.

Recoverability results

- Recht, Fazel and Parrilo, 2007
- Candès and Recht, 2008
- (add more)

Quadratic penalty framework

- Unconstrained Nuclear Norm Minimization:

$$\min \; F(X) := \mu\|X\|_* + \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2.$$

- Optimality condition:

$$\mathbf{0} \in \mu\partial\|X^*\|_* + \mathcal{A}^*(\mathcal{A}(X^*) - b),$$

where $\partial\|X\|_* = \{UV^\top + W : U^\top W = 0, WV = 0, \|W\|_2 \leq 1\}$.

- Linearization approach (g is the gradient of $\frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$):

$$\begin{aligned} X^{k+1} \quad &:= \quad \arg\min_X \; \mu\|X\|_* + \left\langle g^k, X - X^k \right\rangle + \frac{1}{2\tau}\|X - X^k\|_F^2 \\ &= \quad \arg\min_X \; \mu\|X\|_* + \frac{1}{2\tau}\|X - (X^k - \tau g^k)\|_F^2 \end{aligned}$$

Matrix Shrinkage Operator

For a matrix $Y \in \mathbb{R}^{m \times n}$, consider:

$$\min_{X \in \mathbb{R}^{m \times n}} \ \nu \|X\|_* + \frac{1}{2}\|X - Y\|_F^2.$$

The optimal solution is:

$$X := S_\nu(Y) = U \mathrm{Diag}(s_\nu(\sigma)) V^\top,$$

- SVD: $Y = U \mathrm{Diag}(\sigma) V^\top$
- Thresholding operator:

$$s_\nu(x) := \bar{x}, \text{ with } \bar{x}_i = \begin{cases} x_i - \nu, & \text{if } x_i - \nu > 0 \\ 0, & \text{o.w.} \end{cases}$$

Fixed Point Method

Fixed Point Iterative Scheme

$$\begin{cases} Y^k = X^k - \tau\mathcal{A}^*(\mathcal{A}(X^k) - b) \\ X^{k+1} = S_{\tau\mu}(Y^k). \end{cases}$$

Lemma: Matrix shrinkage operator is non-expansive. i.e.,

$$\|S_\nu(Y_1) - S_\nu(Y_2)\|_F \le \|Y_1 - Y_2\|_F.$$

Theorem: The sequence $\{X^k\}$ generated by the fixed point iterations converges to some $X^* \in \mathcal{X}^*$, where $\mathcal{X}^*$ is the optimal solution set.

SVT

Linearized Bregman method:

$$
\begin{aligned}
V^{k+1} &:= V^k - \tau \mathcal{A}^*(\mathcal{A}(X^k) - b) \\
X^{k+1} &:= S_{\tau\mu}(V^{k+1})
\end{aligned}
$$

Convergence to

$$\min \ \tau \|X\|_* + \frac{1}{2}\|X\|_F^2, \text{ s.t. } \mathcal{A}(X) = b$$

Accelerated proximal gradient (APG) method

Complexity of the fixed point method:

$$F(X^k) - F(X^*) \le \frac{L_f \|X^0 - X^*\|^2}{2k}$$

APG algorithm ($t^{-1} = t^0 = 1$):

$$\begin{aligned} Y^k &= X^k + \frac{t^{k-1} - 1}{t^k}(X^k - X^{k-1}) \\ G^k &= Y^k - (\tau^k)^{-1}\mathcal{A}^*(\mathcal{A}(Y^k) - b) \\ X^{k+1} &= S_{\tau^k}(G^k), \quad t^{k+1} = \frac{1 + \sqrt{1 + 4(t^k)^2}}{2} \end{aligned}$$

Complexity:

$$F(X^k) - F(X^*) \le \frac{2L_f \|X^0 - X^*\|^2}{(k+1)^2}$$

Low-rank factorization model

- Finding a low-rank matrix W so that $\|\mathcal{P}_\Omega(W - M)\|_F^2$ or the distance between W and $\{Z \in \mathbb{R}^{m\times n}, Z_{ij} = M_{ij}, \forall (i,j) \in \Omega\}$ is minimized.
- Any matrix $W \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(W) \le K$ can be expressed as $W = XY$ where $X \in \mathbb{R}^{m\times K}$ and $Y \in \mathbb{R}^{K\times n}$.

New model

$$\min_{X,Y,Z} \frac{1}{2}\|XY - Z\|_F^2 \ \text{ s.t. } \ Z_{ij} = M_{ij}, \forall (i,j) \in \Omega$$

- Advantage: SVD is no longer needed!
- Related work: the solver `OptSpace` based on optimization on manifold

Nonlinear Gauss-Seideal scheme

First variant of alternating minimization:

$$\begin{aligned} X_+ &\leftarrow ZY^\dagger \equiv ZY^\top(YY^\top)^\dagger, \\ Y_+ &\leftarrow (X_+)^\dagger Z \equiv (X_+^\top X_+)^\dagger(X_+^\top Z), \\ Z_+ &\leftarrow X_+Y_+ + \mathcal{P}_\Omega(M - X_+Y_+). \end{aligned}$$

Let $\mathcal{P}_A$ be the orthogonal projection onto the range space $\mathcal{R}(A)$

- $X_+Y_+ = \left(X_+(X_+^\top X_+)^\dagger X_+^\top\right) Z = \mathcal{P}_{X_+}Z$
- One can verify that $\mathcal{R}(X_+) = \mathcal{R}(ZY^\top)$.
- $X_+Y_+ = \mathcal{P}_{ZY^\top}Z = ZY^\top(YZ^\top ZY^\top)^\dagger(YZ^\top)Z.$
- idea: modify X_+ or Y_+ to obtain the same product X_+Y_+

Nonlinear Gauss-Seideal scheme

Second variant of alternating minimization:

$$\begin{aligned} X_+ &\leftarrow ZY^\top, \\ Y_+ &\leftarrow (X_+)^\dagger Z \equiv (X_+^\top X_+)^\dagger (X_+^\top Z), \\ Z_+ &\leftarrow X_+ Y_+ + \mathcal{P}_\Omega(M - X_+ Y_+). \end{aligned}$$

Third variant of alternating minimization: $V = \mathtt{orth}(ZY^\top)$

$$\begin{aligned} X_+ &\leftarrow V, \\ Y_+ &\leftarrow V^\top Z, \\ Z_+ &\leftarrow X_+ Y_+ + \mathcal{P}_\Omega(M - X_+ Y_+). \end{aligned}$$

Nonlinear SOR

- The nonlinear GS scheme can be slow
- Linear SOR: applying extrapolation to the GS method to achieve faster convergence

The first implementation:

$$
\begin{aligned}
X_+ &\leftarrow ZY^\top(YY^\top)^\dagger, \\
X_+(\omega) &\leftarrow \omega X_+ + (1-\omega)X, \\
Y_+ &\leftarrow (X_+(\omega)^\top X_+(\omega))^\dagger (X_+(\omega)^\top Z), \\
Y_+(\omega) &\leftarrow \omega Y_+ + (1-\omega)Y, \\
Z_+(\omega) &\leftarrow X_+(\omega)Y_+(\omega) + \mathcal{P}_\Omega(M - X_+(\omega)Y_+(\omega)),
\end{aligned}
$$

Nonlinear SOR

- Let $S = \mathcal{P}_\Omega(M - XY)$. Then $Z = XY + S$
- Let $Z_\omega \triangleq XY + \omega S = \omega Z + (1-\omega)XY$
- Assume Y has full row rank, then

$$\begin{aligned} Z_\omega Y^\top (YY^\top)^\dagger &= \omega Z Y^\top (YY^\top)^\dagger + (1-\omega) XYY^\top (YY^\top)^\dagger \\ &= \omega X_+ + (1-\omega) X, \end{aligned}$$

Second implementation of our nonlinear SOR:

$$\begin{aligned} X_+(\omega) &\leftarrow Z_\omega Y^\top \text{ or } Z_\omega Y^\top (YY^\top)^\dagger, \\ Y_+(\omega) &\leftarrow (X_+(\omega)^\top X_+(\omega))^\dagger (X_+(\omega)^\top Z_\omega), \\ \mathcal{P}_{\Omega^c}(Z_+(\omega)) &\leftarrow \mathcal{P}_{\Omega^c}(X_+(\omega) Y_+(\omega)), \\ \mathcal{P}_\Omega(Z_+(\omega)) &\leftarrow \mathcal{P}_\Omega(M). \end{aligned}$$

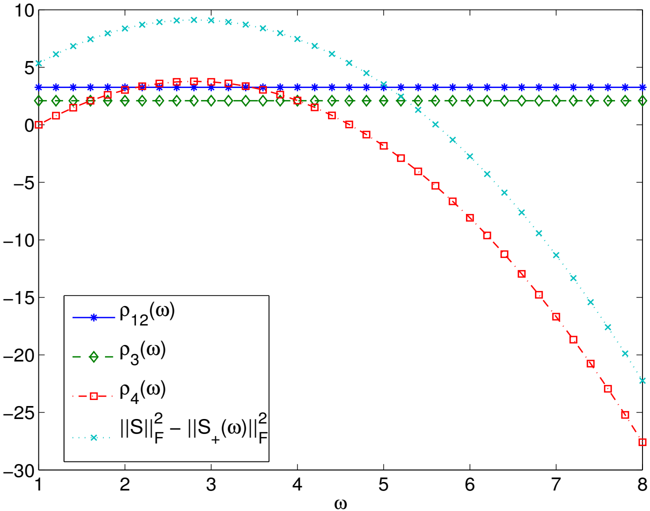
Reduction of the residual $\|S\|_F^2 - \|S_+(\omega)\|_F^2$

Assume that $\mathrm{rank}(Z_\omega) = \mathrm{rank}(Z), \forall\, \omega \in [1, \omega_1]$ for some $\omega_1 \geq 1$. Then there exists some $\omega_2 \geq 1$ such that

$$\|S\|_F^2 - \|S_+(\omega)\|_F^2 = \rho_{12}(\omega) + \rho_3(\omega) + \rho_4(\omega) > 0, \quad \forall\, \omega \in [1, \omega_2].$$

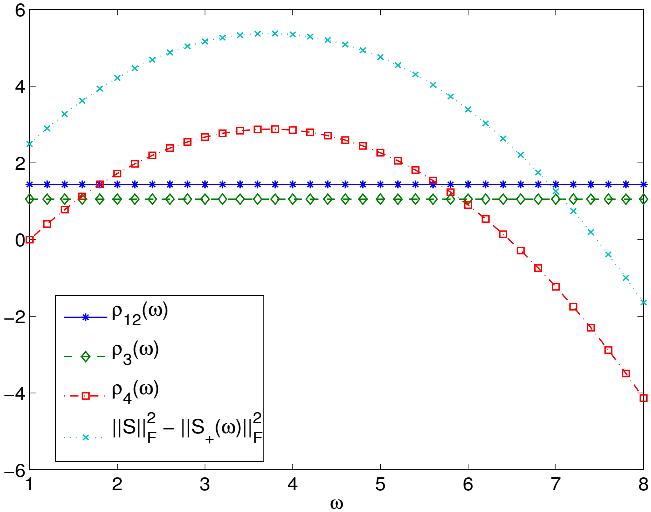
- $\rho_{12}(\omega) \triangleq \|SP\|_F^2 + \|Q(\omega)S(I-P)\|_F^2 \geq 0$
- $\rho_3(\omega) \triangleq \|\mathcal{P}_{\Omega^c}(SP + Q(\omega)S(I-P))\|_F^2 \geq 0$
- $\rho_4(\omega) \triangleq \frac{1}{\omega^2}\|S_+(\omega) + (\omega-1)S\|_F^2 - \|S_+(\omega)\|_F^2$
- Whenever $\rho_3(1) > 0$ ($\mathcal{P}_{\Omega^c}(X_+(1)Y_+(1) - XY) \neq 0$) and $\omega_1 > 1$, then $\omega_2 > 1$ can be chosen so that $\rho_4(\omega) > 0, \forall\, \omega \in (1, \omega_2]$.

Reduction of the residual $\|S\|_F^2 - \|S_+(\omega)\|_F^2$



(a)

Reduction of the residual $\|S\|_F^2 - \|S_+(\omega)\|_F^2$



(b)

Nonlinear SOR: convergence guarantee

Problem: how can we select a proper weight ω to ensure convergence for a nonlinear model?
Strategy: Adjust ω dynamically according to the change of the objective function values.

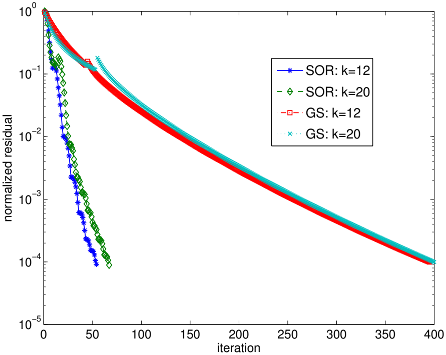
- Calculate the residual ratio $\gamma(\omega) = \frac{\|S_+(\omega)\|_F}{\|S\|_F}$
- A small $\gamma(\omega)$ indicates that the current weight value ω works well so far.
- If $\gamma(\omega) < 1$, accept the new point; otherwise, ω is reset to 1 and this procedure is repeated.
- ω is increased only if the calculated point is acceptable but the residual ratio $\gamma(\omega)$ is considered "too large"; that is, $\gamma(\omega) \in [\gamma_1, 1)$ for some $\gamma_1 \in (0, 1)$.

Nonlinear SOR: complete algorithm

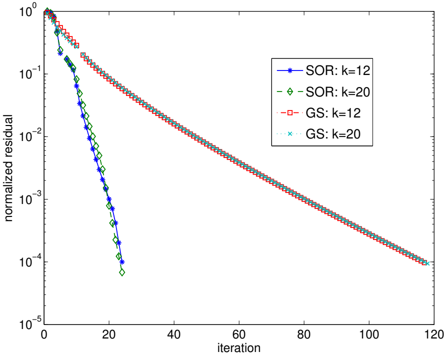
Algorithm 1: A low-rank matrix fitting algorithm (`LMaFit`)

1 Input index set Ω, data $\mathcal{P}_\Omega(M)$ and a rank overestimate $K \geq r$.
2 Set Y^0, Z^0, $\omega = 1$, $\tilde{\omega} > 1$, $\delta > 0$, $\gamma_1 \in (0, 1)$ and $k = 0$.
3 **while** *not convergent* **do**
4   Compute $(X_+(\omega), Y_+(\omega), Z_+(\omega))$.
5   Compute the residual ratio $\gamma(\omega)$.
6   **if** $\gamma(\omega) \geq 1$ **then** set $\omega = 1$ and go to step 4.
7   Update $(X^{k+1}, Y^{k+1}, Z^{k+1})$ and increment k.
8   **if** $\gamma(\omega) \geq \gamma_1$ **then**
9   set $\delta = \max(\delta, 0.25(\omega - 1))$ and $\omega = \min(\omega + \delta, \tilde{\omega})$.

nonlinear GS .vs. nonlinear SOR



(a) n=1000, r=10, SR = 0.08



(b) n=1000, r=10, SR=0.15

Sparse covariance selection (A. d'Aspremont)

We estimate a covariance matrix Σ from empirical data

- Infer independence relationships between variables
- Given $m+1$ observations $x_i \in \mathbb{R}^n$ on n random variables, compute $S := \frac{1}{m}\sum_{i=1}^{m+1}(x_i - \bar{x})(x_i - \bar{x})$
- Choose a symmetric subset I of matrix coefficients and denote by J the complement
- Choose a covariance matrix $\hat{\Sigma}$ such that
  - $\hat{\Sigma}_{ij} = S_{ij}$ for all $(i,j) \in I$
  - $\hat{\Sigma}_{ij}^{-1} = 0$ for all $(i,j) \in J$
- Benefits: maximum entropy, maximum likelihood, existence and uniqueness
- Applications: Gene expression data, speech recogintion and finance

Maximum likelihood estimation

Consider estimation:

$$\max_{X \in S^n} \log\det X - \mathrm{Tr}(SX) - \rho\|X\|_0$$

Convex relaxations:

$$\max_{X \in S^n} \log\det X - \mathrm{Tr}(SX) - \rho\|X\|_1,$$

whose dual problem is:

$$\max\ \log\det W \quad \text{s.t. } \|W - S\|_\infty \leq \lambda$$

Block coordinate method

Given $W \succ 0$, we can partition W and S as

$$W = \begin{pmatrix} \xi & y^\top \\ y & B \end{pmatrix} \text{ and } S = \begin{pmatrix} \xi_S & y_S^\top \\ y_S & B_S \end{pmatrix},$$

Fix B and note that $\log\det W = \log(\xi - y^\top B^{-1} y) \det B$, then

$$\min_{[\xi;y]} \; y^\top B^{-1} y - \xi, \quad \text{s.t.} \quad \|[\xi; y] - [\xi_S; y_S]\|_\infty \le \lambda, \; \xi \ge 0.$$

- Set $\xi = \xi_S + \lambda$. (check first-order optimality)
- Update y by solving:

$$y := \arg\min_y \; y^\top B^{-1} y, \quad \text{s.t.} \quad \|y - y_S\|_\infty \le \lambda,$$

whose dual problem is $\min_x \; x^\top B x - y_S^\top x + \lambda \|x\|_1$, which is

$$x := \arg\min_x \left\| B^{\frac{1}{2}} x - \frac{1}{2} B^{-\frac{1}{2}} y_S \right\|_2^2 + \lambda \|x\|_1.$$

Relationship: $y = Bx$.

Zhaosong Lu (*smooth optimization approach for sparse covariance selection*) consider

$$\begin{aligned} \max \ & \log\det X - \mathrm{Tr}(SX) - \rho\|X\|_1 \\ \text{s.t.} \ & \mathcal{X} := \{X \in S^n : \beta I \succeq X \succeq \alpha I\}, \end{aligned}$$

which is equivalent to ($\mathcal{U} := \{U \in S^n : |U_{ij}| \le 1, \forall ij\}$)

$$\max_{X\in\mathcal{X}} \min_{U\in\mathcal{U}} \log\det X - \langle S + \rho U, X\rangle$$

Let $f(U) := \max_{X\in\mathcal{X}} \log\det X - \langle S + \rho U, X\rangle$

- $\log\det X$ is strongly concave on $\mathcal{X}$
- $f(U)$ is continuous differentiable
- $\nabla f(U)$ is Lipschitz cont. with $L = \rho\beta^2$

Therefore, APG can be applied to the dual problem

$$\min_{U\in\mathcal{U}} f(U)$$

Extension

Consider

$$\max_{x \in \mathcal{X}} \; g(x) := \min_{u \in \mathcal{U}} \phi(x, u)$$

Assume:

- $\phi(x, u)$ is a cont. fun. which is strictly concave in $x \in \mathcal{X}$ for every fixed $u \in \mathcal{U}$, and convex diff. in $u \in \mathcal{U}$ for every fixed $x \in \mathcal{X}$. Then $f(u) := \max_{x \in \mathcal{X}} \phi(x, u)$ is diff.
- $\nabla f(u)$ is Lipschitz cont.

Then

- the primal and the dual $\min_{u \in U} f(u)$ are both solvable and have the same optimal value;
- Nesterov's smooth minimization approach can be applied to the dual

Nesterov's smoothing technique

Consider

$$\max_{x \in \mathcal{X}} \min_{u \in \mathcal{U}} \phi(x, u)$$

Question: What if the assumptions do not hold?

- Add a strictly convex function $\mu d(u)$ to the obj. fun.

$$g(u) := \arg\min_{u \in \mathcal{U}} \phi(x, u) + \mu d(u)$$

- $g(u)$ is differentiable
- Apply Nesterov's smooth minimization
- Complexity of finding a ϵ-suboptimal point: $O(\frac{1}{\epsilon})$ iterations
- Other smooth technique?