Fundamentals of Mathematical and Computing Sciences:
Applied Mathematical Science

Part III: Low rank matrix estimation (Lecture 3) Computational methods

2014-7-11

Taiji Suzuki (Room W707, post W8-46)
e-mail: suzuki.t.ct@m.titech.ac.jp

Today's topics:

- ADMM (Alternating Direction Method of Multiplier) for trace norm regularization.
- Gibbs sampling for Bayes estimator.

1 ADMM (Alternating Direction Method of Multiplier)

1.1 Procedure of ADMM

We want to solve the following problem:

$$\min_{A\in\mathbb{R}^{M\times N}} \quad \|Y-\mathcal{X}(A)\|^2 + C\|A\|_{\mathrm{Tr}}, \tag{1}$$

where $Y=[y_1,\dots,y_n]^\top\in\mathbb{R}^n$, $\mathcal{X}(A)=(\langle X_i,A\rangle)_{i=1}^n\mathbb{R}^n$.

ADMM is a method to solve the following linearly constrained optimization problem [4, 2]:

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}^{n'}} \quad f(x)+g(y) \tag{2a}$$

$$\text{s.t.} \quad B_1x+B_2y=b. \tag{2b}$$

The optimization problem (1) can be rewritten as

$$\min_{A,A'\in\mathbb{R}^{M\times N}} \quad \|Y-\mathcal{X}(A)\|^2 + C\|A'\|_{\mathrm{Tr}} \tag{3a}$$

$$\text{s.t.} \quad A-A'=O, \tag{3b}$$

which is a special form of the problem (2).

Consider the **augmented Lagrangian** defined as

$$\mathcal{L}(x,y,\lambda)=f(x)+g(y)-\lambda^\top(B_1x+B_2y-b)+\frac{\rho}{2}\|B_1x+B_2y-b\|^2.$$

Here, ρ is a parameter (usually $\rho=1$ is chosen). Then, ADMM procedure is given as follows:

ADMM:
Initialize y^0, λ^0. For $k = 0, 1, 2, \ldots$

$$\begin{aligned} x^{k+1} &= \arg\min_x \mathcal{L}(x, y^k, \lambda^k) \\ y^{k+1} &= \arg\min_y \mathcal{L}(x^{k+1}, y, \lambda^k) \\ \lambda^{k+1} &= \lambda_k - \rho(B_1 x + B_2 y - b) \end{aligned}$$

If the update of (x, y) is replaced by the joint minimizer $(x^{k+1}, y^{k+1}) = \arg\min_{x,y} \mathcal{L}(x, y, \lambda^k)$, then it gives **Hestens-Powell's multiplier method** [6, 8, 9]. ADMM minimizes the augmented Lagrangian $\mathcal{L}$ with respect to x and y alternatively.

1.2 Intuition of ADMM

If the optimization problem (2) is solvable, the optimal solution (x^*, y^*, λ^*) satisfies

$$f(x^*) + g(y^*) = \max_\lambda \min_{x,y} \mathcal{L}(x, y, \lambda).$$

Now let λ^* be the optimal dual variable of the RHS, then

$$\begin{aligned} &\max_\lambda \min_{x,y} \mathcal{L}(x, y, \lambda) \\ &= f(x^*) + g(y^*) - \lambda^{*\top}(B_1 x^* + B_2 y^* - b) + \frac{\rho}{2}\|B_1 x^* + B_2 y^* - b\|^2 \\ &(= f(x^*) + g(y^*) \quad (\because B_1 x^* + B_2 y^* - b = 0)). \end{aligned}$$

Since (x^*, y^*) minimizes $\mathcal{L}(x, y, \lambda^*)$, by (sub-)differentiating $\mathcal{L}$, we have

$$\begin{aligned} \nabla f(x^*) - B_1^\top \lambda^* + \rho B_1^\top (B_1 x^* + B_2 y^* - b) = 0, \\ \nabla g(y^*) - B_2^\top \lambda^* + \rho B_2^\top (B_1 x^* + B_2 y^* - b) = 0. \end{aligned}$$

Since $B_1 x^* + B_2 y^* - b = 0$, this gives the following **KKT condition**:

$$\begin{aligned} \nabla f(x^*) - B_1^\top \lambda^* = 0, \\ \nabla g(y^*) - B_2^\top \lambda^* = 0. \end{aligned}$$

On the other hand, $y_{k+1} = \arg\min_y \mathcal{L}(x^{k+1}, y, \lambda^k)$ indicates

$$\begin{aligned} &\nabla g(y^{k+1}) - B_2^\top \lambda^k + \rho B_2^\top (B_1 x^{k+1} + B_2 y^{k+1} - b) = 0 \\ \Rightarrow &\nabla g(y^{k+1}) - B_2^\top \lambda^{k+1} = 0. \end{aligned}$$

Therefore λ_{k+1} is updated so that the optimality condition is satisfied.

1.3 Computation of ADMM update

Let $\mathcal{X}^*$ be the conjugate of the linear operator $\mathcal{X}$, that is, $\mathcal{X}^* : \mathbb{R}^n \to \mathbb{R}^{M \times N}$ that satisfies $\langle Y, \mathcal{X}(A) \rangle = \langle \mathcal{X}^*(Y), A \rangle$ for all $A \in \mathbb{R}^{M \times N}$. Then the update of A is given by

$$A^{k+1} = \left(\mathcal{X}^* \mathcal{X} + \frac{\rho I}{2} \right)^{-1} \left(\frac{\lambda^k}{2} + \mathcal{X}^*(Y) + \rho A'^k \right).$$

The update of A' is given by

$$A'^{k+1} = \underset{A' \in \mathbb{R}^{M \times N}}{\arg\min} \left\{ \frac{C}{\rho} + \frac{1}{2} \|A' - (A^{k+1} - \frac{\lambda^k}{\rho})\|_F^2 \right\}.$$

This is a special case of so called **proximal mapping**. The proximal mapping associated with a convex function ψ is defined by

$$\mathrm{prox}(q|\psi) = \underset{x}{\arg\min} \left(\psi(x) + \frac{1}{2}\|x - q\|^2 \right).$$

Here let

$$\mathrm{ST}_{C'}(\sigma) = \max\{\sigma - C', 0\},$$

for $\sigma > 0,\ C' > 0$. Then the update of A' is explicitly given as follows.

Lemma 1. *Let $Q^k := A^{k+1} - \frac{\lambda^k}{\rho}$ and its SVD be $Q^k = U\mathrm{Diag}(\sigma_1, \ldots, \sigma_p)V^\top$. Then*

$$A'^{k+1} = U \begin{pmatrix} \mathrm{ST}_{\frac{C}{\rho}}(\sigma_1) & & \\ & \ddots & \\ & & \mathrm{ST}_{\frac{C}{\rho}}(\sigma_p) \end{pmatrix} V^\top,$$

In summary, the ADMM procedure for the trace norm regularization problem (3) is given as follows.

ADMM for trace norm regularization:

$$A^{k+1} = \left(\mathcal{X}^*\mathcal{X} + \frac{\rho I}{2} \right)^{-1} \left(\frac{\lambda^k}{2} + \mathcal{X}^*(Y) + \rho A'^k \right),$$
$$A'^{k+1} = U \begin{pmatrix} \mathrm{ST}_{\frac{C}{\rho}}(\sigma_1) & & \\ & \vdots & \\ & & \mathrm{ST}_{\frac{C}{\rho}}(\sigma_p) \end{pmatrix} V^\top,$$
$$\lambda^{k+1} = \lambda^k - \rho(A^{k+1} - A'^{k+1}).$$

Note that ADMM can be applied to many other regularized sparse estimation problems.

1.4 Convergence of ADMM

Theorem 2. *If B_1 and B_2 have full column rank and the optimization problem (2) is solvable, then there exists an optimal variable (x^*, y^*) such that*

$$f(x^k) + g(y^k) \to f(x^*) + g(y^*)$$
$$(x^k, y^k) \to (x^*, y^*).$$

Theorem 3 (Linear convergence of ADMM)**.** *If g is strongly convex, ∇g is Lipschitz continuous, B_1 has full column rank, B_2 has full row rank, and there exists (x^*, y^*, λ^*) satisfying*

the KKT condition, then $\exists C_0,\ \delta > 0$ *such that*

$$\left\| \begin{bmatrix} x^k \\ y^k \\ \lambda^k \end{bmatrix} - \begin{bmatrix} x^* \\ y^* \\ \lambda^* \end{bmatrix} \right\| \leq (1-\delta)^k C_0.$$

[7] gives a proof of Theorem 2, [5] showed $O(1/k)$ convergence of ADMM, and [3] showed the linear convergence (Theorem 3).

2 Bayes estimator

Assume that the rank of A^* is known, say d. We consider the prior distribution of U, V for $A = UV^\top$ which is given by

$$\pi(U, V|d) = \prod_{i=1}^{M} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{\mathrm{p}}^2}} \exp\left(-\frac{U_{i,j}^2}{2\sigma_{\mathrm{p}}^2}\right) \times \prod_{i=1}^{N} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{\mathrm{p}}^2}} \exp\left(-\frac{V_{i,j}^2}{2\sigma_{\mathrm{p}}^2}\right).$$

Q: How to sample A from the posterior distribution?
A: Gibbs sampling.

Let $D_n : \{(X_i, Y_i)\}_{i=1}^n$. Here we assume that the noise $\{\epsilon_i\}_{i=1}^n$ is i.i.d. Gaussian $N(0, \sigma^2)$. Then the likelihood function is given by

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \langle X_i, A\rangle)^2}{2\sigma^2}\right).$$

Basically the Gibbs sampling is given by the following procedure.

Gibbs sampling: Iterate

$$U^{k+1} \sim \pi(U|V^k, D_n)$$
$$V^{k+1} \sim \pi(V|U^{k+1}, D_n)$$

Let vec be the vectorization of a matrix. Then the posterior distribution of U conditioned by V is given by

$$\pi(U|V, D_n) \simeq \exp\left(-\frac{1}{2}\|\mathrm{vec}(U) - G_V^{-1} q_V\|_{G_V}^2\right),$$

where $q_V = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{vec}(Y_i V X_i)$ and $G_V = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{vec}(VX_i)\mathrm{vec}(VX_i)^\top + \frac{I}{\sigma_{\mathrm{p}}^2}$. This is the density function of $N(G_V^{-1}q_V, G_V^{-1})$.

Similarly, we have

$$\pi(V|U, D_n) \simeq \exp\left(-\frac{1}{2}\|\mathrm{vec}(V) - H_U^{-1} r_U\|_{H_U}^2\right),$$

where $r_U = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{vec}(Y_i U X_i^\top)$ and $H_U = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{vec}(UX_i^\top)\mathrm{vec}(UX_i^\top)^\top + \frac{I}{\sigma_{\mathrm{p}}^2}$. This is the density function of $N(H_U^{-1}r_U, H_U^{-1})$.

Important point: We employed the Gaussian distribution as the prior distribution. Then the (conditional) posterior distribution is also Gaussian distribution. Prior distributions which give posterior distributions in the same class as the prior are called **conjugate prior**.

Therefore the Gibbs sampling procedure is summarized as follows.

Gibbs sampling for low rank matrix estimation: Iterate

$$\mathrm{vec}(U^{k+1}) \sim N(G_{V^k}^{-1} q_{V^k}, G_{V^k}^{-1})$$
$$\mathrm{vec}(V^{k+1}) \sim N(H_{U^{k+1}}^{-1} r_{U^{k+1}}, H_{U^{k+1}}^{-1})$$

It is known that the distribution of (U_k, V_k) converges to the posterior $\pi(U, V|D_n)$.

Based on Gibbs sampling, the posterior mean is estimated by

$$\hat{A} = \frac{1}{K - K_0 + 1} \sum_{k=K_0}^{K} U^k V^{k\top},$$

for sufficiently large $K_0 \ll K$.

For more information about Bayes estimator, see [1].

References

[1] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos. Sparse bayesian methods for low-rank matrix. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 60(8), 2012.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2010.

[3] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. Technical report, Rice University CAAM TR12-14, 2012.

[4] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Computers & Mathematics with Applications*, 2:17–40, 1976.

[5] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numerical Analisis*, 50(2):700–709, 2012.

[6] M. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory & Applications*, 4:303–320, 1969.

[7] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Püschel. A proof of convergence for the alternating direction method of multipliers applied to polyhedral-constrained functions. Technical report, 2011. arXiv:1112.2295.

[8] M. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, London, New York, 1969.

[9] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1:97–116, 1976.