Fundamentals of Mathematical and Computing Sciences:
Applied Mathematical Science

Part III: Low rank matrix estimation (Lecture 2) Estimation methods

2014-7-3

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

A basic idea to estimate the low rank matrix is given as follows:

$$\min_{A\in\mathbb{R}^{M\times N}} \quad \sum_{i=1}^{n}(y_i - \langle X_i, A\rangle)^2 \tag{1a}$$

$$\text{s.t.} \quad \mathrm{rank}(A) \le d. \tag{1b}$$

In this lecture, three approaches are introduced.

- Singular value thresholding
- Trace norm regularization
- Bayes estimator

1 Singular value thresholding

Singular value thresholding is the most simple method which can be used in the setting that all elements of A^*_{ij} are observed with observation noise. In that setting, Eq. (1) is reformulated as

$$\min_{A\in\mathbb{R}^{M\times N}} \quad \sum_{i=1}^{n}(Y_{ij} - A_{ij})^2, \tag{2a}$$

$$\text{s.t.} \quad \mathrm{rank}(A) \le d. \tag{2b}$$

Here, $Y_{ij} = A^*_{ij} + \epsilon_{ij}$ where ϵ_{ij} is observation noise. This problem can be solved analytically by using **singular value decomposition**.

Let $p = \min\{M, N\}$.

Theorem 1 (Singular Value Decomposition, SVD)**.** *For arbitrary $A \in \mathbb{R}^{M\times N}$, there exist orthonormal matrices $U \in \mathbb{R}^{M\times p}$ and $V \in \mathbb{R}^{N\times p}$ ($U^\top U = I$ and $V^\top V = I$), and a diagonal matrix $\Sigma \in \mathbb{R}^{p\times p}$, such that*

$$A = U\Sigma V^\top,$$

where $\Sigma \succeq O$.

This decomposition is called *Singular Value Decomposition (SVD)*, and the diagonal elements $\sigma_1, \sigma_2, \ldots, \sigma_p$ in Σ are called *singular values.*

A symmetric matrix can be diagonalized as follows.

Lemma 2. *For a real symmetric matrix $A \in \mathbb{R}^{M\times M}$, there exist an orthogonal matrix $U \in \mathbb{R}^M$ and a diagonal matrix $\Sigma \in \mathbb{R}^M$ such that*

$$A = U\Sigma U^\top.$$

Σ is not necessarily positive semi-definite. But, by setting $V^\top = \mathrm{Diag}(\mathrm{sign}(\sigma_1), \ldots, \mathrm{sign}(\sigma_p))U^\top$, we have SVD of A as $A = U|\Sigma|V^\top$.

Remark 3. *$A = U\Sigma U^\top$ (U is orthogonal, Σ is diagonal) if and only if A is normal, that is, $A^\top A = AA^\top$.*

Theorem 4. *Let $A, B \in \mathbb{R}^{M\times M}$ be symmetric matrices, and $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ be the Frobenius norm. If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M$ are the eigenvalues of A and $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_M$ are the eigenvalues of B, then*

$$\sum_{i=1}^{M} (\sigma_i - \gamma_i)^2 = \min_{\tau:\text{permutation}} \sum_{i=1}^{M} (\sigma_i - \gamma_{\tau(i)})^2 \leq \|A - B\|_F^2.$$

Proof. See Corollary 6.3.8 of [1] and its proof. □

We are ready to obtain the solution of the problem (2).

Lemma 5. *For arbitrary $A \in \mathbb{R}^{M\times N}$ with SVD $A = U\Sigma V^\top$, it holds that*

$$\begin{bmatrix} O & A \\ A^\top & O \end{bmatrix} = \begin{bmatrix} U & -U \\ V & V \end{bmatrix} \begin{bmatrix} \Sigma & O \\ O & -\Sigma \end{bmatrix} \begin{bmatrix} U & -U \\ V & V \end{bmatrix}^\top.$$

One can easily check that $\begin{bmatrix} U & -U \\ V & V \end{bmatrix}$ is an orthonormal matrix. Thus, the lemma shows that the eigenvalues of the symmetric matrix $\begin{bmatrix} O & A \\ A^\top & O \end{bmatrix}$ is given by $\sigma_1 \geq \cdots \geq \sigma_p \geq 0 = \cdots = 0 \geq -\sigma_p \geq \cdots \geq -\sigma_1$ where $\{\sigma_i\}$ are the singular values of A.

Theorem 6 (Low rank approximation of an arbitrary real matrix)**.** *Let $A \in \mathbb{R}^{M\times N}$ be an arbitrary real matrix. Then the minimum of*

$$\min_{B\in\mathbb{R}^{M\times N}} \quad \|A - B\|_F^2, \quad \textit{s.t.}\ \ \mathrm{rank}(B) \leq d,$$

is attained by

$$B = U\mathrm{Diag}(\sigma_1, \ldots, \sigma_d, 0, \ldots, 0)V^\top$$

where $A = U\mathrm{Diag}(\sigma_1, \ldots, \sigma_p)V^\top$ is the SVD of A. The optimal objective is given by $\sum_{j=d+1}^{p} \sigma_j^2$.

Proof. Note that

$$\|A - B\|_F^2 = \frac{1}{2} \left\| \begin{bmatrix} O & A \\ A^\top & O \end{bmatrix} - \begin{bmatrix} O & B \\ B^\top & O \end{bmatrix} \right\|_F^2.$$

By Theorem 4, the RHS is lower bounded by $\sum_{j=1}^{p} (\sigma_j - \gamma_j)^2$, where $\{\sigma_j\}$ and $\{\gamma_j\}$ are the singular values of A and B in decreasing order. This lower bound is minimized by $\gamma_j = \sigma_j$ ($j = 1, \ldots, d$) and $\gamma_j = 0$ ($j > d$) (note that $\mathrm{rank}(B)$ is at most d). This minimum objective is attained by $B = U\mathrm{Diag}(\sigma_1, \ldots, \sigma_d, 0, \ldots, 0)V^\top$. □

This theorem gives the solution of the problem (2):

$$\text{(Singular value thresholding)} \qquad \hat{A} = U\mathrm{Diag}(\sigma_1, \ldots, \sigma_d, 0, \ldots, 0)V^\top,$$

where $Y = U\mathrm{Diag}(\sigma_1, \ldots, \sigma_p)V^\top$ is SVD of Y.

Finally, the following corollary gives low rank approximation of a symmetric matrix.

Corollary 7 (Low rank approximation of a symmetric matrix)**.** *Let $A \in \mathbb{R}^{M\times M}$ be a symmetric matrix. Then the minimum of*

$$\min_{B\in\mathbb{R}^{M\times M}:\text{symmetric}} \quad \|A - B\|_F^2, \quad \textit{s.t.} \;\; \mathrm{rank}(B) \le d,$$

is attained by $B = U\mathrm{Diag}(\sigma_1, \ldots, \sigma_d, 0, \ldots, 0)U^\top$ where $\sigma_1, \ldots, \sigma_p$ are the eigenvalues of A such that $|\sigma_1| \ge |\sigma_2| \ge \cdots \ge |\sigma_p|$. The optimal objective is given by $\sum_{j=d+1}^p \sigma_j^2$.

2 Trace norm regularization

Singular value thresholding can be applied just a simple case. In general settings, the optimization problem can not be analytically solved. Moreover the problem is not convex.

The trace norm regularization technique gives a computationally tractable alternative of the problem (1). It is a convex relaxation of the original problem.

Trace norm regularization:

$$\min_{A\in\mathbb{R}^{M\times M}} \quad \|Y - \mathcal{X}(A)\|^2 \;\; \text{s.t.} \;\; \|A\|_{\mathrm{Tr}} \le C,$$

or

$$\min_{A\in\mathbb{R}^{M\times M}} \quad \|Y - \mathcal{X}(A)\|^2 + \lambda\|A\|_{\mathrm{Tr}}.$$

Here $\|A\|_{\mathrm{Tr}} = \mathrm{Tr}[(A^\top A)^{\frac{1}{2}}]$ is called *trace norm*. Note that

$$\begin{aligned}\|A\|_{\mathrm{Tr}} &= \mathrm{Tr}[(A^\top A)^{\frac{1}{2}}] = \mathrm{Tr}[(U\Sigma(A)V^\top V\Sigma(A)U^\top)^{\frac{1}{2}}] = \mathrm{Tr}[(U\Sigma(A)^2U^\top)^{\frac{1}{2}}] = \mathrm{Tr}[U\Sigma(A)U^\top] \\ &= \sum_{j=1}^p \sigma_j.\end{aligned}$$

$\star$ Trace norm is the sum of singular values.

Theorem 8.

- $\|cA\|_{\mathrm{Tr}} = |c|\|A\|_{\mathrm{Tr}} \quad (\forall c \in \mathbb{R})$,
- $\|A + B\|_{\mathrm{Tr}} \le \|A\|_{\mathrm{Tr}} + \|B\|_{\mathrm{Tr}}$,
- $\|A\|_{\mathrm{Tr}} = 0 \;\Leftrightarrow\; A = O$.

Proof. See Corollary 4.3.27 of [1]. □

This theorem says that trace "norm" is actually norm.

Remark 9. *Every orthogonal invariant norm, $\|A\|_M$ ($\|A\|_M = \|UAV\|_M$ for all orthogonal matrices U, V), satisfies*

$$\|A - B\|_M \ge \|\Sigma(A) - \Sigma(B)\|_M, \tag{5}$$

where $\Sigma(A)$ and $\Sigma(B)$ are diagonal matrices such that the singular values of A and B are on the diagonal elements in decreasing order (see Theorem 7.4.51 of [1]).

We have already seen that $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{Tr}}$ satisfy Eq. (5).

Q: Why trace norm?
A: Because it is the tight convex envelope of the rank function.

Theorem 10. *Trace norm is the tight convex envelope of the rank function in the set of* $\{A \in \mathbb{R}^{M\times N} \mid \|A\|_\infty \leq 1\}$*, where* $\|A\|_\infty$ *is the maximum singular value.*

Proof. Let $\Psi^* : \mathbb{R}^{M\times N} \to \mathbb{R}\cup\{\pm\infty\}$ be the convex conjugate of a function $\Psi : \mathbb{R}^{M\times N} \to \mathbb{R}\cup\{\pm\infty\}$, that is,

$$\Psi^*(Z) := \sup_{A\in\mathbb{R}^{M\times N}} \{\langle A, Z\rangle - \Psi(A)\}.$$

It is known that Ψ^{**} is the convex envelope of Ψ (Theorem 12.2 of [2]). By setting

$$\Psi(A) := \begin{cases} \|A\|_{\mathrm{Tr}} & (\|A\|_\infty \leq 1), \\ 0 & (\text{otherwise}), \end{cases}$$

we can check the assertion. □

By extending $\|\cdot\|_{\mathrm{Tr}}$ to outside of the box $\{A \in \mathbb{R}^{M\times N} \mid \|A\|_\infty \leq 1\}$, we can see that $\|\cdot\|_{\mathrm{Tr}}$ is a nice convex approximation of the rank function.

⋆ The solution of the trace norm regularized minimization problem is actually low rank (c.f. L_1-regularization).

3 Bayes estimator

- Construct a *prior distribution* $\pi(A)$.

↓
- Compute the likelihood of $D_n = \{(X_i, Y_i)\}_{i=1}^n$: $\prod_{i=1}^n p(Y_i|X_i, A)$.

↓
- Obtain the posterior distribution:

$$\pi(A|D_n) = \frac{\prod_{i=1}^n p(Y_i|X_i, A)\pi(A)}{\int \prod_{i=1}^n p(Y_i|X_i, A)\pi(A)\mathrm{d}A}.$$

The *posterior mean* is obtained by

$$\hat{A} = \int A\pi(A|D_n)\mathrm{d}A.$$

How to obtain the estimator based on the posterior distribution is determined by which loss are considered. More precisely, the Bayes estimator should minimize the Bayes risk:

$$\int \mathrm{E}_{D_n|A}[\ell(\delta(D_n), A)]\pi(A)\mathrm{d}A,$$

where $\delta(D_n)$ is an estimator constructed from the data D_n and ℓ is a loss function that measures how $\delta(D_n)$ is close to A (e.g., KL-divergence and Frobenius norm). The posterior mean corresponds to $\ell(\delta(D_n), A) = \|\delta(D_n) - A\|_F^2$.

The followings are examples of prior distributions of low rank matrices.

Let $0 < \xi < 1$ and $\sigma_{\mathrm{p}} > 0$ be hyper parameters.

$$
\begin{aligned}
d \sim \mathrm{Mult}(\pi(1), \ldots, \pi(p)) \text{ where } \pi(d) = \xi^d \left(\frac{1-\xi}{\xi - \xi^{p+1}} \right), \\
U_{i,j} | d \sim N(0, \sigma_{\mathrm{p}}^2) \quad (i = 1, \ldots, N,\ j = 1, \ldots, d), \\
V_{i,j} | d \sim N(0, \sigma_{\mathrm{p}}^2) \quad (i = 1, \ldots, M,\ j = 1, \ldots, d). \\
\text{Set} \quad A = UV^\top.
\end{aligned}
$$

Let $0 < a, b$ and $\sigma_{\mathrm{p}} > 0$ be hyper parameters.

$$
\begin{aligned}
\gamma_j \sim \Gamma(a, b) \quad (j = 1, \ldots, p), \\
U_{i,j} \sim N(0, \sigma_{\mathrm{p}}^2) \quad (i = 1, \ldots, N,\ j = 1, \ldots, p), \\
V_{i,j} \sim N(0, \sigma_{\mathrm{p}}^2) \quad (i = 1, \ldots, M,\ j = 1, \ldots, p). \\
\text{Set} \quad A = U \begin{pmatrix} \gamma_1^{-1} & & \\ & \ddots & \\ & & \gamma_p^{-1} \end{pmatrix} V^\top.
\end{aligned}
$$

References

[1] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge University Press, New York, 1985.

[2] G. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.