Factor Group-Sparse Regularization for Efficient Low-Rank Matrix Recovery

Jicong Fan, Lijun Ding, Yudong Chen, and Madeleine Udell
Cornell University

Low-rank matrix completion (LRMC)

Recover a low-rank $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ from partial observation $\mathcal{P}_\Omega(\boldsymbol{M})$.

(a) Low-rank matrix

(b) Matrix completion[1]

Applications

- collaborative filtering (recommendation system)
- classification (especially on incomplete data)
- image inpainting
- other data pre-processing tasks

Rank minimization and Schatten-p "norm" surrogate

1. Rank Minimization:

$$\underset{\boldsymbol{X}}{\text{minimize}}\ \text{rank}(\boldsymbol{X}),\ \text{subject to}\ \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M})$$

2. Need certain surrogate $R(\boldsymbol{X})$ of rank$(\boldsymbol{X})$:

$$\underset{\boldsymbol{X}}{\text{minimize}}\ R(\boldsymbol{X}),\ \text{subject to}\ \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}).$$

3. Schatten-p "norm", $0 < p \leqslant 1$:

$$R(\boldsymbol{X}) = \|\boldsymbol{X}\|_{S_p} := \Big(\sum_{i=1}^{\min(m,n)} \sigma_i^p(\boldsymbol{X})\Big)^{1/p}.$$

1. **Approximation advantage**: smaller value p leads to better approximation of rank$(\boldsymbol{X})$
2. **Computational challenge**: full SVD computation

Factorization Approaches for $\|\boldsymbol{X}\|_{S_p}$

Factorize $\boldsymbol{X} = \boldsymbol{AB}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times d}$, $\boldsymbol{B} \in \mathbb{R}^{d\times n}$

1. Factored nuclear (F-nuclear) norm[1]: $\|\boldsymbol{X}\|_* = \min_{\boldsymbol{AB}=\boldsymbol{X}} \frac{1}{2}\big(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2\big)$
2. Bi-nuclear norm[2]: $\|\boldsymbol{X}\|_{S_{1/2}} = \min_{\boldsymbol{AB}=\boldsymbol{X}} \|\boldsymbol{A}\|_*\|\boldsymbol{B}\|_* = \min_{\boldsymbol{AB}=\boldsymbol{X}} \big(\frac{\|\boldsymbol{A}\|_*+\|\boldsymbol{B}\|_*}{2}\big)^2$
3. F^2+nuclear norm[3]: $\|\boldsymbol{X}\|_{S_{\frac{2}{3}}^2} = \min_{\boldsymbol{AB}=\boldsymbol{X}} \Big[\frac{2}{3}\big(\|\boldsymbol{A}(\|_* + \frac{1}{2}\|\boldsymbol{B})\|_F^2\big)\Big]^{3/2}$

Problematic for larger d

1. Full rankness of factors $\boldsymbol{A}, \boldsymbol{B}$ during optimization procedures
2. (Possible) SVD computation of factors $\boldsymbol{A}, \boldsymbol{B}$

Notations

We factor $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ as $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_d] \in \mathbb{R}^{m\times d}$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_d]^T \in \mathbb{R}^{d\times n}$, where $d \geqslant r := \text{rank}(\boldsymbol{X})$, and $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$ are column vectors. WLOG, we assume $m \leqslant n$.

Our Methods: (i) group-sparse regularizers (FGSR)

We suggest using the following surrogates $R(\boldsymbol{X})$:

$$\text{FGSR}_{1/2}(\boldsymbol{X}) := \frac{1}{2}\min_{\boldsymbol{AB}=\boldsymbol{X}} \|\boldsymbol{A}\|_{2,1} + \|\boldsymbol{B}^T\|_{2,1}, \quad \text{FGSR}_{2/3}(\boldsymbol{X}) := \frac{2}{3\alpha^{1/3}}\min_{\boldsymbol{AB}=\boldsymbol{X}} \|\boldsymbol{A}\|_{2,1} + \frac{\alpha}{2}\|\boldsymbol{B}\|_F^2,$$

where $\|\boldsymbol{A}\|_{2,1} := \sum_{j=1}^d \|\boldsymbol{a}_j\|$.

Theorem 1 (Relation to $\|\boldsymbol{X}\|_{S_p}$)

$$\text{FGSR}_{1/2}(\boldsymbol{X}) = \sum_{j=1}^r \sigma_j^{1/2}(\boldsymbol{X}) = \|\boldsymbol{X}\|_{S_{1/2}}^{1/2}, \qquad \text{FGSR}_{2/3}(\boldsymbol{X}) = \sum_{j=1}^r \sigma_j^{2/3}(\boldsymbol{X}) = \|\boldsymbol{X}\|_{S_{2/3}}^{2/3}$$

In general, we have

Theorem 2 Fix $\alpha > 0$, and choose $q \in \{1, \frac{1}{2}, \frac{1}{4}, \cdots\}$. Set $p = \frac{2q}{(2+q)}$. For any matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ with rank$(\boldsymbol{X}) = r \leqslant d \leqslant \min(m,n)$, we have

$$\text{FGSR}_p(\boldsymbol{X}) := \frac{1}{(1/2+1/q)\alpha^{q/(q+2)}} \min_{\boldsymbol{X}=\sum_{j=1}^d \boldsymbol{a}_j\boldsymbol{b}_j^T} \sum_{j=1}^d \frac{1}{q}\|\boldsymbol{a}_j\|^q + \frac{\alpha}{2}\|\boldsymbol{b}_j\|^2 = \|\boldsymbol{X}\|_{S_p}^p. \tag{1}$$

Our Methods: (ii) optimization templates and algorithms

LRMC

- Optimization template:

$$\underset{\boldsymbol{X}}{\text{minimize}}\ \text{FGSR}_p(\boldsymbol{X}), \quad \text{subject to}\ P_\Omega(\boldsymbol{X}) = P_\Omega(\boldsymbol{M}). \tag{2}$$

Take $\text{FGSR}_{2/3}$ as an example, and rewrite (2) as

$$\underset{\boldsymbol{X},\boldsymbol{A},\boldsymbol{B}}{\text{minimize}}\ \|\boldsymbol{A}\|_{2,1} + \frac{\alpha}{2}\|\boldsymbol{B}\|_F^2, \quad \text{subject to}\ \boldsymbol{X} = \boldsymbol{AB},\ P_\Omega(\boldsymbol{X}) = P_\Omega(\boldsymbol{M}).$$

- Algorithm: Alternating Direction Method of Multipliers (ADMM) [4] with linearization.

Noisy LRMC

Observe $P_\Omega(\boldsymbol{M}_e)$ with $\boldsymbol{M}_e = \boldsymbol{M} + \boldsymbol{E}$ with noise $\boldsymbol{E}$.

- Optimization template:

$$\underset{\boldsymbol{A},\boldsymbol{B}}{\text{minimize}} \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{M}_e - \boldsymbol{AB})\|_F^2 + \gamma\Big(\frac{1}{q}\|\boldsymbol{A}\|_{2,q}^q + \frac{\alpha}{2}\|\boldsymbol{B}^T\|_F^2\Big), \tag{3}$$

where $q \in \{1, \frac{1}{2}, \frac{1}{4}, \cdots\}$ and $\|\boldsymbol{A}\|_{2,q} := \Big(\sum_{j=1}^d \|\boldsymbol{a}_j\|^q\Big)^{1/q}$.

- Algorithm: Proximal alternating linearized minimization (PALM) [5] for better efficiency.

Our methods also extends to Robust PCA (see section 6 of the paper for detail)

Generalization error bound for LRMC

Each entry of a rank-r ($r \leqslant d$) matrix $\boldsymbol{M}$ is observed with probability ρ plus some noise $E_{ij} \sim \mathcal{N}(0, \epsilon^2)$. Observation is $P_\Omega(\boldsymbol{M} + \boldsymbol{E})$. Consider solving

$$\underset{\|\boldsymbol{X}\|_{S_p}^p \leqslant R_p, \text{rank}(\boldsymbol{X}) \leqslant d}{\text{minimize}} \|\mathcal{P}_\Omega(\boldsymbol{M} + \boldsymbol{E} - \boldsymbol{X})\|_F^2. \tag{4}$$

This is equivalent to (3) for certain choice of γ and R_p.

Theorem 3

Suppose $\|\boldsymbol{M}\|_{S_p}^p \leqslant R_p$, $\hat{\boldsymbol{M}}$ is the optimal solution of (4), and $|\Omega| \geqslant \frac{32}{3} n \log^2 n$. Denote $\zeta := \max\{\|\boldsymbol{M}\|_\infty, \|\hat{\boldsymbol{M}}\|_\infty\}$. Then there exist numerical constants c_1 and c_2 such that the following inequality holds with probability at least $1 - 5n^{-2}$

$$\|\boldsymbol{M} - \hat{\boldsymbol{M}}\|_F^2 \leqslant \max\left\{c_1\zeta^2\frac{n\log n}{|\Omega|}, (5.5 + \sqrt{10})R_p\left((4\sqrt{3}\epsilon_0 + c_2\zeta)^2\frac{n\log n}{|\Omega|}\right)^{1-p/2}\right\}. \tag{5}$$

In sum, Theorem 3 shows it is possible to reduce the matrix completion error by using a smaller p in (4) or a smaller q in (3) (as the second term in the brace of (5) is the dominant term for sufficiently large $|\Omega|$, which decreases as p decreases).

Advantage of our methods

- **Tighter rank approximation.** Compared to the nuclear norm, the spectral quantities in Theorem 1 are tighter approximations to the rank of $\boldsymbol{X}$.
- **Robust to rank initialization.** The iterative algorithms we proposed to minimize $\text{FGSR}_{2/3}$ and $\text{FGSR}_{1/2}$ quickly force some of the columns of $\boldsymbol{A}$ and $\boldsymbol{B}^T$ to zero, where they remain. Hence the number of nonzero columns is reduced dynamically, and converges to r quickly in experiments: these methods are *rank-revealing*.
- **Low computational cost and SVD-free!** The per iteration complexity of algorithms LRMC can be as low as $O(d'\text{card}(\Omega))$.

Numerics for LRMC

Noiseless synthetic data

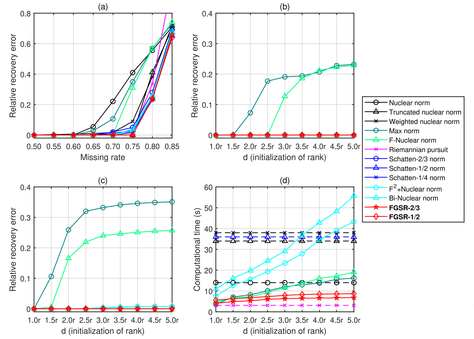

Figure: Matrix completion on noiseless synthetic data ($r = 50$): (a) the effect of missing rate on recovery error; (b)(c) the effect of rank initialization on recovery error (missing rate $= 0.6$ or 0.7); (d) the effect of rank initialization on computational cost (missing rate $= 0.7$)

Noisy synthetic data

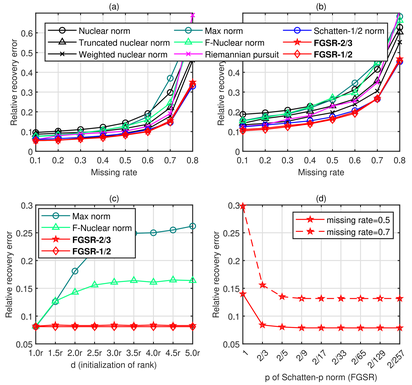

Figure: Matrix completion on noisy synthetic data: (a)(b) recovery error when SNR $= 10$ or 5; (c) the effect of rank initialization on recovery error (SNR $= 10$, missing rate $= 0.5$); (d) the effect of p in Schatten-p norm (using FGSR when $p < 1$).

Similar results for the MovieLens-1M dataset (shown in the paper).

Reference

[1] Srebro, Nathan, Jason Rennie, and Tommi S. Jaakkola. NIPS, 2005.
[2] Shang, Fanhua, Yuanyuan Liu, and James Cheng. AISTATS, 2016.
[3] Shang, Fanhua, James Cheng, Yuanyuan Liu, Zhi-Quan Luo, and Zhouchen Lin. IEEE TPAMI, 2017
[4] Qinghua Liu, Xinyue Shen, and Yuantao Gu. arXiv preprint arXiv, 2017
[5] Bolte, Jérôme, Shoham Sabach, and Marc Teboulle. Mathematical Programming 146, no. 1-2 (2014): 459-494.