Clipped Matrix Completion: A Remedy for Ceiling Effects

Takeshi Teshima,[1,2] Miao Xu,[2] Issei Sato,[1,2] Masashi Sugiyama[2,1]
[1]The University of Tokyo [2]RIKEN

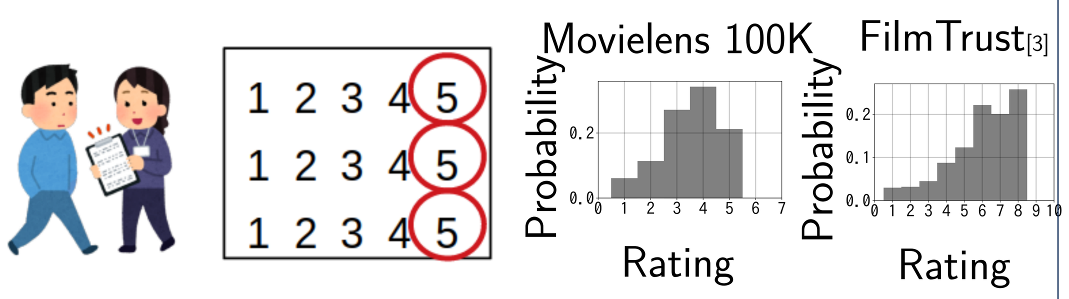
Quick summary

- Novel **problem setting**: Clipped matrix completion.
  - In empirical science, ceiling effect (= Data is clipped during observation) is widely seen.
  - On the other hand, there is matrix completion (MC) to recover matrix data.
- **Goal**: Recover matrix data from clipping (and missing).
- **Results**:
  1. **CMC is possible**: Under sufficient condition, an exact recovery is possible.
  2. **How to recover**: Minimize Squared hinge loss + regularization term. (The regularization has theoretical guarantee)
  3. **Experiments**: Robustness to ceiling effect may benefit recommender systems!

Background

Ceiling effect

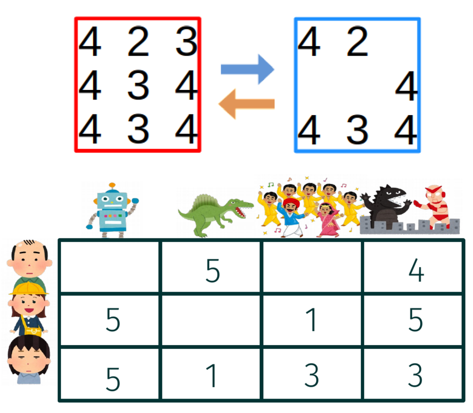
- In natural/social sciences, ceiling effect is widely seen [1].
- It is often modeled as a clipping phenomenon [2].



- Right-truncated histogram: typical for ceiling effects [4].
- One explanation: A user may rate Item A with ★5 and later find better Item B. ⇒ The true rating for Item B should be above 5, but the recorded value is still ★5.

Low-rank matrix completion

- Complete a deficient matrix (e.g., missing, noise, discretization).
- Ex. Recommender systems
  - Organize ratings of movies by users in a matrix.
  - By filling in the blanks, predict ratings for unwatched movies.



Problem setting

Problem (clipped matrix completion; CMC)

Low-rank $\mathbf{M}$

4	7	4	7	4
0	3	6	15	12
4	6	2	2	0
2	6	7	16	12
8	13	6	9	4

Obs. ⇒

Observed $\mathbf{M}^{\mathrm{c}}_{\Omega}$

	7	4	7	4
0	3	6	10	10
4	6	2	2	0
2	6	7	10	10
8	10	6	9	4

CMC ⇒

Restored $\widehat{\mathbf{M}}$

4.0	7.0	4.0	7.0	4.0
-0.0	3.0	6.0	14.9	11.9
4.0	6.0	2.0	2.0	0.0
2.0	6.0	7.0	15.9	11.9
8.0	13.0	6.0	9.0	4.0

- Missing and clipping (from above with threshold C).
- Goal of CMC: Recover $\mathbf{M}$ from $\mathbf{M}^{\mathrm{c}}_{\Omega}$ and C.

When is it possible?

Theorem: Feasibility of CMC (informal)

- Message: CMC is feasible under sufficient conditions.
- Algorithm: Trace-norm minimization

$$\arg\min_{\mathbf{X}} \|\mathbf{X}\|_{\mathrm{tr}} \text{ s.t. } \begin{cases} \mathcal{P}_{\Omega\setminus\mathcal{C}}(\mathbf{X}) = \mathcal{P}_{\Omega\setminus\mathcal{C}}(\mathbf{M}^{\mathrm{c}}_{\Omega}), \\ \mathcal{P}_{\mathcal{C}}(\mathbf{M}^{\mathrm{c}}_{\Omega}) \leq \mathcal{P}_{\mathcal{C}}(\mathbf{X}), \end{cases} \quad (1)$$

where $\Omega := \{(i,j) : \text{observed}\}$ and $\mathcal{C} := \{(i,j) \in \Omega : M^{\mathrm{c}}_{ij} = C\}$

- Assumptions:
  1. The "info. loss" due to clipping is small enough (not necessarily ignorable).
  2. $\mathbf{M}$ is low-rank.
  3. $\mathbf{M}$ is "incoherent" (each entry holds some information of the entire matrix)
  4. Elements are observed independently with high enough probability p.
- Statement: with high probability, the output of (1) is unique and matches the true matrix $\mathbf{M}$ completely.

References

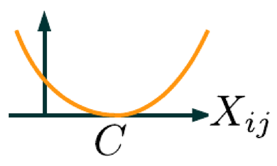
[1] Neil J. Salkind, editor. *Encyclopedia of Research Design*. SAGE Publications, Thousand Oaks, Calif, 2010.
[2] Peter C Austin and Lawrence J Brunner. Type I error inflation in the presence of a ceiling effect. *The American Statistician*, 57(2):97–104, 2003.
[3] G. Guo, J. Zhang, and N. Yorke-Smith. A novel Bayesian similarity measure for recommender systems. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2619–2625, 2013.
[4] William H. Greene. *Econometric Analysis*. Prentice Hall, Boston, 7th edition, 2012.
[5] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6:615–640, 2010.
[6] Haim Avron, Satyen Kale, Shiva Prasad Kasiviswanathan, and Vikas Sindhwani. Efficient and practical stochastic subgradient descent for nuclear norm regularization. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.
[7] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pages 665–674, 2013.
[8] Emmanuel J. Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, June 2010.

How to recover?

General idea

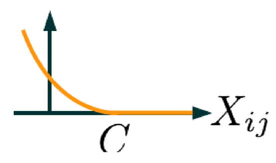
- Ordinary MC (prior work): squared loss [5]

$$\arg\min_{\mathbf{X}} \tfrac{1}{2}\textstyle\sum_{ij\in\Omega}(M^{\mathrm{c}}_{ij} - X_{ij})^2 + \mathcal{R}(\mathbf{X})$$



- CMC (proposed): squared hinge loss

$$\arg\min_{\mathbf{X}} \tfrac{1}{2}\textstyle\sum_{ij\in\Omega:\, M^{\mathrm{c}}_{ij} < C}(M^{\mathrm{c}}_{ij} - X_{ij})^2 + \tfrac{1}{2}\sum_{ij\in\Omega:\, M^{\mathrm{c}}_{ij} = C}\max(0, M^{\mathrm{c}}_{ij} - X_{ij})^2 + \mathcal{R}(\mathbf{X})$$



Design of regularizers

- Double trace-norm regularization (DTr-CMC) (proposed)
  $\mathcal{R}(\mathbf{X}) = \lambda_1\|\mathbf{X}\|_{\mathrm{tr}} + \lambda_2\|\mathrm{Clip}(\mathbf{X})\|_{\mathrm{tr}}$ $\mathrm{Clip} = \min(\cdot, C)$
  - Induces low-rankness in $\mathbf{X}$ and $\mathrm{Clip}(\mathbf{X})$.
  - Optimization: (approximate) subgradient descent [6].
- Trace-norm regularization (Tr-CMC) [5]
  $\mathcal{R}(\mathbf{X}) := \lambda\|\mathbf{X}\|_{\mathrm{tr}}$ $\|\mathbf{X}\|_{\mathrm{tr}} = \sum_{l=1}^{\min(n_1,n_2)}\sigma_l$ (σ_l: l-th singular value)
  - Induces low-rankness in $\mathbf{X}$ (cf. rank is the count of nonzero σ_i's).
  - Optimization: accelerated proximal gradient descent (APG) [5].
- Frobenius norm regularization (Fro-CMC) [7]
  $\mathcal{R}(\mathbf{P},\mathbf{Q}) := \lambda_1\|\mathbf{P}\|_{\mathrm{F}}^2 + \lambda_2\|\mathbf{Q}\|_{\mathrm{F}}^2$ $\mathbf{X} = \mathbf{P}\mathbf{Q}^\top$
  - Induces low-rankness in $\mathbf{X}$.
  - Optimization: (approximate) alternating least squares (ALS) [7].

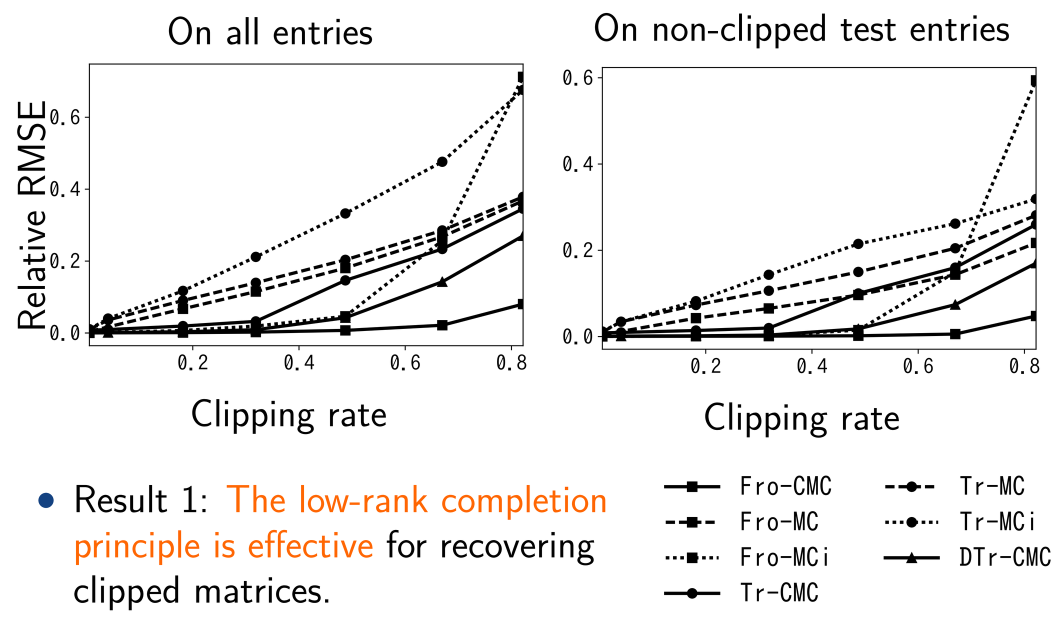
Experiments

Compared methods

- *-MC: Ordinary MC (squared loss).
- *-MCi: Ordinary MC (squared loss) with clipped entries ignored.

Synthetic data

- Controlled experiment with known true values.



- Result 1: The low-rank completion principle is effective for recovering clipped matrices.
  - Proposed method recovers the matrices with a relative error of 10^{-2} order.
- Result 2: Proposed method is more robust to the clipping.
  - Ordinary MC: Recovery error on non-clipped entries increased with the clipping rate. Indicative of the disturbance of the clipped entries.

Challenge in experiments with real-world data

- "True values" for real-world data with ceiling effect are rarely available. → Evaluations on the ability to "predict which entries should be above threshold."

Real-world data experiment 1/2

- Training with artificially clipped data (e.g., ★5 → ★4)
- Task: classify entries into "(true rating) ≥ (threshold)" or not.
- (Baseline always predicts as "above threshold.")

f_1 score	**DTr-CMC**	**Fro-CMC**	Fro-MC	**Tr-CMC**	Tr-MC	(Baseline)
Film Trust	**0.47** (0.01)	0.35 (0.01)	0.27 (0.01)	0.36 (0.00)	0.22 (0.00)	0.41 (0.00)
Movielens 100K	0.39 (0.00)	**0.41** (0.00)	0.21 (0.01)	0.40 (0.00)	0.12 (0.00)	0.35 (0.00)

- Improvement is seen by the change from -MC to -CMC.
- Proposed CMC methods give better estimates of the true values of the clipped entries than MC counterparts.

Real-world data experiment 2/2

- Training with original real-world data (e.g., ★1 - ★5).
- Task: classify entries into "(true rating) ≥ (max rating)" or not.
- (Baseline always predicts as "above threshold.")

f_1 score	**DTr-CMC**	**Fro-CMC**	Fro-MC	**Tr-CMC**	Tr-MC	(Baseline)
Film Trust	**0.46** (0.01)	0.40 (0.01)	0.35 (0.01)	0.39 (0.00)	0.35 (0.01)	0.41 (0.00)
Movielens 100K	0.38 (0.00)	**0.41** (0.01)	0.38 (0.01)	0.40 (0.00)	0.38 (0.00)	0.35 (0.00)

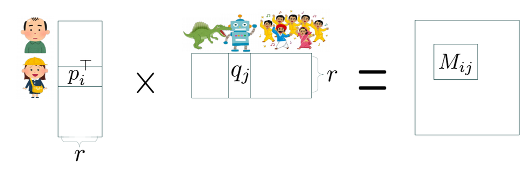
- Improvement is seen by the change from -MC to -CMC.
- Enhanced performance on the prediction of "high rating" by being robust to the ceiling effect.

More details

Background: Matrix completion (MC)

The low-rank completion principle

- For a low-rank and *incoherent* matrix, the recovery may be possible [8].
- Low-rank: Entries are dot products of low-dim. row/col features.
- To complete: Estimate latent vectors → Take inner products.



Limitations of ordinary MC methods

1. Info deficit depends on the true value.
   → Existing theory does not apply.
2. Potentially large gap b/w true value and the observed.
   → Learning is disturbed.

4.8	7.0	4.5	6.7	4.0
-0.0	3.1	6.1	9.7	10.2
4.5	5.4	1.8	2.5	-0.2
2.2	5.6	6.8	10.6	9.7
7.6	10.5	5.9	8.7	4.2

Result of ordinary MC from $\mathbf{M}^{\mathrm{c}}_{\Omega}$.

Def: Incoherence

- Let $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (skinny singular value decomposition).
- Coherence $\mu_0 := \max\left\{\frac{n_1}{r}\mu^{\mathrm{U}}(\mathbf{M}), \frac{n_2}{r}\mu^{\mathrm{V}}(\mathbf{M})\right\}$
  - where $\mu^{\mathrm{U}}(\mathbf{M}) := \max_{i\in[n_1]}\|\mathbf{U}_{i,\cdot}\|^2$, $\mu^{\mathrm{V}}(\mathbf{M}) := \max_{j\in[n_2]}\|\mathbf{V}_{j,\cdot}\|^2$, and $r = \mathrm{rank}(\mathbf{M})$.
- Joint coherence $\mu_1 := \sqrt{\frac{n_1 n_2}{r}}\|\mathbf{U}\mathbf{V}^\top\|_\infty$
- $\mathbf{M}$ is *incoherent* iff. μ_0 and μ_1 are small.

Detailed condition of exact recovery

Def: The information loss

- $\mathcal{B} := \{(i,j) : M_{ij} < C\}$
- $T := \mathrm{span}\left(\{u_k y^\top : k\in[r], y\in\mathbb{R}^{n_2}\}\cup\{x v_k^\top : k\in[r], x\in\mathbb{R}^{n_1}\}\right)$
- $(\mathcal{P}^*(\mathbf{Z}))_{ij} := 1\{M_{ij} < C\}Z_{ij} + 1\{M_{ij} = C\}(Z_{ij})_+$
- $\rho_{\mathrm{F}} := \sup_{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\|\mathbf{Z}\|_{\mathrm{F}}\leq\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}} \frac{\|\mathcal{P}_T\mathcal{P}^*(\mathbf{Z})-\mathbf{Z}\|_{\mathrm{F}}}{\|\mathbf{Z}\|_{\mathrm{F}}}$
- $\rho_\infty := \sup_{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\|\mathbf{Z}\|_\infty\leq\|\mathbf{U}\mathbf{V}^\top\|_\infty} \frac{\|\mathcal{P}_T\mathcal{P}^*(\mathbf{Z})-\mathbf{Z}\|_\infty}{\|\mathbf{Z}\|_\infty}$
- $\rho_{\mathrm{op}} := \sqrt{r}\mu_1\left(\sup_{\substack{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\\ \|\mathbf{Z}\|_{\mathrm{op}}\leq\sqrt{n_1 n_2}\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{op}}}} \frac{\|\mathcal{P}^*(\mathbf{Z})-\mathbf{Z}\|_{\mathrm{op}}}{\|\mathbf{Z}\|_{\mathrm{op}}}\right)$
- $\nu_{\mathcal{B}} := \|\mathcal{P}_T\mathcal{P}_{\mathcal{B}}\mathcal{P}_T - \mathcal{P}_T\|_{\mathrm{op}}$

Theorem: Exact recovery guarantee for CMC

- Assume
  - $\rho_{\mathrm{F}} < 1/2, \rho_{\mathrm{op}} < 1/4, \rho_\infty < 1/2, \nu_{\mathcal{B}} < 1/2$.
  - Entries are independently observed with probability p.
  - $n_1, n_2 \geq 2$, $p \geq 1/(n_1 n_2)$
- If $p \geq \min\left\{1, c_\rho \max(\mu_1{}^2, \mu_0) r f(n_1,n_2)\right\}$ is satisfied, then $\widehat{\mathbf{M}}$ is unique and equal to $\mathbf{M}$ with probability at least $1-\delta$, where
  - $c_\rho = \max\left\{\frac{24}{(1/2-\rho_{\mathrm{F}})^2}, \frac{8}{(1/4-\rho_{\mathrm{op}})^2}, \frac{8}{(1/2-\rho_\infty)^2}, \frac{8}{(1/2-\nu_{\mathcal{B}})^2}\right\}$
  - $f(n_1,n_2) = \mathcal{O}\left(\frac{(n_1+n_2)(\log(n_1 n_2))^2}{n_1 n_2}\right)$
  - $\delta = \mathcal{O}\left(\frac{\log(n_1,n_2)}{n_1+n_2}\right)(n_1+n_2)^{-1}$

DTr-CMC theoretical guarantee

Theorem: DTr-CMC estimation error bound

- Message: Even under clipping, if the assumptions are satisfied, an accurate estimation is possible by DTr-CMC.
- DTr-CMC is a Lagrange-relaxation of the following problem with a convex-relaxation of the objective function.
- $\widehat{\mathbf{M}} \in \arg\min_{\mathbf{X}\in G}\sum_{(i,j)\in\Omega}(M^{\mathrm{c}}_{ij} - \mathrm{Clip}(X_{ij}))^2$
  $G = \left\{\mathbf{X}\in\mathbb{R}^{n_1\times n_2} : \|\mathbf{X}\|_{\mathrm{tr}}^2 \leq \beta_1\sqrt{k n_1 n_2}, \|\mathrm{Clip}(\mathbf{X})\|_{\mathrm{tr}}^2 \leq \beta_2\sqrt{k n_1 n_2}\right\}$
- $\mu(\mathbf{X}) = \max\{\mu^{\mathrm{U}}(\mathbf{X}), \mu^{\mathrm{V}}(\mathbf{X})\}$
- Theorem: Suppose $\mathbf{M}\in G$. Let $\mu_G = \sup_{\mathbf{X}\in G}\mu(\mathrm{Clip}(\mathbf{X}))$. Then $\exists C_0, C_1 > 0$ s.t. with probability at least $1 - C_1/(n_1+n_2)$,

$$\sqrt{\frac{1}{n_1 n_2}\|\widehat{\mathbf{M}}-\mathbf{M}\|_{\mathrm{F}}^2} \leq \underbrace{\frac{\|\mathbf{M}-\mathbf{M}^{\mathrm{c}}\|_{\mathrm{F}}}{\sqrt{n_1 n_2}}}_{=B_1:\text{Complexity of data}} + \underbrace{\frac{\|\widehat{\mathbf{M}}-\mathrm{Clip}(\widehat{\mathbf{M}})\|_{\mathrm{F}}}{\sqrt{n_1 n_2}}}_{=B_2:\text{Complexity of hypothesis}} + \underbrace{\frac{\|\mathrm{Clip}(\widehat{\mathbf{M}})-\mathrm{Clip}(\mathbf{M})\|_{\mathrm{F}}}{\sqrt{n_1 n_2}}}_{=B_3:\text{Estimation error}},$$

where

$$B_1, B_2 \leq (\sqrt{\beta_1}+\sqrt{\beta_2})k^{\frac{1}{4}}(n_1 n_2)^{-\frac{1}{4}},$$

$$B_3 \leq \sqrt{C_0\frac{2\mu_G{}^2\beta_2}{p}\left(\frac{pk(n_1+n_2)+k\log(n_1+n_2)}{n_1 n_2}\right)^{\frac{1}{4}}}.$$

Source code (GitHub)

Repo: clipped-matrix-completion-py

Longer paper (ArXiv)

arXiv:1809.04997