Research on Restoring Clipped Low-rank Matrices
(打ち切りを受けた低ランクな行列の復元の研究)

手嶋毅志

2019 年 11 月 16 日

@第 4 回 統計・機械学習若手シンポジウム

東京大学 (新領域) 博士課程

研究内容

- 「天井効果」に対処する為の機械学習（今回の発表）

  Teshima, T., Xu, M., Sato, I., and Sugiyama, M.,

  Clipped matrix completion: a remedy for ceiling effects. AAAI-19.

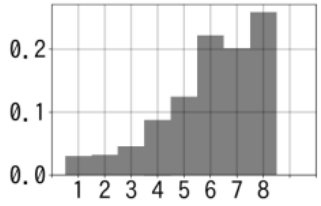
  (Miao Xu 博士，佐藤一誠先生，杉山将先生との共同研究)

打ち切り行列補完：全体像

① 問題：天井効果

＝計測に上限がある.

1	2	3	4	5
1	2	3	4	5
1	2	3	4	5



② 方法：行列補完

＝低ランク行列を復元する.

4	2	3
4	3	4
4	3	4

⇄

4	2	
		4
4	3	4

- 欠測
- ノイズ

など

③ 今回の研究内容

打ち切りを受けた低ランク行列の補完

低ランク

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

⇄

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

打ち切り
（+欠測）

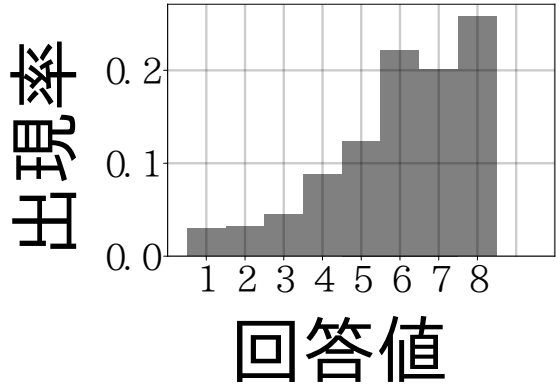
①天井効果（重要！）

天井効果

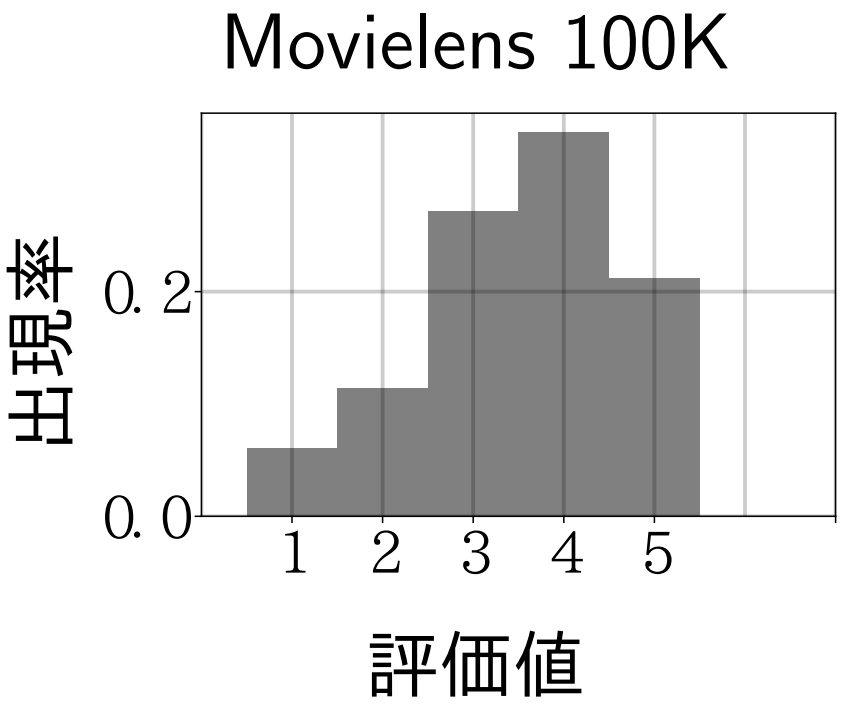
計測可能な値に上限があるという状況のこと [1].
上限を超えた値は，上限値に打ち切られて観測される.

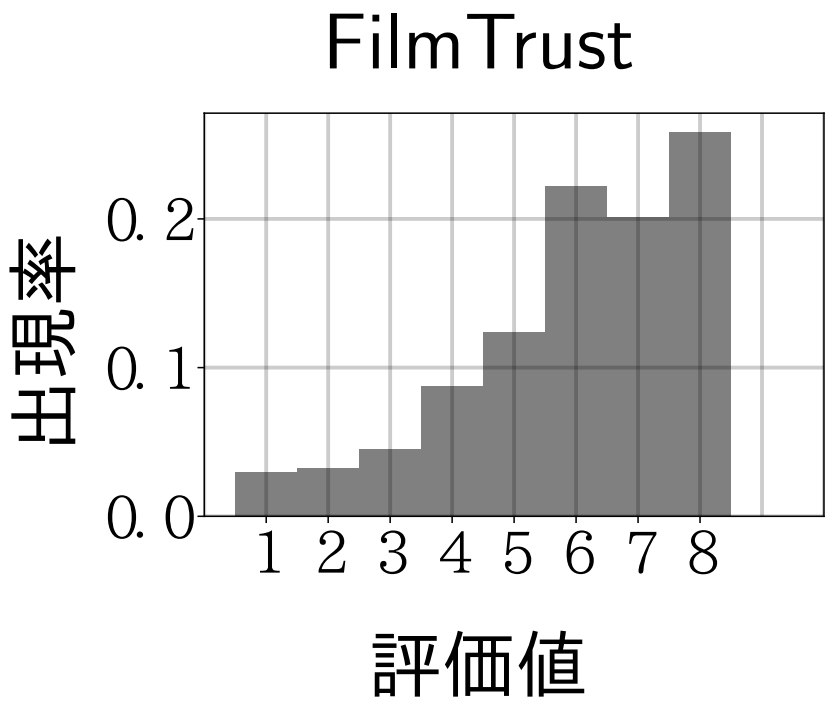
- 例. 5 段階評価の質問

- “5” の人がやたらと多い場合は，「質問が悪く，測定対象を正確に測れなかった」と考えられる.



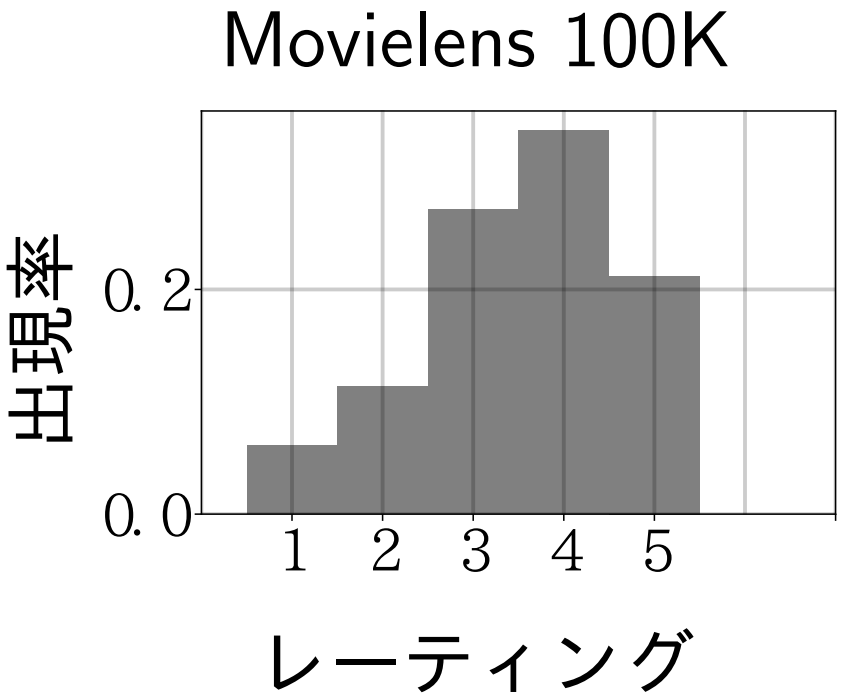
天井効果は機械学習のデータにも

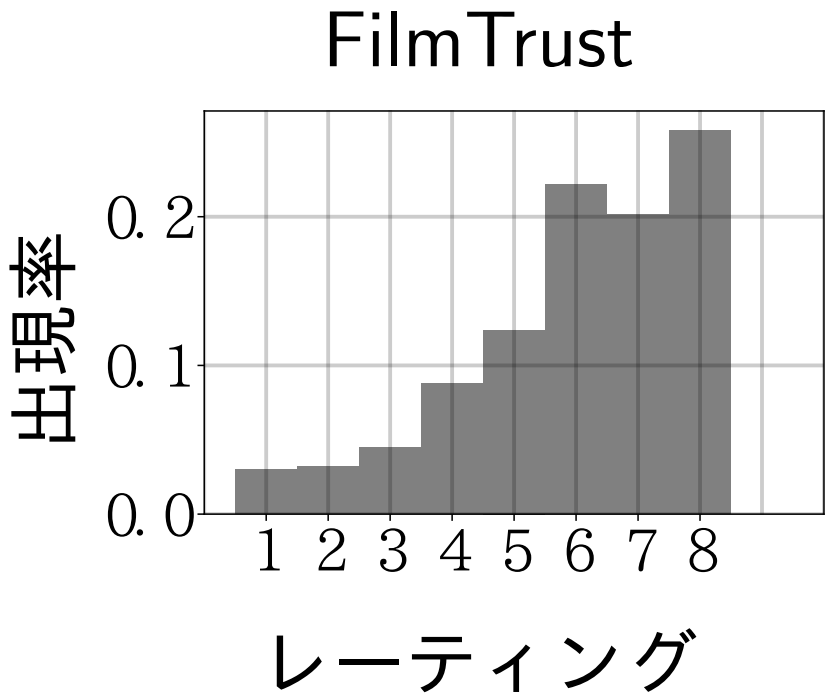




- 推薦システムのベンチマークデータ.
- 右側が切られた形＝天井効果の典型例

どうすれば天井効果があるデータの真値を調べられるだろうか？



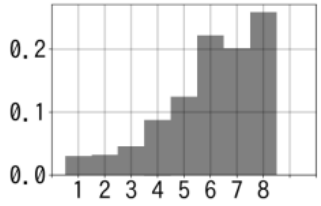


打ち切りを受けた低ランク行列の補完

① 問題：天井効果

＝計測に上限がある.

1	2	3	4	5
1	2	3	4	5
1	2	3	4	5



② 方法：行列補完

＝低ランク行列を復元する.

4	2	3
4	3	4
4	3	4

⇄

4	2	
		4
4	3	4

- 欠測
- ノイズ

など

③ 今回の研究内容

打ち切りを受けた低ランク行列の補完

低ランク

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

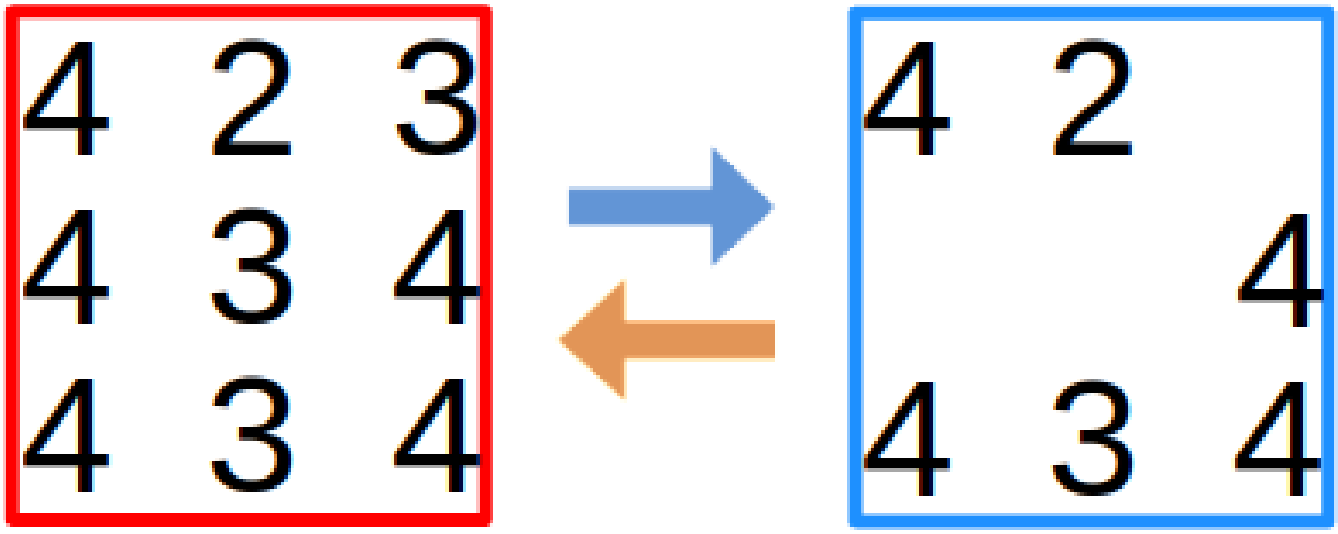
⇄

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

打ち切り
（+欠測）

②行列補完

- 行列の欠測成分などを復元する技術 [3]

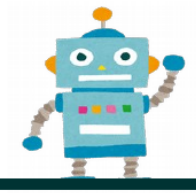
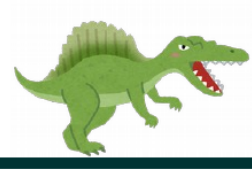


- ゴール：虫食いの穴埋め
- 欠測，ノイズ，離散化，. . . それぞれに手法が開発

②行列補完の応用例

- 応用例：映画推薦システム

映画


ユーザー

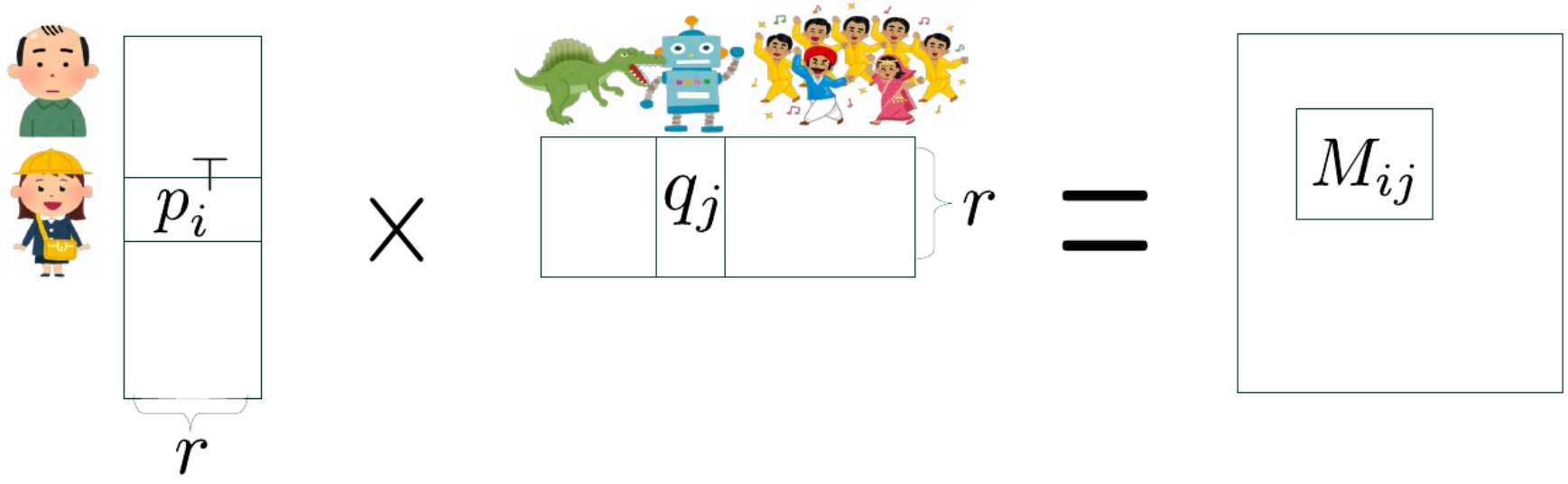
	5		4
5		1	5
5	1	3	3

②行列補完の仕組み

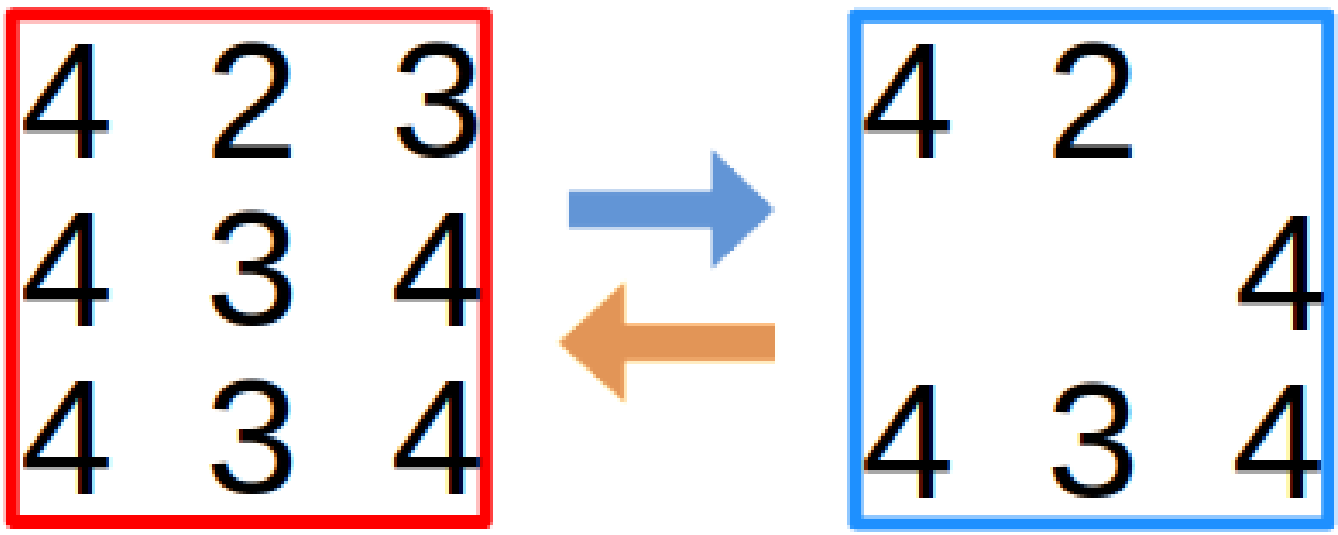
- 穴埋めできるには条件がある
- 低ランクかつ好性質な行列の場合，復元可能 [3]

低ランク補完の原理

- 低ランク…低次元ベクトルの内積で各成分が表せる
- 穴埋め時：潜在ベクトルを推定→内積を計算



低ランク行列補完の原理：欠測などから行列を復元

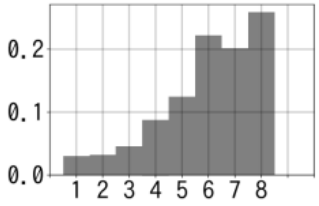


打ち切り行列補完：全体像

① 問題：天井効果

＝計測に上限がある.

1	2	3	4	5
1	2	3	4	5
1	2	3	4	5



② 方法：行列補完

＝低ランク行列を復元する.

4	2	3
4	3	4
4	3	4

⇄

4	2	
		4
4	3	4

- 欠測
- ノイズ

など

③ 今回の研究内容

打ち切りを受けた低ランク行列の補完

低ランク

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

⇄

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

打ち切り
（+欠測）

打ち切りを受けた低ランク行列の補完

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

観測値

$\Rightarrow$

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

真の行列

Problem (打ち切り行列補完；CMC)

観測値と打ち切り閾値（既知）から，真の行列を精度よく復元せよ.

結果・貢献の一覧

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

真値（未知）

観測 ⇒

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

観測データ
（上限 10）

提案法 ⇒

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

補完後

1. 問題設定の提案→天井効果を受けた行列の復元
2. 復元は可能？→真の行列次第では完全復元も可能！
3. 復元手法は？→二乗ヒンジ損失＋正則化項を最小化
4. 実験的には？→天井効果への耐性は，推薦システムでも役に立つ可能性がある！

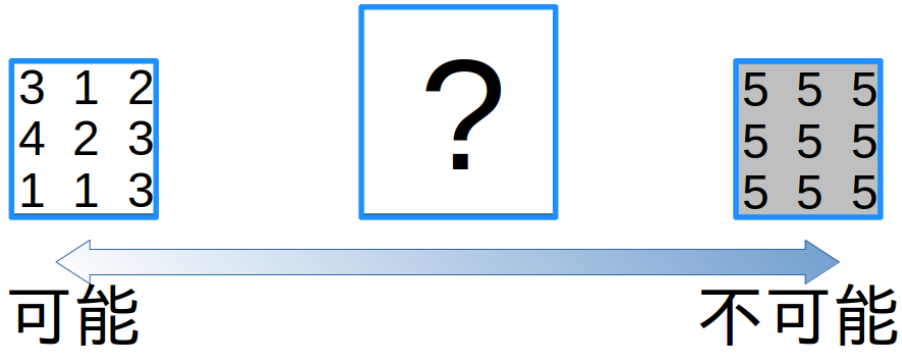
（※提案した正則化項は理論保証付き）

技術的詳細の一覧

1. 主定理：補完はいつ可能なのか？
2. 提案法：実際，どう補完するか？
3. 提案法の性能保証 (省略)
4. 実験結果

技術的詳細①：補完はいつ可能なのか？

- 理論分析の動機：打ち切りのある行列は，いつでも復元できるとは限らない．



  - 復元が明らかに不可能な場合もある．
  - 復元がそもそも不要な場合もある．

- 本研究では復元可能性の十分条件を与えた

主定理:打ち切り補完可能性の定理

仮定 (インフォーマル)

1. 真の行列 $\mathbf{M}$ は低ランク
2. $\mathbf{M}$ は “Incoherent” (少ない観測で全体を推測できる)
3. $\mathbf{M}$ の各成分は独立に十分大きい確率 p で観測
4. 打ち切りの「情報損失」がある絶対定数より小さい

定理 (CMC の完全復元可能性)

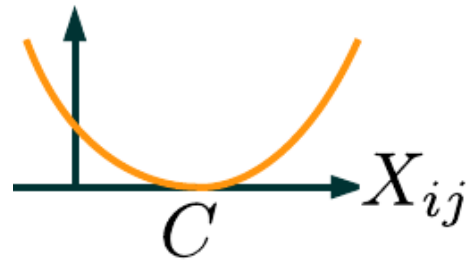
高い確率で,「トレースノルム最小化」というアルゴリズムの出力が真の行列に一致する.

(ここでの「高い確率」は各成分の観測／非観測の確率)

技術的詳細②：実際，どう補完するか？

- 通常の行列補完 (既存法)[5] ：二乗損失

$$\arg\min_{\mathbf{X}} \frac{1}{2} \sum_{ij：観測} (観測値_{ij} - X_{ij})^2 + \mathcal{R}(\mathbf{X})$$

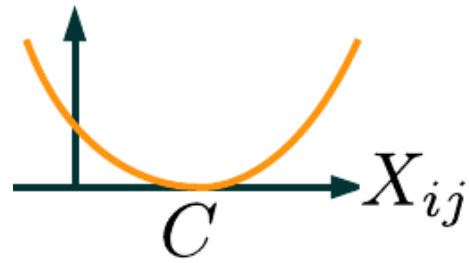


- 打ち切りを受けた場所では閾値 C を復元してしまう

技術的詳細②：実際，どう補完するか？

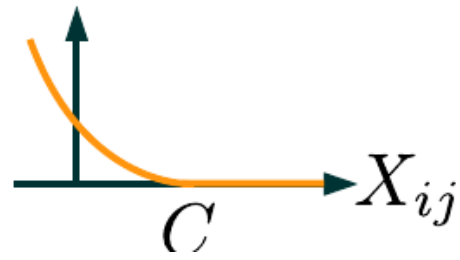
- 通常の行列補完 (既存法)[5] ：二乗損失

$$\arg\min_{\mathbf{X}} \frac{1}{2} \sum_{ij：観測} (観測値_{ij} - X_{ij})^2 + \mathcal{R}(\mathbf{X})$$



- 打ち切りを受けた場所では閾値 C を復元してしまう
- 提案法：二乗ヒンジ損失

$$\arg\min_{\mathbf{X}} \frac{1}{2} \sum_{ij：非打ち切り観測} (観測値_{ij} - X_{ij})^2 + \frac{1}{2} \sum_{ij：打ち切り観測} \max(0, 観測値_{ij} - X_{ij})^2 + \mathcal{R}(\mathbf{X})$$



正則化項の設計：低ランクな解を誘導

1. Tr-CMC: 核ノルム正則化 [5]
   - 効果：低ランクな解を誘導

$\mathcal{R}(\mathbf{X}) := \lambda \|\mathbf{X}\|_{\mathrm{tr}}$ $\|\mathbf{X}\|_{\mathrm{tr}} = \sum_{l=1}^{\min(n_1,n_2)} \sigma_l$

(σ_l: l-th singular value)

2. Fro-CMC: フロベニウスノルム正則化 [4]
   - 効果：低ランクな解を誘導

$\mathcal{R}(\mathbf{P},\mathbf{Q}) := \lambda_1 \|\mathbf{P}\|_{\mathrm{F}}^2 + \lambda_2 \|\mathbf{Q}\|_{\mathrm{F}}^2$ $\mathbf{X} = \mathbf{P}\mathbf{Q}^\top$

3. DTr-CMC: 二重核ノルム正則化 (今回提案)
   - 効果：$\mathbf{X}$ と $\mathrm{Clip}(\mathbf{X})$ の両方に低ランク性を誘導
   - 理論保証付き (詳細略)

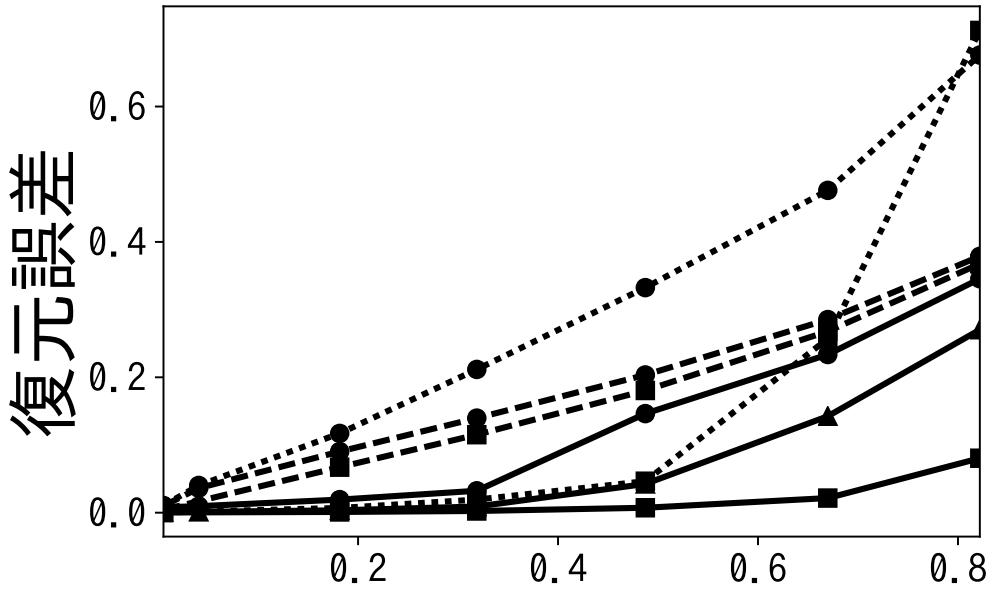
$\mathcal{R}(\mathbf{X}) = \lambda_1 \|\mathbf{X}\|_{\mathrm{tr}} + \lambda_2 \|\mathrm{Clip}(\mathbf{X})\|_{\mathrm{tr}}$ $\mathrm{Clip} = \min(\cdot, C)$

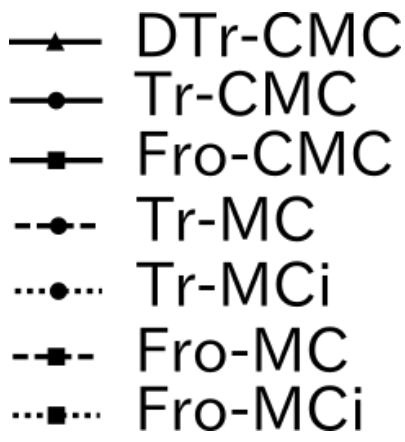
技術的詳細③：実験

1. 人工データによる実験
   - 真値が分かる状況で復元の性能を評価できる

1. 実データによる実験
   - 真値が不明なので実データでは復元精度の評価は難しい
     （天井効果を受けた実データの「打ち切り前の値」は不明）
   - 評価方法に工夫：真値が「閾値以上かどうか」を予測する二値分類タスクで評価

実験 1/2 人工データによる実験





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

観測値

⇒

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

補完後

- 実線は提案法，点線は既存法.
- 打ち切り閾値を変化→復元のテスト誤差を評価.
- 提案法 (実線) は 70%程度の打ち切りに対しても 10^{-2} のオーダーの誤差で精度良く推定できている.

実験 2/2 実データによる実験

f_1 値	DTr-CMC	Fro-CMC	Fro-MC	Tr-CMC	Tr-MC	(ベースライン)
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)

- 実データ (★ 1〜★ 5) の行列から学習
- 真の評価値を「★ 5 以上」と「★ 4 以下」に分類する
- 天井効果への頑健性が「高評価」の識別性能を改善
- （ベースラインは，常に $+1$ を出力する識別器）

まとめ

- 問題設定は？ → 天井効果を受けた行列の復元
- 復元は可能？ → 真の行列次第では完全復元も可能！
- 復元手法は？ → 二乗ヒンジ損失＋正則化項を最小化
- 実験的には？ → 天井効果への耐性は，推薦システムでも役に立つ可能性がある！

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

真値（未知）

観測 ⇒

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

観測データ（上限 10）

提案法 ⇒

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

補完後

主定理:打ち切り補完可能性の定理

トレースノルム最小化は以下のアルゴリズム

$$\arg\min_{\mathbf{X}} \|\mathbf{X}\|_{\mathrm{tr}} \text{ s.t. } \begin{cases} \mathbf{X}_{ij} = \mathbf{M}^{\mathrm{c}}_{ij} \ (ij \text{ は非打ち切り観測}), \\ \mathbf{X}_{ij} \geq \mathbf{M}^{\mathrm{c}}_{ij} \ (ij \text{ は打ち切り観測}), \end{cases}$$

ここで $\mathbf{M}^{\mathrm{c}}$ は $\mathbf{M}$ を成分ごとに閾値 C で打ち切った行列 (i.e. 観測値).

定理 (CMC の完全復元可能性)

高い確率で，トレースノルム最小化の解は真の行列に一致する.

打ち切り低ランク行列補完の今後の展望

- トレースノルム最小化で復元が可能となるための必要条件の特徴付け
- 復元を不可能にするように打ち切りを人為的に施すアルゴリズムの開発

復元が可能か判断するにはどうするか

- 真の行列を見なくては決められない（モデルの仮定）
- しかし，どのような行列は復元が可能か，という直観はいくつかある
  - 低ランク＝行ごと・列ごとの特徴ベクトルが低次元
  - （打ち切り無しに観測できた場合に）似ている行・列がある
  - $\mathbf{M}$ の特異ベクトルがスパースでない
    - 「特定の添字だけ値が大きい」ということが起きない

天井効果の存在を確かめるには？

- ドメイン知識の一種．データだけから何も仮定を置かずに存在を確かめる方法はない．
  - 真の分布をある程度仮定できるなら分かる場合もある
  - ヒストグラムの形＋背景知識で「恐らくある」と想定できるときもある
- 確実にあると主張するには（天井効果を回避できる方法で）再計測して比較する．

天井効果だけではなくて床効果も扱える？

- はい，今回の結果は床効果にも拡張可能です（論文中には示しています）.
- さらに，成分ごとに異なる閾値を用いることもできます.

離散化からの復元手法はもうある？

- 離散化の場合を研究した先行研究はあります.
- 但し，全成分が離散化されているときにのみ使えます.
- 全成分が離散化しているわけではなく，打ち切りだけ受ける応用例も多々想定できるので，打ち切り単体で研究する価値があると考えます.

ハイパーパラメータ選択の方法

- 復元問題に対しては理論的に正当化された方法はまだない.
  - 人工データ実験では，復元後の行列を打ち切ったものとデータとの誤差（事前に取り置きした検証用添字上で計算）が最小のものを選択した
- 実データ実験で用いた添字の分類問題では，最終的な精度が計算可能なので，検証用添字上での精度が最大のものを選択した.
  - 同様に，推薦システム応用においては精度評価をハイパーパラメータ選択に用いることができる場合が多いと思われる.

正則化項の設計：最適化

1. DTr-CMC: 二重核ノルム正則化 (今回提案)
   - $\mathcal{R}(\mathbf{X}) = \lambda_1 \|\mathbf{X}\|_{\mathrm{tr}} + \lambda_2 \|\mathrm{Clip}(\mathbf{X})\|_{\mathrm{tr}}$ $\mathrm{Clip} = \min(\cdot, C)$
   - 最適化法：(近似的な) 劣勾配降下法 [2]
2. Tr-CMC: 核ノルム正則化 $\mathcal{R}(\mathbf{X}) := \lambda \|\mathbf{X}\|_{\mathrm{tr}}$

   $\|\mathbf{X}\|_{\mathrm{tr}} = \sum_{l=1}^{\min(n_1,n_2)} \sigma_l$ (σ_l: l-th singular value)
   - 最適化法：加速勾配降下法 [5]
3. Fro-CMC: フロベニウスノルム正則化 $\mathcal{R}(\mathbf{P}, \mathbf{Q}) := \lambda_1 \|\mathbf{P}\|_{\mathrm{F}}^2 + \lambda_2 \|\mathbf{Q}\|_{\mathrm{F}}^2$ $\mathbf{X} = \mathbf{P}\mathbf{Q}^\top$
   - 最適化法：(近似的な) 交互最小二乗法 [4]

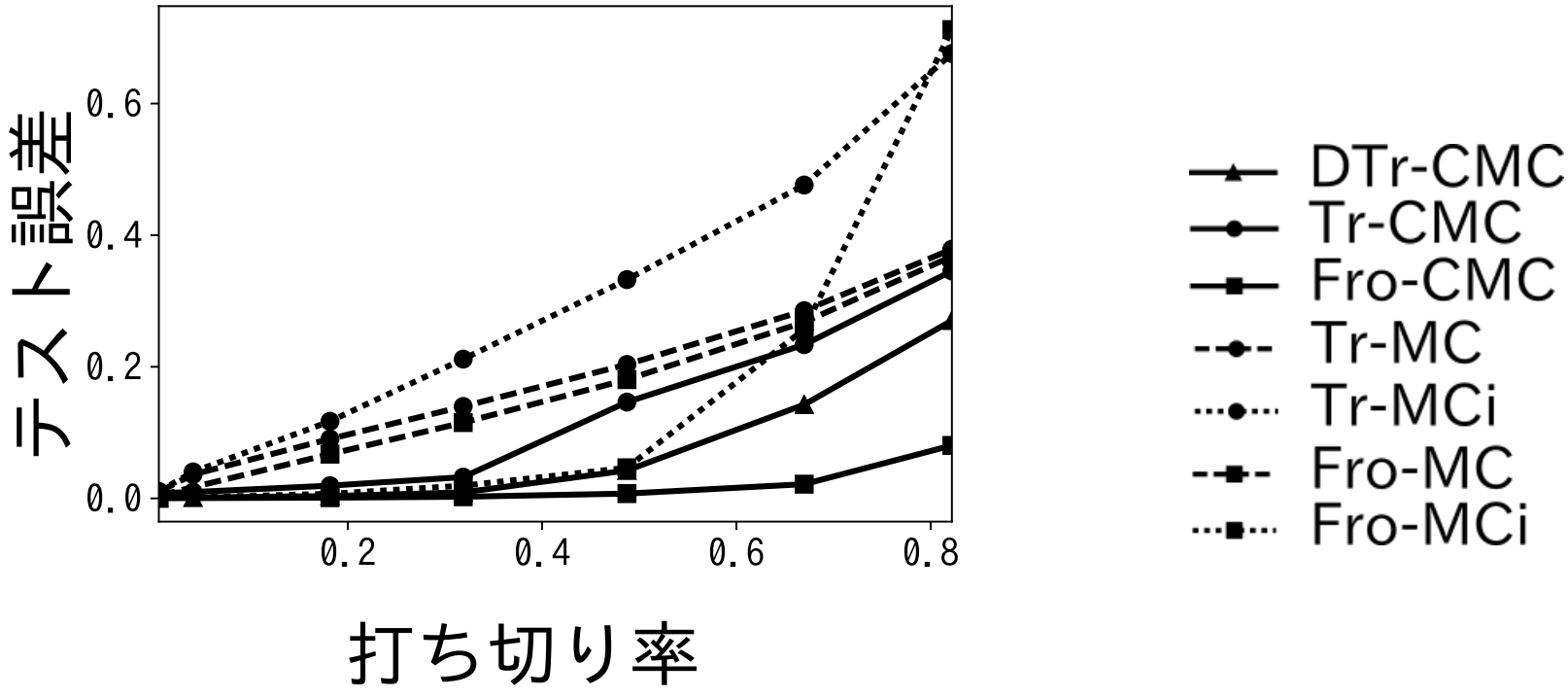
Coherence の定義

- $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ とする (特異値分解).
- Coherence は $\mu_0 := \max\left\{\frac{n_1}{r}\mu^{\mathrm{U}}(\mathbf{M}), \frac{n_2}{r}\mu^{\mathrm{V}}(\mathbf{M})\right\}$ で定義
  - ここで $\mu^{\mathrm{U}}(\mathbf{M}) := \max_{i\in[n_1]} \|\mathbf{U}_{i,\cdot}\|^2$,
    $\mu^{\mathrm{V}}(\mathbf{M}) := \max_{j\in[n_2]} \|\mathbf{V}_{j,\cdot}\|^2$, $r = \mathrm{rank}(\mathbf{M})$.
- Joint coherence は $\mu_1 := \sqrt{\frac{n_1 n_2}{r}}\|\mathbf{U}\mathbf{V}^\top\|_\infty$ で定義
- μ_0 および μ_1 が小さいとき $\mathbf{M}$ は incoherent であるという.

打ち切りによる情報損失の定義

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

実験 1/2 人工データによる実験詳細



- 打ち切り閾値を変化→ $\frac{\|\mathcal{P}_{\mathrm{test}}(\widehat{\mathbf{M}}-\mathbf{M})\|_{\mathrm{F}}}{\|\mathcal{P}_{\mathrm{test}}(\mathbf{M})\|_{\mathrm{F}}}$ を評価
- $*$-MCi は，打ち切られた観測値を欠測とみなして従来の行列補完を適用したもの

実験 2/2 実データによる実験詳細

- $\mathrm{precision}$：予測が「yes」のうち，真値が「yes」であるものの割合 （正確に予測できた割合）
- recall：真値が「yes」のうち，予測が「yes」であるものの割合 （真の「yes」のうち，正しく呼び戻せた割合）
- $\mathrm{f}_1 = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$
- f_1 を用いたのは，この 2 値分類タスクでは，「小さめの値を入力し，大きめの値を予測する」という外延問題を解いているため，recall の部分に難しさがあると考えられるから．

どの辺りが重要で，どの辺りが難しいか

重要性・新規性

- 天井効果は経験科学でよく現れる
- 従来，行列補完の分野では考えられてこなかった種類の欠損

技術的困難

1. 打ち切りは従来の方法の（理論的／実用的に）適用範囲外
2. とくに，問題の可解性を表す定理を証明
   - 従来の行列補完の理論的保証はそのままでは適用できなかった
   - 打ち切りの影響を測る理論的な量を提案

参考資料

[1] Encyclopedia of research design.

[2] 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, pages 1231–1238.

[3] Emmanuel J. Cands and Yaniv Plan.
Matrix completion with noise.
98(6):925–936.

[4] 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.

[5] Kim-Chuan Toh and Sangwoon Yun.
An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems.
6:615–640.