打ち切り行列の補完：天井効果の治療法

手嶋毅志 (東京大学)

1　背景

本稿では，天井効果と呼ばれる計測限界によって情報が欠損した観測値のみから，低ランクな行列を復元する手法の研究結果 [6] を紹介する．

天井効果とは，計測方法の制約によりデータの観測値に上限が生じる状況のことである．上限を超えた値は閾値で打ち切られてから観測されるため，真の値は観測できない．天井効果は統計分析に悪影響を与えることが知られているため，教育学や生体医学など幅広い科学分野において議論の対象となってきた [1]．同様の現象は機械学習の応用分野，例えば推薦システムのベンチマークデータにおいてもみられる (図 2)．

一方，低ランク行列補完は，低ランクな行列の観測値がさまざまに欠損している場合に元の行列を精度良く推定する技術である．欠測やノイズ，離散化といった情報欠損に対して有効であることが知られている [2]．

低ランク行列補完の原理を応用すれば，天井効果を受けた行列から，天井効果を受ける前の行列を復元できる可能性がある．しかし，天井効果のように真値に依存した情報欠損は，既存の行列補完の理論および手法の適用範囲外である．そこで本研究では閾値による打ち切りを受けた行列を復元する手法を提案し，その有効性を理論，実験の両面から示した．

2　問題の定式化と復元可能性解析

以下では一般に行列を太字で表し，その成分は対応する細字に添字を付けて表す (例えば行列 $\mathbf{X}$ の第 (i,j)-成分は X_{ij} で表す)．

問題の定式化．　行列 $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ を一つ固定し，真の行列と呼ぶ．そのランクを r とする．打ち切りの閾値を $C \in \mathbb{R}$ とし，行列 $\mathbf{M}^{\mathrm{c}}$ を $M^{\mathrm{c}}_{ij} = \min(C, M_{ij})$ で定義する．この $\mathbf{M}^{\mathrm{c}}$ のうち，成分の一部が一様な確率 p で独立に観測されるとする．観測された添字の集合を $\Omega \subset [n_1] \times [n_2]$ で表す．ここで自然数 $n \in \mathbb{N}$ に対して $[n] := \{1, 2, \ldots, n\}$ とした．打ち切られた行列の補完問題 (Clipped matrix completion) は，$\{M^{\mathrm{c}}_{ij}\}_{ij\in\Omega}$ および C の値から $\mathbf{M}$ を精度よく復元せよという問題である（図 1）．

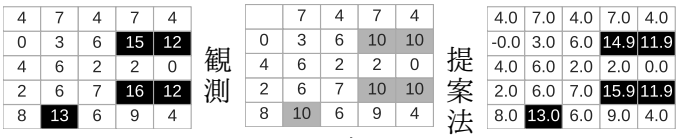

図 1: 本研究の問題設定．復元結果は実際に提案法の一つ「Fro-CMC」を適用した結果である．真の行列を精度良く復元できていることが分かる．

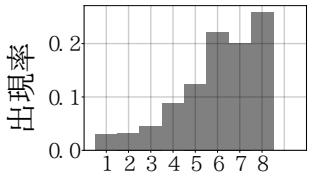

図 2: 映画推薦のベンチマークデータ FilmTrust の評価値頻度分布．右に打ち切られた形状が天井効果の存在を示唆する．

復元可能性の解析．　打ち切られた行列の補完問題においては，仮に全ての成分が観測されたとしても復元が可能であるとは限らない．特に自明な例として，全ての成分が打ち切りを受けている場合は復元が不可能である．そこで本研究では問題の可解性を特徴づけるために，真の行列の完全な補完が高い確率で可能になるための十分条件を提案した．この十分条件は，打ち切りの影響を適切に測る理論的な量 [6] を用いて表わされる．この量がある絶対定数より小さければ，トレースノルム最小化と呼ばれるアルゴリズム

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

によって真の行列を高い確率で復元できる（ここで $\mathcal{C} := \{(i,j) \in \Omega : M^{\mathrm{c}}_{ij} = C\}$ であり $\mathcal{P}_A$ は A に含まれない添字の成分を 0 に置き換える作用素）．(なおここでの確率は，観測のランダム性についてのものである．)

3　提案手法

従来の行列補完では，以下の正則化付き最小二乗法の解を復元値とする方法が標準的に用いられてきた [5, 7]：

$$\arg\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} \frac{1}{2}\sum_{ij\in\Omega}(M^{\mathrm{c}}_{ij} - X_{ij})^2 + \mathcal{R}(\mathbf{X}). \qquad (1)$$

ここで $\mathcal{R}(\cdot)$ は解の低ランク性を誘導する正則化項である．しかし打ち切り行列補完においては，打ち切られた観測値に対して二乗損失を用いるのは適切ではない．なぜなら，打ち切られた観測値よ

本稿の内容は Miao Xu 博士 (理研)，佐藤一誠講師 (東大/理研)，杉山将教授 (理研/東大) との共同研究 [6] に基づく．

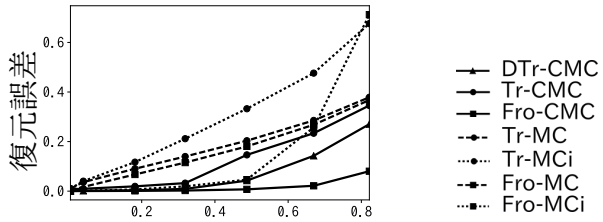

図 3: 人工データ実験の結果．実線は提案法（*-CMC），破線は通常の行列補完手法（*-MC），点線は打ち切られた成分を欠測と扱い通常の補完手法を適用した結果（*-MCi）．

りも真値はかなり大きい可能性があるが，二乗損失を用いると予測値が閾値を上回った場合に罰則が生じるためである．そこで本研究では，打ち切られた観測値に対しては二乗ヒンジ損失を用いることを提案した：

$$\begin{aligned}\arg\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} \frac{1}{2}\sum_{ij\in\mathcal{C}}\max(0, M^{\mathrm{c}}_{ij} - X_{ij})^2 \\ + \frac{1}{2}\sum_{ij\in\Omega\setminus\mathcal{C}}(M^{\mathrm{c}}_{ij} - X_{ij})^2 + \mathcal{R}(\mathbf{X})\end{aligned} \tag{2}$$

ここで $\mathcal{C} := \{(i,j) \in \Omega : M^{\mathrm{c}}_{ij} = C\}$ である．これにより予測値が観測値を上回った場合にも過剰罰則が生じないことが期待される．

正則化項の設計． 本研究では本問題設定に適した二重トレースノルム正則化 (DTr-CMC) を提案し，これに基づく推定量の平均二乗誤差に関する理論解析を行い，その確率的上界を得た．また，従来の行列補完で用いられてきたトレースノルム正則化 (Tr-CMC) およびフロベニウスノルム正則化 (Fro-CMC) はいずれも解を低ランクにする効果を持ち，本問題設定においても有効であると考えられる．いずれも詳細は [6] にある．

4 計算機実験

人工データ実験． 真の値が分かる状況で提案法の復元性能を評価するために，人工データでの実験を行った．人工的に生成した低ランクな行列を真の行列とし，打ち切り閾値を徐々に小さくしながら，復元誤差を評価した（相対平方平均二乗誤差）．図 3 はその結果である．既存法（点線）は打ち切られた成分の割合に応じて復元誤差が拡大しているのに対し，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真の行列を推定できていることが分かる．

実データ実験． 実データ実験では，打ち切り前の真値が得られないため，単純な復元精度の評価が難しい．そこで本研究では評価方法を工夫し，各添字に対して「真値は閾値以上かどうか」を予測する二値分類問題にて提案手法を評価した．評価には映画推薦のベンチマークデータを用いた [3, 4]．表 1 はその結果であり，提案法は対応する（正則化が同一な）通常の行列補完法と比較して性能が向上している．これは，提案法は天井効果の存在による撹乱への頑健性が高いためであると考えられる．

表 1: 実データ実験の結果．ベースラインは常に「閾値以上」クラスに分類する分類器．

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

f_1 値	**Tr-CMC**	Tr-MC	(ベースライン)
Film Trust	0.39 (0.00)	0.35 (0.01)	0.41 (0.00)
Movielens 100K	0.40 (0.00)	0.38 (0.00)	0.35 (0.00)

5 結論と将来の展望

本研究では，打ち切りを受けた低ランクな行列の復元という課題の研究成果を報告した．打ち切りは経験科学に広く見られる情報欠損であり，提案された復元手法は幅広い応用を持つと期待できる．

更なる研究の方向性として，人為的な打ち切りを施すことで復元を不可能にし情報を保護するアルゴリズムを開発すること，トレースノルム最小化による完全復元可能性定理が成立するための必要条件を特徴づけること，また本研究の理論解析の結果を他の種類の情報欠損に拡張することなどが考えられる．

参考文献

[1] P. C. Austin, L. J. Brunner, “Type I Error Inflation in the Presence of a Ceiling Effect”In: *The American Statistician* 57.2 (2003), pp. 97–104.

[2] E. J. Cands, B. Recht, “Exact Matrix Completion via Convex Optimization”In: *Foundations of Computational mathematics* 9.6 (2009), pp. 717–772.

[3] G. Guo, J. Zhang, N. Yorke-Smith, “A Novel Bayesian Similarity Measure for Recommender Systems”In: *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2013, pp. 2619–2625.

[4] F. M. Harper, J. A. Konstan, “The Movielens Datasets: History and Context”In: *ACM Transactions on Interactive Intelligent Systems* 5.4 (Dec. 2015), pp. 1–19.

[5] P. Jain, P. Netrapalli, S. Sanghavi, “Low-Rank Matrix Completion Using Alternating Minimization”In: *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, 2013, pp. 665–674.

[6] T. Teshima, M. Xu, I. Sato, M. Sugiyama, “Clipped Matrix Completion: A Remedy for Ceiling Effects”In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, July 2019, pp. 5151–5158.

[7] K.-C. Toh, S. Yun, “An Accelerated Proximal Gradient Algorithm for Nuclear Norm Regularized Linear Least Squares Problems”In: *Pacific Journal of Optimization* 6 (2010), pp. 615–640.