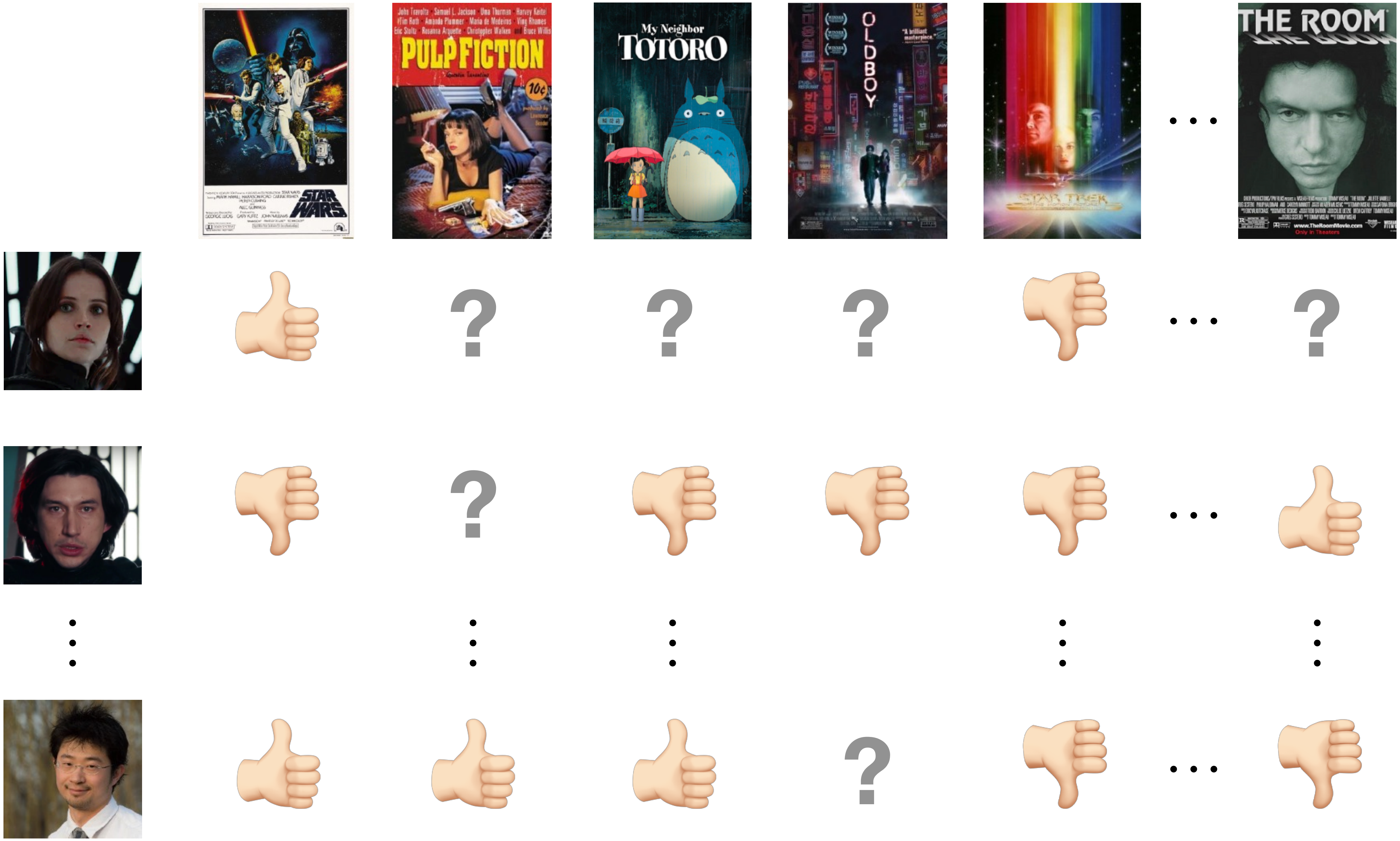
Missing Not at Random in Matrix Completion

The Effectiveness of Estimating Missingness Probabilities Under a Low Nuclear Norm Assumption

Wei Ma* George H. Chen*

*equal contribution & both from Carnegie Mellon University

Matrix Completion



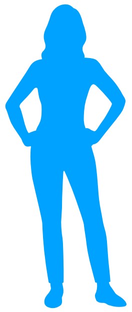
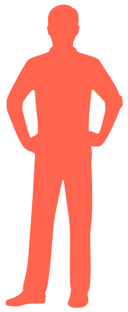
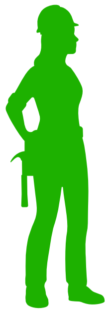
Goal: Fill in question marks (subject to constraints)

Largely popularized by the Netflix Prize (Bennett & Lanning 2007)

Application: Prediction with Missing Values

Feature vectors

Labels to predict

	Gluten allergy	Immuno-suppressant	Low resting heart rate	Irregular heart beat	High BMI	Time of death
	✗	?	✓	✗	?	?
	✓	?	✗	?	✓	?
	✗	✗	?	✓	✗	?

Common approach:

1. Impute missing features with **matrix completion**
2. Use imputed feature vectors to solve prediction task

Missing Not at Random (MNAR) in MC

MNAR: missingness is not uniform at random and can depend on value of entry (if it were forced to be revealed)

- Restaurant ratings: a vegan is unlikely to go to & rate a BBQ restaurant
- Movie ratings: some people refuse to watch horror movies
- Health care: doctor chooses measurements to take for a patient

The vast majority of existing literature on matrix completion assumes entries are missing with equal probability independent of everything else (Candès & Recht 2009, Recht 2009, Cai et al 2010, Keshavan et al 2010, …)

- Many methods rely on this missing-completely-at-random (MCAR) assumption and produce biased predictions when the data are MNAR

This paper: new approach to MNAR matrix completion with
(1) finite sample debiasing guarantees & (2) competitive empirical accuracy

Example of Bias in MC (Steck 2010)

True ratings matrix $S \in \mathbb{R}^{m \times n}$

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} \text{Horror movies} & & & \text{Romance movies} & \end{array} \\
\begin{array}{c} \text{Horror lovers} \\ \\ \\ \text{Romance lovers} \\ \\ \end{array} &
\left[\begin{array}{ccc|cc}
+1 & +1 & +1 & -1 & -1 \\
+1 & +1 & +1 & -1 & -1 \\
+1 & +1 & +1 & -1 & -1 \\
\hline
-1 & -1 & -1 & +1 & +1 \\
-1 & -1 & -1 & +1 & +1
\end{array}\right]
\end{array}
$$

Revealed ratings matrix X

Ω : set of revealed indices

$$
\left[\begin{array}{ccc|cc}
+1 & +1 & ? & ? & ? \\
? & +1 & +1 & ? & ? \\
+1 & ? & +1 & ? & ? \\
\hline
? & ? & ? & +1 & +1 \\
? & ? & ? & +1 & ?
\end{array}\right]
$$

Goal: Given X, construct estimate $\widehat{S}$ of S

Predict all 1's (set $\widehat{S}$ to all 1's)

Ideally, minimize: $$L_{\text{ideal-MSE}}(\widehat{S}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (S_{i,j} - \widehat{S}_{i,j})^2 = 1.92$$

In practice, minimize: $$L_{\text{naive-MSE}}(\widehat{S}) = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} (X_{i,j} - \widehat{S}_{i,j})^2 = 0$$

If every entry revealed with equal probability:

$L_{\text{naive-MSE}}(\widehat{S})$ is unbiased estimate of $L_{\text{ideal-MSE}}(\widehat{S})$

Model

True ratings matrix $S \in \mathbb{R}^{m \times n}$

$$\begin{array}{c|c}
 & \begin{array}{cc} \text{Horror movies} & \text{Romance movies} \end{array} \\
\begin{array}{c} \text{Horror lovers} \\ \\ \text{Romance lovers} \end{array} &
\left[\begin{array}{ccc|cc}
+1 & +1 & +1 & -1 & -1 \\
+1 & +1 & +1 & -1 & -1 \\
+1 & +1 & +1 & -1 & -1 \\
\hline
-1 & -1 & -1 & +1 & +1 \\
-1 & -1 & -1 & +1 & +1
\end{array}\right]
\end{array}$$

Revealed ratings matrix X

Ω : set of revealed indices

$$\left[\begin{array}{ccc|cc}
-1 & -1 & ? & ? & ? \\
? & +1 & -1 & ? & ? \\
+1 & ? & +1 & ? & ? \\
\hline
? & ? & ? & -1 & +1 \\
? & ? & ? & +1 & ?
\end{array}\right]$$

Goal: Given X, construct estimate $\widehat{S}$ of S

Probabilities of entries being revealed $P \in [0,1]^{m \times n}$

$$\left[\begin{array}{ccc|cc}
0.5 & 0.5 & 0.5 & 0.0 & 0.0 \\
0.5 & 0.5 & 0.5 & 0.0 & 0.0 \\
0.5 & 0.5 & 0.5 & 0.0 & 0.0 \\
\hline
0.0 & 0.0 & 0.0 & 0.5 & 0.5 \\
0.0 & 0.0 & 0.0 & 0.5 & 0.5
\end{array}\right]$$

Generative process:

1. Reveal entries of S based on P
2. Add noise to revealed entries

Debiasing MC with Propensity Scores

Goal: Given X, construct estimate $\widehat{S}$ of S

Probabilities of entries being revealed $P \in [0,1]^{m \times n}$

$$\left[\begin{array}{ccc|cc} 0.5 & 0.5 & 0.5 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.5 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.5 & 0.0 & 0.0 \\ \hline 0.0 & 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.5 & 0.5 \end{array}\right]$$

Matrix of *propensity scores*

Will need probabilities > 0

Think of revealing an entry as a "treatment"
(Schnabel et al 2016)

Use inverse propensity score weighting
(Horvitz & Thompson 1952, ...)

1. Construct estimate $\widehat{P}$ of P
2. Minimize:

$$L(\widehat{S}|\widehat{P}) = \frac{1}{mn} \sum_{(i,j)\in\Omega} \frac{(X_{i,j} - \widehat{S}_{i,j})^2}{\widehat{P}_{i,j}}$$

Unbiased estimate of $L_{\text{ideal-MSE}}(\widehat{S})$ if $\widehat{P} = P$

(Other weighting schemes are also possible)

Debiasing MC

1. Construct estimate $\widehat{P}$ of P

Typically done via parametric model (logistic regression, naive Bayes)
(for MC: Liang et al 2016, Schnabel et al 2016, Wang et al 2018/2019, ...)

- Usually requires auxiliary information (on rows/cols, some MCAR data)
- Unclear what error is for estimating propensity scores

2. Solve modified version of standard MC problem:

$$\widehat{S} = \underset{Z \in [-1,1]^{m \times n}}{\operatorname{argmin}} \{L(Z|\widehat{P}) + \lambda \underbrace{\|Z\|_*}_{\text{nuclear norm}}\}$$

Convex program

nuclear norm
(encourages low rank)

where

$$L(Z|\widehat{P}) = \frac{1}{mn} \sum_{(i,j)\in\Omega} \frac{(X_{i,j} - Z_{i,j})^2}{\widehat{P}_{i,j}}$$

Standard approach uses $L_{\text{naive-MSE}}(Z)$ instead of $L(Z|\widehat{P})$ (Mazumder et al 2010)

Debiasing MC

1. Construct estimate $\widehat{P}$ of P

Main contribution: New strategy to estimating $\widehat{P}$ with

- Finite sample bounds for $\|\widehat{P}-P\|_F$ & $|L(\widehat{S}|\widehat{P})-L_{\text{ideal-MSE}}(\widehat{S})|$
- Competitive empirical performance

No auxiliary information on rows or columns needed!

2. Solve modified version of standard MC problem:

$$\widehat{S} = \operatorname*{argmin}_{Z\in[-1,1]^{m\times n}} \{L(Z|\widehat{P}) + \lambda\boxed{\|Z\|_*}\}$$

Convex program

nuclear norm
(encourages low rank)

where

$$L(Z|\widehat{P}) = \frac{1}{mn}\sum_{(i,j)\in\Omega}\frac{(X_{i,j}-Z_{i,j})^2}{\widehat{P}_{i,j}}$$

Standard approach uses $L_{\text{naive-MSE}}(Z)$ instead of $L(Z|\widehat{P})$ (Mazumder et al 2010)

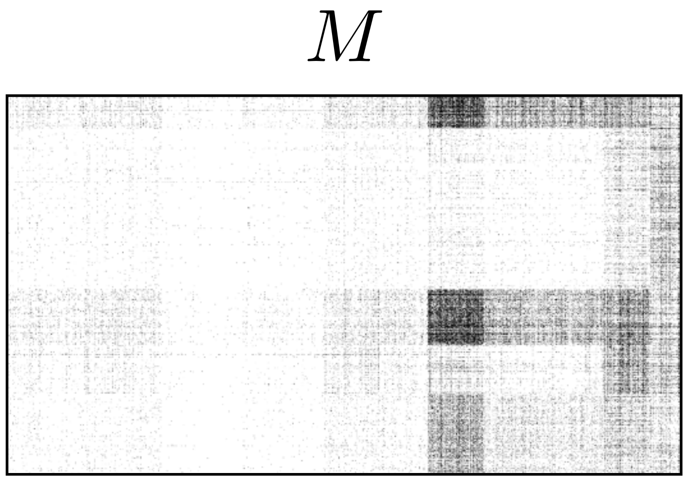
What do missingness patterns look like?

$$\left[\begin{array}{ccc|cc} -1 & -1 & ? & ? & ? \\ ? & +1 & -1 & ? & ? \\ +1 & ? & +1 & ? & ? \\ \hline ? & ? & ? & -1 & +1 \\ ? & ? & ? & +1 & ? \end{array}\right] \xrightarrow[\substack{\text{1 if not missing,} \\ \text{0 if missing}}]{\qquad\qquad\qquad} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

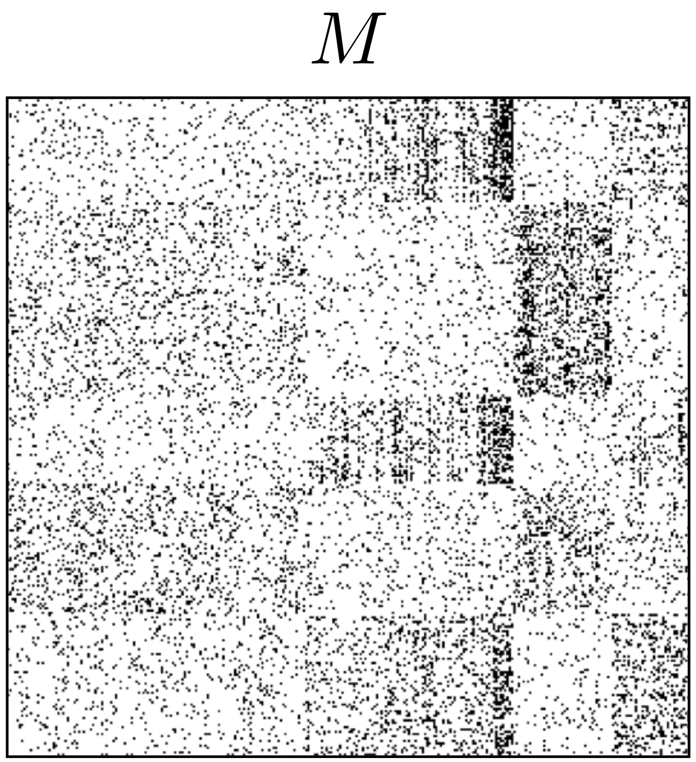
Revealed ratings matrix X

Missingness matrix M

Missingness Matrices in Real Data



MovieLens (Harper and Konstan 2015)



Coat (Schnabel et al 2016)

There is block structure → Low rank

(Can also show that there is *topic modeling* structure) → Low rank

Goal: Given M, estimate P under *low nuclear norm structure* (will in some sense also cover low rank)

General Low Nuclear Norm Structure
(Davenport et al 2014)

Parameterize P with user-specified link function $\sigma : \mathbb{R} \to [0, 1]$

$$P_{i,j} = \sigma(A_{i,j})$$

Example: standard logistic function
$\sigma(x) = 1/(1 + e^{-x})$

for parameter matrix $A \in \mathbb{R}^{m \times n}$

Idea: impose constraints on A instead of P
(helpful for theoretical analysis)

Assumption **A1**: There exists $\theta > 0$ s.t. $\|A\|_* \leq \theta\sqrt{mn}$ (low nuclear norm)

Assumption **A2**: There exists $\alpha > 0$ s.t. $\max_{i,j} |A_{i,j}| \leq \alpha$

(bounded probabilities $P_{i,j} \in [\sigma(-\alpha), \boxed{\sigma(\alpha)}])$

Block structure, clustering, topic models are all special cases!

Any bounded low rank A satisfies **A1** and **A2**

Technical detail: with some changes to theory & algorithm, can make upper bound 1

Algorithm: 1bitMC (Davenport et al 2014)

1. Solve a nuclear-norm-regularized maximum likelihood estimation problem:

standard Bernoulli log likelihood

$$\widehat{A} = \underset{A \in \mathbb{R}^{m \times n}}{\text{argmax:}} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} \boxed{M_{i,j} \log \sigma(A_{i,j}) + (1 - M_{i,j}) \log(1 - \sigma(A_{i,j}))}$$

$$\text{subject to:} \quad \boxed{\|A\|_* \leq \theta\sqrt{mn}, \ \max_{i,j} |A_{i,j}| \leq \alpha}$$

constraints correspond to Assumptions **A1** & **A2**

Convex program depending on choice of σ

2. Estimate propensity scores as follows:

$$\widehat{P}_{i,j} = \sigma(\widehat{A}_{i,j})$$

Davenport et al developed this algorithm for *binary matrix completion with MCAR entries*

Key idea: apply *matrix completion* algorithm to *fully-observed* matrix M to estimate P

We are debiasing matrix completion *with more matrix completion*!

Can also use other algorithms designed for matrix completion aside from 1bitMC to estimate P, such as collaborative filtering

(Technically, we are doing matrix *denoising* not matrix completion for M)

Theoretical Guarantees

Theorem (1bitMC): Choose link $\sigma(x) = 1/(1+e^{-x})$.
Under assumptions **A1** and **A2**, if # rows m & # cols n are large enough, then with high probability, we simultaneously have:

$$\frac{1}{mn}\sum_{i,j}(\widehat{P}_{i,j} - P_{i,j})^2 \leq \mathcal{O}\Big(\theta\Big[\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\Big]\Big)$$

$$|L(\widehat{S}|\widehat{P}) - L_{\text{ideal-MSE}}(\widehat{S})| \leq \mathcal{O}\Big(\frac{\sqrt{\theta}}{[\sigma(-\alpha)]^2}\boxed{\Big[\frac{1}{m^{1/4}} + \frac{1}{n^{1/4}}\Big]}\Big)$$

$m^{-1/4}$
if
$m \asymp n$

Theorem (CF): If there's *clustering structure* (across rows/columns), we can get faster debiasing rate $m^{-1/2}$ instead of $m^{-1/4}$ (uses collaborative filtering to estimate P instead of 1bitMC)

(The collaborative filtering results are in a forthcoming longer version of the paper)

Matrix Completion (MovieLens, Coat)

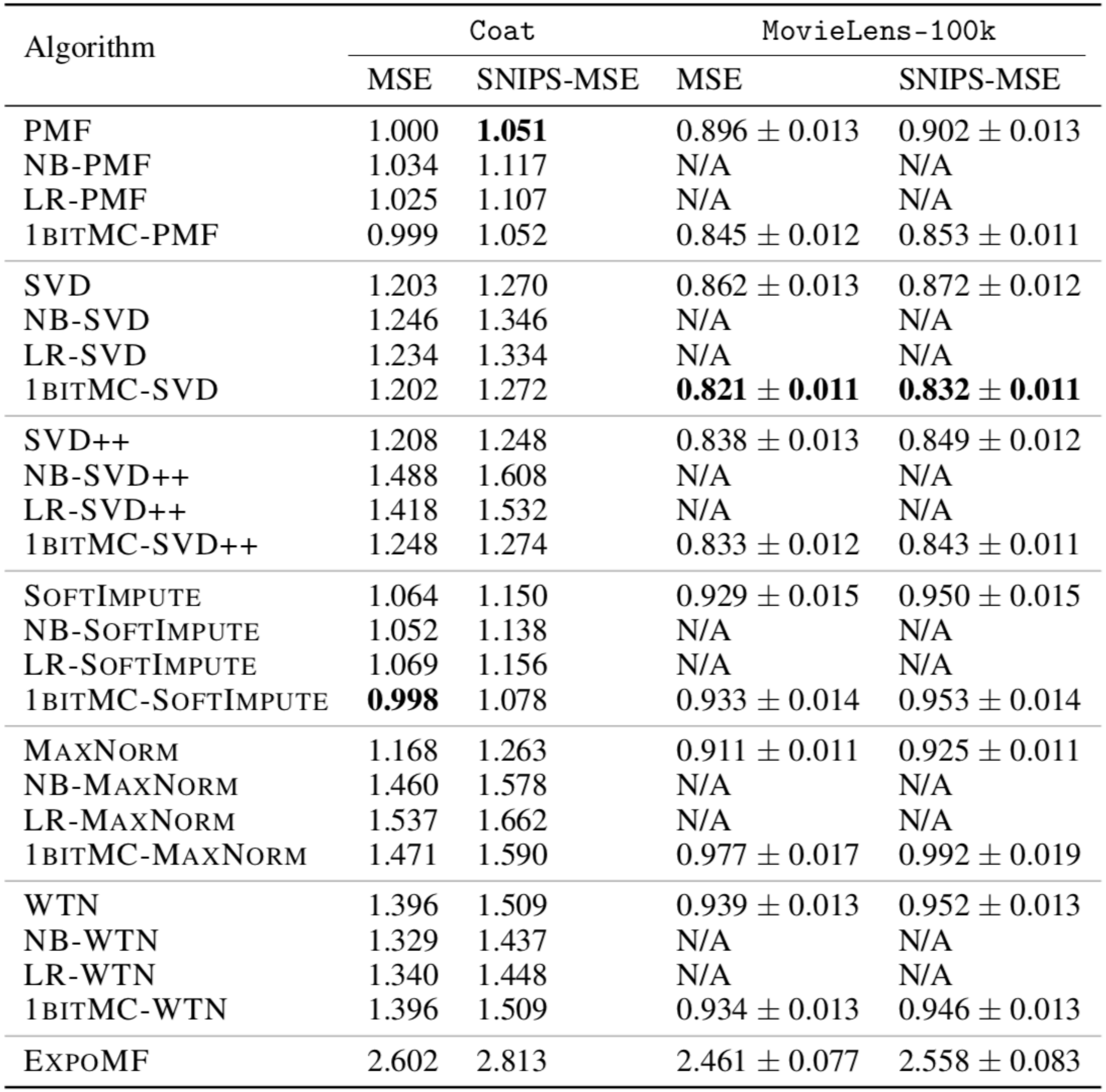
Experiment (per dataset):

- Separate revealed entries into train/test split
- 5-fold cross-validation for hyperparameter selection
- Evaluate prediction error on test entries

Coat has its own train/test split
MovieLens: 90/10 split with 10 experimental repeats

Main findings:

- 1bitMC debiasing tends to outperform naive Bayes and logistic regression debiasing
- 1bitMC debiasing often improves existing methods, at times yielding the best or nearly the best accuracies



Algorithm	Coat		MovieLens-100k	
	MSE	SNIPS-MSE	MSE	SNIPS-MSE
PMF	1.000	**1.051**	0.896 ± 0.013	0.902 ± 0.013
NB-PMF	1.034	1.117	N/A	N/A
LR-PMF	1.025	1.107	N/A	N/A
1bitMC-PMF	0.999	1.052	0.845 ± 0.012	0.853 ± 0.011
SVD	1.203	1.270	0.862 ± 0.013	0.872 ± 0.012
NB-SVD	1.246	1.346	N/A	N/A
LR-SVD	1.234	1.334	N/A	N/A
1bitMC-SVD	1.202	1.272	**0.821 ± 0.011**	**0.832 ± 0.011**
SVD++	1.208	1.248	0.838 ± 0.013	0.849 ± 0.012
NB-SVD++	1.488	1.608	N/A	N/A
LR-SVD++	1.418	1.532	N/A	N/A
1bitMC-SVD++	1.248	1.274	0.833 ± 0.012	0.843 ± 0.011
SoftImpute	1.064	1.150	0.929 ± 0.015	0.950 ± 0.015
NB-SoftImpute	1.052	1.138	N/A	N/A
LR-SoftImpute	1.069	1.156	N/A	N/A
1bitMC-SoftImpute	**0.998**	1.078	0.933 ± 0.014	0.953 ± 0.014
MaxNorm	1.168	1.263	0.911 ± 0.011	0.925 ± 0.011
NB-MaxNorm	1.460	1.578	N/A	N/A
LR-MaxNorm	1.537	1.662	N/A	N/A
1bitMC-MaxNorm	1.471	1.590	0.977 ± 0.017	0.992 ± 0.019
WTN	1.396	1.509	0.939 ± 0.013	0.952 ± 0.013
NB-WTN	1.329	1.437	N/A	N/A
LR-WTN	1.340	1.448	N/A	N/A
1bitMC-WTN	1.396	1.509	0.934 ± 0.013	0.946 ± 0.013
ExpoMF	2.602	2.813	2.461 ± 0.077	2.558 ± 0.083

Our paper also has results on synthetic data

Conclusions & Future Work

Main takeaways:

- We recommend using 1bitMC to estimate propensity scores if:
    1. You don't want parametric assumptions
    2. Your data matrix is sufficiently large (e.g., at least hundreds of rows/cols)
- Other MNAR matrix completion methods lack debiasing guarantees; some do not estimate the propensity score matrix (possibly useful for other tasks)

Future directions:

- More robust way to debias MC using propensity score estimates (that neatly handles propensity scores that are 0)
- Handling the case entries are not revealed independently (revealing one entry makes another more/less likely to be revealed)
- Debiasing guarantees for prediction tasks using MNAR imputed features
- Any benefits to using this approach in causal reasoning context?