Carnegie Mellon University
Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

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)

Code: https://github.com/georgehc/mnar_mc

Introduction

MNAR: probability of entry being missing is unknown (can relate to the entry's value)

- 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, Cai et al 2010, Keshavan et al 2010a/b, Recht 2011, Chatterjee 2015, …)

- 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

Debiasing Matrix Completion

Bias in matrix completion, illustrated using an example by Steck (2010)

True ratings matrix $S \in \mathbb{R}^{m \times n}$ (Horror movies | Romance movies)

Horror lovers	+1	+1	+1	−1	−1
	+1	+1	+1	−1	−1
	+1	+1	+1	−1	−1
Romance lovers	−1	−1	−1	+1	+1
	−1	−1	−1	+1	+1

Revealed ratings matrix X (Ω : set of revealed indices)

+1	+1	?	?	?	
?	+1	+1	?	?	
+1	?	+1	?	?	
?	?	?	+1	+1	
?	?	?	+1	?	

Goal: Given X, construct estimate $\widehat{S}$ of S — Predict all 1's (set $\widehat{S}$ to all 1's)

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

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

If every entry revealed with equal probability: $L_{\text{naive-MSE}}(\widehat{S})$ is unbiased estimate of $L_{\text{ideal-MSE}}(\widehat{S})$

Model (heterogeneous missingness probabilities)

Assume that there is an unknown *propensity score matrix* $P \in [0,1]^{m \times n}$

1. Reveal entry (i,j) of S with probability $P_{i,j}$ independent of everything else
2. Add noise to revealed entries to obtain X

Debias matrix completion using inverse probability weighting (Schnabel et al 2016)

1. Somehow construct estimate $\widehat{P}$ of P
2. For user-specified matrix completion algorithm, minimize debiased loss:

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

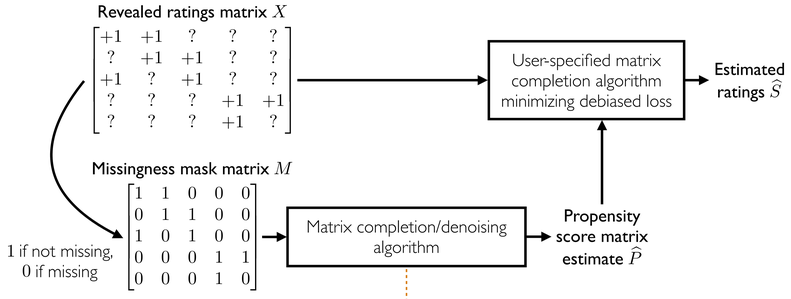
Proposed Approach: Debias Matrix Completion Using More Matrix Completion

Problem: Propensity score matrix P is usually estimated using logistic regression or naive Bayes (Liang et al 2016, Schnabel et al 2016, Wang et al 2018/2019, …)

- Usually requires auxiliary information (e.g., row/column features, some missing-at-random ratings)
- Unclear what error is for estimating propensity scores

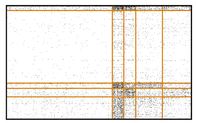
Strategy: Estimate propensity score matrix P using matrix completion/denoising algorithm instead

No auxiliary information needed, and we get finite sample bounds for $\|\widehat{P} - P\|_F$ & $|L(\widehat{S}|\widehat{P}) - L_{\text{ideal-MSE}}(\widehat{S})|$

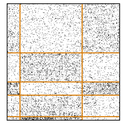


We use the *1-bit matrix completion* (1bitMC) algorithm by Davenport et al (2014), which relies on low nuclear norm structure in P

Why should the proposed strategy work? In real data, missingness mask matrix M is low rank



MovieLens-100k (Harper and Konstan 2015)



Coat (Schnabel et al 2016)

Possible explanation: M is generated from low rank P

Algorithm 1bitMC (Davenport et al 2014). Given M, constructs estimate for P as follows.

Step 1. Solve Bernoulli maximum likelihood problem with constraints:

$$\widehat{A} = \operatorname*{argmax}_{\widetilde{A}\in\mathbb{R}^{m\times n}} \sum_{i,j}[M_{i,j}\log\sigma(\widetilde{A}_{i,j}) + (1-M_{i,j})\log(1-\sigma(\widetilde{A}_{i,j}))]$$

$$\text{subject to: } \|\widetilde{A}\|_* \le \theta\sqrt{mn},\ \max_{i,j}|\widetilde{A}_{i,j}| \le \alpha$$

Step 2. Compute $\widehat{P}_{i,j} = \sigma(\widehat{A}_{i,j})$

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

Results

Theoretical analysis

Assumptions:

- **General low nuclear norm structure** (Davenport et al 2014):
  There is a true parameter matrix $A \in \mathbb{R}^{m\times n}$, with $P_{i,j} = \sigma(A_{i,j})$, satisfying:
  - *Low nuclear norm.* There exists $\theta > 0$ s.t. $\|A\|_* \le \theta\sqrt{mn}$
  - *Bounded probabilities.* There exists $\alpha > 0$ s.t. $\max_{i,j}|A_{i,j}| \le \alpha$ (i.e., $P_{i,j} \in [\sigma(-\alpha), \sigma(\alpha)]$)

  Any bounded low rank A satisfies these
- **Bounded ratings.** The true ratings S and estimated ratings $\widehat{S}$ are bounded in entry-wise max norm

Theorem. Using 1bitMC to estimate P, for large enough m and n, w.h.p. we have:

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

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

Longer version of paper (forthcoming): faster debiasing $m^{-1/2}$ is possible with stronger structure (clustering) using collaborative filtering algorithm to estimate P

Numerical experiments

Experiment (per dataset):

- Separate revealed entries into train/test split
  - MovieLens: 90/10 split with 10 experimental repeats
  - Coat comes with its own train/test split
- 5-fold cross-validation for hyperparameter selection
- Evaluate prediction error on test entries

Main findings:

- Our proposed strategy debiases better than naive Bayes & logistic regression baselines
- Our debiased matrix completion methods can achieve the best prediction accuracy per dataset

See paper for experiments on synthetic data

Test Set Mean Squared Error

Algorithm	MovieLens-100k	Coat
PMF (Mnih & Salakhutdinov 2008)	0.896 ± 0.013	1.000
NB-PMF	N/A	1.034
LR-PMF	N/A	1.025
Our proposed debiased PMF	0.845 ± 0.012	0.999
SVD (Funk 2006)	0.862 ± 0.013	1.203
NB-SVD	N/A	1.246
LR-SVD	N/A	1.234
Our proposed debiased SVD	**0.821 ± 0.011**	1.202
SVD++ (Koren 2008)	0.838 ± 0.013	1.208
NB-SVD++	N/A	1.488
LR-SVD++	N/A	1.418
Our proposed debiased SVD++	0.833 ± 0.012	1.248
SoftImpute (Mazumder et al 2010)	0.929 ± 0.015	1.064
NB-SoftImpute	N/A	1.052
LR-SoftImpute	N/A	1.069
Our proposed debiased SoftImpute	0.933 ± 0.014	**0.998**
MaxNorm (Cai & Zhou 2016)	0.911 ± 0.011	1.168
WTN (Srebro & Salakhutdinov 2010)	0.939 ± 0.013	1.396
ExpoMF (Liang et al 2016)	2.461 ± 0.077	2.602

Baselines that already handle entries missing with different probabilities

Code available at https://github.com/georgehc/mnar_mc