A Comparative Study of Matrix Completion for Different Missing Data Patterns

Ningyuan Huang (nh1724), Yi Li (liy31)

***Abstract*—Matrix completion and imputation is a common problem and an important cornerstone in data science applications. We compare the performance of three matrix imputation methods (SoftImpute, SoftImpute-concat, Multiple Imputation) on different types of missing data patterns using both synthetic and real datasets. Although no method is found to perform significantly better under all circumstances, SoftImpute-concat indeed outperforms original SoftImpute when missing is not at random (NMAR) on our data. Besides, SoftImpute-concat has more robust performance than SoftImpute when missing rate increases. However, no theoretical proof of SoftImpute-concat superior performance under NMAR setting is given so far. Therefore, we discuss possible reasons based on a toy example.**

I. Introduction

Missing data is commonly present in data science research and applications. Most state-of-art imputation methods rely on the assumption that missing data is random. In reality, this assumption might not hold and recent studies have proposed approaches to jointly estimate the observed data and missing mechanism [1] [2]. Our study aims to compare the performance of various matrix imputation methods on different types of missing data structures. Rubin [3] introduced three types of missing-data mechanisms: (i) missing completely at random (MCAR), where no systematic differences between observed and missing values (ii) missing at random (MAR), where the missing data may only depend on the observable data, and (iii) not missing at random (NMAR), when the missing data depends on other variables and its own value. Popular matrix imputation methods usually develop their theoretical guarantees based on MCAR or MAR assumption, but this could lead to biased estimator if the real data is NMAR [2]. Thus, our goal is to compare and recommend the best method for different types of missing data, based on prediction accuracy, missing percentage and data nature.

II. State of the art

In our study, we compare three methods: Multiple Imputation [4], nuclear-norm minimization [5] [6], implicit joint modeling of observed and missing data [1]. The first two methods assume MAR data while the third method is specifically designed for NMAR data. In this section, we describe their algorithm, assumption and implication.

A. Multiple Imputation ("MI")

Multiple Imputation estimates the incomplete matrix several times separately and uses the average imputed values to fill the missing entries. For each imputation, a series of regression models are run whereby each incomplete variable is modeled conditional upon the other variables in the data [4]. Note that the missing entries are filled by an initial guess and subsequently updated at each iteration by the regression result. It assumes MAR but extensions are proposed to cater to MCAR or NMAR scenario [7]. It doesn't have a low-rank assumption and performs well for mixed input matrices. Since it is an iterative procedure on a per variable-basis, its convergence rate is slow compared to matrix spectral methods. Note that Multiple Imputation is a popular method for missing data in clinical research so it is chosen as our benchmark.

B. Nuclear-norm minimization ("SoftImpute")

Candès and Tao [5] showed that nuclear-norm minimization can achieve exact matrix recovery with high probability, under the assumption of MCAR/MAR. Inspired by this, Mazumder et.al [6] proposed an algorithm SoftImpute to iteratively estimate the underlying matrix by thresholding its

singular values. It solves the following minimization problem:

$$\min_M \frac{1}{2}||P_\Omega(Y-M)||_F^2 + \lambda||M||_*$$

where Y is the original incomplete matrix, $P_\Omega(X)$ is the projection of X with the observed elements preserved, and the missing entries replaced with 0. The algorithm follows two steps iterated until convergence:

1) Replace the missing values in Y with the corresponding entries in current estimated $\hat{M}$

$$\hat{Y} \leftarrow P_\Omega(Y) + P_\Omega^\perp(\hat{M})$$

2) Update $\hat{M}$ by computing the soft-threshold SVD of Y

$$\begin{aligned} \hat{Y} &= USV^T \\ \hat{M} &\leftarrow U\mathcal{S}_\lambda V^T \end{aligned} \tag{1}$$

where the $\mathcal{S}_\lambda$ sets all singular values less than λ to 0 and replace all other singular values with $(D_i - \lambda)_+$.

A crucial assumption of SoftImpute is that the underlying matrix is low-rank. Compared to MI, SoftImpute is significantly faster with similar accuracy performance. Therefore it is widely used on huge sparse matrix like movie ratings [8].

C. Implicit joint modeling of observed and missing data ("SoftImpute-concat")

Motivated to solve matrix completion under NMAR setting, Sportisse et.al [1] proposed a method to implicitly model the joint distribution between observed data and the missing mechanism. They suggested to concatenate the original data matrix with its missing-data mask (i.e., the binary encoding matrix to indicate missing entries), and then use SoftImpute or other state-of-art methods to impute this new input matrix. They assumed that this concatenated matrix is low-rank, so as to effectively model the relationship between variables and missing mechanism. They used simulated datasets to demonstrate that this SoftImpute-concat method has slightly better performance than traditional SoftImpute. Yet no theoretical proof has been given to testify this low-rank assumption, and no real-world datasets are used to verify this method.

III. Methodology

We compared three methods discussed above on simulated data and real-world datasets (movie rating and clinical trial data).

A. Datasets

Our first synthetic movie rating dataset is obtained by sampling from a matrix factorization model, similar to the strategy of Hernandez-Lobato et.al [2]. We generate two 100 x 10 matrices U, V, with standard Gaussian i.i.d entries. We then obtain $C = UV^T$ which is a rank-10 matrix, and partition its values in continuous interval $[-\infty, -6, -2, 2, 6, \infty]$. For each C_{ij}, we create a value Y_{ij} from $\{1, 2, 3, 4, 5\}$ corresponding to the interval C_{ij} lies. Y is thus our simulated fully-observed movie rating matrix with approximately rank=10.

To generate a MCAR pattern, we randomly sample entries in Y with equal probability and set them to be missing. To simulate MAR, note that the missing pattern should only depend on the observed entries. So our strategy is to calculate average movie ratings based on the first 10% users, and classify movies into low-rating versus high-rating. We then sample the remaining 90% user ratings, with higher missing probability for low-rating movies and vice versa for high-rating movies. We expect MAR has a similar missing pattern as MCAR. To simulate NMAR, our assumption is high-rating movies are rated more times than low-rating movies. Its missingness is encoded in hidden factors or missing values themselves. So we set the entry to missing with probability $\sigma(e_i f_j^T + \sum_{l=1}^5 z_i I[r_{ij} = 1])$, where σ is the logistic function, e_i and f_j are the i-th and j-th rows of U and V, $(z1, ..., z5) = \{-3, -3, -3, 1, 1\}$. Note that e_i and f_j represent the latent factors and z_i is the effect of missing values magnitude. Thus we expect to see more ratings with value 4 and 5. The rating distribution for MCAR, MAR and NMAR pattern is shown below, which is in line with our expectation.

We also consider two real-world datasets. The MovieLens small dataset [9] includes 100,000 ratings (ranging from 0.5 to 5) of 9000 movies by 600 users, from which we create two subsets consist of

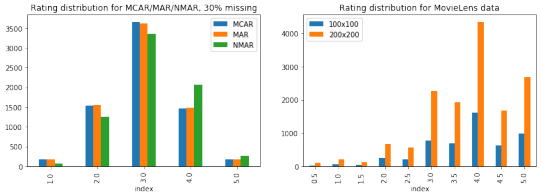

Fig. 1. Rating distribution: synthetic data and MovieLens data

top movies rated by top active users (100 x 100, 200 x 200). Their rating distributions illustrate the NMAR pattern, which are skewed to the higher ratings (Figure 1).

The third data set is from EMBARC study, an 8-week randomized placebo-controlled clinical trial of sertraline enrolled about 300 patients with Major Depressive Disorder.[10] Baseline data set includes assessments of brain structure, function and connectivity(fMRI) along with electrophysiological(EEG), biological, behavioral and clinical features(Clinical) from 287 subjects. Three data sets (Clinical, EEG and fMRI) have different size and proportion of data missing (Table 1). For each data set, we hold a subset of fully observed samples to generate different missing patterns for method comparison. To generate MCAR data, we simply sample missing entries for each variable randomly with the same missing rate as the full data has. For MAR data, we generate missing entries for each variable randomly but with higher missing rate within treatment group (SER/CIT) compared to the placebo group (PLA). As in clinical trials, there might be different missing rate within different treatment groups.[11] Here we assume higher missing rate due to the side effect of treatment. Note that the treatment group indicator is also an input column of the matrix and thus satisfies the MAR assumption. Finally, we set larger values of each variable to be missing to generate NMAR pattern, as data is more likely to be missing for extreme values in medical research [11]. As the second part of our analysis on empirical data, we compare the performance of 3 methods with different missing rates on fMRI data under 3 missing mechanism assumptions.

TABLE I
Missing characteristics of EMBARC data

Data	Max(mean)	MCAR	MAR	NMAR
Clinical (240,32)	10%(1%)	5%	5% PLA 10% TRT	Top 5%
EEG (213,16)	21%(12%)	20%	15% PLA 25% TRT	Top 20%
fMRI (146, 208)	40%(31%)	40%	35% PLA 45% TRT	Top 40%

B. Experiment Protocol

We evaluate these methods based on their imputation accuracy on the ground truth entries in synthetic data and EMBARC fully observed subset, while using a separate hold-out test set on the MoviesLens data. The accuracy is measured by normalized test error [6]:

$$\frac{||P_{\Omega}^{\perp}(Y-M)||_F^2}{||P_{\Omega}^{\perp}(Y)||^2}$$

Note that for the synthetic rating data and EMBARC subset, we can explicitly compare the performance of 3 methods under 3 missing patterns. For the MovieLens data, we cannot explicitly model the missing mechanism as the original sample has missing entries, so we assume an NMAR pattern based on its actual rating distribution and cross-validate with the performance result from 3 methods.

IV. Results

When applying Multiple Imputation for all datasets, we set iteration as 5 cycles for each single imputation and use the average from 10 imputation multiples as the imputed result. For SoftImpute and SoftImpute-concat, we double-normalize the input matrix and set the maximum rank based on the SVD result if given fully observed data, or vary the rank to test the robustness of the method.

For the synthetic data, it is approximately rank-10 from its SVD (6) so we set maximum rank as 10 in SoftImpute and SoftImpute-concat. We vary the missing percentage ranging from 10% to 30%.

The results show that MCAR/MAR are very similar, due to their similar missing data distribution. SoftImpute achieves the lowest test errors for all different missing rates, while Multiple Imputation

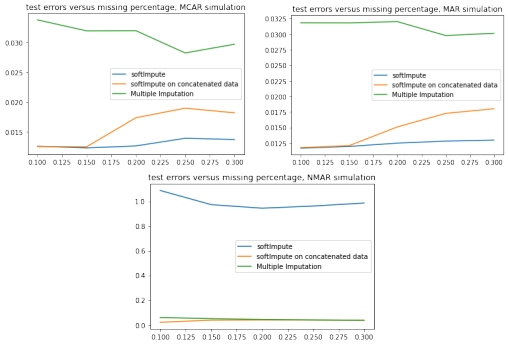

Fig. 2. Test errors of 3 methods on synthetic data under different missing mechanisms

performs the worst. However, in NMAR setting, SoftImpute has the highest test errors, while the other two methods give significantly better results.

For the real movie rating datasets, recall we observe their empirical NMAR pattern, so our hypothesis is that SoftImpute-concat would outperform MI and SoftImpute, under low-rank assumption on movie rating data. The test errors on different low-rank models confirm our hypothesis, which is also consistent with the findings from Sportisse et.al [1].

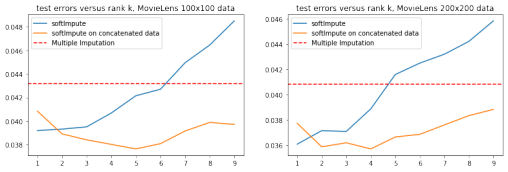

Fig. 3. Test errors of 3 methods on MovieLens data under different missing mechanisms

From the SVD of three EMBARC datasets(7), we can observe that all of them satisfy the low-rank assumption, so we set maximum rank to be 5 for SoftImpute and SoftImpute-concat. As figure 3 shows, test errors increase for all three methods on all three datasets when the missingness changed from completely random to not random. Although SoftImpute-concat looks quite similar to without concatenation in clinical data, it indeed outperforms SoftImpute when missingness is not completely random on EEG and fMRI data. However, we

TABLE II
Test errors of 3 imputation methods under different missing mechanisms on EMBARC data

Data	Method	MCAR	MAR	NMAR
Clinical	SoftImpute	0.1144	0.4325	0.4834
Clinical	SoftImpute-Concat	0.1143	0.4345	0.4823
Clinical	Multiple Imputation	0.1831	0.4186	0.4986
EEG	SoftImpute	0.0551	0.0535	0.2700
EEG	SoftImpute-Concat	0.0593	0.0556	0.2101
EEG	Multiple Imputation	0.0492	0.0487	0.2970
fMRI	SoftImpute	0.0832	0.1286	0.1847
fMRI	SoftImpute-Concat	0.0843	0.1015	0.1871
fMRI	Multiple Imputation	0.0972	0.1047	0.1858

observe no significant outperformance of SoftImpute and SoftImpute-concat compared to Multiple Imputation.

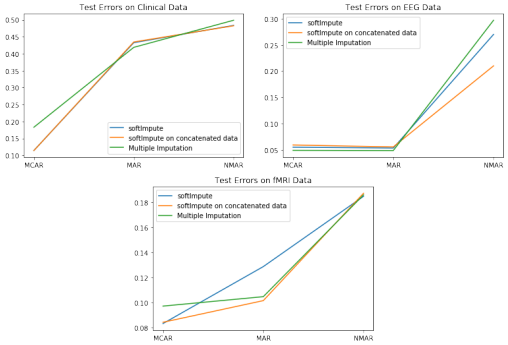

Fig. 4. Test errors of 3 methods on EMBARC data under different missing mechanisms

From the analysis of performance under different missing rates on fMRI data (Figure 5), we can see that the performance of all three methods become worse when missing rate increases under MCAR and NMAR. However, the performance of SoftImpute-concat and Multiple Imputation is quite stable under MAR. Besides, under all 3 assumptions about missingness, SoftImpute-concat outperforms SoftImpute, it even performs better than Multiple Imputation under MAR and NMAR with significantly faster convergence.

V. Discussion

From the empirical results above, it seems that SoftImpute-concat generally works better than Soft-

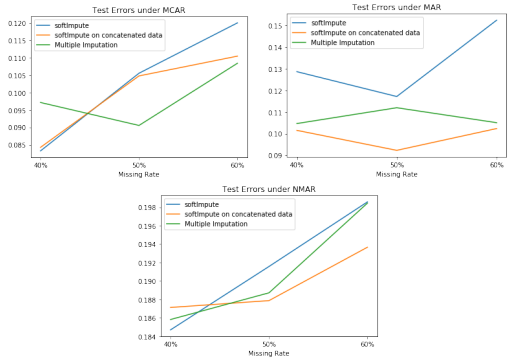

Fig. 5. Test errors of 3 methods under different missing rate generated on fMRI data.

Impute under NMAR scenario. Intuitively, if missingness is not at random, encoding this information in the matrix is helpful if the missing mechanism is an important latent factor. In the machine learning community, it is a common practice to add missing entry indicator columns as auxiliary features for the prediction task. We demonstrate the effect of concatenating missing mask using a simple example.

Consider the following 5x4 matrix and its SVD:

$$M = \begin{bmatrix} 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 3 & -3 & 3 & -3 \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 8 & 0 \\ 0 & 6 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 \end{bmatrix}$$

We can decompose M into two rank-1 matrices. Notice that the second left singular vector is 1-sparse, which captures all information of the last row. If we have any missing entries in the last row, it is impossible to exactly recover the matrix. Now suppose M has one missing entry at each row, and consider concatenate M with its missing mask (1 indicates missing entry):

$$M_{con} = \begin{bmatrix} 2 & 2 & 2 & 2 & 1 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 & 1 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 & 0 & 1 & 0 \\ 2 & 2 & 2 & 2 & 0 & 0 & 0 & 1 \\ 3 & -3 & 3 & -3 & 0 & 1 & 0 & 0 \end{bmatrix}; U = \begin{bmatrix} -0.50 & 0.02 & -0.00 & 0.82 & -0.29 \\ -0.50 & -0.01 & 0.00 & 0.00 & 0.87 \\ -0.50 & 0.02 & -0.71 & -0.41 & -0.29 \\ -0.50 & 0.02 & 0.71 & -0.41 & -0.29 \\ -0.02 & -1.00 & -0.00 & 0.00 & -0.02 \end{bmatrix}$$

The SVD of $M_{con} = UDV^T$ has five non-zero singular values $\{8.06, 6.08, 1, 1, 0.99\}$ (while the first two are significantly larger so we can still approximate it with a low-rank model). Inspect the left singular vectors in U and none of them are sparse. The second left singular vector $U_{:2}$ still places some weights on the first four rows. So even if we have missing entries in the last row, we can improve the imputation accuracy if the missingness is controlled by latent factors applied to the whole matrix. In other words, the missing mask helps to spread out the extremely localized information in the problematic sparse singular vectors.

To conclude, adding the missing mask in SoftImpute-concat achieves better imputation performance when missing at random assumption is violated. Multiple Imputation is popular due to its robust performance and no assumption about low-rank data structure, so it is considered as a benchmark in medical data research. Moreover, although Multiple Imputation typically assumes missing at random, it still performs better than SoftImpute under NMAR. However, the lowest imputation error is observed for SoftImpute-concat when the missing rate is higher and missing is not at random. One thing to notice is that in terms of execution time, SoftImpute and SoftImpute-concat are much faster than Multiple Imputation, especially for larger data sets. Therefore, we recommend the following strategy for matrix completion under different scenarios:

TABLE III
Matrix completion method comparison: The number of + indicates the performance, from weak (+) to strong (+++)

Methods	**Low-rank**	**NMAR**	**High missing rate**
SoftImpute	++	+	+
SoftImpute-concat	++	+++	++
Multiple Imputation	+	++	++

As for directions of future work, it would be important to provide a theoretical guarantee for SoftImpute-concat. One could evaluate these methods for matrices with mixed data types, longitudinal data, and experiment with larger sparse matrix. In addition, it could be interesting to derive the confidence interval of the estimated missing values and investigate the robustness of imputation methods.

REFERENCES

[1] A. Sportisse, C. Boyer, and J. Josse, "Imputation and low-rank estimation with Missing Non At Random data," pp. 1–29, 2018. [Online]. Available: http://arxiv.org/abs/1812.11409 I, II, II-C, IV

[2] N. H. Hernández-Lobato, José Miguel and Z. Ghahramani, "Probabilistic Matrix Factorization with Non-random Missing Data," 2014. I, III-A

[3] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976. I

[4] S. van Buuren and K. Groothuis-oudshoorn, "MICE : Multivariate Imputation by Chained Equation in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. [Online]. Available: http://www.jstatsoft.org/v45/i03 II, II-A

[5] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010. II, II-B

[6] T. H. Rahul Mazumder and R. Tibshirani., "Spectral regularization algorithms for learning large incomplete matrices," p. 11:2287â2322, 2010. II, II-B, III-B

[7] D. M. Tompsett, F. Leacy, M. Moreno-Betancur, J. Heron, and I. R. White, "On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice," *Statistics in Medicine*, vol. 37, no. 15, pp. 2338–2353, 2018. II-A

[8] T. Hastie, R. Mazumder, J. Lee, and R. Zadeh, "Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares," pp. 1–41, 2014. [Online]. Available: http://arxiv.org/abs/1410.2596 II-B

[9] F. M. H. Konstan. and J. A., "The MovieLens Datasets: History and Context," 2015. III-A

[10] M. H. Trivedi, P. J. McGrath, M. Fava, R. V. Parsey, B. T. Kurian, M. L. Phillips, M. A. Oquendo, G. Bruder, D. Pizzagalli, M. Toups, C. Cooper, P. Adams, S. Weyandt, D. W. Morris, B. D. Grannemann, R. T. Ogden, R. Buckner, M. McInnis, H. C. Kraemer, E. Petkova, T. J. Carmody, and M. M. Weissman, "Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design," *Journal of Psychiatric Research*, vol. 78, pp. 11–23, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jpsychires.2016.03.001 III-A

[11] S. Kaushal, "Missing data in clinical trials: Pitfalls and remedies," *International journal of applied & basic medical research*, vol. 4, no. Suppl 1, pp. 6–7, 2014. III-A

APPENDIX A
LOW-RANK ASSUMPTION CHECK ON REAL DATA

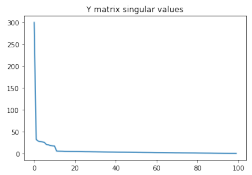

Fig. 6. singular values of synthetic data

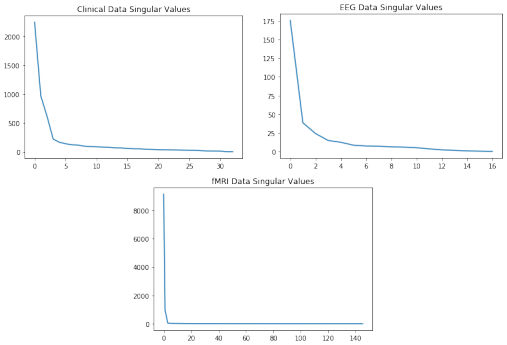

Fig. 7. singular values of EMBARC data