A Comparative Study of Matrix Completion for Different Missing Data Patterns

Ningyuan Huang
Yi Li

Missing data mechanism

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Not missing at random (NMAR)

Evaluation: test error

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

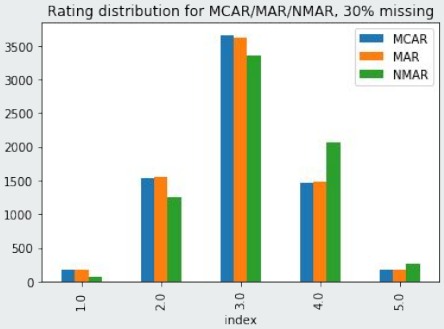
State-of-art methods

- Multiple Imputation
- Nuclear-norm minimization (SoftImpute)
  - Objective: $\min_M \frac{1}{2}||P_\Omega(Y-M)||_F^2 + \lambda||M||_*$
  - Algorithm: $\hat{Y} \leftarrow P_\Omega(Y) + P_\Omega^\perp(\hat{M})$

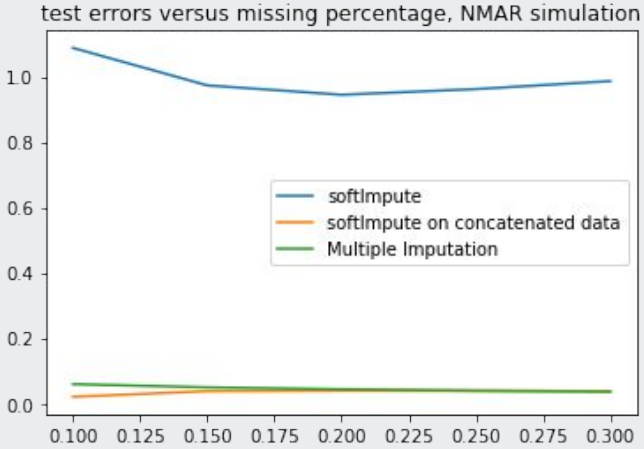
    $\hat{Y} = USV^T$

    $\hat{M} \leftarrow U\mathcal{S}_\lambda V^T$
- Implicit joint modeling (SoftImpute-concat)

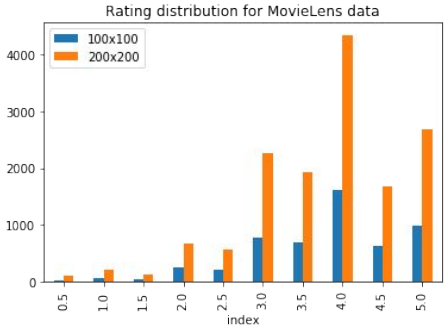
Synthetic rating data



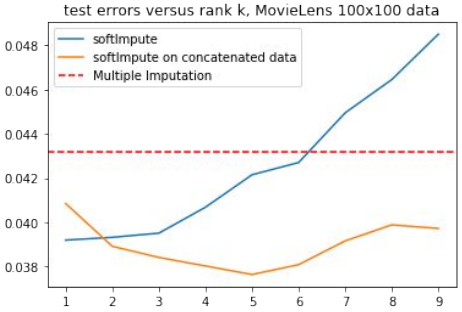
- 100 x 100
- True rank = 10
- Simulated from matrix factorization

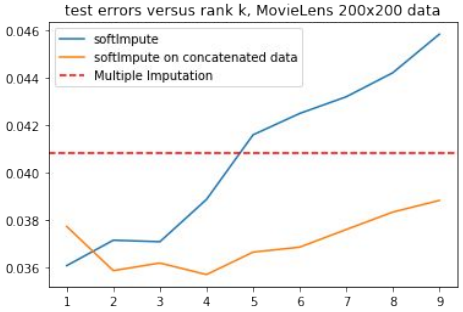


MovieLens data



- 100 x 100, 200 x 200
- NMAR pattern (skewed to higher rating)





EMBARC Data

- 8-week RCT of Sertraline enrolled 287 patients with Major Depressive Disorder.
- Baseline variables were divided to 3 data sets (Clinical, EEG and fMRI) with different size and proportion of data missing.
- MCAR, MAR and NMAR generated from fully observed subsets.
- Different missing rates generated on fMRI.

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%

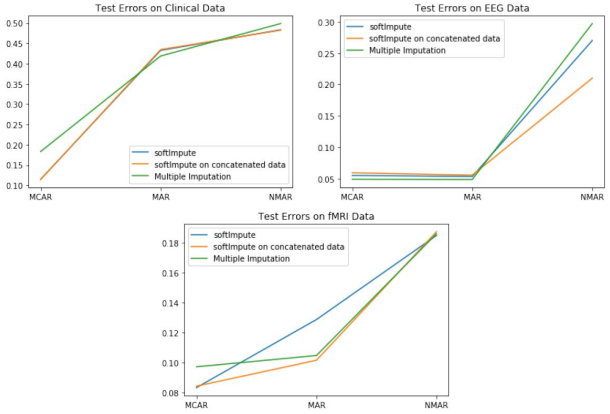


Fig. 4. Test errors of 3 methods on EMBARC data under different missing mechanisms

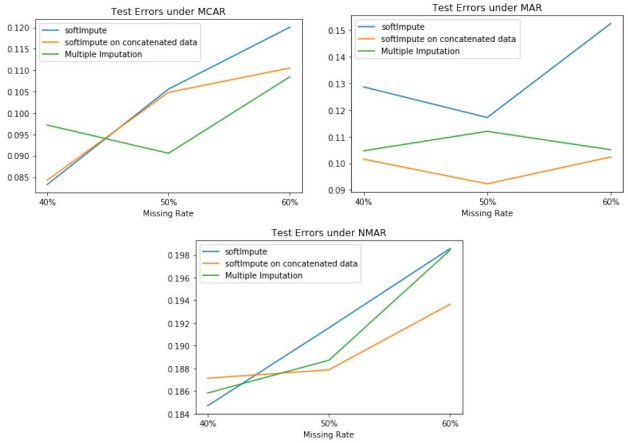


Fig. 5. Test errors of 3 methods under different missing rate generated on fMRI data.

Why SoftImpute-Concat Works?

Consider M and its SVD:

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

Can we recover it?

$$M = \begin{bmatrix} \text{🙈} & 2 & 2 & 2 \\ 2 & \text{🙈} & 2 & 2 \\ 2 & 2 & \text{🙈} & 2 \\ 2 & 2 & 2 & \text{🙈} \\ 3 & \text{🙈} & 3 & -3 \end{bmatrix}$$

M concatenated with mask

$$M_{con} = \begin{bmatrix} 2 & 2 & 2 & 2 & 1 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 & 1 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 & 0 & 1 & 0 \\ 2 & 2 & 2 & 2 & 0 & 0 & 0 & 1 \\ 3 & -3 & 3 & -3 & 0 & 1 & 0 & 0 \end{bmatrix} = UDV^T$$

$$D = \mathrm{diag}\{8.06, 6.08, 1, 1, 0.99\}$$

$$U = \begin{bmatrix} -0.50 & 0.02 & -0.00 & 0.82 & -0.29 \\ -0.50 & -0.01 & 0.00 & 0.00 & 0.87 \\ -0.50 & 0.02 & -0.71 & -0.41 & -0.29 \\ -0.50 & 0.02 & 0.71 & -0.41 & -0.29 \\ -0.02 & -1.00 & -0.00 & 0.00 & -0.02 \end{bmatrix}$$

Conclusion

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	+	++	++

Future Work

- ➔ Theoretical guarantee for SoftImpute-concat
- ➔ Evaluate mixed data matrix, longitudinal data, large-sparse matrix
- ➔ Derive confidence interval for estimated missing values

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 Miss- ing 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 reg- ularization algorithms for learning large incomplete matri- ces,” 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 medi- cal research, vol. 4, no. Suppl 1, pp. 6–7, 2014. III-A