

Information Availability and Missingness in One-pass Principal Component Completion

Ponnala Vimal Mosahari

vimalmosahari@gmail.com

Tattva Foundation <https://orcid.org/0000-0001-9507-0033>

Riddhi Bora

Sarla Birla Gyan Jyoti, Guwahati

Eklavya Sharma

Thapar University

Jon Jyoti Kalita

The Assam Royal Global University <https://orcid.org/0000-0002-7139-265X>

Research Article

Keywords: principal component analysis, missing data, matrix completion, cross-validation, reproducibility

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11501554/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

This study examines how cross-column structure, observation density, and missing data influence one-pass principal component completion. The study conducted a controlled simulation with 2,400 datasets across 12 scenarios, using either rank-two or independent structure, two observation fractions, and three missingness mechanisms. The study assessed column means, six PCA ranks, and two nuclear-norm regularisation settings on known masked values. With rank-two structure and random observation, rank-2 completion reduced mean absolute error compared to column means by 0.162 and 0.317 at observation fractions of 0.35 and 0.70. For independent columns, error increased by 0.017 and 0.041. Selective observation changed error levels, but improvement over means did not indicate unbiased recovery. The analysis used a sparse published study table and five standard MovieLens 100K splits as empirical checks. The study-table error was 0.221 for column means and 0.200 for rank-3 completion. For MovieLens, error was 0.779 for rank-10 completion and 0.749 for a regularised user/item baseline. Full-rank completion returned column means by design. This work provides a controlled view of when this procedure improves prediction. The findings depend on the tested designs and do not confirm recovery of naturally missing clinical outcomes.

1 Introduction

Drawing conclusions from incomplete data depends heavily on the assumptions made about both the dataset and the mechanism leading to missing values [1]. Completing missing entries does not, by itself, ensure valid statistical inference [2]. Common techniques include principal component analysis (PCA) on missing matrices [3], multivariate imputation [4], singular value decomposition (SVD)-based imputation [5], and matrix completion methods that rely on explicit assumptions regarding sampling and underlying structure [6].

Moreno-García and colleagues [7] applied PCA to sparse study-by-outcome tables. The authors were unable to reproduce their reported completion using the present implementation. This analysis examines how available information and observation methods affect a specific one-pass procedure. The simulation compares correlated signals with independent columns that share the same marginal distributions. The analysis uses two empirical examples to test if improvements persist against simple baselines. The main contribution is a controlled evaluation of an existing method.

2 Methods

2.1 One-pass completion

For an $n \times q$ training matrix Y , let Ω be the set of observed cells. Each column has a training mean μ_j and a scale s_j , which is its observed range, sample standard deviation, or one. The working matrix W has entries $W_{ij} = (Y_{ij} - \mu_j)/s_j$ on Ω and zero elsewhere. Range scaling and centering is the same as min-max scaling followed by centering. If the scale is zero or undefined, a value of one is used. A column with only

one value or a constant value has zero-centered variation, and an empty column cannot be estimated. Requested k is capped at the number of estimable columns; entirely unobserved columns remain missing.

The leading right singular vectors V_k provide the rank- k least-squares approximation [8] of the fixed, zero-filled working matrix. The authors chose a single pass for transparency because the working matrix stays fixed, while iterative methods update missing entries during fitting.

$$A_k \in \underset{\text{rank}(A) \leq k}{\text{argmin}} \|W - A\|_F^2, \quad A_k = WV_kV_k^T$$

This loss includes every entry of W , not just the observed ones. Missing values are back-transformed, as $\mu_i + s_i A_{ij}$, and observed entries are restored. Covariance is calculated as $W_k^T W_k / (n - 1)$. Using n instead of $n-1$ rescales eigenvalues but does not change eigenvectors or retained proportions. At full rank, missing working entries stay zero, so completion returns column means. This serves as an implementation check, not an empirical finding.

2.2 Source data and concordance

The obesity transcription has 68 observations and 136 empty cells across 34 rows and six outcomes [7]. These outcomes are body-mass index, obesity prevalence, physical activity, sedentary activity, fruit intake, and snack intake. The source score definitions do not provide clinically comparable units, so pooled raw errors are used as numerical diagnostics. The analysis reports outcome-specific and training-scale-normalized errors. For inputs, the analysis uses 0.21 for row 7 sedentary activity in Table 2 (p. 12), and 0.18 for row 22 obesity prevalence in historical Table 3 (p. 14).

The analysis checked historical agreement at ranks 1–6 using range scaling, observed-standard-deviation scaling, no scaling, and correlation PCA of the zero-filled matrix. The analysis also checked if pairwise-complete covariance could be defined. The analysis tested hypothetical changes to 0.13 and 0.19 separately, without claiming these as verified alternative transcriptions. Agreement means every originally missing cell is within 0.005 of the rounded historical value. Historical completions are not considered ground truth.

2.3 Validation and comparators

Leave-one-cell-out validation hides each observed cell once, resulting in 68 predictions for each method. Means, scales, and fitted matrices are recalculated within each training fold. The results separate 14 cells from single-outcome studies and 54 from multiple-outcome studies. Whole-study holdouts hide all recorded outcomes for each row and score the same 68 cells across 34 folds. These methods test what happens when row information is missing, not just selective outcome reporting.

The analysis evaluated all PCA ranks 1–6 using three types of scaling. The comparators are training column means, SOFT-IMPUTE [9], and a three-neighbor predictor. SOFT-IMPUTE minimizes the following objective, where the nuclear norm is the sum of singular values: Missing entries are updated by singular-

value soft thresholding, starting from zero. Training range scaling stays fixed within each fold. λ is set as α times the largest singular value of the zero-filled training matrix, with α values of 0.1, 0.5, and 1. Iteration stops when the relative Frobenius change is below 10^{-8} or after 20,000 iterations.

$$\frac{1}{2} \sum_{(i,j) \in \Omega} (W_{ij} - A_{ij})^2 + \lambda \|A\|_*$$

The neighbor predictor uses training-standardized mean squared distance over mutually observed other outcomes. It averages up to three eligible donor values, breaks distance ties by row order, and uses the column mean if there is no eligible donor. This is a set baseline, not multiple imputation. Regularized iterative PCA [10] and fast alternating least squares [11] are other alternatives not used here.

The analysis calculates mean absolute error (MAE) and root mean squared error (RMSE) on held-out cells. Supplementary results include equally weighted study-level MAE, errors divided by the training column standard deviation or range, and counts of wins, ties, and losses compared to column means. Positive reductions mean improvement over the baseline. Ranks and penalties are explored in a grid; the best observed setting does not estimate the performance of a model chosen independently, as this would introduce selection bias. The authors do not report empirical significance tests or population confidence intervals. Simulation Monte Carlo error is reported separately.

2.4 Supplementary benchmark and reproducibility

MovieLens 100K [12] provided a secondary benchmark using the five supplied 80%/20% splits. Ratings and timestamps were checked against the full data; test sets were disjoint. All 943 users and 1,682 items were kept without popularity filtering. Training-only predictors included global, item and user means, sequential user/item biases [13], jointly fitted ridge biases at fixed penalties 1, 10 and 100, and range-scaled one-pass PCA ranks 5, 10 and 20. Each setting was evaluated both unclipped and clipped to the 1–5 rating range; no test-based model selection was done. PCA and item means used the training global mean for unseen items, while bias models kept any available user effect. Results are reported for all test ratings and separately for items represented in training. Earlier masked experiments are kept separately for historical comparison. Online Resource A includes code, transcribed inputs, fold results, convergence logs, transformations, and environment details. New validation folds are deterministic, and historical random masks and seeds are documented separately. ChatGPT/Codex (OpenAI) assisted with the analysis code and verification. Grammarly was used for paraphrasing, grammar checking, and plagiarism checks. The authors are responsible for the simulation design, analysis, source attribution, and the final text.

2.5 Controlled simulation

The authors set the design before running the simulation, generating 200 independent datasets for each of 12 scenarios. Each matrix had 80 rows and six dimensionless columns. The simulation compared a rank-two Gaussian signal with noise standard deviation 0.3 to independent Gaussian columns with the same marginal variance, 1.09. Column 1 was always observed. Target columns 2–6 had expected

observation fractions of 0.35 or 0.70, under missing completely at random (MCAR), missing at random (MAR, based on column 1), or missing not at random (MNAR, based on the value itself). Logistic selection on the relevant value divided by $\sqrt{1.09}$ used a slope of -1.5 , with intercepts set to reach the target fraction. Full generation equations are in Online Resource A.

The analysis scored training-only column means, PCA ranks 1–6, and SOFT-IMPUTE at $\alpha = 0.1$ and 0.5 on the same masked cells. The authors did not select settings using hidden truth, and no datasets were redrawn. The main measure was expected per-dataset MAE; paired reduction was mean-baseline MAE minus method MAE. Monte Carlo standard error (MCSE) was calculated as the standard deviation of paired reductions divided by $\sqrt{200}$ [14]. Mean ± 1.96 MCSE shows simulation precision, not clinical or population uncertainty. The root seed was 20261022; protocol, seeds, failures, and complete grids are provided.

3 Results

3.1 Historical concordance

None of the tested computable variants reproduced all historical completed cells within rounding tolerance. Range-scaled concordance MAE at ranks 1–6 was 0.0557, 0.0513, 0.0546, 0.0437, 0.0356, and 0.0340. The lowest MAE among the four main variants was 0.0287 for correlation PCA at rank 1. Pairwise covariance could not be defined for three outcome pairs with only one shared observation, and no arbitrary covariance fill was used.

The source reports 99.23% retained variance at rank 4 [7] (p. 15, Section 4.2). The present implementation gives 87.94% with range scaling, 90.30% with observed-standard-deviation scaling, 95.38% without scaling, and 85.92% for zero-filled correlation PCA. None of these explain the source figure. The reason is unclear, and the full-rank identity does not match the source's reported rank-4 result.

3.2 Withheld-cell performance

Column-mean MAE was 0.221. Range-scaled rank-3 PCA gave 0.200, a descriptive reduction of 0.021 (9.5%). SOFT-IMPUTE at $\alpha = 0.1$ gave 0.192 and the neighbour baseline 0.237 (Table 1). The lowest PCA MAE occurred at rank 3 with range or standard-deviation scaling, and rank 2 without scaling. These settings were not independently selected. The complete grid is supplied in Online Resource A.

Table 1
Deterministic leave-one-cell-out results on 68 observed cells

Method	MAE	RMSE	Reduction	Wins / ties / losses
Column mean	0.221	0.323	0.000	0 / 68 / 0
One-pass PCA k = 1	0.224	0.327	-0.003	24 / 14 / 30
One-pass PCA k = 2	0.202	0.306	0.019	36 / 14 / 18
One-pass PCA k = 3	0.200	0.305	0.021	32 / 14 / 22
One-pass PCA k = 4	0.212	0.315	0.009	31 / 14 / 23
One-pass PCA k = 5	0.219	0.321	0.002	26 / 14 / 28
One-pass PCA k = 6	0.221	0.323	0.000	0 / 68 / 0
SOFT-IMPUTE $\alpha = 0.1$	0.192	0.294	0.029	35 / 14 / 19
SOFT-IMPUTE $\alpha = 0.5$	0.200	0.305	0.021	36 / 14 / 18
SOFT-IMPUTE $\alpha = 1$	0.221	0.323	0.000	0 / 68 / 0
Three-neighbour baseline	0.237	0.340	-0.016	24 / 14 / 30

PCA and SOFT-IMPUTE use range scaling in this analysis. Reduction is calculated as column-mean MAE minus method MAE. Absolute error differences within a tolerance of 10^{-10} are counted as ties. All comparisons are descriptive.

For rank-3 range-scaled PCA, all 14 single outcome cells were ties. Of the remaining 54 cells, 32 were wins and 22 were losses. SD-normalized MAE was 0.916 compared to 0.983 for means, and 0.287 compared to 0.308 for range-normalized values. These ratios do not show clinical importance. Eleven held-out values were outside the range of their training column. Outcome-specific MAE for BMI was very similar (0.152 compared to 0.153), and the overall improvement was not evenly spread across outcomes.

Whole study holdouts gave column means for all methods tested (MAE 0.221). A row with no information cannot provide a study-specific latent score. All 306 regularized fits met the stopping rule, with a maximum of 7,088 iterations.

3.3 Standard MovieLens benchmark

For 100,000 test ratings, unclipped rank-10 PCA had MAE 0.779 and pooled RMSE 0.983. Sequential biases gave 0.749 and 0.950, and joint ridge biases at $\lambda = 10$ gave 0.749 and 0.945. The joint ridge biases had lower MAE than any tested PCA rank in every fold. Penalties of 1 and 100 gave MAE 0.743 and 0.803, showing sensitivity to regularization. All 15 ridge fits converged. There were 167 test ratings for unseen items, and the warm-item subset kept the comparison. Full fold-level and clipping results are in Online Resource 1.

3.4 Simulation results

With rank-two MCAR generation, rank-2 PCA reduced MAE by 0.162 (MCSE 0.002) at 35% target observation and 0.317 (0.003) at 70%. For independent columns, it increased MAE by 0.017 (0.001) and 0.041 (0.002). SOFT-IMPUTE at $\alpha = 0.1$ also helped in correlated scenarios but increased error in independent scenarios (see Table 2). MNAR led to larger absolute errors than MCAR in these cases, even when a method improved on means. All 4,800 regularized fits converged, with no prediction failures or empty columns. Full-rank PCA matched means in every dataset.

Table 2
Simulation mean absolute errors and paired reductions

Structure	Observed fraction	Mechanism	Mean MAE	PCA 2 reduction	Soft 0.1 reduction
Rank two	0.35	MCAR	0.852	0.162	0.247
Rank two	0.35	MAR	0.971	0.133	0.169
Rank two	0.35	MNAR	1.207	0.102	0.145
Rank two	0.7	MCAR	0.839	0.317	0.420
Rank two	0.7	MAR	0.969	0.283	0.372
Rank two	0.7	MNAR	1.224	0.287	0.415
Independent	0.35	MCAR	0.849	-0.017	-0.019
Independent	0.35	MAR	0.845	-0.014	-0.009
Independent	0.35	MNAR	1.215	-0.007	-0.009
Independent	0.7	MCAR	0.845	-0.041	-0.009
Independent	0.7	MAR	0.841	-0.037	-0.005
Independent	0.7	MNAR	1.220	-0.018	-0.004

The fractions refer to the five target columns; the anchor column is always observed. Positive reductions mean completion is favored. Rank 2 was chosen based on the generating structure, not from test results. All ranks, penalties, signed errors, and Monte Carlo intervals are in Online Resource A.

4 Discussion

The simulation uses one sample size, six columns, one noise level, and one selection slope. The results depend on these choices, including the fully observed anchor used for MAR. Better prediction under the specified MNAR mechanism does not mean that mechanism is identified in real data. Empirical folds overlap, and observed ratings come from selection [15]. Neither empirical benchmark confirms recovery

of naturally missing study outcomes. The simulation uses known synthetic values for evaluation, but does not show clinical recovery.

The clinical context comes from the source table. Completed effect scores do not provide sampling variances, valid study weights or protection against outcome-reporting bias. Multivariate meta-analysis combines correlated evidence under explicit assumptions [16], but this table lacks the information required to evaluate that use. The reported errors therefore support no pooled treatment effect, clinical recommendation or policy conclusion.

5 Conclusion

One-pass completion improved prediction when the simulated columns shared a signal, but increased error when they did not. Observation mechanisms changed absolute errors, and simple baselines stayed competitive in the empirical examples. Full-rank and empty-row identities are useful for checking implementation. The findings support evaluating completion against likely missingness mechanisms and simple comparators. Historical replication and recovery of naturally missing outcomes remain unresolved.

Declarations

Funding

No funding was received. The work was undertaken independently by the authors.

Competing interests

The authors declare no relevant financial or non-financial conflicts of interest.

Author contributions

Ponnala Vimal Mosahari: conceptualization and study planning, methodology, project administration and writing; Eklavya Sharma: software, formal analysis and computational validation; Riddhi Bora: data curation, reference checking and manuscript review; Jon Jyoti Kalita: methodological review, interpretation and manuscript review.

Ethics and consent

The study used a published aggregate table, publicly available MovieLens ratings and computer-generated data. No participants were recruited and no new participant data were collected. MovieLens demographic files were not analysed.

Acknowledgements

The authors acknowledge the GroupLens Research Project at the University of Minnesota for MovieLens 100K, described by Harper and Konstan. The obesity example is transcribed from Moreno-García and colleagues.

Data and code availability

Online Resource A provides the supplementary methods, transcribed aggregate inputs, analysis scripts and numerical outputs. MovieLens ratings are obtained separately from GroupLens under its terms and are not redistributed.

Independent work

The findings and interpretations are those of the authors and do not represent the institutional positions of their affiliated organisations.

References

1. RUBIN DB (Dec. 1976) Inference and missing data. *Biometrika* 63(3):581–592. <https://doi.org/10.1093/biomet/63.3.581>
2. Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. *Psychol Methods* 7(2):147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
3. Ilin A, Raiko T (2010) Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *J Mach Learn Res* 11(66):1957–2000
4. Josse J, Husson F, missMDA (2016) A Package for Handling Missing Values in Multivariate Data Analysis. *J Stat Softw* 70(1):1–31. <https://doi.org/10.18637/jss.v070.i01>
5. Troyanskaya O et al (Jun. 2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17(6):520–525. <https://doi.org/10.1093/bioinformatics/17.6.520>
6. Candès EJ, Recht B (Dec. 2009) Exact Matrix Completion via Convex Optimization. *Found Comput Math* 9(6):717–772. <https://doi.org/10.1007/s10208-009-9045-5>
7. Moreno-García CF, Aceves-Martins M, Serratos F (Mar. 2016) Unsupervised Machine Learning Application to Perform a Systematic Review and Meta-Analysis in Medical Research. *Comput Syst* 20(1):7–17. <https://doi.org/10.13053/cys-20-1-2360>
8. Eckart C, Young G (Sep. 1936) The Approximation of One Matrix by Another of Lower Rank. *Psychometrika* 1(3):211–218. <https://doi.org/10.1007/BF02288367>
9. Mazumder R, Hastie T, Tibshirani R (2010) Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J Mach Learn Res* 11(80):2287–2322
10. Josse J, Husson F (2012) Handling missing values in exploratory multivariate data analysis methods. *J Société Fr Stat* 153(2):79–99
11. Hastie T, Mazumder R, Lee JD, Zadeh R (2015) Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *J Mach Learn Res* 16(104):3367–3402

12. Harper FM, Konstan JA (2015) The MovieLens Datasets: History and Context, ACM Trans. Interact. Intell. Syst. TiiS, vol. 5, no. 4, p. 19:1–19:19, Dec. <https://doi.org/10.1145/2827872>
13. Koren Y, Bell R, Volinsky C (Aug. 2009) Matrix Factorization Techniques for Recommender Systems. Computer 42(8):30–37. <https://doi.org/10.1109/MC.2009.263>
14. Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Stat Med 38(11):2074–2102. <https://doi.org/10.1002/sim.8086>
15. Steck H Training and testing of recommender systems on data missing not at random, in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 713–722. <https://doi.org/10.1145/1835804.1835895>
16. Riley RD et al (Sep. 2017) Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. BMJ 358:j3932. <https://doi.org/10.1136/bmj.j3932>

Table 3

Table 3 is not available with this version

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMethodsandResults.pdf](#)