

Supplementary methods and complete results

Information availability and missingness in one-pass principal component completion. The supplement reports empirical validation, the controlled simulation, complete method grids and reproduction settings.

S1 Complete rank and scaling grid

Each row summarises 68 leave-one-cell-out predictions. SD MAE divides each absolute error by its training column sample standard deviation before averaging. Study MAE first averages within study, then across 34 studies. CSVs also contain range-normalised errors and separate single-outcome and multiple-outcome strata. All ranks and penalties are exploratory.

Method	MAE	RMSE	SD MAE	Study MAE
knn_zscore_n3	0.237	0.340	1.075	0.215
mean	0.221	0.323	0.983	0.199
pca_minmax_k1	0.224	0.327	1.000	0.202
pca_minmax_k2	0.202	0.306	0.915	0.187
pca_minmax_k3	0.200	0.305	0.916	0.186
pca_minmax_k4	0.212	0.315	0.955	0.193
pca_minmax_k5	0.219	0.321	0.975	0.198
pca_minmax_k6	0.221	0.323	0.983	0.199
pca_none_k1	0.211	0.315	0.956	0.193
pca_none_k2	0.208	0.312	0.948	0.191
pca_none_k3	0.210	0.314	0.930	0.191
pca_none_k4	0.217	0.317	0.960	0.196
pca_none_k5	0.219	0.320	0.971	0.198
pca_none_k6	0.221	0.323	0.983	0.199
pca_zscore_k1	0.222	0.324	0.990	0.200
pca_zscore_k2	0.208	0.311	0.940	0.190
pca_zscore_k3	0.202	0.305	0.919	0.187
pca_zscore_k4	0.216	0.319	0.963	0.195
pca_zscore_k5	0.220	0.322	0.977	0.198
pca_zscore_k6	0.221	0.323	0.983	0.199
soft_minmax_a0.1	0.192	0.294	0.879	0.183
soft_minmax_a0.5	0.200	0.305	0.914	0.186
soft_minmax_a1	0.221	0.323	0.983	0.199

S2 Historical concordance

Errors compare 136 completed values with the rounded source Table 3. Correlation PCA divides zero-filled working columns by their full-matrix SD before projection, then reverses that step. Observed-only SD scaling is distinct. Pairwise covariance centres each pair on its paired subset: three distinct pairs have only one joint observation, so those covariances are undefined. No arbitrary fill is used.

Variant	k	Retained %	MAE
minmax / zero_cov	1	32.56	0.05570
minmax / zero_cov	2	59.85	0.05126

Variant	k	Retained %	MAE
minmax / zero_cov	3	76.50	0.05461
minmax / zero_cov	4	87.94	0.04365
minmax / zero_cov	5	96.44	0.03559
minmax / zero_cov	6	100.00	0.03396
zscore / zero_cov	1	37.46	0.05167
zscore / zero_cov	2	62.00	0.04579
zscore / zero_cov	3	80.68	0.04968
zscore / zero_cov	4	90.30	0.03956
zscore / zero_cov	5	96.75	0.03572
zscore / zero_cov	6	100.00	0.03396
none / zero_cov	1	39.96	0.04193
none / zero_cov	2	68.22	0.04328
none / zero_cov	3	83.96	0.05367
none / zero_cov	4	95.38	0.03643
none / zero_cov	5	98.46	0.03292
none / zero_cov	6	100.00	0.03396
minmax / correlation	1	31.71	0.02871
minmax / correlation	2	57.10	0.06230
minmax / correlation	3	73.50	0.05386
minmax / correlation	4	85.92	0.05232
minmax / correlation	5	95.80	0.03537
minmax / correlation	6	100.00	0.03396

None agrees at every cell within rounding tolerance. Hypothetical perturbations are in hypothetical_cell_perturbations.csv; these do not authenticate previously suggested PDF/HTML differences.

S3 Algorithm details and checks

SOFT-IMPUTE forms $B = P_{\Omega}(W) + P_{\Omega^c}(A_t)$, computes $B = U D V^T$, and updates $A_{t+1} = U \max(D - \lambda I, 0) V^T$. Means, scales and λ stay fixed within the fold. Starting A is zero. Relative change is $\|A_{t+1} - A_t\|_F / \max(1, \|A_t\|_F)$. Observed entries are restored only in the returned table. This is not missMDA's specific regularisation or multiple imputation. All 306 fits converged within 7088 iterations with a 20,000 cap. Four initial $\alpha = 0.1$ fits exceeded an initial 5,000 cap; the cap was increased without changing tolerance. Objective increases were checked against 10^{-8} tolerance. Full-rank mean equivalence and whole-study mean fallback were checked at 10^{-10} and 10^{-9} tolerances.

S4 Original obesity masks

The original experiment sampled ten observed cells without replacement in each of 20 repeats using default_rng(20260915), in row-major observed-cell order. No mask was redrawn or dropped and none depleted a column. Of 200 masking events, 162 belonged to originally multiple-outcome rows; 146 retained another observation after masking. These repeated events are not independent studies. Current claims use the deterministic analysis; former bootstrap intervals are not retained.

Method	MAE	Reduction	Wins / repeats
column_mean	0.238	0.000	0/20
pca_k1	0.238	0.000	10/20
pca_k2	0.229	0.009	17/20
pca_k3	0.227	0.011	17/20

S5 All historical MovieLens results

MovieLens 100K contains 943 users, 1,682 items and 100,000 ratings. Earlier filters retained 939 items with ≥ 20 ratings and 603 with ≥ 50 ratings. Filters preceded masking, partly defining the population using future test ratings. Random masks hid 2,000 ratings for ten primary repeats (seed 20260916) and eight denser repeats (20260918). Partial-column masks chose 15 items, hiding 30% of each item's observed ratings, rounded down with minimum one, for ten repeats (20260917). All preprocessing was refitted per mask.

The sequential bias predictor estimates global mean g , item bias $b_j = \text{mean}(Y_{ij} - g)$, then user bias $a_i = \text{mean}(Y_{ij} - g - b_j)$. It predicts $g + a_i + b_j$, with unavailable biases zero, without iteration, regularisation or clipping. PCA uses min-max scaling and centred zero filling. Only clip15 restricts predictions to [1,5]. Positive reduction is bias MAE minus method MAE; signed error is prediction minus rating. Reported RMSE is mean repeat RMSE, not pooled RMSE.

items ge20 random cell

Method	MAE	RMSE	Signed error	Reduction
global_mean	0.925	1.105	-0.005	-0.189
item_mean	0.804	1.004	-0.004	-0.068
pca_k10	0.746	0.943	0.016	-0.010
pca_k10_clip15	0.745	0.943	0.015	-0.009
pca_k20	0.747	0.944	0.020	-0.011
pca_k5	0.753	0.948	0.009	-0.017
user_item_bias	0.736	0.931	-0.002	0.000
user_mean	0.825	1.032	-0.007	-0.089

items ge20 column block

Method	MAE	RMSE	Signed error	Reduction
global_mean	0.938	1.114	-0.009	-0.211
item_mean	0.793	1.001	-0.004	-0.066
pca_k10	0.752	0.954	0.005	-0.025
pca_k10_clip15	0.752	0.954	0.005	-0.025
pca_k20	0.751	0.955	0.008	-0.025
pca_k5	0.758	0.960	0.002	-0.031
user_item_bias	0.727	0.927	0.002	0.000
user_mean	0.844	1.046	-0.012	-0.118

items ge50 random cell

Method	MAE	RMSE	Signed error	Reduction
global_mean	0.905	1.085	-0.012	-0.169
item_mean	0.800	1.001	-0.007	-0.065
pca_k10	0.739	0.933	0.015	-0.003
pca_k10_clip15	0.738	0.933	0.014	-0.002
pca_k20	0.742	0.940	0.016	-0.006
pca_k5	0.746	0.941	0.008	-0.010
user_item_bias	0.736	0.932	-0.003	0.000
user_mean	0.816	1.018	-0.007	-0.080

The preceding tables re-summarise previously verified per-repeat outputs. These historical experiments are distinct from the new standard-split benchmark below. No published scores are imported for comparison, and no confidence intervals or significance claims are made.

S6 Standard MovieLens five-fold benchmark

The user-supplied files were checked for 100,000 unique full-data user/item pairs, valid identifier and rating ranges, 80,000 training and 20,000 test rows per fold, no within-fold overlap, disjoint test sets, complete full-data coverage, and identical ratings and timestamps. These checks establish internal consistency; source hashes document the supplied files rather than independently authenticate their provenance. No popularity filter or demographic predictor was used.

All preprocessing used training ratings. PCA uses observed item range scaling, item-mean centering and zero filling of missing centred entries. Exact thin SVD gives the rank-5, rank-10 and rank-20 projections. Singleton and constant items use their training means; entirely unseen items use the training global mean as an explicitly added operational fallback, not a PCA estimate. User means retain the available user mean; joint and sequential biases set the unavailable item bias to zero and retain the user effect. The secondary warm-item analysis excludes these cold-item cases consistently for every method.

Joint biases minimise $\sum_{(i,j) \in \Omega} (y_{ij} - g - a_i - b_j)^2 + \lambda \left(\sum_i a_i^2 + \sum_j b_j^2 \right)$, with g fixed at the training global mean and λ in $\{1, 10, 100\}$. Alternating exact coordinate updates start at zero and stop when the largest absolute coefficient change is below 10^{-9} , with a cap of 10,000 iterations. Objective increases above 10^{-7} or nonconvergence abort the run. All 15 fits converged in 16–525 iterations. The tested ranks and penalties were fixed before these data were scored; no setting is presented as an independently validated optimum.

Table S6 reports pooled errors across the five disjoint tests. Each method scores all 100,000 ratings; mean fold MAE therefore equals pooled MAE. Pooled RMSE is the square root of the weighted mean squared error, not mean fold RMSE. Both RMSE summaries, signed errors and all fold metrics are supplied as CSVs. Test sets are disjoint but training sets overlap; no independent-fold confidence intervals are claimed.

Table S6 Standard-split errors on all test ratings

Method	MAE raw	RMSE raw	MAE clipped	RMSE clipped
Global mean	0.9447	1.1257	0.9447	1.1257
Item mean	0.8174	1.0246	0.8174	1.0246
User mean	0.8362	1.0438	0.8362	1.0438
Sequential bias	0.7489	0.9496	0.7475	0.9487
Ridge bias $\lambda=1$	0.7434	0.9423	0.7425	0.9419
Ridge bias $\lambda=10$	0.7485	0.9448	0.7484	0.9447
Ridge bias $\lambda=100$	0.8028	0.9940	0.8028	0.9940
PCA rank 5	0.7846	0.9888	0.7846	0.9888
PCA rank 10	0.7793	0.9834	0.7791	0.9833
PCA rank 20	0.7797	0.9845	0.7796	0.9844

The five tests contained 32, 36, 36, 27 and 36 cold-item ratings, respectively (167 total). On the 99,833 warm-item ratings, unclipped rank-10 PCA had MAE 0.7781, compared with 0.7483 for ridge biases at $\lambda = 10$. Clipping produced small changes and did not reverse that comparison. Aggregate fit times ranged from 7.42 to 8.99 seconds per fold; the full run, including checks and output, took 41.91 seconds in the recorded execution environment. Timing is descriptive and hardware dependent. These results concern prediction of held-out observed ratings, not ranking quality, future behaviour or naturally missing outcomes.

S7 Outcome-specific obesity errors

These raw errors retain the source score scales and cannot be compared as clinical effect magnitudes across outcomes. BMI denotes body-mass index; obesity prevalence, physical activity, sedentary activity, fruit and snack consumption follow the source column definitions. The complete by-outcome file contains all methods and normalised errors. Sparse outcome counts limit interpretation.

Table S7 Outcome-specific leave-one-cell-out mean absolute errors

Outcome	Cells	Mean	PCA rank 3	Soft $\alpha=0.1$
BMI	23	0.153	0.152	0.145
ObesityPrev	9	0.144	0.137	0.125
PhysAct	12	0.259	0.216	0.217
Sedentary	6	0.174	0.161	0.179
Fruit	9	0.268	0.223	0.216
Snack	9	0.406	0.367	0.332

S8 Simulation design and uncertainty

The design follows explicit aims, generation, estimands, methods and performance measures. The protocol was fixed internally before execution, with no external registration. There are 12 scenarios and 200 independent datasets per scenario. $N=80$ permits finite-sample estimation; six columns match the empirical table width. The Gaussian rank-two case deliberately favours low-rank methods; independent columns with the same marginal variance provide a negative control. Noise $SD=0.3$, observation fractions 0.35/0.70 and selection slope 1.5 are fixed stress-test choices, not clinically calibrated parameters or parameter distributions.

For the correlated case, independent standard normal entries form Z (80×2) and E (80×6):

$$Y = Z L^T + 0.3 E, L_j = (\cos(\pi j / 6), \sin(\pi j / 6)), j = 0, \dots, 5$$

The independent case has standard normal E and uses:

$$Y = \sqrt{1.09} E$$

Column 1 is always observed and is available to all methods. Columns 2–6 are observed independently conditional on the complete generated matrix. MCAR uses probability q . Selective observation uses:

$$Pr(R_{ij}=1)=\text{logit}^{-1}(a-1.5d_{ij})$$

For MAR, d is column 1 divided by $\sqrt{1.09}$; for MNAR it is the target cell divided by $\sqrt{1.09}$. An 80-node Gauss–Hermite integral and 100 bisection steps on $[-20,20]$ choose a so marginal expected observation equals q . An independent 160-node calculation verified the calibration within 10^{-10} . The complete matrix enters generation and evaluation only. All model means, scales and singular values use observed entries. The anchor means the whole-matrix observation fraction is $(1+5q)/6$, rather than q .

Generator(SeedSequence([20261022, scenario_index, replicate_index])) supplies independent scenario/replicate streams. The same mask pairs all methods within a replicate. Scenario order is structure (rank2, independent), q (0.35,0.70), mechanism (MCAR,MAR,MNAR). Each replication scores every masked target value. The primary summaries average per-dataset MAE equally, not individual cells pooled across differently sized masks. RMSE likewise averages per-dataset RMSE. Signed error is prediction minus truth. No method is chosen using evaluation truth.

For each method, paired reduction is the mean-baseline MAE minus method MAE. MCSE is the replicate SD of this reduction divided by $\sqrt{200}$; the supplied limits are mean ± 1.96 MCSE. They express Monte Carlo precision conditional on this simulation design. They are not empirical confidence intervals or evidence of clinical efficacy. The 200-replicate count was fixed without adaptive stopping. Actual MCSE is reported rather than claiming a prespecified precision target.

Table S8 Paired simulation reductions and Monte Carlo standard errors

Scenario	PCA rank 2 reduction (MCSE)	Soft $\alpha=0.1$ reduction (MCSE)
rank2_0.35_MCAR	0.1619 (0.0016)	0.2475 (0.0029)
rank2_0.35_MAR	0.1334 (0.0016)	0.1690 (0.0024)
rank2_0.35_MNAR	0.1021 (0.0012)	0.1445 (0.0023)
rank2_0.7_MCAR	0.3171 (0.0026)	0.4198 (0.0035)
rank2_0.7_MAR	0.2833 (0.0026)	0.3715 (0.0035)
rank2_0.7_MNAR	0.2866 (0.0025)	0.4152 (0.0042)
independent_0.35_MCAR	-0.0166 (0.0009)	-0.0186 (0.0011)
independent_0.35_MAR	-0.0137 (0.0008)	-0.0091 (0.0007)
independent_0.35_MNAR	-0.0075 (0.0006)	-0.0085 (0.0006)
independent_0.7_MCAR	-0.0405 (0.0022)	-0.0088 (0.0009)
independent_0.7_MAR	-0.0370 (0.0022)	-0.0055 (0.0008)
independent_0.7_MNAR	-0.0183 (0.0019)	-0.0037 (0.0008)

The complete 108-row summary reports every method in every scenario; 21,600 method/replicate records and 4,800 SOFT-IMPUTE convergence records are supplied. All regularised fits converged within 3,624 iterations; the maximum objective increase was 7.11×10^{-15} . No missing predictions or empty columns occurred, and no dataset was redrawn. Full-rank mean equivalence passed in every replicate. All aggregate means, MCSEs and intervals were recomputed from replicate records. Selective-observation signed errors are retained; relative gains do not establish unbiased recovery.

The design covers a narrow family of Gaussian matrices, one sample size and one selection strength. It does not represent a comprehensive comparison, identify the MNAR process, model clinical outcome-reporting bias, or validate the original historical algorithm. Any conclusions are conditional on these settings.

S9 Reproduction

Run `python3 scripts/06_revision.py` from the package directory for the obesity analysis and historical summaries. Run `python3 scripts/07_movielens_standard.py --data /path/to/ml-100k` for the standard benchmark, setting `OPENBLAS_NUM_THREADS=1` and `OMP_NUM_THREADS=1`. Python 3.12.14, NumPy 2.3.5 and pandas 2.2.3 were used. `AUDIT_ROOT` optionally overrides the package root. Obtain `u.data` and `u1.base/u1.test` through `u5.base/u5.test` from GroupLens; raw ratings are not redistributed. The supplied `u(3).data` was locally copied to `u.data` without content changes. Data hashes, runtime and thread settings are recorded in `outputs/movielens_standard/run.json`. Current benchmark settings and splits are deterministic, with no random seed required. Legacy scripts are retained for historical reproduction. Checksums identify release files. Run `python3 scripts/08_simulation.py` for the complete simulation. Simulation records are in `outputs/simulation/`; runtime was 33.54 seconds with the same single-thread settings. Protocol and code hashes are recorded in `run.json`.