The robust singular value decomposition and the problem of incomplete two-way data

Sergio Arciniegas-Alarcón[1], Camilo Rengifo[1], Marisol García-Peña[2], Wojtek Krzanowski[3]

[1] Universidad de La Sabana

[2] Pontificia Universidad Javeriana

[3] University of Exeter

Abbreviated abstract: We describe imputation strategies resistant to outliers, through modifications of the simple imputation method proposed by Krzanowski and assess their performance. The strategies use a robust singular value decomposition, do not depend on distributional or structural assumptions and have no restrictions as to the pattern or missing data mechanisms. They are tested through the simulation of contamination and unbalance in a matrix of real data from an experiment with genotype-by-environment interaction.

Related publications:

– Arciniegas-Alarcón *et al*, Appl. Syst. Innov. 62 (4), (2021)

– García-Peña *et al*, Crop Science 61 (5), 3288-3300 (2021)

Problem, Data, Previous Works

- SVD88 (SVD published in 1988 by Krzanowski).
- Consider a matrix $\mathbf{Y_{n\times p}}$ with possible missing values.
- Missing values are initially filled with columns mean + standardization.
- $\hat{y}_{ij}^{(m)} = \sum_{h=1}^{m} \tilde{u}_{ih}\left(\tilde{d}_h\sqrt{p/(p-1)}\right)^{1/2}\overline{v}_{jh}\left(\overline{d}_h\sqrt{n/(n-1)}\right)^{1/2}$.
- An iterative updated of the imputations via the above equation.
- The effect of outliers on the quality of imputations of SVD88 has not been considered.

MAIN IDEAS:

- New Robust Imputation Methods → Two-way data.
- Robust SVD → Imputation.
- Methods without structural assumptions.

REAL DATA:

- Ontario winter wheat dataset:
  - 18 genotypes + 9 environments

Table 1. Mean yield (Mg ha^{-1}) of 18 winter wheat cultivars (G1 to G18) tested at nine Ontario locations (E1 to E9) in 1993.

Geno-types	Test Environments									
	E1	E2	E3	E4	E5	E6	E7	E8	E9	Mean
G1	4.46	4.15	2.85	3.08	5.94	4.45	4.35	4.04	2.67	4.00
G2	4.42	4.77	2.91	3.51	5.70	5.15	4.96	4.39	2.94	4.31
G3	4.67	4.58	3.10	3.46	6.07	5.03	4.73	3.90	2.62	4.24
G4	4.73	4.75	3.38	3.90	6.22	5.34	4.23	4.89	3.45	4.54
G5	4.39	4.60	3.51	3.85	5.77	5.42	5.15	4.10	2.83	4.40
G6	5.18	4.48	2.99	3.77	6.58	5.05	3.99	4.27	2.78	4.34
G7	3.38	4.18	2.74	3.16	5.34	4.27	4.16	4.06	2.03	3.70
G8	4.85	4.66	4.43	3.95	5.54	5.83	4.17	5.06	3.57	4.67
G9	5.04	4.74	3.51	3.44	5.96	4.86	4.98	4.51	2.86	4.43
G10	5.20	4.66	3.60	3.76	5.94	5.35	3.90	4.45	3.30	4.46
G11	4.29	4.53	2.76	3.42	6.14	5.25	4.86	4.14	3.15	4.28
G12	3.15	3.04	2.39	2.35	4.23	4.26	3.38	4.07	2.10	3.22
G13	4.10	3.88	2.30	3.72	4.56	5.15	2.60	4.96	2.89	3.80
G14	3.34	3.85	2.42	2.78	4.63	5.09	3.28	3.92	2.56	3.54
G15	4.38	4.70	3.66	3.59	6.19	5.14	3.93	4.21	2.93	4.30
G16	4.94	4.70	2.95	3.90	6.06	5.33	4.30	4.30	3.03	4.39
G17	3.79	4.97	3.38	3.35	4.77	5.30	4.32	4.86	3.38	4.24
G18	4.24	4.65	3.61	3.91	6.64	4.83	5.01	4.36	3.11	4.48
Mean	4.36	4.44	3.14	3.49	5.68	5.06	4.24	4.36	2.90	4.19

Methods:

Imputation Alternatives:

- SVD88: classic method, Krzanowski 88.
- rSVD84: Gabriel & Odoroff. García-Peña *et al.*
- M5RobSVD: In SVD88 consider Robust standardization of the initial matrix + rSVD84 + just singular vectors.
- GOKImputation
  - 1. rSVD84 gives rise a complete matrix 2. Imputation via SVD88.
- M5GOKImputation: GOKImputation just singular vectors.
- MissForest: random forest, randomized regression trees.

The comparison was based on:

1. Prediction error: P_e
2. $GF_1 = 1 - {\|O-I\|^2}/{\|I\|^2}$
3. $GF_2 = \cos^2(O,I) = \frac{tr^2(O^T I)}{tr(O^T O)tr(I^T I)}$.

O is the Ontario original data matrix, and I is the imputation matrix after cross-validation under the modified Ontario matrix (deleted values and contaminated).

Results and Conclusions

1. Consider robust versions of the method that will allow for outliers.
2. Suggest which specific method to use in each practical application requiring imputation.

Method	GF2	GF1	Pe
SVD88	0.3202	-2.0001	7.4491
M5RobSVD	0.9836	0.9812	0.5891
rSVD84	0.9843	0.9821	0.5756
GOKImputation	0.9842	0.9819	0.5788
M5GOKImputation	0.9833	0.9808	0.5958
missForest	0.8730	0.3719	3.4083

There are four possible robust version of SVD88, all improved substantially of SVD88. In a practical situation, if the matrix is small then rSVD84 is recommended , but for large for larger matrices either M5RobSVD or M5GOKImputation would be preferable.