Missing-value imputation using the robust singular-value decomposition: Proposals and numerical evaluation

Marisol García-Peña[1], Sergio Arciniegas-Alarcón[2], Wojtek Krzanowski[3], Diego Duarte[2]

[1] Pontificia Universidad Javeriana

[2] Universidad de La Sabana

[2] University of Exeter

Abbreviated abstract: A common problem in the analysis of data from multi-environment trials is imbalance caused by missing observations. To get around this problem, Yan proposed a method for imputing the missing values based on the singular-value decomposition (SVD) of a matrix. However, this SVD can be affected by outliers and produce low quality imputations. We propose four robust SVD extensions of the Yan method that are resistant to outliers. We evaluate these methods, using exclusively numerical criteria in a cross-validation study based on real data. We conclude that in the presence of outliers, the best alternatives are the robust SVD methods based on sub-sampling when the percentage of contamination is less than 2% following a completely random missing data mechanism.

Related publications:

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

3rd Conference on Statistics and Data Science
Salvador, Brazil (online)
October 28-30, 2021

Problem, Data, Previous Works

- SVD13 (SVD published in 2013 by Yan).
- Consider a matrix $\mathbf{Y}_{(nxp)}$ with some elements missing.
- Missing observations are initially filled with an estimation.
- An iterative SVD is calculated, and the imputations are updated until achieve a convergence criterion.
- The effect of outliers on the quality of the imputations of the method proposed by Yan (2013) has not been considered in the literature.

Core Ideas

- New robust imputation methods → two-way data matrices.
- Robust SVD → imputations.
- Methods without structural assumptions, also applied in multi-environment trials.

Complete datasets

Reference	Species	No. of genotypes	No. of environments	Response variable
Yan et al. (2007)	Wheat (*Triticum aestivum* L.)	18	9	Mean yield
Lavoranti et al. (2007)	Eucalyptus (*Eucalyptus grandis* W.)	20	7	Mean tree height
Calinski et al. (2009)	Rye	18	15	Mean yield
Flores et al. (1998)	Beans (*Vicia faba* L.)	15	12	Mean yield
Rad et al. (2013)	Wheat	36	6	Mean yield

Methods

rSVD84 Method

- rSVD – robust SVD, Gabriel and Odoroff.
- Trimmed means or methods for detecting univariate outliers.

Consider a matrix $\mathbf{Y}_{(nxp)}$ with possible missing entries.

a. Using observed information, calculate the vectors of trimmed means (at 10% or 20%) by columns and by rows, $\mathbf{b_1}_{(1xp)}$ and $\mathbf{a_1}_{(nx1)}$.

b. Determine presence of outliers in $\mathbf{a_1}$ and replace it with a trimmed mean of the elements of $\mathbf{a_1}$.

c. Update $\mathbf{b_1}$ and $\mathbf{a_1}$: $b_c = med\left\{\left|\frac{y_{r,c}}{a_r}\right|; r = 1, \cdots, n\right\}$ and $a_r = med\left\{\left|\frac{y_{r,c}}{b_c}\right|; c = 1, \cdots, p\right\}$

- Go back to step b. with $\mathbf{b_1}$ and $\mathbf{a_1}$ updated → new update according step c. until reach some specified convergence criterion.
- Then a robust lower rank approximation of **Y** is obtained by $\mathbf{a_1}\,\mathbf{b_1}^T$.
- First singular value and first right and left singular vector of robust SVD → standard SVD to $\mathbf{a_1}\,\mathbf{b_1}^T$.
- Second singular value and second right and left singular vector → same procedure above but instead of **Y** as initial matrix, use the deflated **Y** - $\mathbf{a_1}\,\mathbf{b_1}^T$.
- Iterative strategy, deflated the matrix, continues step by step until the desired number of components has been reached.

* SVD13 method * rSVD01 method

* rSVD13 method * rSVD17 method

Results and Conclusions

- No outliers → Yan's (2013) SVD13 should be used and very good imputation quality will be obtained.
- Outliers → SVD13 → low quality imputations. Is recommended to use robust singular-value decompositions.
- If an MCAR missing data mechanism with contamination below 2% can be assumed → rSVD with subsampling method.
- In any other case, procedures that minimize the L_2 (rSVD84) norm or fit L_1 (rSVD01) regressions should be used.

Method	$Cos^{2\dagger}$	GF_1	P_e	Method	$Cos^{2\dagger}$	GF_1	P_e
Yan et al. (2007)				Lavoranti et al. (2007)			
SVD13	0.1466	−5.1486	10.6642	SVD13	0.3436	−3.4168	36.4317
rSVD17	0.9866	0.9853	0.5219	rSVD17	0.9960	0.9957	1.1323
rSVD84	0.9821	0.9794	0.6172	**rSVD84**	**0.9966**	**0.9963**	**1.0569**
rSVD01	0.9858	0.9851	0.5243	rSVD01	0.9965	0.9962	1.0743
rSVD13	**0.9875**	**0.9867**	**0.4955**	rSVD13	0.9954	0.9951	1.2184
Calinski et al. (2009)				Flores et al. (1998)			
SVD13	0.0568	−159.9183	996.4870	SVD13	0.0257	−46620.65	762736.0
rSVD17	0.2951	−4.1090	177.5567	rSVD17	0.9843	0.9826	465.4740
rSVD84	**0.9968**	**0.9967**	**4.5114**	**rSVD84**	**0.9853**	0.9831	458.5970
rSVD01	0.9953	0.9949	5.5879	**rSVD01**	0.9848	**0.9837**	**450.8791**
rSVD13	0.1228	−40.0221	503.1271	rSVD13	0.3232	0.3230	2906.5550
Rad et al. (2013)							
SVD13	0.0220	−191.4041	129.0142				
rSVD17	**0.9723**	**0.9683**	**1.6553**				
rSVD84	0.9607	0.9549	1.9760				
rSVD01	0.9634	0.9591	1.8816				
rSVD13	0.6187	0.2336	8.1423				

$^{\dagger}Cos^2 = GF_2$.

‡The values in bold represent the methods with maximum Cos^2, GF_1 and minimum prediction error (P_e).