MethodsX 11 (2023) 102289

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Missing value imputation in a data matrix using the regularised singular value decomposition

Sergio Arciniegas-Alarcón [a], Marisol García-Peña [b,*], Wojtek J. Krzanowski [c], Camilo Rengifo [a]

[a] Universidad de La Sabana, Facultad de Ingeniería, Chía, Colombia
[b] Pontificia Universidad Javeriana, Departamento de Matemáticas, Bogotá, Colombia
[c] University of Exeter, College of Engineering, Mathematics and Physical Sciences, Exeter, UK

ARTICLE INFO

Method name:
GabrielEigen imputation system

Keywords:
Eigenvalues
Eigenvectors
Iterative computational scheme
Cross-validation
Genotype-by-environment interaction
Overfitting

ABSTRACT

Some statistical analysis techniques may require complete data matrices, but a frequent problem in the construction of databases is the incomplete collection of information for different reasons. One option to tackle the problem is to estimate and impute the missing data. This paper describes a form of imputation that mixes regression with lower rank approximations. To improve the quality of the imputations, a generalisation is proposed that replaces the singular value decomposition (SVD) of the matrix with a regularised SVD in which the regularisation parameter is estimated by cross-validation. To evaluate the performance of the proposal, ten sets of real data from multienvironment trials were used. Missing values were created in each set at four percentages of missing not at random, and three criteria were then considered to investigate the effectiveness of the proposal. The results show that the regularised method proves very competitive when compared to the original method, beating it in several of the considered scenarios. As it is a very general system, its application can be extended to all multivariate data matrices.

- The imputation method is modified through the inclusion of a stable and efficient computational algorithm that replaces the classical SVD least squares criterion by a penalised criterion. This penalty produces smoothed eigenvectors and eigenvalues that avoid overfitting problems, improving the performance of the method when the penalty is necessary. The size of the penalty can be determined by minimising one of the following criteria: the prediction errors, the Procrustes similarity statistic or the critical angles between subspaces of principal components.

Specifications table

Subject area:	Agricultural and Biological Sciences
More specific subject area:	*Biometry, Statistics*
Name of your method:	GabrielEigen imputation system
Name and reference of original method:	*K.R. Gabriel. Le biplot – outil d'exploration de données multidimensionnelles. Journal de la Societé Française de Statistique. (143) (2002) 5–55.* *M. García-Peña, S. Arciniegas-Alarcón, W. Krzanowski, D. Barbin. Multiple imputation procedures using the GabrielEigen algorithm. Communications in Biometry and Crop Science. (11) (2016) 149–163.* *M. García-Peña, S. Arciniegas-Alarcón, W.J. Krzanowski. Missing value imputation using least squares techniques in contaminated matrices. MethodsX. (9) (2022)* https://doi.org/10.1016/j.mex.2022.101683
Resource availability:	https://www.researchgate.net/profile/Sergio-Arciniegas-Alarcon

* Corresponding author.
E-mail addresses: marisolgarcia@javeriana.edu.co, luzmara@gmail.com (M. García-Peña).

https://doi.org/10.1016/j.mex.2023.102289
Received 2 May 2023; Accepted 16 July 2023
Available online 17 July 2023

Method details

In many different areas of knowledge, data are collected and arranged in a matrix form for further analysis. This happens with multivariate data where the rows represent the individuals or independent entities being studied, while the columns represent the variables obtained from each entity. This structure is also widely used in plant genetic improvement and specifically in multienvironment trials, in which the variables of interest (for example, yield) are arranged in two-way tables in which the genotypes are found in rows and the test environments are found in the columns [1,2].

Possibly one of the most prevalent problems in these datasets is incomplete information. For example, if the experimental units are plants or animals, some of them may die before all the variables of interest are collected. In archaeology, artefacts or skulls may be damaged during excavations, or specimens may be incomplete when excavated, and in opinion polls, a respondent may fail to answer some of the questions. In all the situations described above, many multivariate analysis techniques require complete records, so two options can be considered: i) units or variables that contain missing data have to be excluded from the analysis or ii) one or several estimates can be found for each missing value. This latter option is also known as imputation [3].

Recently, García-Peña et al. [4,5] described an imputation method free of distributional and structural assumptions for any dataset that can be arranged in matrix form and that uses a mix between regression and lower rank approximations (LRA). The method was called GabrielEigen because it is based on the predictive properties of Gabriel's cross-validation method [6] and uses the eigenvectors and eigenvalues of the singular value decomposition (SVD) to obtain the corresponding LRA's. This paper will be focused exclusively on GabrielEigen.

To the best of our knowledge, the literature does not yet provide a study that evaluates the performance of a regularised version of the GabrielEigen approach. Regularisation is a technique used widely in statistics and data science to correct the problem of overfitting, thus avoiding poor quality imputations and problematic parameter estimation [7,8,9]. Taking this into account, our proposal is to replace the default SVD of the original imputation system with a regularised LRA or equivalently a regularised SVD - regSVD [10].

GabrielEigen method

Suppose that the $(n \times p)$ matrix $\mathbf{X}$ contains elements x_{ij} ($i = 1,\ldots,n$; $j = 1,\ldots,p$), some of which are missing. The rows represent genotypes and the columns the environments. Step 1: Start by inserting into each missing entry the mean of its column, thereby obtaining a completed matrix $\mathbf{X}$. Step 2: The columns of $\mathbf{X}$ are standardised by subtracting m_j from each element and dividing the result by s_j (where m_j and s_j are respectively the mean and the standard deviation of the jth column). Step 3: Using the standardised matrix, each original missing entry x_{ij} is replaced by

$$x_{ij}^{(m)} = \boldsymbol{x}_{1\bullet}^{T} \boldsymbol{V} \boldsymbol{D}^{+} \boldsymbol{U}^{T} \boldsymbol{x}_{\bullet 1} \tag{1}$$

where $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T}$ represents the SVD of $\boldsymbol{X}_{11}$, $\boldsymbol{D}^{+}$ is the Moore-Penrose generalised inverse of $\boldsymbol{D}$ and $\boldsymbol{V}\boldsymbol{D}^{+}\boldsymbol{U}^{T}\boldsymbol{x}_{\bullet 1}$ is the regression of the first column omitting the first row in (2).

Here the vectors $\boldsymbol{x}_{1\bullet}^{T}$, $\boldsymbol{x}_{\bullet 1}$ and the matrices $\boldsymbol{V}$, $\boldsymbol{D}$ and $\boldsymbol{U}$ are obtained from the partition

$$\boldsymbol{X} = \begin{bmatrix} x_{ij} & \boldsymbol{x}_{\bullet 1}^{T} \\ \boldsymbol{x}_{\bullet 1} & \boldsymbol{X}_{11} \end{bmatrix} \tag{2}$$

with $\boldsymbol{X}_{11} = \sum_{k=1}^{m} \boldsymbol{u}_{(k)} d_k \boldsymbol{v}_{(k)}^{T} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T}$, where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m]$, $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]$, $\boldsymbol{D} = diag(d_1, \ldots, d_m)$ and $m \leq \min\{n-1, p-1\}$. Also note that for each missing observation the components of the considered partition will be different, and this partition is obtained through elementary operations on the rows and columns of $\mathbf{X}$. Step 4**:** This imputation process depends on the choice of the value for m in Step 3 and it is usual to choose m to be the smallest value satisfying

$$\frac{\sum_{k=1}^{m} d_k^2}{\sum_{k=1}^{\min\{n-1,p-1\}} d_k^2} \geq 0.75. \tag{3}$$

Step 5: Finally, the imputed values $\hat{x}_{ij}^{(m)}$ must be returned to their original scale, $x_{ij} = m_j + s_j \hat{x}_{ij}^{(m)}$, replacing them in the matrix $\mathbf{X}$. Steps 2 to 5 are then iterated until the imputations achieve stability. This process assumes that $n>p$. If this is not the case, then the matrix should first be transposed before conducting the iterations.

Proposed modifications

To obtain a regularised version of GabrielEigen it is necessary to replace the standard SVD with a regSVD. Fortunately, both the statistical and data science literature provide many such algorithms [10,11,12,13]. However, taking into account the characteristics of the imputation system, which works with a different matrix partition for each missing observation, a simple, fast and stable algorithm is needed that does not lose much of the calculation speed of the original method nor the convergence of the regularised imputations. For these reasons, in this study we adopted the proposal by Zheng et al. [14] to obtain the regSVD of $\boldsymbol{X}_{11}$ and we now describe the steps to achieve it.

Table 1
Datasets chosen to perform the cross-validation study.

References	Species	No. of genotypes	No. of environments	Response variable	AMMI model to explain the interaction $G \times E$
Yan et al. [16]	Wheat	18	9	Mean yield	AMMI2
Lavoranti [17]	Eucalyptus	20	7	Mean tree height	AMMI2
Calinski et al. [18]	Pea	18	9	Mean yield	AMMI1
Calinski et al. [19]	Rye	18	15	Mean yield	AMMI2
Farias [20]	Cotton	15	27	Mean yield	AMMI1
Filho et al. [21]	Cotton	17	23	Mean yield	AMMI5
Flores et al. [22]	Bean	15	12	Mean yield	AMMI4
Mattos et al. [23]	Sugarcane	22	5	Mean yield	AMMI1
Rad et al. [24]	Wheat	36	6	Mean yield	AMMI3
Yang [25]	Barley	6	18	Yields	AMMI1

Let $\boldsymbol{X}_{11}$ denote the matrix $(n-1 \times p-1)$ and m the desired rank. i) Initially, a $\boldsymbol{V}$ $(p-1 \times m)$ matrix is obtained with random entries from a uniform distribution (0,1). ii) The matrix $\boldsymbol{U}$ $(n-1 \times m)$ is calculated as follows $\boldsymbol{U} = \boldsymbol{X}_{11}\boldsymbol{V}(\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V} + \lambda \boldsymbol{I}_m)^{+}$ where $\boldsymbol{I}_{\boldsymbol{m}}$ represents the identity matrix $(m \times m)$, $(\bullet)^{+}$ represents a generalised inverse and λ is the regularisation parameter (in the original algorithm the ordinary inverse is used, but in previous tests on matrices of real data some singularities have appeared). iii) The matrix $\boldsymbol{V}$ is updated through $\boldsymbol{V} = \boldsymbol{X}_{11}^{T}\boldsymbol{U}(\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{U} + \lambda \boldsymbol{I}_m)^{+}$. iv) The value of the regularised objective function is calculated, that is, $J = \|\boldsymbol{X}_{11} - \boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}\|_F^2 + \lambda\|\boldsymbol{U}\|_F^2 + \lambda\|\boldsymbol{V}\|_F^2$. v) Steps ii), iii) and iv) are repeated iteratively until reaching convergence in the value of J. vi) The standard SVD is calculated over $\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}$ to obtain the corresponding regularised eigenvalues and eigenvectors. In this way, the imputation equation of the regularised GabrielEigen becomes $x_{ij}^{(m)} = \boldsymbol{x}_{1\bullet}^{T}\boldsymbol{V}_{\boldsymbol{reg}}\boldsymbol{D}_{reg}^{+}\boldsymbol{U}_{reg}^{T}\boldsymbol{x}_{\bullet 1}$, where $\boldsymbol{U}_{\boldsymbol{reg}}\boldsymbol{D}_{\boldsymbol{reg}}\boldsymbol{V}_{reg}^{T}$ represents the regSVD of $\boldsymbol{X}_{11}$ where λ can be chosen by a direct search as follows. For each value of λ from 0 to 1 in steps of 0.1, the above process is repeated and the value of λ that yields the minimum value of J is noted. If this value occurs at λ strictly less than 1, then the process ends. If the minimising value is 1, however, the whole process is repeated for λ in steps of 0.1 from 1 to 2. Again, if the minimising value of λ occurs at strictly less than 2, that is the chosen value of λ. Otherwise, the process is repeated for λ in steps of 0.1 from 2 to 3, and so on until the minimising value occurs strictly within one of these intervals. Note that when λ is zero, the standard SVD is obtained, in which case the original GabrielEigen is obtained.

Validation of the method

To evaluate the regularised GabrielEigen, a cross-validation study based on real data was performed using the methodology and open access datasets presented by García-Peña et al. [5,15]. Table 1 presents basic information of the ten complete matrices considered with the corresponding references in case the reader needs additional information. According to García-Peña et al. [5], the multi-environment data have different interaction structures that were determined by the Eigenvector method to choose the best additive main effects model with multiplicative interaction – AMMIk (k components to explain the genotype × environment interaction - $G \times$ E*)*. In this case, there is a simple interaction for those data that can be explained by an AMMI1 model, an intermediate interaction when explained with an AMMI2 model and a complex interaction when more than two components are needed.

Missing values were introduced into each experimental matrix $(\boldsymbol{Y})$ via four percentages (P = 0, 5, 10 e 20%) of missing not at random (MNAR) deletions, which is very common in this class of studies. When the missing percentage was zero, one observation was removed at a time and the regularised GabrielEigen was applied to obtain a matrix with the corresponding imputations of all positions. When the missing percentage was different from zero, the Pth percentile was initially found in each environment or column of $\boldsymbol{Y}$ and data that were smaller than said percentile were considered missing. Subsequently, on each incomplete matrix $(\boldsymbol{Y}_I)$ each element was removed in turn and the regularised GabrielEigen was applied to obtain the corresponding imputation. The positions that were missing from the beginning were imputed once with the regularised method. In this way, cross-validation matrices were obtained that contained the imputations for all positions and which were called $\boldsymbol{I}_{CV}$.

To compare $\boldsymbol{Y}$ and $\boldsymbol{I}_{CV}$, matrices, three criteria were chosen: the prediction error [26], the Procrustes similarity statistic [27] and the critical angle between two subspaces of principal components [28,29]. The prediction error P_e is defined by the square root of the mean squared error between the true values (removed) and the corresponding imputations. An alternative way to compare the matrices was using the Procrustes M^2 statistic where $M^2 = trace(\boldsymbol{Y}\boldsymbol{Y}^{\boldsymbol{T}} + \boldsymbol{I}_{\boldsymbol{CV}}\boldsymbol{I}_{\boldsymbol{CV}}^{T} - 2\boldsymbol{Y}\boldsymbol{Q}\boldsymbol{I}_{CV}^{T})$ and $\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{O}^{\boldsymbol{T}}$ represents the rotation matrix calculated from the SVD elements of the matrix $\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{I}_{CV} = \boldsymbol{O}\boldsymbol{\Sigma}\boldsymbol{R}^{\boldsymbol{T}}$. Finally, to obtain the critical angle (θ), the SVD´s of $\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{G}\boldsymbol{P}^{\boldsymbol{T}}$ and $\boldsymbol{I}_{CV} = \boldsymbol{W}\boldsymbol{J}\boldsymbol{K}^{\boldsymbol{T}}$ were calculated, then the matrices $\boldsymbol{Y}$ and $\boldsymbol{I}_{CV}$ with retained k components, $\boldsymbol{Y}_{(k)}$ and $\boldsymbol{I}_{CV(k)}$, were compared and the critical angle calculated as $\theta = cos^{-1}(d)$, where d is the smallest element of $\boldsymbol{L}$ in the SVD of matrix $\boldsymbol{Y}_{(k)}^{T}\boldsymbol{I}_{CV(k)} = \boldsymbol{S}\boldsymbol{L}\boldsymbol{A}^{\boldsymbol{T}}$. The best λ using the regularised method will be the one that minimises each criterion, with $\lambda = 0$ being the standard comparison value as it represents the original GabrielEigen. We do not expect that in all cases a single λ will minimise the three statistics simultaneously, because each of them represents a different feature in evaluating performance over matrices that have different structures for the $G \times$ E interaction. For this reason, being a very flexible and adaptive method, the selection of λ was carried out by repeating the above procedure for each value of λ from 0 to 1 in steps of 0.1, each criterion in turn in place of J and recording the value that minimised

Table 2
Summary of the cross-validation study on the Yan et al. [16] data.

P	λ	P_e	M^2	θ
0	0	0.3888	12.7456	0.0814
	0.3	0.3908	12.4800	0.0787
0.05	0	0.4060	31.8679	0.5286
	0.4	0.4023	28.7060	0.4022
0.1	0	0.5173	136.8122	1.4028
	0.4	0.6348	210.0024	0.7066
	0.6	0.4878	102.7738	1.1851
0.2	0	0.5849	182.9877	1.0119
	0.6	0.5902	156.7244	0.7005
	0.9	0.5849	220.3691	0.5925
	1	0.5842	218.3659	0.7377

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Table 3
Summary of the cross-validation study on the Lavoranti [17] data.

P	λ	P_e	M^2	θ
0	0	0.8491	25.7167	1.1074
	0.6	0.8486	27.2383	0.8544
	1	0.8685	21.8352	0.2287
	1.2	0.8748	21.9613	0.2281
0.05	0	0.9125	36.7293	1.3471
	1.1	0.8828	28.8920	1.1073
	2.1	0.9357	25.7175	0.3664
	2.3	0.9668	26.0377	0.2887
0.1	0	1.0594	57.7581	1.3484
	1.7	0.9520	38.2263	0.5742
	1.8	0.9661	37.4250	0.6785
0.2	0	1.1275	105.1459	0.7153
	1.1	1.0561	65.2960	1.3609
	1.2	1.0544	69.4481	1.2718
	1.4	1.6390	90.8105	0.5738

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Table 4
Summary of the cross-validation study on the Calinski et al. [18] data.

P	λ	P_e	M^2	θ
0	0	3.3340	71.8047	1.5708
	0.3	3.3172	60.9821	1.5708
0.05	0	3.1381	116.4965	1.5708
	1.3	3.1951	108.6089	1.5708
0.1	0	3.2787	145.2247	1.5708
	0.5	3.2472	148.2420	1.5708
	1.1	3.2937	132.0886	1.5708
0.2	0	3.4189	147.1461	1.5708
	0.1	3.4280	146.5964	1.5708
	0.2	3.4116	162.3310	1.5708

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

each of the three criteria. If the minimising value occurred at 1, the interval of search was extended to 2 in further steps of 0.1, and similarly beyond 2 if necessary.

Tables 2-11 present the summaries of the cross-validation studies carried out on each data set respectively, the complete results are available in the Supplementary material section. In each table, the percentage P considered and the values of λ that minimised the comparison criteria are found. When only $\lambda = 0$ appears in some percentage, it indicates that the best method was the original GabrielEigen. For example, in Tables 3 and 8, it is observed that regularised GabrielEigen was superior in all percentages because it always minimised the performance statistics with a penalty value, λ, different from zero.

Table 5
Summary of the cross-validation study on the Calinski et al. [19] data.

P	λ	P_e	M^2	θ
0	0	4.1180	44.3537	0.4627
	0.5	4.1421	42.9701	0.3926
0.05	0	4.1683	50.7815	0.6590
	0.5	4.1358	47.2441	0.5100
	5.5	6.6672	45.4674	0.4813
0.1	0	4.1930	63.3659	0.8554
	1	4.1719	49.7279	0.5656
	3.5	4.9132	48.2016	0.5642
	5	7.0407	50.1386	0.5544
0.2	0	4.6765	87.9293	1.2874
	1	4.5474	62.6246	0.7142

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Table 6
Summary of the cross-validation study on the Farias [20] data.

P	λ	P_e	M^2	θ
0	0	313.3214	11.6758	1.5708
	0	628.3844	79.0002	1.5708
0.05				
0.1	0	655.0411	86.0643	1.5708
	0.1	654.4184	86.2050	1.5708
0.2	0	791.5721	127.0425	1.5708

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Table 7
Summary of the cross-validation study on the Filho et al. [21] data.

P	λ	P_e	M^2	θ
0	0	494.3442	35.8798	0.3167
	0.1	494.4528	35.9373	0.3142
0.05	0	503.5711	46.2577	0.4505
	0.4	501.1091	45.7147	0.4994
	1	505.1621	45.4245	0.6029
0.1	0	510.6541	49.1562	0.9724
	0.2	507.5024	48.8906	1.0174
0.2	0	805.1187	149.3996	1.2661

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

In the four datasets that had a simple $G \times E$ interaction structure (Tables 4, 6, 9 and 11) the critical angle was not effective as an evaluation criterion, since the value of d was zero, therefore, the value of θ was the same in the four datasets studied. Thus, for these sets the comparison was based on P_e and M^2. Of these four datasets, only one (Table 6) showed that GabrielEigen ($\lambda = 0$) was superior to the regularised version minimising the statistics in almost all percentages. In the three remaining data sets, the regularised version outperformed the original method, minimising the statistics in at least two of the percentages studied (Tables 4, 9 and 11).

Cross-validation studies on data matrices with intermediate $G \times E$ interaction structure (Tables 2, 3 and 5) showed that the regularised system always minimised the critical angles, similarity statistics and prediction errors between the real and the matrices that contained the imputations when the percentage of missing data was different from 0%. In this specific case ($P = 0\%$), when there was only one missing observation, the original method outperformed the regularised version in two of the three data sets (Tables 2 and 5).

Finally, considering the three data sets with complex $G \times E$ interaction (Tables 7, 8 and 10), the regularised version minimised the statistics in most of the considered scenarios, but there were some exceptions in which the original method outperformed the regularised one. For example, in Table 7, when the percentage of missing data was 0 and 20%, the original method outperformed the regularised one by minimising P_e and M^2, but θ was minimised using GabrielEigen ($\lambda \neq 0$) when $P = 0\%$. Tables 8 and 10 show the best performance of the regularised version proposed in this research for complex interactions.

Table 8
Summary of the cross-validation study on the Flores et al. [22] data.

P	λ	P_e	M^2	θ
0	0	408.6810	80.5414	1.1105
	0.3	408.3536	61.3890	0.8755
	1.8	428.0701	38.9867	0.4631
	2	411.2996	39.1653	0.2926
0.05	0	397.5333	60.3773	1.3677
	0.5	392.8433	55.0385	1.1930
	1.2	406.2281	54.2030	0.7543
	1.4	411.5636	54.8298	0.7373
0.1	0	410.4625	95.9522	1.4797
	0.1	410.1320	84.5007	1.4695
	0.6	404.0759	96.7766	0.8229
	1.4	426.5184	106.0734	0.6046
0.2	0	449.0753	147.7002	1.3157
	0.6	455.4980	142.6922	0.6836
	1	437.3100	136.0719	1.4614
	1.1	438.0416	134.5970	1.2774

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Table 9
Summary of the cross-validation study on the Mattos et al. [23] data.

P	λ	P_e	M^2	θ
0	0	1.7100	37.5537	1.5708
	0.1	1.7094	37.2896	1.5708
0.05	0	1.7786	59.9372	1.5708
	0.3	1.7733	62.5957	1.5708
	0.9	1.7924	39.5298	1.5708
0.1	0	1.8450	66.4930	1.5708
	0.8	1.8183	59.7910	1.5708
	1	1.8191	51.5971	1.5708
0.2	0	1.8946	125.9495	1.5708
	0.1	1.8932	98.6778	1.5708

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Table 10
Summary of the cross-validation study on the Rad et al. [24] data.

P	λ	P_e	M^2	θ
0	0	0.9865	90.5471	1.3099
	1.6	1.3846	29.7955	0.5241
0.05	0	1.0282	119.5269	1.2888
	0.1	1.0201	123.7278	1.2921
	1.2	1.3977	57.2260	0.1416
	1.6	1.3756	49.2441	0.1794
0.1	0	1.2426	68.8992	0.1288
	0.6	1.1709	57.6604	0.1131
0.2	0	1.5559	130.5261	0.0859
	0.1	1.5511	130.0996	0.0850
	0.2	1.5486	130.3880	0.0875

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Table 11
Summary of the cross-validation study on the Yang [25] data.

P	λ	P_e	M^2	θ
0	0	0.4773	1.1489	1.5708
0.05	0	0.5713	4.2444	1.5708
	0.3	0.5586	3.7725	1.5708
	0.6	0.5811	3.4985	1.5708
0.1	0	0.5834	5.1723	1.5708
	0.3	0.5711	4.6265	1.5708
0.2	0	1.3669	41.3616	1.5708

In red, the minimised values of the statistics by regularised GabrielEigen in each percentage considered. $\lambda = 0$ represents the original GabrielEigen.

Summary comments

A generalization of the GabrielEigen imputation method has been proposed, using a regSVD. The penalised version of the system shows that the original formulation should not be applied to any data set without first performing an adequate study to choose the regularisation parameter λ. In this article, the choice of this value proved to be quite flexible regardless of the type of interaction, the dimension of the matrix and the percentage of missing data in matrices with $G \times$ E interaction. Likewise, given that the system can be applied to any data matrix, it can be an alternative for non-parametric imputation and without structural assumptions for multivariate data.

For the choice of the appropriate λ we have proposed three criteria that should be minimised. In general, this λ will depend on the data set studied, but we recommend starting the study in the interval [0;2.5] because in nine of the ten data sets studied in this interval, P_e, M^2 and θ were minimised. Naturally, additional research will be needed, for example, to investigate the performance of regularised GabrielEigen under other absence mechanisms or to study the effect that different probability distributions can have on λ. Lastly, the method also looks promising for wrapping around methodologies that have already been proposed to obtain robust and multiple imputations [5,15].

Ethics statements

Not applicable

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Sergio Arciniegas-Alarcón: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Marisol García-Peña:** Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Wojtek J. Krzanowski:** Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Camilo Rengifo:** Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

The authors of this paper acknowledge the High-Performance Computing Center–ZINE of Pontificia Universidad Javeriana for assistance during the cross-validation study. The authors acknowledge the support of the Pontificia Universidad Javeriana and the Universidad de La Sabana, through projects 10756 (research group Física Matemática) and ING-309-2023 (research group Física y Matemáticas Aplicadas) respectively.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2023.102289.

References

[1] C.T.S. Dias, Methods for choosing components in the additive main effect and multiplicative interaction (AMMI) model, 2005, doi:10.11606/T.11.2006.tde-27042006-064514.
[2] P.M. Amoêdo, S.M.S. Piedade, C.T.S. Dias, S. Arciniegas-Alarcón, Multiple imputation MIGAMMI algorithm, Brazilian J. Biometrics (40) (2022) 1–20.
[3] W.J. Krzanowski, Statistical Principles and Techniques in Scientific and Social Research, Oxford University Press, pp 241 + xiv, 2007.
[4] M. García-Peña, S. Arciniegas-Alarcón, W.J. Krzanowski, D. Barbin, Multiple imputation procedures using the GabrielEigen algorithm, Commun. Biometry Crop Sci. (11) (2016) 149–163.
[5] M. García-Peña, S. Arciniegas-Alarcón, W.J. Krzanowski, Missing value imputation using least squares techniques in contaminated matrices, MethodsX (9) (2022) 101683.
[6] K.R. Gabriel, Le biplot–outil d'exploration de données multidimensionnelles, J. Soc. Franç. Statist. (143) (2002) 5–55.
[7] P.J. Bickel, B. Li, Regularization in statistics, Test (15) (2006) 271–344.
[8] J. Josse, F. Husson, Handling missing values in exploratory multivariate data analysis methods, J. Soc. Franç. Statist. (153) (2012) 79–99.
[9] T. Hastie, Ridge regularization: an essential concept in data science, Technometrics (62) (2020) 426–433.
[10] Z. Hu, F. Nie, R. Wang, X. Li, Low rank regularization: a review, Neural Netw. (136) (2021) 218–232.
[11] H. Shen, J.Z. Huang, Sparse principal component analysis via regularized low rank matrix approximation, J. Multivar. Anal. (99) (2008) 1015–1034.
[12] L. Zhang, H. Shen, J.Z. Huang, Robust regularized singular value decomposition with application to mortality data, Ann. Appl. Stat. (7) (2013) 1540–1561.
[13] Y. Ji, W. Hong, Y. Shangguan, H. Wuang, J. Ma, Regularized singular value decomposition in news recommendation system, in: Proceedings of the 11th International Conference on Computer Science & Education (ICCSE 2016), Nagoya University, Japan, 2016, pp. 621–626.
[14] S. Zheng, C. Ding, F. Nie, Regularized singular value decomposition and application to recommender system, arXiv:1804.05090, 2018.
[15] M. García-Peña, S. Arciniegas-Alarcón, W.J. Krzanowski, D. Duarte, Missing value imputation using the robust singular-value decomposition: proposals and numerical evaluation, Crop Sci. (61) (2021) 3288–3300.
[16] W. Yan, M.S. Kang, B. Ma, S. Woods, P.L. Cornelius, GGE biplot vs. AMMI analysis of genotype-by-environment data, Crop Sci. (47) (2007) 643–653.
[17] O.J. Lavoranti. Phenotypic stability and adaptability via ammi model with bootstrap re-sampling, 2003, doi:10.11606/T.11.2003.tde-22102003-160700.
[18] T. Calinski, S. Czajka, Z. Kaczmarek, P. Krajewski, W. Pilarczyk, Analyzing the genotype-by-environment interactions under a randomization-derived mixed model, J. Agricult., Biol., Environ. Statistics (14) (2009) 224–241 a.
[19] T. Calinski, S. Czajka, Z. Kaczmarek, P. Krajewski, W. Pilarczyk, A mixed model analysis of variance for multi-environment variety trials, Stat. Papers (50) (2009) 735–759.
[20] F.J.C. Farias. Selection index in upland cotton cultivars, 2005, doi:10.11606/T.11.2005.tde-12012006-162727.
[21] J.L.S. Filho, C.L. Morello, F.J.C. Farias, F.M. Lamas, M.B. Pedrosa, J.L. Ribeiro, Comparison of methods for the evaluation of adaptability and stability for yield in cotton genotypes, Pesq. Agropec. Bras. (43) (2008) 349–355.
[22] F. Flores, M.T. Moreno, J.I. Cubero, A comparison of univariate and multivariate methods to analyze G ×E interaction, Field Crops Res. (56) (1998) 271–286.
[23] P.H.C. Mattos, R.A.J. Oliveira, C.B. Filho, E. Daros, M.A.A. Veríssimo, Evaluation of sugarcane genotypes and production environments in Paraná by GGE biplot and AMMI analysis, Crop Breed. Appl. Biotechnol. (13) (2013) 83–90.
[24] M.R.N. Rad, M.A. Kadir, M.Y. Rafii, H.Z.E. Jaafar, M.R. Naghavi, F. Ahmadi, Genotype ×environment interaction by AMMI and GGE biplot analysis in three consecutive generations of wheat (Triticum aestivum) under normal and drought stress conditions, Aus. J. Crop. Sci. (7) (2013) 956–961.
[25] R.C. Yang, Mixed model analysis of crossover genotype-environment interactions, Crop. Sci. (47) (2007) 1051–1062.
[26] W. Yan, Biplot analysis of incomplete two-way data, Crop Sci. (53) (2013) 48–57.
[27] W.J. Krzanowski, Principles of Multivariate Analysis: A User's Perspective Oxford, University Press, Oxford, UK, 2000.
[28] W.J. Krzanowski, Between-group comparison of principal components–some sampling results, J. Stat. Comput. Simul. (15) (1982) 141–154.
[29] W.J. Krzanowski, Cross-validation in principal component analysis, Biometrics (43) (1987) 575–584.