




Article

Cross-Validation for Lower Rank Matrices Containing Outliers

Sergio Arciniegas-Alarcón ¹, Marisol García-Peña ^{2,*} and Wojtek J. Krzanowski ³

¹ Facultad de Ingeniería, Universidad de La Sabana, Campus Puente del Común, Km. 7 Autopista Norte, Chía 140013, Colombia; sergio.arciniegas@unisabana.edu.co

² Departamento de Matemáticas, Pontificia Universidad Javeriana, Carrera 7 40-62, Bogotá 110231, Colombia

³ College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK; w.j.krzanowski@exeter.ac.uk

* Correspondence: luzmara@gmail.com

Abstract: Several statistical techniques for analyzing data matrices use lower rank approximations to these matrices, for which, in general, the appropriate rank must first be estimated depending on the objective of the study. The estimation can be conducted by cross-validation (CV), but most methods are not designed to cope with the presence of outliers, a very common problem in data matrices. The literature suggests one option to circumvent the problem, namely, the elimination of the outliers, but such information removal should only be performed when it is possible to verify that an outlier effectively corresponds to a collection or typing error. This paper proposes a methodology that combines the robust singular value decomposition (rSVD) with a CV scheme, and this allows outliers to be taken into account without eliminating them. For this, three possible rSVD's are considered and six resistant criteria are proposed for the choice of the rank, based on three classic statistics used in multivariate statistics. To test the performance of the various methods, a simulation study and an analysis of real data are described, using an exclusively numerical evaluation through Procrustes statistics and critical angles between subspaces of principal components. We conclude that, when data matrices are contaminated with outliers, the best estimation of rank is the one that uses a CV scheme over a robust lower rank approximation (RLRA) containing as many components as possible. In our experiments, the best results were obtained when this RLRA was calculated using an rSVD that minimizes the L_2 norm.

Keywords: outliers; resistant statistics; singular value decomposition



Citation: Arciniegas-Alarcón, S.; García-Peña, M.; Krzanowski, W.J. Cross-Validation for Lower Rank Matrices Containing Outliers. *Appl. Syst. Innov.* **2022**, *5*, 69. <https://doi.org/10.3390/asi5040069>

Academic Editor: Patricia Ramos

Received: 18 June 2022

Accepted: 15 July 2022

Published: 19 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The singular value decomposition (SVD) is a mathematical result that allows the calculation of a low-rank approximation of any matrix and serves as the basis of many statistical methods used in data analysis [1]. The application areas are diverse, for example, SVD can be used in principal component analysis [2], the imputation of missing data [3], the graphical representation of multivariate data [4], the formulation of models to explain the interaction in two-way tables [5,6] and non-parametric analysis of time series [7], to mention just a few.

A problem inherent in the use of SVD is the determination of the appropriate rank of the approximation, i.e., the appropriate number of components to be retained, and a very convenient way to solve this problem uses the resampling technique known as cross-validation. Standard cross-validation consists of subdividing the study matrix into a certain number of groups, deleting each group in turn, evaluating the parameters of a chosen predictor from the remaining data and using the result to predict the deleted values [8].

A numerical measure of agreement between predicted and actual values can then be computed for each possible rank, and the best rank to choose is the one providing the best calculated measure. Readers interested in classic references, alternative procedures, method comparisons and new developments of the subject can refer to the works of

Bro et al. [9], Owen and Perry [10], Josse and Husson [11], Camacho and Ferrer [12] and Saccenti and Camacho [13].

Among the wide variety of methods for conducting the cross-validation, one considered as classic and influential is the Eastment and Krzanowski [14] (EK) scheme proposed in 1982, and which currently (April 2022) has 452 citations in Google Scholar. The method depends on the SVD, so does not rest on any distributional or structural assumptions. Moreover, it provides exact calculations, so does not have the convergence problems that can occur with expectation maximization (EM) approximations [15].

A frequent problem in data analysis is the presence of outliers [16] and the performance of the EK method can be affected by them because of its use of SVD, which is a least squares technique. To circumvent the problem, an option used by Krzanowski [17] was to compare the results obtained for the complete data with those for the reduced data, after eliminating the outliers. To the best of our knowledge, no specific study has been conducted to evaluate the effect of outliers on the EK method when applied to lower rank matrices; therefore, the aim of the current paper is to propose a methodology that combines robust singular value decomposition (rSVD) with the EK method and that does not require the elimination of outliers. This proposed methodology would therefore be applicable for any dataset that has some suspicion of contamination and the outliers do not correspond to errors in the collection or typing, as in that case the elimination of information is perfectly sensible.

This article is organized as follows: the EK method is presented first, followed by the various proposed methods along with associated statistics that robustify the criterion for choosing the rank of a matrix. Subsequently, three rSVD's are presented that can be used as a way to reduce the effect of outliers, but without eliminating them. Then, a simulation study is described for evaluating possible cross-validation strategies for different levels of contamination using different resistant statistics. A study of a real dataset is also described to compare the proposed alternatives, and finally the results are presented together with a relevant discussion.

2. EK Method

Consider a standardized data matrix $Y (n \times p)$ with elements $y_{ij} (i = 1, \dots, n; j = 1, \dots, p)$ and $n \geq p$ (if $n < p$ the matrix should be first transposed), for which we wished to determine the best lower rank approximation. For this, Eastment and Krzanowski's [14] cross-validation scheme can be used, which quantifies the idea of "acceptable accuracy" in terms of predicting the elements of Y . The scheme is based on the fact that element y_{ij} in the i -th row and in the j -th column of Y can be written as a multiplicative model using SVD, that is:

$$y_{ij} = \sum_{t=1}^p u_{it} s_t v_{tj} \tag{1}$$

Krzanowski [8] used this representation as a basis for determining the dimensionality of a multivariate dataset: if the data structure is essentially m -dimensional, then the variation in the remaining $(p-m)$ dimensions can be treated as random noise. For this reason, it can be assumed that the main characteristics of the data are found in the space of the first m components. This means that the data can be written according to an m -component model, such as:

$$y_{ij} = \sum_{t=1}^m u_{it} s_t v_{tj} + \varepsilon_{ij} \tag{2}$$

where ε_{ij} is a residual term and setting ε_{ij} equal to zero, the expression results in a predictor of y_{ij}

$$\hat{y}_{ij}^{(m)} = \sum_{t=1}^m u_{it} s_t v_{tj} \tag{3}$$

Since the calculations of u_{it} , s_t and v_{tj} involve all values of Y , the predictor of y_{ij} uses the value of y_{ij} itself. This is an undesirable feature in cross-validation because it

can be a source of bias. To avoid this bias, Eastment and Krzanowski [14] suggested the following scheme: suppose we are looking for the prediction of y_{ij} in Y , then, the i -th row from Y is deleted and the SVD for the $((n - 1) \times p)$ resulting matrix $Y^{(-i)}$ is calculated as $Y^{(-i)} = \overline{U}\overline{D}\overline{V}^T$, $\overline{U} = (\overline{u}_{sh})$, $\overline{V} = (\overline{v}_{sh})$, $\overline{D} = (\overline{d}_1, \dots, \overline{d}_p)$. The next step is to delete the j -th column from Y and obtain the SVD for the $(n \times (p - 1))$ matrix $Y_{(-j)}$ as $Y_{(-j)} = \tilde{U}\tilde{D}\tilde{V}^T$, $\tilde{U} = (\tilde{u}_{sh})$, $\tilde{V} = (\tilde{v}_{sh})$, $\tilde{D} = (\tilde{d}_1, \dots, \tilde{d}_{p-1})$. The matrices \overline{U} , \overline{V} , \tilde{U} , and \tilde{V} are orthonormal, while \tilde{D} and \overline{D} are diagonal. By combining the two SVDs, $Y^{(-i)}$ and $Y_{(-j)}$, a predictor of Y_{ij} based on the m component model is given by

$$\hat{y}_{ij}^{(m)} = \sum_{h=1}^m \tilde{u}_{ih} \left(\tilde{d}_h \sqrt{p/(p-1)} \right)^{\frac{1}{2}} \overline{v}_{jh} \left(\overline{d}_h \sqrt{n/(n-1)} \right)^{\frac{1}{2}} \tag{4}$$

This predictor is slightly modified relative to the one proposed by Eastment and Krzanowski [14] following the work of Bro et al. [9] and Arciniegas-Alarcón et al. [3] because the inclusion of the constants $\sqrt{p/(p-1)}$ and $\sqrt{n/(n-1)}$ improves the quality of the predictions. On the other hand, in order to avoid computational problems, a parity check should be done in each prediction by matching the sign of $\left(\tilde{u}_{ih} \left(\tilde{d}_h \sqrt{p/(p-1)} \right)^{\frac{1}{2}} \right) \left(\overline{v}_{jh} \left(\overline{d}_h \sqrt{n/(n-1)} \right)^{\frac{1}{2}} \right)$ in (4) to the sign of $u_{ih}d_hv_{jh}$ obtained from the SVD of the Y matrix for each m .

After establishing the predictor of the observations, an overall measure of predictive accuracy of the m -component model is given by

$$PRESS(m) = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \left(\hat{y}_{ij}^{(m)} - y_{ij} \right)^2 \tag{5}$$

This function can be computed for each value of m , where $m = 1, \dots, p - 1$ and an optimal choice of m (the rank of Y) can then be based on some appropriate function of these values. The suggestion made by Krzanowski [17] and Krzanowski and Kline [18] was to calculate the statistic

$$W_m = \frac{PRESS(m-1)}{n+p-2m} \div \frac{PRESS(m)}{D_R} \tag{6}$$

for each m , where $D_R = (n - m - 1)(p - m)$. To calculate W_1 , it is necessary to define $PRESS(0)$; in this case, you can use $\hat{y}_{ij}^{(m)} = 0$, Forkman and Piepho [19], or $\hat{y}_{ij}^{(m)}$ can be the mean of j -th column without the element (i,j) (Carlos Tadeu dos Santos Dias, personal communication, 3 August 2021). Finally, the m -th component of the SVD can be considered "important" if $W_m > 0.9$ and the total number of important components constitutes an estimation of the optimal rank of the matrix Y taking into account predictive criteria. In those cases where $W_m < 0.9$ for all components, the rank estimation can be determined by the component with the highest value of W_m . The acronym EK82 can be used to identify the method described above.

3. Proposed Methodology

The choice of the optimal rank for a lower rank approximation using the EK method basically depends on two aspects: the predictions of the elements of Y and the criterion to determine if a component is important or not. It is known that the quality of predictions using the standard SVD decreases in the presence of outliers in the data matrix [20]. For that reason, our first attempt at circumventing this problem was to create a robust version of the EK schema. For this, the usual standardization of Y was replaced by a robust standardization and, in the calculation of the elements of the predictor described in (4), the SVDs of $Y^{(-i)}$ and $Y_{(-j)}$ were replaced by an rSVD (details for calculating it are described in the next section).

Although this approach produced good-quality predictions in small test matrices (for instance, 20×5), the main disadvantage was computational because the implementation for larger matrices became too heavy due to the inclusion of rSVDs. In this way, the problem of outliers was circumvented, but the computational speed of the EK method was lost. This could later be a problem in the future for large arrays. Thus, we now have a more complex problem: how do we improve the predictions of the EK scheme in the presence of outliers, without eliminating them while at the same time trying to preserve the computational speed of the method as much as possible?

To resolve this issue, we decided to use the work of Arciniegas-Alarcón et al. [21] and our proposal was to conduct the cross-validation in two stages: (i) calculate an rSVD on the original matrix Y and obtain a robust lower rank approximation Y_{RLRA} with the maximum possible number of components, that is, with the smallest number between n and p . With this Y_{RLRA} , we can obtain a good approximation of the original values, and in the case of outliers we can obtain a robust approximation without eliminating any information and following the maximum-data precept. (ii) The EK method is then applied to Y_{RLRA} , and the optimal rank is determined. Such a combination of methodologies (rSVD + EK) was not found in our literature review so becomes a possible way to perform a robust cross-validation on contaminated lower rank matrices that is computationally efficient as the size of the study matrix increases.

Determining the rank depends on the chosen criterion; therefore, in addition to W_m , we could use several criteria based on trimmed PRESS [22]. The following is a list of the possible criteria we considered:

1. PRESS: According to Equation (5), the optimal rank is the one that minimizes the statistic. Bro et al. [9] found that, in some cases, this criterion may be more effective than W_m .
2. PRESS75: A resistant PRESS statistic is constructed by averaging the 75% of the smallest squared errors $(\hat{y}_{ij}^{(m)} - y_{ij})^2$. The optimal rank is the one that minimizes PRESS75.
3. PRESS50: A resistant PRESS statistic is constructed by averaging the 50% of the smallest squared errors $(\hat{y}_{ij}^{(m)} - y_{ij})^2$. The optimal rank is the one that minimizes PRESS50.
4. W_m : According to Equation (6), the optimal rank is the total number of important components.
5. $W_{m(Max)}$: The optimal rank is the number of the largest important component using W_m . Krzanowski [8] found that, in some data sets, the W_m statistic does not always show a monotonically decreasing behaviour. For example, if $W_1 = 26.22$, $W_2 = 5.95$, $W_3 = 0.19$ and $W_4 = 1.03$, the optimal rank is 4, even though component 3 is not important.
6. W_{m75} : In Equation (6), PRESS is replaced by PRESS75 and the optimal rank is the total number of important components.
7. $W_{m75(Max)}$: The optimal rank corresponds to the number of the largest important components using W_{m75} .
8. W_{m50} : In Equation (6), PRESS is replaced by PRESS50 and the optimal rank is the total number of important components.
9. $W_{m50(Max)}$: The optimal rank corresponds to the number of the largest important component using W_{m50} .

4. Robust Singular Value Decompositions (rSVDs)

The cross-validation proposed in this paper depends directly on the rSVDs used initially; therefore, to delimit the research, we considered the proposals of Gabriel and Odoroff [23], Hawkins et al. [24] and Zhang et al. [25] for the actual computational procedures. Below, we provide a brief outline of each proposal, but a complete algorithmic description is available from García-Peña et al. [20].

Gabriel and Odoroff's [23] proposal cyclically used fits of rank-one approximations and obtained the residuals after each of these adjustments. They inserted medians and trimmed means into the technique of reciprocal averaging used to minimize the L_2 norm. A computational implementation in the R statistical environment [26] is found in a study by Arciniegas-Alarcón et al. [21]. The acronym EK84 will be used to describe the methodology that mixes this rSVD with the EK schema.

On the other hand, Hawkins et al. [24] found the eigenvalues and eigenvectors for each component of the rSVD through an iterative procedure minimizing the L_1 norm. The drawbacks are that this rSVD can be affected in the case of the presence of leverage points, the eigenvectors can be non-orthogonal and the eigenvalues do not always have a decreasing order. A computational implementation is in the `pcaMethods` package of R [26]. The acronym EK01 is used to describe the methodology that mixes this rSVD with the EK schema.

The rSVD by Zhang et al. [25] computed a sequence of robust rank-one approximations. Robust estimates are obtained by minimizing the Huber function in an iterative re-weighted least squares algorithm. The procedure is implemented in the R `RobRSVD` package [26]. The acronym EK13 is used to describe the methodology that mixes this rSVD with the EK schema.

5. Simulation Study

To evaluate the performance of the cross-validation schemes (i.e., EK01, EK13, EK82, and EK84) on contaminated low-rank matrices, one hundred matrices of dimension (100×8) were simulated using the following process: "Clean" observations were generated as $Y = T + E$, where $T = AB^T = [t_{ij}]$ is a rank four matrix with A (100×4), B (8×4), and E (100×8) is "pure noise". The elements of A , B and E were iid $N(0, 1)$. Outliers were produced on each Y matrix in different percentages (0, 5, 10 and 20%), their positions were chosen randomly and were generated using the normal distribution with mean $\mu_j + 100\sigma_j^2$ and variance σ_j^2 . In this case, μ_j and σ_j^2 represent the mean and variance of the j -th column. Contaminated matrices will be noted by Y_C .

The four cross-validation schemes were applied to the Y_C matrices to determine their rank m (each scheme with each statistic considered can provide different ranks). The quality of this choice was evaluated from the predictive point of view, calculating the corresponding (robust) lower rank approximation with m components (\hat{Y}) and comparing it with the original matrix Y without contamination. For this, the Procrustes M^2 statistic was used [27]. In this case, $M^2 = \text{trace}(\mathbf{Y}\mathbf{Y}^T + \hat{\mathbf{Y}}\hat{\mathbf{Y}}^T - 2\mathbf{Y}\mathbf{Q}\hat{\mathbf{Y}}^T)$ where $\mathbf{Q} = \mathbf{V}\mathbf{U}^T$ is the rotation matrix calculated from elements of the SVD of the matrix $\mathbf{Y}^T\hat{\mathbf{Y}} = \mathbf{U}\Sigma\mathbf{V}^T$. The M^2 statistic measures the difference between two configurations of points, so the (robust) lower rank approximation that minimizes this difference indicates the rank selection method that yields the closest match between clean data and calculated approximations in the presence of outliers.

Another criterion for comparing the proposed methods, following the work of Krzanowski [8,28], was the critical angle (θ) between two subspaces of principal components. For this, both the SVD of $Y = \mathbf{U}\mathbf{D}\mathbf{V}^T$ and the (r)SVD of $Y_C = \mathbf{W}\mathbf{J}\mathbf{K}^T$ were calculated. If the cross-validation shows that the rank is m , the matrices \mathbf{V} and \mathbf{K} with m retained components are compared, that is, $\mathbf{V}_{(m)}$ and $\mathbf{K}_{(m)}$. The calculation of the critical angle is defined by $\theta = \cos^{-1}(d)$, where d is the smallest element of \mathbf{G} in the SVD of the matrix $\mathbf{V}_{(m)}^T\mathbf{K}_{(m)} = \mathbf{M}\mathbf{G}\mathbf{P}^T$. The greater the critical angle, the greater the influence of outliers on the principal coefficients components; therefore, the best cross-validation scheme is the one that provides the lowest value for θ .

In each percentage of outliers considered, one hundred values of the Procrustes statistics and one hundred values of critical angles were obtained. From these values, the means of the 90% with the lowest values were calculated. This average is a robust criterion

that leaves out some flaws that a method may have, but which, in general, presents a good behavior [20,29].

6. Real Data

In addition to the simulation study, cross-validation schemes were applied to the real dataset previously studied by Skov et al. [30] and Bro and Smilde [31], from whose work we attained the main description of the data: “Red wines, 44 samples, produced from the same grape (Cabernet sauvignon) were collected. A Foss WineScan instrument was used to measure 14 characteristic parameters of the wines such as the ethanol content, pH, etc. Hence a dataset consists of 44 samples and 14 variables. The actual measurements can be arranged in a table or a matrix of size (44 × 14)”. The data are available at http://www.models.life.ku.dk/Wine_GCMS_FTIR (accessed on 1 April 2022) and, as it is a multivariate matrix with different measurement scales, we used the standardized matrix in all our analyses.

A difficulty in evaluating schemas on real data is the lack of a priori knowledge of what should be “good behaviour”. Because of this, and following the recommendation of Maronna and Yohai [29], we added outliers in the data set following the same procedure described in the simulation study and using the same comparison criteria, but the difference in the simulations occurred only once.

7. Results

Table 1 presents the trimmed means to evaluate the cross-validation schemes when the M^2 statistic was used in the simulation study. It can be seen that, without contamination (i.e., outliers = 0%), the best method is the EK82 method because it minimizes the value of the criterion. This result is to be expected because the standard SVD provides very good results in the absence of outliers. It can also be seen that, of the resistant criteria considered to choose the rank of the matrices, the one that provided the best results was PRESS50. This indicates that using only 50% of the smallest residuals by cross-validation may be sufficient to select the rank.

Table 1. The 0.9-trimmed means of Procrustes values in the simulation study.

0.9-Trimmed Means of Procrustes Values									
Outliers = 0%					Outliers = 10%				
Criterion	Methods				Criterion	Methods			
	EK01	EK13	EK82	EK84		EK01	EK13	EK82	EK84
PRESS	928	411	405	2326	PRESS	5,929,179	9,858,840	10,533,642	2726
PRESS75	845	367	357	2284	PRESS75	6,503,877	10,051,928	10,592,223	2764
PRESS50	802	357	315	2273	PRESS50	9,915,026	13,550,429	13,794,540	2771
W_m	1773	1547	1563	2364	W_m	16,969,901	23,472,838	23,671,632	2703
$W_{m(Max)}$	1431	1017	1000	2359	$W_{m(Max)}$	16,969,901	23,472,838	23,671,632	2706
W_{m75}	1222	841	826	2295	W_{m75}	6,890,112	10,644,089	11,219,853	2741
$W_{m75(Max)}$	1102	647	614	2292	$W_{m75(Max)}$	8,091,972	11,470,833	11,918,769	2748
W_{m50}	1177	822	806	2280	W_{m50}	10,431,960	15,358,453	15,708,124	2753
$W_{m50(Max)}$	1041	640	619	2279	$W_{m50(Max)}$	13,724,787	18,830,770	19,093,611	2760
Outliers = 5%					Outliers = 20%				
Criterion	Methods				Criterion	Methods			
	EK01	EK13	EK82	EK84		EK01	EK13	EK82	EK84
PRESS	1,349,656	5,347,643	5,517,974	2503	PRESS	16,219,919	20,858,768	23,719,760	3794
PRESS75	1,593,606	5,978,138	6,160,294	2500	PRESS75	16,585,561	20,389,484	23,240,805	3875
PRESS50	2,085,285	7,208,180	7,361,273	2491	PRESS50	24,876,820	26,596,384	26,604,794	3896
W_m	6,121,399	12,296,349	12,339,272	2513	W_m	32,095,002	43,746,109	47,429,575	3491
$W_{m(Max)}$	6,145,795	12,296,349	12,339,272	2511	$W_{m(Max)}$	32,300,042	43,746,109	47,429,575	3529
W_{m75}	1,802,253	6,876,657	7,043,231	2498	W_{m75}	17,200,893	20,911,456	23,301,683	3710
$W_{m75(Max)}$	2,427,846	8,446,853	8,567,883	2500	$W_{m75(Max)}$	18,713,152	21,288,046	23,488,022	3736
W_{m50}	2,659,677	8,260,863	8,433,306	2494	W_{m50}	23,995,906	27,336,626	27,389,422	3769
$W_{m50(Max)}$	4,369,880	10,172,078	10,266,390	2493	$W_{m50(Max)}$	28,428,412	31,084,469	29,862,533	3799

In bold, the method with the lowest statistic value in each percentage of outliers.

On the other hand, when there is some degree of contamination in the simulations, according to Procrustes statistics, methods EK01, EK13 and EK84 always present better performances (lower values) than EK82 using any of the criteria considered (resistant or not). This indicates that the estimation of the rank in the presence of outliers should be performed with one of the robust procedures rather than the standard EK method.

Taking into account predictive criteria and making a comparison between the robust methodologies, it is very clear that the EK84 method always provides the best results, being the most consistent and stable in the simulations, regardless of the percentage of outliers. EK84 shows the highest efficiency (lower M^2 values) using the PRESS50 as a rank selection criterion when the contamination level is low (5%), but when said level is intermediate (10%) or high (20%), the criterion with the best performance is W_m .

Table 2 presents the performance of the schemes in the simulations using critical angles as a criterion. It is observed that the smallest angles (very close to zero) are obtained with the EK82 method in the matrices without outliers, while in the face of contamination in any of the three percentages considered, EK84 is the best method. These results confirm what was found with the Procrustes statistics. As for the criterion for choosing the rank in matrices without outliers, we found that any of them can be used, but according to the critical angles, it is not recommended to use either W_m or $W_{m(Max)}$, while in the presence of outliers the EK84 scheme in combination with W_m or PRESS provides the best results.

Table 2. The 0.9-trimmed means of critical angles in the simulation study.

0.9-Trimmed Means of Critical Angles									
Outliers = 0%		Methods			Outliers = 10%		Methods		
Criterion	EK01	EK13	EK82	EK84	Criterion	EK01	EK13	EK82	EK84
PRESS	0.8547	0.1242	0.0000	0.9791	PRESS	1.5224	1.5708	1.5612	1.1699
PRESS75	0.9473	0.1574	0.0000	1.1286	PRESS75	1.5100	1.5568	1.5652	1.2546
PRESS50	1.0053	0.1935	0.0000	1.1629	PRESS50	1.3777	1.3662	1.4003	1.2339
W_m	0.8177	0.7221	0.7330	1.0202	W_m	1.2399	1.2184	1.2110	1.1507
$W_{m(Max)}$	0.6050	0.2747	0.2094	1.0174	$W_{m(Max)}$	1.2399	1.2184	1.2110	1.1608
W_{m75}	0.6846	0.0950	0.0000	1.1048	W_{m75}	1.4913	1.5103	1.5134	1.2072
$W_{m75(Max)}$	0.7120	0.0695	0.0000	1.0948	$W_{m75(Max)}$	1.4939	1.5216	1.5465	1.2083
W_{m50}	0.6963	0.0956	0.0000	1.1163	W_{m50}	1.3135	1.1962	1.2189	1.2315
$W_{m50(Max)}$	0.7274	0.0824	0.0000	1.1184	$W_{m50(Max)}$	1.3118	1.2863	1.2883	1.2278
Outliers = 5%		Methods			Outliers = 20%		Methods		
Criterion	EK01	EK13	EK82	EK84	Criterion	EK01	EK13	EK82	EK84
PRESS	1.5524	1.5680	1.5674	1.0965	PRESS	1.5331	1.5397	1.5462	1.2667
PRESS75	1.5361	1.5136	1.5256	1.1683	PRESS75	1.5506	1.5619	1.5650	1.2965
PRESS50	1.4639	1.4059	1.4189	1.2177	PRESS50	1.3658	1.3845	1.4704	1.3262
W_m	1.2733	1.2049	1.1807	1.1067	W_m	1.3682	1.1988	1.1768	1.1356
$W_{m(Max)}$	1.2754	1.2049	1.1807	1.1059	$W_{m(Max)}$	1.3618	1.1988	1.1768	1.1540
W_{m75}	1.4223	1.2850	1.2965	1.1224	W_{m75}	1.5051	1.5372	1.5628	1.2210
$W_{m75(Max)}$	1.4773	1.3853	1.3826	1.1292	$W_{m75(Max)}$	1.5164	1.5613	1.5680	1.2497
W_{m50}	1.1589	1.2008	1.2478	1.1582	W_{m50}	1.3567	1.3086	1.3888	1.2847
$W_{m50(Max)}$	1.3298	1.2757	1.2939	1.1666	$W_{m50(Max)}$	1.3588	1.3698	1.4449	1.2950

In bold, the method with the lowest statistic value in each percentage of outliers.

The simulation study was complemented with the analysis based on the real matrix from the multivariate characterization of wines. Table 3 presents the M^2 values and it is observed that without outliers again the EK82 method works very well with PRESS50 and, unlike the simulations, $W_{m50(Max)}$ becomes an alternative criterion. With these criteria, the rank chosen was nine and, according to the analysis by Bro and Smilde [31], with nine components, the explained variation is above 90%, but some of these components can explain very little variation.

Table 3. Procrustes statistics for the real dataset.

Procrustes Statistics for the Real Dataset																	
Outliers = 0%				Methods				Outliers = 10%				Methods					
Criterion	EK01	R	EK13	R	EK82	R	EK84	R	Criterion	EK01	R	EK13	R	EK82	R	EK84	R
PRESS	185	9	66	7	162	4	493	4	PRESS	598	1	79,846	1	124,788	1	500	5
PRESS75	185	9	66	7	53	7	493	4	PRESS75	598	1	79,846	1	124,788	1	514	6
PRESS50	280	5	66	7	24	9	492	7	PRESS50	598	1	79,846	1	124,788	1	500	5
W_m	300	4	469	1	455	1	491	3	W_m	188,330	5	250,402	4	343,767	4	500	5
$W_{m(Max)}$	280	5	231	3	222	3	491	3	$W_{m(Max)}$	188,330	5	250,402	4	343,767	4	500	5
W_{m75}	280	5	129	5	162	4	493	4	W_{m75}	611	2	79,846	1	124,788	1	499	5
$W_{m75(Max)}$	280	5	66	7	53	7	493	4	$W_{m75(Max)}$	317,830	8	79,846	1	124,788	1	499	5
W_{m50}	300	4	94	6	82	6	493	4	W_{m50}	147,451	4	79,846	1	124,788	1	499	5
$W_{m50(Max)}$	300	4	38	9	24	9	493	4	$W_{m50(Max)}$	317,830	8	79,846	1	124,788	1	499	5
Outliers = 5%				Methods				Outliers = 20%				Methods					
Criterion	EK01	R	EK13	R	EK82	R	EK84	R	Criterion	EK01	R	EK13	R	EK82	R	EK84	R
PRESS	564	1	50,690	1	63,605	1	475	4	PRESS	190,811	1	218,985	1	354,334	1	1236	4
PRESS75	564	1	50,690	1	63,605	1	475	4	PRESS75	190,811	1	218,985	1	354,334	1	1662	5
PRESS50	564	1	50,690	1	63,605	1	475	4	PRESS50	190,811	1	218,985	1	354,334	1	1662	5
W_m	130,404	7	291,105	11	247,466	6	475	4	W_m	190,811	1	849,997	7	354,334	1	915	3
$W_{m(Max)}$	130,404	7	291,105	11	247,466	6	475	4	$W_{m(Max)}$	498,098	3	849,997	7	626,980	3	915	3
W_{m75}	564	1	99,950	2	63,605	1	475	4	W_{m75}	378,913	2	218,985	1	503,121	2	1662	5
$W_{m75(Max)}$	564	1	130,657	3	63,605	1	475	4	$W_{m75(Max)}$	498,098	3	218,985	1	626,980	3	1662	5
W_{m50}	30,753	2	130,657	3	123,006	2	475	4	W_{m50}	587,876	4	218,985	1	503,121	2	1662	5
$W_{m50(Max)}$	101,359	6	252,100	8	173,123	3	475	4	$W_{m50(Max)}$	778,336	7	218,985	1	626,980	3	1662	5

In bold, the method with the lowest statistic values and the rank used. R: Rank.

According to the cross-validation study by Bro and Smilde [31], in the original data matrix, a maximum of three or four components is sufficient. This same result was obtained with the EK82 method using $W_{m(Max)}$ and W_{m75} ; however, if W_{m75} is used as a criterion, the EK13 method with five components would be preferable as it presents better predictions in the absence of contamination.

For the wine matrix contaminated at three levels, it can be verified that the EK84 methodology always presents the best results, minimizing the Procrustes statistics. In this case, with 5, 10 and 20% contaminations, the ranks chosen were four, five and three, respectively, but the selection criteria were different depending on the number of outliers. With a low level of outliers (5%), EK84 produced the same rank with all criteria (resistant or not), while with an intermediate level of outliers the only criterion that did not work very well was PRESS75 and with a high level of contamination the best performance criteria were W_m and $W_{m(Max)}$.

Finally, Table 4 presents the results of the wine matrix taking into account only the critical angles. As expected without contamination, the EK82 method was the best with most criteria presenting angles close to zero, but there was a wide multiplicity of ranks with an approximately equal result. For this reason, in this specific situation, it is recommended to take the lowest rank (three) that corresponds to the $W_{m(Max)}$ criterion.

With the change in criterion, at the maximum level of contamination (20%), EK84 (with W_m and $W_{m(Max)}$) again showed the best performance selecting rank three to obtain the lowest critical angle. Results that we did not expect occurred at the other contamination levels (5 and 10%), EK84 was surpassed by EK82 (with $W_{m50(Max)}$) and EK01 (with W_{m75}) with the smallest angles. These last results compared to those obtained from the simulations suggest that the methods can become unstable if the rank is selected taking into account these angles. Because of this situation, it is suggested in practice to repeat the procedure several times when the contamination level is between 5 and 10%. We repeated the procedure ten times for these two percentages of outliers (not shown) and found that, for the wine matrix, the lowest average of the critical angles was obtained with EK01 (using W_{m75}).

A short comment now follows regarding the distributional and computational aspects. The methods considered in this research depend on (r)SVDs; therefore, they are distribution-free and only require that the dataset under study can be written in a matrix form. On the other hand, the computational characteristics of the algorithms depend on several factors, such as implementation used, computer characteristics, matrix dimension, correlation

structure and outliers’ quantity and location. This last aspect is important because, in some cases, the location and quantity of outliers may or may not favor the calculations and the rapid convergence of the iterative procedures that involve the calculation of an (r)SVD.

Table 4. Critical angles for the real dataset.

Critical Angles for the Real Dataset																	
Outliers = 0%								Outliers = 10%									
Criterion	Methods							Criterion	Methods								
	EK01	R	EK13	R	EK82	R	EK84		R	EK01	R	EK13	R	EK82	R	EK84	R
PRESS	0.9468	9	0.2145	7	0.0000	4	1.5706	4	PRESS	1.5708	1	1.5708	1	1.5708	1	1.5419	5
PRESS75	0.9468	9	0.2145	7	0.0000	7	1.5706	4	PRESS75	1.5708	1	1.5708	1	1.5708	1	1.4672	6
PRESS50	1.1845	5	0.2145	7	0.0000	9	1.3454	7	PRESS50	1.5708	1	1.5708	1	1.5708	1	1.5419	5
W_m	0.9548	4	1.5708	1	1.5708	1	1.5160	3	W_m	1.5340	5	1.5212	4	1.4336	4	1.5419	5
$W_{m(Max)}$	1.1845	5	0.1481	3	0.0000	3	1.5160	3	$W_{m(Max)}$	1.5340	5	1.5212	4	1.4336	4	1.5419	5
W_{m75}	1.1845	5	0.2507	5	0.0000	4	1.5706	4	W_{m75}	0.8734	2	1.5708	1	1.5708	1	1.5419	5
$W_{m75(Max)}$	1.1845	5	0.2145	7	0.0000	7	1.5706	4	$W_{m75(Max)}$	1.2729	8	1.5708	1	1.5708	1	1.5419	5
W_{m50}	0.9548	4	0.2297	6	0.0000	6	1.5706	4	W_{m50}	1.4129	4	1.5708	1	1.5708	1	1.5419	5
$W_{m50(Max)}$	0.9548	4	0.2196	9	0.0000	9	1.5706	4	$W_{m50(Max)}$	1.2729	8	1.5708	1	1.5708	1	1.5419	5
Outliers = 5%								Outliers = 20%									
Criterion	Methods							Criterion	Methods								
	EK01	R	EK13	R	EK82	R	EK84		R	EK01	R	EK13	R	EK82	R	EK84	R
PRESS	1.5708	1	1.5708	1	1.5708	1	1.5674	4	PRESS	1.5708	1	1.5708	1	1.5708	1	1.4772	4
PRESS75	1.5708	1	1.5708	1	1.5708	1	1.5674	4	PRESS75	1.5708	1	1.5708	1	1.5708	1	1.4279	5
PRESS50	1.5708	1	1.5708	1	1.5708	1	1.5674	4	PRESS50	1.5708	1	1.5708	1	1.5708	1	1.4279	5
W_m	1.4977	7	1.4511	11	1.5510	6	1.5674	4	W_m	1.5708	1	1.3250	7	1.5708	1	1.2653	3
$W_{m(Max)}$	1.4977	7	1.4511	11	1.5510	6	1.5674	4	$W_{m(Max)}$	1.4304	3	1.3250	7	1.4549	3	1.2653	3
W_{m75}	1.5708	1	1.5394	2	1.5708	1	1.5674	4	W_{m75}	1.4678	2	1.5708	1	1.2746	2	1.4279	5
$W_{m75(Max)}$	1.5708	1	1.3051	3	1.5708	1	1.5674	4	$W_{m75(Max)}$	1.4304	3	1.5708	1	1.4549	3	1.4279	5
W_{m50}	1.2620	2	1.3051	3	1.3082	2	1.5674	4	W_{m50}	1.4821	4	1.5708	1	1.2746	2	1.4279	5
$W_{m50(Max)}$	1.5166	6	1.5696	8	1.2432	3	1.5674	4	$W_{m50(Max)}$	1.3989	7	1.5708	1	1.4549	3	1.4279	5

In bold, the method with the lowest statistic values and the rank used. R: Rank.

To present the reader with an idea of the times of each algorithm, we analyzed one of the previously simulated matrices and the real data matrix of red wines on a personal computer with an Intel(R) Core(TM) i7-8550U, CPU 1.80 GHz 1.99 GHz and 8 GB of RAM installed. Table 5 shows the comparison of times of the different cross-validation methods. According to the information in the table, the best average times both in the simulated lower rank matrix and in the real multivariate data were provided by EK82, which can be considered as an expected result because, unlike the proposed methods, this method does not include any extra treatment in the presence of outliers. Clearly, the robust methods need a little more time, which by construction could also be expected. However, for the dimensions of the matrices studied, the robust algorithms present good average times of less than twenty seconds, which in practice makes them quite competitive, taking into account that they adequately treat all observations, including outliers.

Table 5. Time in seconds.

Simulated Matrix with $n = 100, p = 8, \text{Rank} = 4$				
Outliers	EK01	EK13	EK82	EK84
0%	15.05	10.36	9.94	17.43
5%	10.52	11.40	8.45	12.94
10%	11.36	10.47	8.69	13.98
20%	14.89	12.75	11.58	20.47
Mean	12.96	11.25	9.67	16.21
SD	2.35	1.11	1.43	3.43
Real matrix with $n = 44$ and $p = 14$				
0%	23.93	13.84	16.37	17.55
5%	15.69	23.40	14.34	18.17
10%	16.98	21.34	19.11	19.48
20%	16.67	16.86	16.65	19.17
Mean	18.32	18.86	16.62	18.59
SD	3.78	4.32	1.95	0.89

SD: Standard deviation.

8. Discussion and Conclusions

The simulation study and the analysis on the real dataset provide some very clear conclusions. In lower rank matrices without suspected contamination, classic EK82 can be used without problems with a simple statistic, such as PRESS50 that uses only 50% of the smallest residuals. Thus, being a resistant criterion, it presented good results in scenarios that did not present outliers. According to the results of our analysis, other competitive criteria were $W_{m50(Max)}$ (which also depends on PRESS50) and $W_{m(Max)}$.

However, when there is any suspicion of contamination, EK82 should be replaced by one of the proposed robust schemes. Thus, if the objective is to choose the rank that produces good predictions with robust lower rank approximations, the EK84 method presented the best performance together with the W_m criterion in most of the situations studied.

On the other hand, if the objective is to minimize the critical angle between two principal component subspaces, EK84 (with PRESS or W_m) can be used if the contamination level is high (20%). From our analysis on real data, we can conclude that at lower contamination levels (5 or 10%) EK84 (with PRESS or W_m) should be compared with EK01 (with W_{m75}), repeating the analysis several times (say, 10 or more) to gain consistent results. This can be achieved, for example, by adding a small number of outliers at random positions or by applying the methods on submatrices obtained from the original study matrix.

The methodologies proposed in this paper proved to be efficient in solving the contamination problem without having to use any method of detecting outliers in the original data and without eliminating information, but there are still some aspects that are worth discussing. One of these refers to the simulation study. Similar to any simulation study, more features can be added to make it more complex, but in our case it was enough to show the performance of robust procedures taking the EK82 as the standard method. Our idea of fixing the matrix rank and the level of “pure noise” was to detect the strengths and weaknesses of the schemes when the contamination levels were variable, following ideas similar to those used by Maronna and Yohai [29] and Rodrigues et al. [32]. Future research from the computational point of view can be conducted to determine the robustness of the procedures under other conditions.

A common problem in data analysis is performing cross-validation from incomplete matrices. In this case, two options are suggested, the first is to use EK84. This method is based on the rSVD of Gabriel and Odoroff [23], and the fit of rank-one approximations leaves out the missing positions and does not need to perform data imputation for this calculation, but if an imputation of the missing data is required, a robust lower rank approximation can be a very useful alternative if contamination is suspected. The second suggestion to get around the problem of missing data in cross-validation without suspected contamination is to apply iterative EK82, in which case there are already algorithms available [3,33] that can be easily adapted for optimal selection of the rank of a matrix. Arciniegas-Alarcon et al. [3] showed that an iterative algorithm using parity check provides better quality imputations than the expectation-maximization method proposed by Bro et al. [9].

In the dimensions of the matrices that the cross-validation schemes were tested in this study, computational time was not a problem, but if the matrix is of a much larger dimension, there are ways to circumvent this potential problem. For example, Krzanowski and Kline [18] used several random samples from the rows of a multivariate matrix to achieve consistency in the rank selection and reduce computational time. Another alternative is to delete some random positions from the matrix and perform the imputation following one of the previously mentioned schemes. In this case, between 10 and 30% of the original data can be considered as missing data and a comparison between the real data and the imputations can be made to choose the optimal rank. This approximation is used in the R *imputation* package (<https://github.com/jeffwong/imputation>) (accessed on 1 April 2022). In the literature, there is also the option to perform leave-group-out elimination in the EK82 method [34]; this alternative computationally reduces calculation times, but the

computational gain is accompanied by a decrease in accuracy in determining the optimal rank of a matrix.

Finally, future research could be conducted to continue testing the methods proposed here. For example, more types of contamination could be considered. In the works of Maronna and Yohai [29] and González-Cebrián et al. [35], there are more alternatives to produce outliers. However, there is still no robust cross-validation procedure in the literature to choose the best additive main effects model with a robust multiplicative interaction or robust AMMI model [32], so the EK84 (or EK01) scheme could be tested for the choice of that model. The methodologies presented here can also be compared with those that detect outliers [36] and assume the outliers as missing values for later imputation by EM algorithms or with multiple imputations [37]. Finally, the EK84 and EK01 schemes are based on the EK method that could be replaced by Gabriel's method [4] or by the Eigenvector method [9], which can be computationally faster for large matrices.

Author Contributions: Conceptualization, S.A.-A., M.G.-P. and W.J.K.; methodology, S.A.-A., M.G.-P. and W.J.K.; software, S.A.-A., M.G.-P. and W.J.K.; validation, S.A.-A., M.G.-P. and W.J.K.; formal analysis, S.A.-A., M.G.-P. and W.J.K.; investigation, S.A.-A., M.G.-P. and W.J.K.; resources, S.A.-A., M.G.-P. and W.J.K.; data curation, S.A.-A., M.G.-P. and W.J.K.; writing—original draft preparation, S.A.-A., M.G.-P. and W.J.K.; writing—review and editing, S.A.-A., M.G.-P. and W.J.K. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors of this paper acknowledge the High-Performance Computing Center—ZINE of Pontificia Universidad Javeriana for assistance during the simulation study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Good, I.J. Some applications of the singular value decomposition of a matrix. *Technometrics* **1969**, *11*, 823–831. [\[CrossRef\]](#)
2. Geladi, P.; Linderholm, J. Principal component analysis. In *Comprehensive Chemometrics 2nd Edition: Chemical and Biochemical Data Analysis*; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2020; pp. 17–37.
3. Arciniegas-Alarcón, S.; García-Peña, M.; Krzanowski, W.J. Imputation using the singular-value decomposition: Variants of existing methods, proposed and assessed. *Int. J. Innov. Comput. Inf. Control* **2020**, *16*, 1681–1696.
4. Gabriel, K.R. Le biplot—outil d'exploration de données multidimensionnelles. *J. Soc. Française Stat.* **2002**, *143*, 5–55.
5. Gauch, H. A simple protocol for AMMI analysis of yield trials. *Crop Sci.* **2013**, *53*, 1860–1869. [\[CrossRef\]](#)
6. Yan, W. *Crop Variety Trials: Data Management and Analysis*; Wiley Blackwell: Hoboken, NJ, USA, 2014.
7. Rodrigues, P.C.; Lourenço, V.; Mahmoudvand, R. A robust approach to singular spectrum analysis. *Qual. Reliab. Eng. Int.* **2018**, *34*, 1437–1447. [\[CrossRef\]](#)
8. Krzanowski, W.J. Cross-validation in principal component analysis. *Biometrics* **1987**, *43*, 575–584. [\[CrossRef\]](#)
9. Bro, R.; Kjelldahl, K.; Smilde, A.K.; Kiers, H.A.L. Cross-validation of component models: A critical look at current methods. *Anal. Bioanal. Chem.* **2008**, *390*, 1241–1251. [\[CrossRef\]](#)
10. Owen, A.B.; Perry, P. Bi-cross-validation of the svd and the nonnegative matrix factorization. *Ann. Appl. Stat.* **2009**, *3*, 564–594. [\[CrossRef\]](#)
11. Josse, J.; Husson, F. Selecting the number of components in principal component analysis using cross-validation approximations. *Comput. Stat. Data Anal.* **2012**, *56*, 1869–1879. [\[CrossRef\]](#)
12. Camacho, J.; Ferrer, A. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Theoretical aspects. *J. Chemom.* **2012**, *26*, 361–373. [\[CrossRef\]](#)
13. Saccenti, E.; Camacho, J. On the use of the observation-wise k-fold operation in PCA cross-validation. *J. Chemom.* **2015**, *29*, 467–478. [\[CrossRef\]](#)
14. Eastment, H.T.; Krzanowski, W.J. Cross-validated choice of the number of components from a principal component analysis. *Technometrics* **1982**, *24*, 73–77. [\[CrossRef\]](#)
15. Dias, C.T.S.; Krzanowski, W.J. Model selection and cross-validation in additive main effect and multiplicative (AMMI) models. *Crop Sci.* **2003**, *43*, 865–873. [\[CrossRef\]](#)

16. Liu, Y.J.; Tran, T.; Postma, G.; Buydens, L.M.C.; Jansen, J. Estimating the number of components and detecting outliers using Angle Distribution of Loading Subspaces (ADLS) in PCA analysis. *Anal. Chim. Acta* **2018**, *1020*, 17–29. [[CrossRef](#)]
17. Krzanowski, W.J. Cross-validatory choice in principal component analysis: Some sampling results. *J. Stat. Comput. Simul.* **1983**, *18*, 299–314. [[CrossRef](#)]
18. Krzanowski, W.J.; Kline, P. Cross-validation for choosing the number of important components in principal component analysis. *Multivar. Behav. Res.* **1995**, *30*, 149–165. [[CrossRef](#)]
19. Forkman, J.; Piepho, H.P. Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. *Biometrics* **2014**, *70*, 639–647. [[CrossRef](#)]
20. García-Peña, M.; Arciniegas-Alarcón, S.; Krzanowski, W.J.; Duarte, D. Missing value imputation using the robust singular-value decomposition: Proposals and numerical evaluation. *Crop Sci.* **2021**, *61*, 3288–3300. [[CrossRef](#)]
21. Arciniegas-Alarcón, S.; García-Peña, M.; Rengifo, C.; Krzanowski, W.J. Techniques for robust imputation in incomplete two-way tables. *Appl. Syst. Innov.* **2021**, *4*, 62. [[CrossRef](#)]
22. Hubert, M.; Engelen, S. Fast cross-validation of high-breakdown resampling methods for PCA. *Comput. Stat. Data Anal.* **2007**, *51*, 5013–5024. [[CrossRef](#)]
23. Gabriel, K.R.; Odoroff, C.L. Resistant lower rank approximation of matrices. In *Data Analysis and Statistics III*; Diday, E., Jambu, M., Lebart, L., Thomassone, Eds.; North-Holland: Amsterdam, The Netherlands, 1984; pp. 23–30.
24. Hawkins, D.M.; Liu, L.; Young, S.S. *Robust Singular Value Decomposition*; Technical Report 122; National Institute of Statistical Sciences: Washington, DC, USA, 2001.
25. Zhang, L.; Shen, H.; Huang, J.Z. Robust regularized singular-value decomposition with application to mortality data. *Tha Ann. Appl. Stat.* **2013**, *7*, 1540–1561. [[CrossRef](#)]
26. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020; ISBN 3-900051-070. Available online: <https://www.r-project.org/> (accessed on 1 April 2022).
27. Krzanowski, W.J. *Principles of Multivariate Analysis: A User's Perspective Oxford*; University Press: Oxford, UK, 2000.
28. Krzanowski, W.J. Between-group comparison of principal components—some sampling results. *J. Stat. Comput. Simul.* **1982**, *15*, 141–154. [[CrossRef](#)]
29. Maronna, R.A.; Yohai, V.J. Robust low-rank approximation of data matrices with elementwise contamination. *Technometrics* **2008**, *50*, 295–304. [[CrossRef](#)]
30. Skov, T.; Ballabio, D.; Bro, R. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. *Anal. Chim. Acta* **2008**, *615*, 18–29. [[CrossRef](#)]
31. Bro, R.; Smilde, A.K. Principal component analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [[CrossRef](#)]
32. Rodrigues, P.C.; Monteiro, A.; Lourenço, V.M. A robust AMMI model for the analysis of genotype \times environment data. *Bioinformatics* **2016**, *32*, 58–66. [[CrossRef](#)]
33. Krzanowski, W.J. Missing value imputation in multivariate data using the singular value decomposition of a matrix. *Biom. Lett.* **1988**, *25*, 31–39.
34. Eshghi, P. Dimensionality choice in principal component analysis via cross-validatory methods. *Chemom. Intell. Lab. Syst.* **2014**, *130*, 6–13. [[CrossRef](#)]
35. González-Cebrián, A.; Arteaga, F.; Folch-Fortuny, A.; Ferrer, A. How to simulate outliers with desired properties. *Chemom. Intell. Lab. Syst.* **2021**, *212*, 104301. [[CrossRef](#)]
36. Grentzelos, C.; Caroni, C.; Barranco-Chamorro, I. A comparative study of methods to handle outliers in multivariate data analysis. *Comput. Math. Methods* **2020**, *3*, e1129. [[CrossRef](#)]
37. Alkan, B.B.; Atakan, C.; Alkan, N. A comparison of different procedures for principal component analysis in the presence of outliers. *J. Appl. Stat.* **2015**, *42*, 1716–1722. [[CrossRef](#)]