Vitale, R.; Westerhuis, JA.; Naes, T.; Smilde, AK.; De Noord, OE.; Ferrer, A. (2017). Selecting the number of factors in principal component analysis by permutation testing Numerical and practical aspects. Journal of Chemometrics. 31(12):1-15. doi:10.1002/cem.2937

Selecting the number of factors in Principal Component Analysis by permutation testing - Numerical and practical aspects

Raffaele Vitale[a,b,c,*], Johan A. Westerhuis[d], Tormod Næs[e], Age K. Smilde[d], Onno E. de Noord[f], Alberto Ferrer[a]

[a]*Grupo de Ingeniería Estadística Multivariante, Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain*
[b]*Molecular Imaging and Photonics Unit, Department of Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001, Leuven, Belgium*
[c]*Laboratoire de Spectrochimie Infrarouge et Raman - UMR 8516, Université de Lille - Sciences et Technologies, Bâtiment C5, 59655, Villeneuve d'Ascq, France*
[d]*Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands*
[e]*Nofima AS, PO Box 210, 1431 Ås, Norway*
[f]*Shell Global Solutions International B.V., Shell Technology Centre Amsterdam, PO Box 38000, 1030 BN, Amsterdam, The Netherlands*

Abstract

Selecting the correct number of factors in Principal Component Analysis (PCA) is a critical step to achieve a reasonable data modelling, where the optimal strategy strictly depends on the objective PCA is applied for. In the last decades, much work has been devoted to methods like Kaiser's eigenvalue greater than 1 rule, Velicer's minimum average partial rule, Cattell's scree test, Bartlett's chi-square test, Horn's parallel analysis, and cross-validation. However, limited attention has been paid to the possibility of assessing the significance of the calculated components via permutation testing. That may represent a feasible approach in case the focus of the study is discriminating relevant from non-systematic sources of variation and/or the aforementioned methodologies cannot be resorted to (e.g. when the analysed matrices do not fulfill specific properties or statistical assumptions).

The main aim of this article is to provide practical insights for an improved understanding of permutation testing, highlighting its pros and cons, mathematically formalising the numerical procedure to be abided by when applying it for PCA factor selection by the description of a novel algorithm developed to this end, and proposing *ad hoc* solutions for optimising computational time and efficiency.

Keywords: permutation testing, Principal Component Analysis (PCA), deflation, projection, eigenvalues

*Corresponding author:
Telephone number: +33769476654
Email address: rvitale86@gmail.com (Raffaele Vitale)

1. Introduction

Principal Component Analysis (PCA) [1, 2] is probably the most commonly used multivariate statistical tool to compress, describe and interpret large sets of data. Its basic principle can be summarised as follows: let $\mathbf{X}$ be a $N \times J$ matrix with J denoting the number of variables (e.g. J *sensor* responses monitored during an industrial process or J metabolites quantified in biological samples) registered for each of N measurements performed, for instance, at N time instants or for N different individuals. In the modern instrumental context, where J might be very large, the useful and meaningful information in $\mathbf{X}$ is usually intercorrelated among various of these variables over the whole set of recordings. Under this assumption, for a chosen degree of acceptable accuracy, it is possible to reduce the J-dimensional space of the original descriptors to an A-dimensional subspace, onto which all the N objects under study can be projected and represented as new points. Mathematically speaking, PCA is based on the bilinear structure model in Eq. 1:

$$\mathbf{X} = \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{1}$ ($N \times 1$) is a vector of ones, $\mathbf{m}$ ($J \times 1$) contains the mean values of the J variables in $\mathbf{X}$, $\mathbf{P}$ ($J \times A$) is an array of so-called *loadings*, which determine the A basis vectors (*components* or *factors*) of the PCA subspace, $\mathbf{T}$ ($N \times A$) defines the projection coordinates or *scores* of all the N rows of $\mathbf{X}$ on this lower-dimensional space and $\mathbf{E}$ ($N \times J$) stands for the matrix of unmodelled residuals, i.e. the portion of $\mathbf{X}$ not *explained* at the chosen rank, A.

When deriving a PCA decomposition a very critical point is how to properly set A. First of all, it is important to notice that, as stated in [3–5], this assessment connotes an ill-posed problem if formulated without taking into account for which objective PCA is resorted to. In [5] Camacho and Ferrer differentiated three different application scenarios: i) when the interest is on the *observable* or *original* variables; ii) when the interest is on the *principal components*; iii) when the interest is on the distributions of the principal components and residuals. i) refers to situations in which the dimensionality of the PCA subspace has to be determined so that the model-based reconstruction of the original variables is the most accurate possible (e.g. for compression or missing value imputation). ii) mainly relates to data exploration, which normally implies the extraction of all the principal components that can be safely interpreted because they are sufficiently different from

noise. iii) basically concerns statistical process monitoring, where the distributions of the principal components and residuals, calculated from a set of data collected under Normal Operating Conditions (NOC), are utilised to assess whether such NOC are maintained over time or a fault is occurring. Here, the main focus will be on ii). However, in all these circumstances, if no *a priori* knowledge about the investigated systems is available, A has to be empirically retrieved [6].

1.1. Strategies for principal component selection

During the last decades many approaches for principal component selection have been developed, which can be classified in three distinct categories: *ad hoc* rules, statistical tests and computational criteria [7]. *Ad hoc* rules (like Kaiser's eigenvalue greater than 1 rule [8], Velicer's minimum average partial rule [9] and Cattell's scree test [10]) and statistical tests (like Bartlett's chi-square test [11] and Tracy-Widom statistic-based test [12]) generally show a particular drawback: they often constitute case-specific strategies, not easily generalisable for handling data structures of various nature, and sometimes are based on distributional assumptions, which are rarely met in modern analytical contexts. On the other hand, computational criteria are completely data-driven and distribution-free. Therefore, they can be regarded as feasible options when *ad hoc* rules and statistical tests cannot be applied (for instance when the considered datasets do not fulfill particular mathematical properties), even if they might sometimes lead to an excessive time and memory consumption.

Computational criteria encompass both cross-validation and permutation test-based techniques (like Horn's parallel analysis [13] and a data dimensionality detection methodology proposed by Dray in [14]). Although cross-validation is probably the most widespread principal component selection approach, its application is not recommended when the objective of the study is discriminating relevant from noisy factors [5]. In fact, as it permits to determine the dimensionality of the PCA subspace by minimising the prediction error between the initial data and their PCA estimates, it is evidently better-suited for the applications covered by the aforementioned scenario i). On the contrary, when the systematic and non-systematic sources of variation in the data have to be differentiated and/or stable loadings and residuals distributions are desired, the focus moves from the original variables to the principal components. In similar contingencies the employment

of cross-validation may not be adequate: procedures aimed at determining the statistical significance of the principal components or at minimising the Overall Type II ($OTII$) risk when process monitoring is concerned may be required.

Permutation test-based techniques rely on the comparison of some attributes of the analysed data matrices with those of arrays characterised by uncorrelated variables. These attributes are conventionally: i) the singular values or the eigenvalues (the square of the singular values) or ii) functions of the eigenvalues (e.g. the difference or the ratio between consecutive eigenvalues). Based on this, it is clear that these methods directly concentrate on the identification of structured interpretable components and might represent appropriate alternatives to cross-validation when PCA is exploited for such an exploratory purpose [14–20].

For all these reasons, in this article, an extensive guideline on how to select the number of PCA factors by permutation tests is provided. Concretely, a novel algorithm designed to this end and originated from the preliminary proposal outlined in [21] is presented. It will be compared to other permutation test-based strategies (i.e. Horn's parallel analysis and Dray's approach, described in Appendices A and B, respectively), which will permit to point out some of their limitations that, as far as the authors are aware, were only partly spotted before in the scientific literature. Solutions for overcoming these limitations will be additionally reported. The practical aspects of this algorithm will be examined for a better understanding of its pros over its primary version and possible adjustments for optimising the efficiency of its computational procedure (in terms of time and memory consumption) will also be discussed.

2. Methods

A new algorithm for PCA component selection by permutation testing is here introduced. Let $\mathbf{X}$ be now a centred data matrix of N rows and J columns with rank $Q = \min\{N-1, J\}$. This novel computational procedure rests on the estimation of the statistical significance of the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (λ) against the *null*-hypothesis (H_0) that the mechanism generating the columns of $\mathbf{X}$ is nothing but random noise[i]. It comprises the following 10 steps grouped in three consecutive phases

[i] As observed before and according to the taxonomy suggested in [7], this algorithm can be regarded as a methodology that takes as input the eigenvalues of the data covariance matrix to discern structured information from noise.

(see also Figure 1):

- Phase I - Singular Value Decomposition of $\mathbf{X}$:

  1. Perform Singular Value Decomposition (SVD) on $\mathbf{X}$:

  $$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} = \mathbf{TP}^{\mathrm{T}} \tag{2}$$

  where $\mathbf{U}$ ($N \times N$) and $\mathbf{V}$ ($J \times J$) contain the left and right singular vectors of $\mathbf{X}$, respectively, and $\mathbf{S}$ ($N \times J$) is a rectangular diagonal array whose non-zero diagonal elements are its singular values ($\sqrt{\lambda_1}$, $\sqrt{\lambda_2}, \ldots,$ $\sqrt{\lambda_Q}$);

  2. Compute for each a-th calculated component the ratio:

  $$F_a = \frac{\lambda_a}{\sum_{q=a}^{Q} \lambda_q} \tag{3}$$

  where λ_a corresponds to the a-th eigenvalue obtained after the decomposition of $\mathbf{X}$. F_a is used for testing the statistical significance of the single factors. It equals the ratio between the amount of variation explained by the a-th component and the total amount of variation captured by the last $Q - (a - 1)$ components.

- Phase II - Test for the first component:

  3. For $a = 1$, randomly and independently permute the order of the entries within every column of $\mathbf{X}$ constructing a new matrix $\mathbf{X}_{perm}$, featuring uncorrelated variables;
  4. Apply SVD to $\mathbf{X}_{perm}$ and calculate the ratio:

  $$F_{1,perm} = \frac{\lambda_{1,perm}}{\sum_{q=1}^{Q} \lambda_{q,perm}} \tag{4}$$

  where $\lambda_{1,perm}$ denotes the first eigenvalue obtained after the decomposition of $\mathbf{X}_{perm}$. Note that the sum of squares of $\mathbf{X}$ and $\mathbf{X}_{perm}$ is exactly the same, despite the permutations;
  5. Iterate step 3 and 4 to generate a *null*-distribution for $F_{1,perm}$[ii]. If F_1 is found to be higher than its $(1 - \alpha) \times 100^{\text{th}}$ percentile (α equals the nominal Overall Type I - *OTI*

[ii] The total number of iterations is a user-defined parameter and should be selected so as to obtain a precise estimation of such a *null*-distribution.

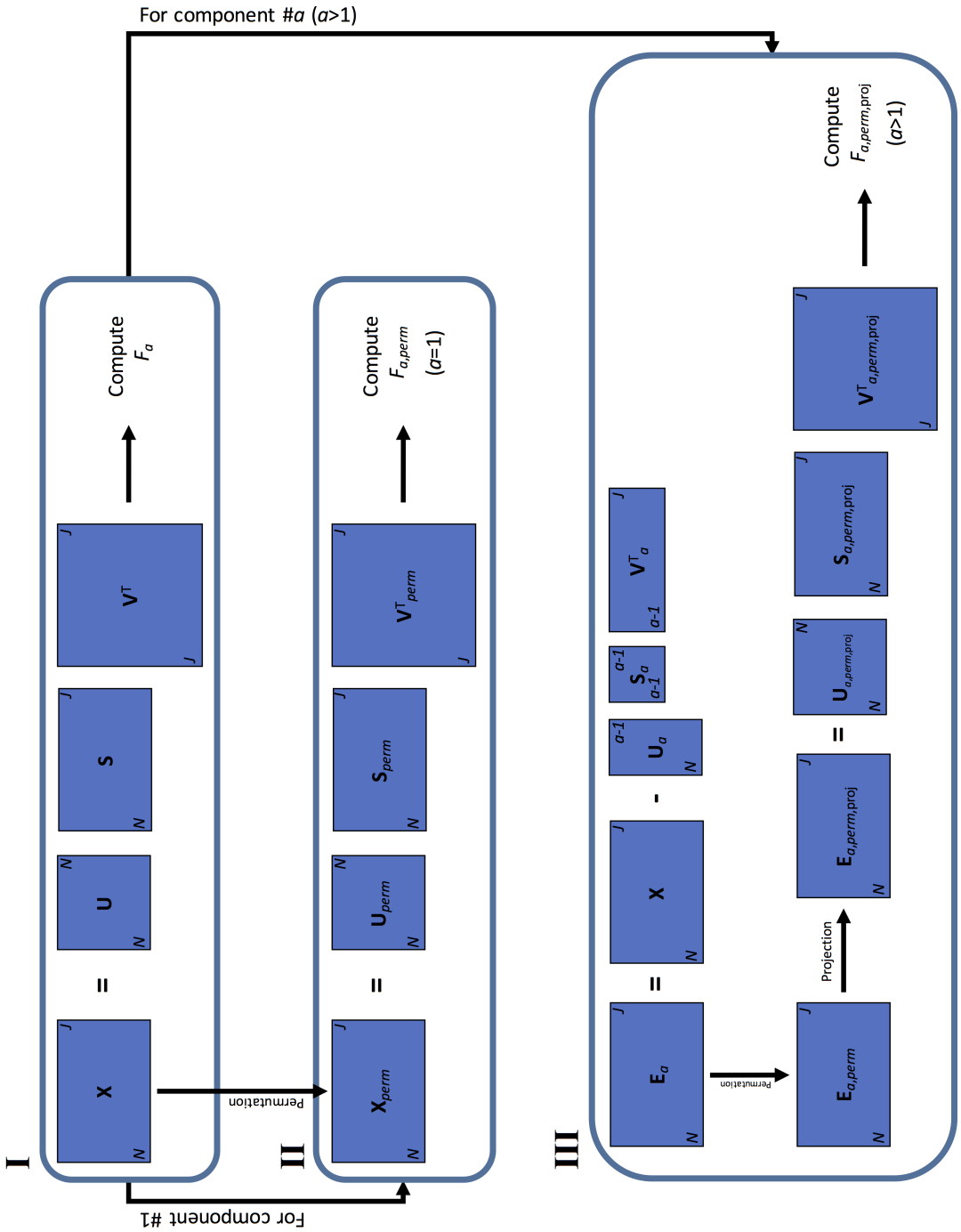

Figure 1 - Schematic representation of the permutation-based algorithmic procedure proposed in this article. A PCA component is considered statistically significant if its respective F_a value is higher than the $(1-\alpha)\times 100^{\text{th}}$ percentile of the associated *null*-distribution, being α the nominal Overall Type I (OTI) risk value imposed to the test, i.e. its false positive rate

- risk value imposed to the test, i.e. its false positive rate), the first component is considered statistically significant.

- Phase III - Test for the a-th component ($a > 1$):

  6. For $a > 1$, calculate the residual matrix:

$$\mathbf{E}_a = \mathbf{X} - \sum_{q=1}^{a-1} \mathbf{u}_q \sqrt{\lambda_q} \mathbf{v}_q^{\mathrm{T}} = \mathbf{X} - \sum_{q=1}^{a-1} \mathbf{t}_q \mathbf{p}_q^{\mathrm{T}} \tag{5}$$

  where $\mathbf{u}_q$, $\mathbf{v}_q$, $\mathbf{t}_q$ and $\mathbf{p}_q$ are the q-th column vectors of $\mathbf{U}$, $\mathbf{V}$, $\mathbf{T}$ and $\mathbf{P}$ (see Eq. 2), respectively[iii]. Note that after each deflation round $\mathbf{E}_a$ has rank $Q - (a - 1)$;

  7. Randomly and independently permute the order of the entries within each column of $\mathbf{E}_a$ constructing a new matrix $\mathbf{E}_{a,perm}$. Unlike $\mathbf{E}_a$, $\mathbf{E}_{a,perm}$ has rank Q (apart from chance deviations), but their total sums of squares are the same;

  8. Calculate the projection of $\mathbf{E}_{a,perm}$ on a subspace of dimensionality $Q-(a-1)$, $\mathbf{E}_{a,perm,\mathrm{proj}}$. The way to carry out this projection represents the main novelty of this study and will be discussed in the next section;

  9. Perform SVD on $\mathbf{E}_{a,perm,\mathrm{proj}}$ and retain the ratio:

$$F_{a,perm,\mathrm{proj}} = \frac{\lambda_{1,perm,\mathrm{proj}}}{\sum_{q=1}^{Q-(a-1)} \lambda_{q,perm,\mathrm{proj}}} \tag{6}$$

  where $\lambda_{1,perm,\mathrm{proj}}$ is the first eigenvalue obtained after the decomposition of $\mathbf{E}_{a,perm,\mathrm{proj}}$;

  10. Iterate step 7, 8 and 9 to generate a *null*-distribution for $F_{a,perm,\mathrm{proj}}$[ii]. If F_a is found to be higher than its $(1-\alpha)\times 100^{\mathrm{th}}$ percentile, the a-th component is considered statistically significant.

Computations are stopped as soon as the first non-significant component is detected.

3. Numerical and practical aspects of the algorithm

Four particular aspects, which constitute the core of the algorithm, are now elucidated from both a numerical and practical perspective: i) Why are the data permuted column-wise? ii) Why

[iii] According to this notation a hypothetical $\mathbf{E}_1$ would correspond to $\mathbf{X}$.

does $\mathbf{X}$ need to be sequentially deflated? iii) Is the projection of $\mathbf{E}_{a,perm}$ necessary? iv) What is the rationale behind the relative index F_a?

For the sake of a comprehensive assessment of the specific implications of how the calculations are performed, all the tests reported in this section were run for all the extractable components, thus not resorting to the aforementioned stopping criterion.

3.1. Permutations

In both Phases II and III of the computational procedure, $\mathbf{X}$ and $\mathbf{E}_a$ are permuted so that the order of the entries within each one of their columns is independently randomised. This breaks their underlying covariance structure, whereas the mean value and the standard deviation of the measured variables are maintained. Variances are then preserved, but the intrinsic mutual relationships among descriptors are cluttered.

3.2. Deflation

When SVD is applied on a dataset, whose element order was permuted such that all correlation among the measured variables is lost, its total variation will be more or less uniformly distributed across all its extractable factors[iv]. This can lead to overlooking the actual statistical significance of some eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ if they do not account for a substantially high amount of variation, which is exactly what happens with Horn's parallel analysis [22], as will be shown in Section 4.1. Testing this significance with consecutive deflation steps allows this limitation to be overcome. This is illustrated in the following example.

Say for instance that $\mathbf{X}$ has rank 10 and contains 80% of systematic variation in 3 components (60%, 15% and 5%, respectively) and 20% of noise with a total sum of squares of 100. Therefore, $\lambda_1 = 60$, $\lambda_2 = 15$ and $\lambda_3 = 5$. The remaining 20% of the variation of $\mathbf{X}$ is roughly uniformly distributed over the other 7 eigenvalues ($\lambda_q \simeq 2.8 \ \forall q = 4, \ldots, 10$). Permuting $\mathbf{X}$ permits to generate a new matrix $\mathbf{X}_{perm}$ with the same sum of squares. However, as no correlation among the

[iv] Theoretically, the eigenvalues associated to the single factors should be identical provided that the sum of squares of all the variables is the same. However, chance correlations generate their typical smooth descending trend observable in e.g. Figure 2.

measured variables is left, the whole amount of variation is distributed across the whole set of 10 eigenvalues ($\lambda_{q,perm} \simeq 10 \ \forall q = 1, \ldots, 10$). As shown in Figure 2 (left), comparing λ_3 with its *null*-distribution will not lead to detect it as statistically significant (λ_3 is clearly smaller than the 99th percentile of the corresponding *null*-distribution[v]). On the other hand, if the first two components

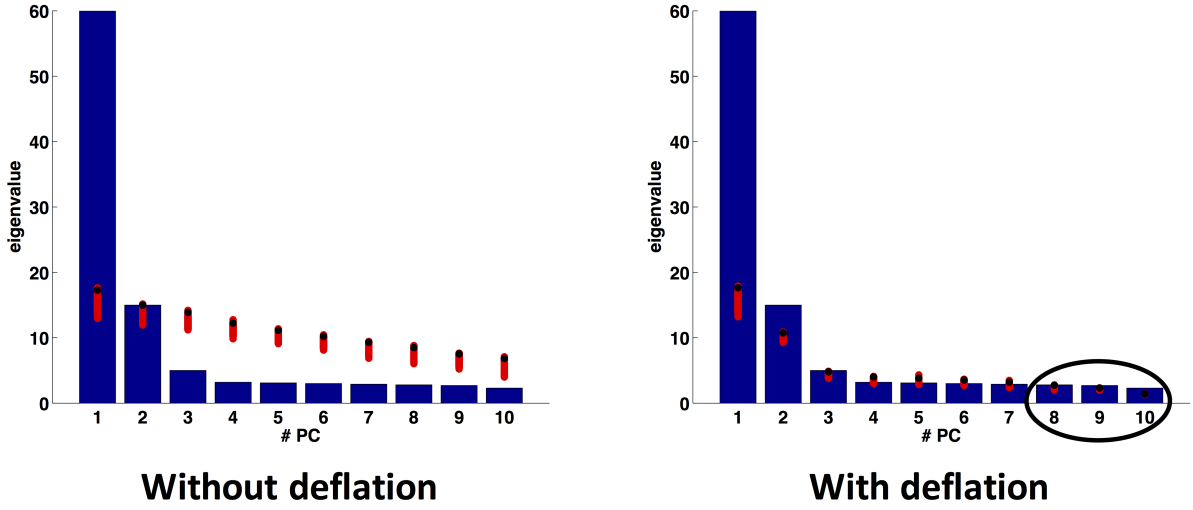

Figure 2 - Eigenvalues (blue bars) of a simulated centred data matrix (100×10) containing 80% of systematic variation in 3 components (60%, 15% and 5%, respectively) and 20% of noise with a total sum of squares of 100, and their empirically estimated *null*-distributions (red dots) obtained by either not deflating (left) or deflating (right) the original array during the execution of the permutation test. Each red dot corresponds to the eigenvalue obtained for the respective component after one of the 300 performed permutation rounds. The black markers denote the 99th percentiles of the aforementioned *null*-distributions

are used to deflate $\mathbf{X}$, the resulting $\mathbf{E}_3$ matrix will be characterised by a sum of squares of 25 and a rank of 8. The first eigenvalue of $\mathbf{E}_{3,perm}$ will be then around $\frac{25}{10} \simeq 2.5$ (shuffling $\mathbf{E}_3$ makes $\mathbf{E}_{3,perm}$ have again rank 10). In this case (see Figure 2 - right), the third factor of $\mathbf{X}$ will be correctly identified as statistically significant (λ_3 is now larger than the 99th percentile of the corresponding *null*-distribution). But, what happens with λ_8, λ_9 and λ_{10} (see black circle in the right subplot of Figure 2)? Are they statistically significant even if λ_4 to λ_7 are not? The root cause of this strange inconsistency will be clarified in the next section.

[v]From now on, for all the reported applications α is set equal to 0.01.

3.3. Projection

As a consequence of the permutations, $\mathbf{E}_a$ and $\mathbf{E}_{a,perm}$ have always different rank ($Q-(a-1)$ and Q, respectively, after each step of deflation for $a > 1$), but equal sum of squares. However, this sum of squares is distributed over $Q-(a-1)$ non-zero eigenvalues in the former array and over Q non-zero eigenvalues in the latter one. Hence, the expected values of λ_a will be on average higher than those of $\lambda_{a,perm}$ as a increases. The projection of $\mathbf{E}_{a,perm}$ on a hyperplane of dimensionality $Q-(a-1)$ can correct for this effect. In the approach proposed in [21] this projection is executed both row-wise and column-wise as:

$$\mathbf{E}_{a,perm,\text{proj}} = (\mathbf{I}_N - \sum_{q=1}^{a-1} \mathbf{u}_q \mathbf{u}_q^{\mathrm{T}}) \mathbf{E}_{a,perm} (\mathbf{I}_J - \sum_{q=1}^{a-1} \mathbf{v}_q \mathbf{v}_q^{\mathrm{T}}) \tag{7}$$

where $\mathbf{I}_N$ is an identity matrix of dimensions $N \times N$ and $\mathbf{I}_J$ is an identity matrix of dimensions $J \times J$. Equation 7 (referred to as P1 from now on) guarantees that both the row and column space of $\mathbf{E}_{a,perm,\text{proj}}$ are identical to those of the original residuals, $\mathbf{E}_a$. Therefore, in [21] it was regarded as the most natural and intuitive way to project $\mathbf{E}_{a,perm}$. P1 proved to be a feasible option when small sets of data were dealt with, but if N and/or J are/is very large, the calculation of the inner-product arrays $\sum_{q=1}^{a-1} \mathbf{u}_q \mathbf{u}_q^{\mathrm{T}}$ ($N \times N$) and/or $\sum_{q=1}^{a-1} \mathbf{v}_q \mathbf{v}_q^{\mathrm{T}}$ ($J \times J$) can be rather expensive in computational terms. An alternative strategy (referred to as P2[vi]) could be projecting $\mathbf{E}_{a,perm}$ onto the hyperplane orthogonal to the first $a-1$ components of $\mathbf{X}$ using their either left or right singular vectors:

$$\mathbf{E}_{a,perm,\text{proj}} = (\mathbf{I}_N - \sum_{q=1}^{a-1} \mathbf{u}_q \mathbf{u}_q^{\mathrm{T}}) \mathbf{E}_{a,perm} \tag{8}$$

P2 would be significantly faster from a computational point of view, but would allow only the row or the column spaces of $\mathbf{E}_{a,perm,\text{proj}}$ and $\mathbf{E}_a$ to be the same.

P1 and P2 lead to different $\mathbf{E}_{a,perm,\text{proj}}$ matrices (a single row-wise or column-wise projection as in P2 reduces the sum of squares of $\mathbf{E}_a$ less than a double projection as in P1), which nevertheless share the same rank, $Q-(a-1)$. This permits to compare them in this particular application

[vi] Formulas 8 and 9 relate to the case in which $N < J$ and define a row-wise projection. If $N > J$, a column-wise projection can be performed resorting to the column vectors of $\mathbf{V}/\mathbf{V}_{\mathbf{E}_{a,perm}}$ instead of those of $\mathbf{U}/\mathbf{U}_{\mathbf{E}_{a,perm}}$, thus preventing excessive time and memory consumption.

scenario and check whether they enable the derivation of reasonable *null*-distributions for the concerned permutation test. Specifically, for each projection approach:

1. 300 matrices of size 51×200 containing random values drawn from the standard normal distribution were simulated;
2. after preprocessing, the algorithm reported in Section 2 was run on each one of these matrices to determine the statistical significance of all their 50 extractable components. 300 permutation rounds per matrix were performed;
3. once every test was completed, a single p-value per component was derived as the ratio between the number of $F_{1,perm}$ (if $a = 1$) or $F_{a,perm,\text{proj}}$ (if $a > 1$) found to be higher than the corresponding F_a and the total number of permutations;
4. the p-values associated to each component were finally averaged over the 300 simulated matrices.

Figure 3 displays the outcomes of this assessment (left and central subplots). Here, both P1 and P2 exhibited very similar performances, generally leading to p-values close to 1, which, as also Figure 4 (left and central subplots) confirms, are rendered by the fact that the $F_{a,perm,\text{proj}}$ values (for $a > 1$) of the various empirical *null*-distributions are systematically larger than expected when quantified by P1 and P2. Such an overestimation is probably a consequence of the fact that $\mathbf{E}_{a,perm}$ is projected onto the subspace spanned by $\mathbf{E}_a$ that still describes part of the systematic structure of the handled data (this structure is spurious but anyway present also in random matrices). As a thought experiment, say one draws a shape (a principal component) in the sand (the original data space) and removes (deflates) all the grains inside its borders. The remainder of the sand (the residual space) will be a *negative* of the shape and will thus keep memory of its contour, inevitably influencing the disposition of hypothetical new grains randomly distributed over it. For the same reason, the aforementioned projection might generate chance covariance in $\mathbf{E}_{a,perm,\text{proj}}$ and then a substantial increase in the corresponding $F_{a,perm,\text{proj}}$ ratios. This should be even more evident when structured data are dealt with. In order to overcome this issue, P3 is proposed. By P3, $\mathbf{E}_{a,perm}$ is

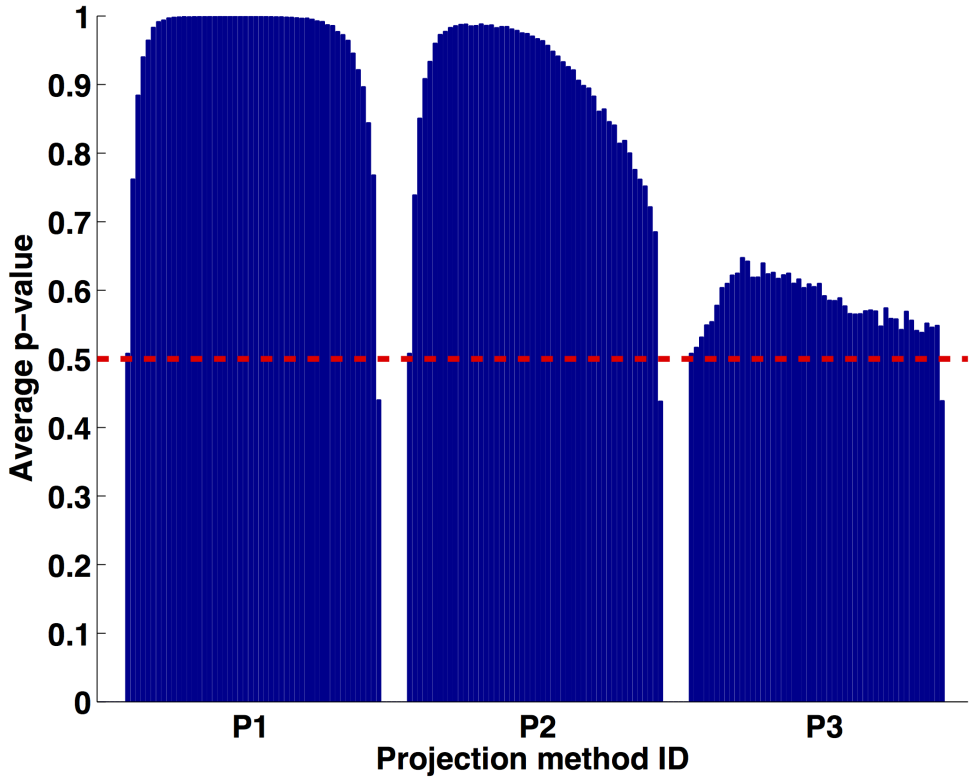

Figure 3 - Mean p-values obtained for the 50 components extractable, after preprocessing, from 300 simulated random value matrices of dimensions 51×200 by exploiting the 3 different projection strategies under comparison. Each bar quantifies the ratio, averaged over the 300 simulations, between the number of $F_{1,perm}$ (if $a = 1$) or $F_{a,perm,\text{proj}}$ (if $a > 1$) found to be higher than the corresponding F_a and the total number of permutation rounds (300 per matrix). The horizontal red dotted line is drawn at $p-\text{value} = 0.5$

projected onto the hyperplane orthogonal to its first $a-1$ components[vi]:

$$\mathbf{E}_{a,perm,\text{proj}} = (\mathbf{I}_N - \sum_{q=1}^{a-1} \mathbf{u}_{q,\mathbf{E}_{a,perm}} \mathbf{u}_{q,\mathbf{E}_{a,perm}}^{\mathrm{T}}) \mathbf{E}_{a,perm} \tag{9}$$

where $\mathbf{E}_{a,perm} = \mathbf{U}_{\mathbf{E}_{a,perm}} \mathbf{S}_{\mathbf{E}_{a,perm}} \mathbf{V}_{\mathbf{E}_{a,perm}}^{\mathrm{T}}$, and being $\mathbf{u}_{q,\mathbf{E}_{a,perm}}$ the q-th column vector of $\mathbf{U}_{\mathbf{E}_{a,perm}}$. At each permutation round, $\mathbf{E}_{a,perm}$ is subjected to SVD and a new subspace is estimated from random residuals for the projection to limit the effect of the chance covariance induced by P1 and P2. As no notable differences were observed between the performance of P1 and P2, P3 was implemented

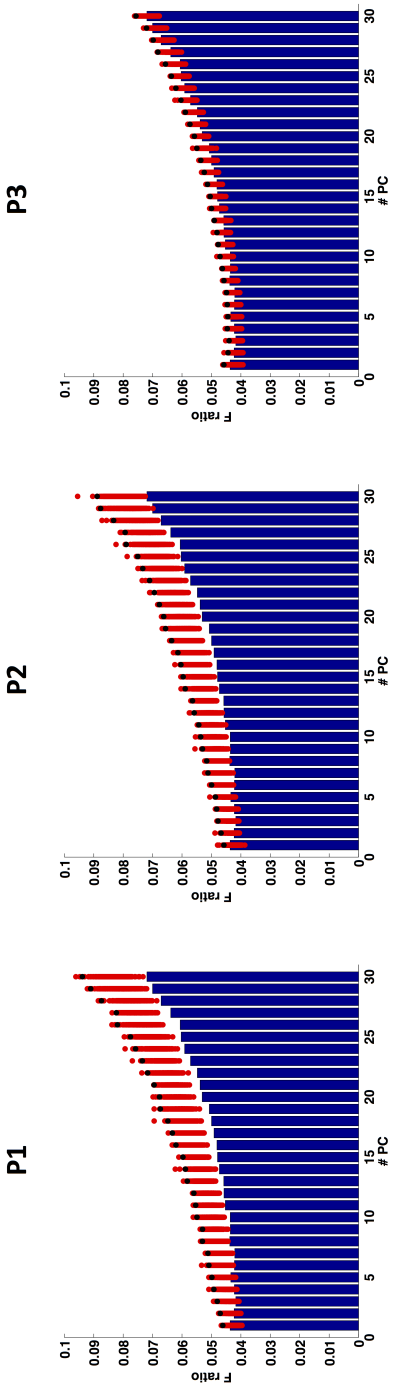

Figure 4 - F_a ratios (blue bars) and related empirical *null*-distributions (red dots) associated to the first 30 components of a simulated 51 × 200 random value matrix and resulting from P1, P2 and P3. Each red dot corresponds to the $F_{1,perm}$ (if $a = 1$) or the $F_{a,perm,\text{proj}}$ (if $a > 1$) estimate obtained for the respective factor after one of the 300 performed permutation rounds. The black markers denote the 99$^{\text{th}}$ percentiles of the aforementioned *null*-distributions

so that the projection is carried out either row-wise or column-wise[vii]. It is also worth noting that P3 yields the largest loss of data variation. Figures 3 and 4 also show the results of the previous study obtained when P3 was used (right subplots): as expected it returned better guesses of the *null*-distributions than those resulting from P1 and P2. In the light of all that, only P1 and P3 will be employed for comparison in the further case-studies illustrated in this article.

The projection of $\mathbf{E}_{a,perm}$ represents the main advantage of this novel algorithm over Dray's approach [14], which progressively deflates $\mathbf{X}$ as new factors are extracted, but does not take into account the change of the rank of the matrices of the permuted residuals. This commonly generates very inconsistent outcomes: as a increases, the *null*-distributions associated to the single components are gradually more underestimated and the method is continuously more prone to detect noisy factors as statistically significant (see Section 4.1 for further details). This issue was originally solved by making the sequential computational procedure stop just after the identification of the first non-significant factor. However, for the properties of the statistic used for the testing procedure, Dray's method generally recognises less significant components than expected (see Appendix B).

3.4. The rationale behind F_a

The projection of $\mathbf{E}_{a,perm}$ yields a decrease in its sum of squares. Then, the eigenvalues of $\mathbf{E}_a$ and $\mathbf{E}_{a,perm,\text{proj}}$ are not directly commensurable. As a solution to this issue, the statistical significance of each specific factor is tested through a relative measure, i.e. the ratio between the respective eigenvalue and its sum with all the smaller ones. In fact, since such a projection modifies at the same time and more or less uniformly the whole set of eigenvalues of $\mathbf{E}_{a,perm}$, this ratio is negligibly affected by the aforementioned decrease in its total sum of squares.

[vii] Nevertheless, P3 allows both the row- and the column-spaces of $\mathbf{E}_{a,perm}$ and $\mathbf{E}_{a,perm,\text{proj}}$ to be the same no matter if the projection is performed either row-wise or column-wise.

4. Performance of the algorithm

4.1. Synthetic datasets

Four synthetic matrices were exploited to verify whether the number of their underlying components (known *a priori*) could be correctly retrieved by the developed methodology. The data generation design was first detailed in [5]: 4, 8, 12 and 15 principal components, simulated independently at random and following a normal distribution with zero mean and unit variance, were respectively exploited to calculate a certain amount of observed variables according to the equations listed in Table 1. All the final arrays, featuring 100 objects, are examples of different correlation structures (from industrial process-like to spectral-like) and were contaminated with measurement noise of diverse magnitude (from 5 to 100% of the global variation of the noise-free data) to get an idea about the robustness of the implemented approach. Table 2 shows how many factors were retained at each noise level for the 4 datasets (for Horn's parallel analysis, Dray's method and both P1 and P3). The displayed values represent the median and the range of the number of selected components estimated over 300 simulation replicates. Clearly, P3 generally enabled an accurate identification of the number of significant components. Nevertheless, from a certain noise level on, depending on the nature of the considered covariance structure, the procedure more often tended to be less sensitive, but this is reasonable considering that noise covers successively more the less predominant factors and prevents them to be correctly pointed out as statistically significant.

On the other hand, regarding the last data matrix, the general overestimation of the number of components was not unexpected. In fact, as stressed in [5], in this particular circumstance and even for very small noise percentages, a notable portion of the variation of the last two factors gets lost in the residuals. Thus, it is difficult to conclude if the actual data dimensionality is either equal to or higher than 15.

Concerning P1, it commonly gave rise to a more conservative selection. In fact, as also evidenced by the comparison reported in Section 3.3, it allows only the major principal components to be appropriately recognised.

Finally, the outcomes resulting from the application of Horn's parallel analysis and Dray's method

Table 1 - Generation scheme of the 4 synthetic datasets. x identifies the observed variables, while pc denotes the principal components exploited for their simulation

Dataset #ID	Number of principal components	Number of original variables	Data generation scheme
1	4	10	$x_i = \sqrt{\frac{i}{5}}pc_1 + \sqrt{\frac{1-i}{5}}pc_2 \;\; \forall i \in 1,\dots,5$ $x_i = \sqrt{0.5}pc_1 + \sqrt{\frac{i}{10-0.5}}pc_2 + \sqrt{\frac{1-i}{10}}pc_3 \;\; \forall i \in 6,\dots,9$ $x_{10} = \frac{\sqrt{0.01}pc_1 + \sqrt{0.01}pc_2 + \sqrt{0.01}pc_3 + pc_4}{\sqrt{1.03}}$
2	8	10	$x_i = \sqrt{0.5}pc_k + \sqrt{0.5}pc_l \;\; \forall i \in 1,\dots,6 \;\; \forall k \neq l \in 1,\dots,4$ $x_i = \sqrt{0.5}pc_k + \sqrt{0.5}pc_l \;\; \forall i \in 7,8,9 \;\; \forall k \neq l \in 5,6,7$ $x_{10} = pc_8$
3	12	27	$x_i = pc_i \;\; \forall i \in 1,\dots,12$ $x_i = \sqrt{0.5}pc_k + \sqrt{0.5}pc_l \;\; \forall i \in 13,\dots,27 \;\; \forall k \neq l \in 1,\dots,6$
4	15	50	$x_i = \sqrt{0.5}pc_k + \sqrt{0.5}pc_l \;\; \forall i \in 1,\dots,45 \;\; \forall k \neq l \in 1,\dots,10$ $x_{46} = pc_{11}$ $x_{47} = pc_{12}$ $x_{48} = \sqrt{0.5}pc_{11} + \sqrt{0.5}pc_{13}$ $x_{49} = \sqrt{0.5}pc_{12} + \sqrt{0.5}pc_{14}$ $x_{50} = pc_{15}$

Table 2 - Median value and range of the number of components retained at each noise level for the 4 synthetic datasets (estimated over 300 simulation replicates for Horn's parallel analysis, Dray's method and both P1 and P3). Their real number of factors is reported in the second column. Bold characters point out a correctly addressed assessment. Dray's computational procedure was stopped just after the detection of the first non-significant factor

Dataset #ID	Noise level*	Real number of components/Number of original variables	Number of estimated components			
			Horn's parallel analysis	Dray's method	P1	P3
1	5%	4/10	1 [1 – 1]	1 [1 – 1]	1 [1 – 4]	**4 [4 – 4]**
	10%	4/10	1 [1 – 1]	1 [1 – 2]	1 [1 – 5]	**4 [4 – 5]**
	15%	4/10	1 [1 – 1]	1 [1 – 4]	1 [1 – 4]	**4 [4 – 4]**
	20%	4/10	1 [1 – 1]	1 [1 – 4]	1 [1 – 4]	**4 [1 – 4]**
	25%	4/10	1 [1 – 1]	1 [1 – 5]	1 [1 – 4]	**4 [1 – 4]**
	50%	4/10	1 [1 – 1]	1 [1 – 4]	1 [1 – 4]	**4 [1 – 4]**
	75%	4/10	1 [1 – 1]	1 [1 – 3]	1 [1 – 3]	2 [1 – 4]
	100%	4/10	1 [1 – 1]	1 [1 – 3]	1 [1 – 2]	1 [1 – 4]
2	5%	8/10	2 [2 – 4]	2 [1 – 5]	2 [2 – 6]	**8 [2 – 8]**
	10%	8/10	2 [2 – 4]	2 [1 – 10]	2 [2 – 6]	7 [2 – 9]
	15%	8/10	2 [2 – 4]	2 [1 – 10]	2 [1 – 6]	6 [2 – 8]
	20%	8/10	2 [1 – 4]	2 [1 – 6]	2 [1 – 6]	6 [2 – 9]
	25%	8/10	2 [1 – 4]	2 [1 – 6]	2 [1 – 5]	5 [1 – 8]
	50%	8/10	2 [1 – 4]	2 [1 – 4]	2 [1 – 3]	3 [1 – 8]
	75%	8/10	2 [0 – 4]	2 [0 – 10]	2 [0 – 3]	2 [0 – 6]
	100%	8/10	2 [0 – 4]	1 [0 – 3]	1 [0 – 3]	2 [0 – 5]
3	5%	12/27	6 [6 – 6]	6 [4 – 6]	6 [6 – 7]	**12 [5 – 12]**
	10%	12/27	6 [6 – 7]	6 [4 – 7]	6 [6 – 7]	**12 [5 – 12]**
	15%	12/27	6 [6 – 7]	6 [5 – 7]	6 [5 – 7]	**12 [5 – 12]**
	20%	12/27	6 [5 – 7]	6 [4 – 7]	6 [5 – 6]	**12 [5 – 12]**
	25%	12/27	6 [5 – 7]	6 [4 – 7]	6 [5 – 7]	**12 [5 – 12]**
	50%	12/27	6 [5 – 7]	6 [3 – 7]	6 [5 – 7]	**12 [5 – 13]**
	75%	12/27	6 [4 – 7]	6 [2 – 6]	6 [4 – 6]	**12 [5 – 14]**
	100%	12/27	6 [4 – 7]	5 [1 – 6]	5 [3 – 6]	10 [5 – 14]
4	5%	15/50	10 [10 – 11]	12 [8 – 17]	12 [10 – 15]	16 [15 – 19]
	10%	15/50	10 [9 – 11]	12 [9 – 17]	12 [10 – 15]	16 [15 – 21]
	15%	15/50	10 [9 – 11]	12 [8 – 16]	12 [10 – 16]	16 [14 – 21]
	20%	15/50	10 [9 – 11]	12 [7 – 18]	12 [10 – 15]	17 [14 – 21]
	25%	15/50	10 [9 – 11]	12 [9 – 16]	12 [9 – 15]	17 [14 – 21]
	50%	15/50	10 [8 – 11]	11 [9 – 14]	10 [9 – 13]	17 [13 – 24]
	75%	15/50	10 [7 – 11]	10 [6 – 13]	10 [7 – 12]	16 [13 – 23]
	100%	15/50	9 [7 – 11]	10 [7 – 13]	10 [7 – 11]	16 [12 – 24]

to the 4 synthetic datasets (also displayed in Figures SM1 and SM2, respectively) corroborate what was stated before about their respective limitations: in fact, both of them always overlooked some components of the original data structures. Furthermore, in all four cases, if the first non-significant component was not used as stopping rule, Dray's method would have been prone to detect their last factors as statistically significant (see Figure SM2)[viii]. In this sense, it can be said that here the proposed permutation test-based procedure (encompassing the P3 step) outperformed these two approaches.

4.2. Real case-studies

The developed algorithm was also applied to 10 real datasets from distinct research fields, from archaeology to food preference. The purpose was to illustrate the different results for the concerned approaches, not to discuss them in great detail. For some of these datasets, based on the findings described in the original publications, a putative number of underlying components could be identified. The outcomes for both P1 and P3 are reported in Figure 5, while those for Horn's parallel analysis and Dray's method are graphed in Figures SM3 and SM4 (see Table 3 for a comprehensive summary). Notice that i) all the methods were run on the auto-scaled data matrices with a 99% confidence level and performing 300 permutation rounds; ii) Dray's computational procedure was stopped just after the detection of the first non-significant factor. When information was available on the actual dimensionality of the data, P3 always permitted to retrieve the correct number of components. On the other hand, Horn's parallel analysis and Dray's method generally led to more conservative selections. In 4 out of 10 situations, P1 returned similar results as P3, probably because the significant components were large enough to limit the effect related to the different projection procedures pointed out in Sections 3.3 and 4.1. However, that is not valid for the other real case-studies where P1 yielded an underestimated number of factors with respect to P3 (see e.g. the performance of the 2 methodologies when the juice sensory array was handled)[ix]. In

[viii] Figures SM1 and SM2 are simply illustrative examples related to a single simulation replicate (noise level: 5%). However, the performance of the two concerned techniques was found to be consistent regardless of both noise percentage and data generation repetition.

[ix] For all the datasets, the 99th percentile front resulting from P1 was found to be higher (as expected) than that obtained when P3 was concerned.

Table 3 - Number of statistically significant components estimated by Horn's parallel analysis, Dray's method and both the P1- and P3-based permutation tests for the 10 real datasets. Their putative number of factors (if known) is reported in the fourth column. Bold characters point out a correctly addressed assessment

Dataset	Size	Reference	Putative number of components	Number of estimated components			
				Horn's parallel analysis	Dray's method	P1	P3
Herring ripening data	180×10	[23]	7	3	3	**7**	**7**
Air pollution spatial data	53×11	[24]	3	2	**3**	**3**	**3**
Train timing data	40×10	[24]	3	1	2	**3**	**3**
Juice sensory data	6×14	[25]	2	1	1	1	**2**
Obsidian trace element data	75×10	[26]	-	2	2	2	6
Solvent physical properties	103×9	[27]	-	2	2	2	3
Blood/urine composition data	65×52	[28, 29]	-	6	5	6	11
Boston area housing data	506×14	[30]	-	3	3	3	7
Ham consumer liking data	8×81	[31]	-	2	3	3	3
Yogurt descriptors	12×200	[32]	-	3	11	6	7

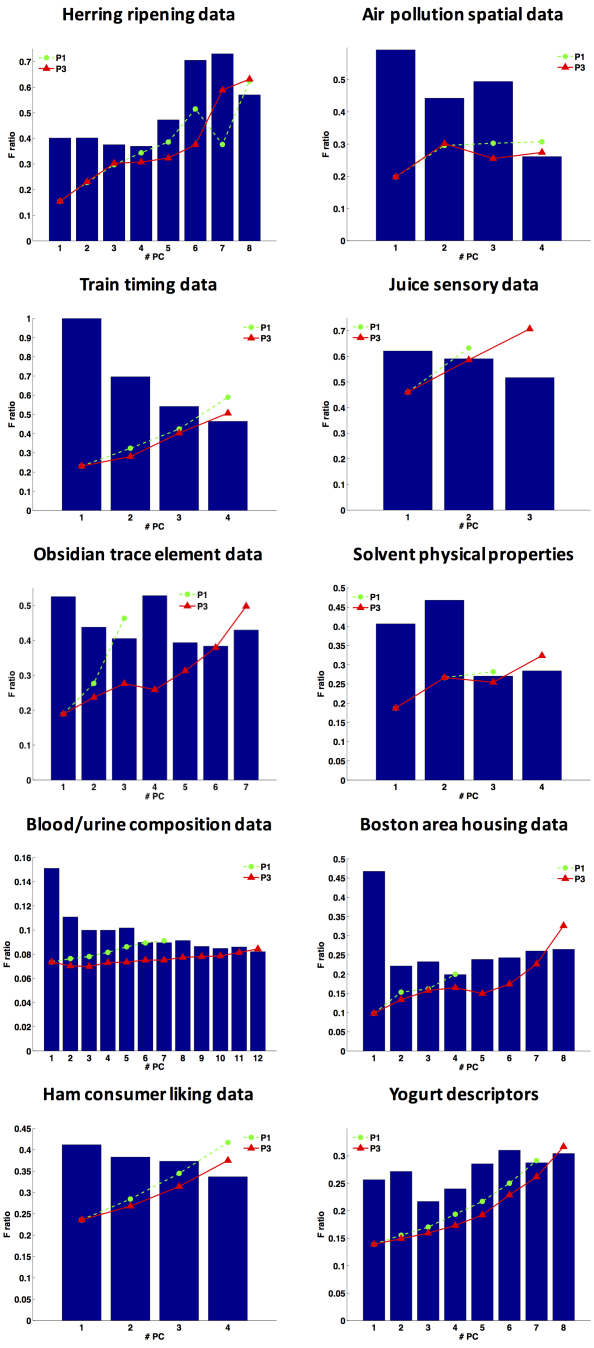

Figure 5 - Results of the application of the P1- and P3-based permutation tests to the 10 real datasets. The blue bars indicate the F_a ratios used for the testing procedure and associated to the single components of the original matrices under study, while the green dots (for P1) and the red triangles (for P3) correspond to the 99th percentiles of their respective *null*-distributions generated after 300 permutations

the light of that and although the actual dimensionality of some of the matrices taken into account was not known, the P3-based permutation test seemed to enable a more appropriate identification of how many principal components to extract in the different scenarios. It is true that sometimes more conservative selections may be safer especially when factors accounting for small amounts of data variation are detected as statistically significant (in this sense, the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, used for the Horn's parallel analysis testing procedure, can be helpful to additionally evaluate this aspect). But rather often phenomena of interest are just captured by such small components, and, thus, a tool being able to systematically unveil them can definitely be of use for many disparate applications.

5. Conclusions

In this paper, an extensive guideline on how to accomplish the selection of PCA components by permutation testing was provided through the description of a novel and efficient algorithm. Its most relevant aspects were discussed and clarified, namely the way the considered covariance structures are randomised, the importance of sequentially deflating the original matrix once every factor is computed, the necessity of a relative measure, the F_a ratio, to estimate their statistical significance and the need of a projection after each permutation round. This also permitted to mathematically formalise all the single numerical operations required when trying to quantify the number of factors underlying particular sets of data in this fashion. Furthermore, the application of the proposed method to both simulated and real case-studies highlighted that it can constitute a feasible and valid alternative to classical permutation test-based approaches such as Horn's parallel analysis and Dray's method, which exhibit specific limitations mainly related to their intrinsic mathematical procedures. The possibility of employing it for effective rank identification prior to multi-set data analysis by means of e.g. Canonical Correlation Analysis [33] or Joint and Individual Variation Explained [34] will be explored in future research.

6. Acknowledgements

This research work was partially supported by the Spanish Ministry of Economy and Competitiveness under the project DPI2014-55276-C5-1R and Shell Global Solutions International B.V.

(Amsterdam, The Netherlands).

7. References

[1] K. Pearson, On lines and planes of closest fit to systems of points in space, Philos. Mag. 2 (1901) 559–572.

[2] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.

[3] R. Bro, K. Kjeldahl, A. Smilde, H. Kiers, Cross-validation of component models: a critical look at current methods, Anal. Bioanal. Chem. 390 (2008) 1241–1251.

[4] J. Camacho, A. Ferrer, Cross-validation in PCA models with the element-wise k-fold (*ekf*) algorithm: theoretical aspects, J. Chemometr. 26 (2012) 361–373.

[5] J. Camacho, A. Ferrer, Cross-validation in PCA models with the element-wise *ekf* algorithm: practical aspects, Chemometr. Intell. Lab. 131 (2014) 37–50.

[6] R. Vitale, A. Zhyrova, J. Fortuna, O. de Noord, A. Ferrer, H. Martens, On-the-fly processing of continuous high-dimensional data streams, Chemometr. Intell. Lab. 161 (2017) 118–129.

[7] E. Saccenti, J. Camacho, Determining the number of components in Principal Components Analysis: a comparison of statistical, crossvalidation and approximated methods, Chemometr. Intell. Lab. 149 (2015) 99–116.

[8] H. Kaiser, The application of electronic computers to Factor Analysis, Educ. Psychol. Meas. 20 (1960) 141–151.

[9] W. Velicer, Determining the number of components from the matrix of partial correlations, Psychometrika 41 (1976) 321–327.

[10] R. Cattell, The scree test for the number of factors, Multivar. Behav. Res. 1 (1966) 245–276.

[11] M. Bartlett, A note on the multiplying factors for various χ^2 approximations, J. Roy. Stat. Soc. B Met. 16 (1954) 296–298.

[12] E. Saccenti, A. Smilde, J. Westerhuis, M. Hendriks, Tracy-widom statistic for the largest eigenvalue of autoscaled real matrices, J. Chemometr. 25 (2011) 644–652.

[13] J. Horn, A rationale and test for the number of factors in factor analysis, Psychometrika 30 (1965) 179–185.

[14] S. Dray, On the number of principal components: a test of dimensionality based on measurements of similarity between matrices, Comput. Stat. Data An. 52 (2008) 2228–2237.

[15] K. Kosanovich, K. Dahl, M. Piovoso, Improved process understanding using multiway Principal Component Analysis, Ind. Eng. Chem. Res. 35 (1996) 138–146.

[16] J. Camacho, J. Picó, A. Ferrer, Data understanding with PCA: structural and variance information plots, Chemometr. Intell. Lab. 100 (2010) 48–56.

[17] J. Camacho, Missing-data theory in the context of exploratory data analysis, Chemometr. Intell. Lab. 103 (2010) 8–18.

[18] J. Camacho, Observation-based missing data methods for exploratory data analysis to unveil the connection between observations and variables in latent subspace models, J. Chemometr. 25 (2011) 592–600.

[19] V. Vieira, Permutation tests to estimate significances on Principal Components Analysis, Comput. Ecol. Softw. 2 (2012) 103–123.

[20] P. Peres-Neto, D. Jackson, K. Somers, How many principal components? Stopping rules for determining the number of non-trivial axes revisited, Comput. Stat. Data An. 49 (2005) 974–997.

[21] I. Endrizzi, F. Gasperi, M. Rødbotten, T. Næs, Interpretation, validation and segmentation of preference mapping models, Food Qual. Prefer. 32 (2014) 198–209.

[22] E. Saccenti, M. Timmerman, Considering Horn's parallel analysis from a random matrix theory point of view, Psychometrika 82 (2017) 186–209.

[23] R. Bro, H. Nielsen, G. Stefánsson, T. Skåra, A phenomenological study of ripening of salted herring. assessing homogeneity of data from different countries and laboratories, J. Chemometr. 16 (2002) 81–88.

[24] R. Henry, E. Park, C. Spiegelman, Comparing a new algorithm with the classical methods for estimating the number of factors, Chemometr. Intell. Lab. 48 (1999) 91–97.

[25] M. Rødbotten, B. Martinsen, B. G.I., H. Mortvedt, S. Knutsen, P. Lea, T. Næs, A cross-cultural study of preference for apple juice with different sugar and acid contents, Food Qual. Prefer. 20 (2009) 277–284.

[26] B. Kowalski, T. Schatzki, F. Stross, Classification of archaeological artifacts by applying pattern recognition to trace element data, Anal. Chem. 44 (1972) 2176–2180.

[27] OpenMV website, `http://openmv.net/`.

[28] PLS_Toolbox Release 7.0.2, Eigenvector Research, Inc., Manson, Washington, USA (2012).

[29] Infometrix, Inc. website, `https://infometrix.com/`.

[30] D. Harrison, D. Rubinfeld, Hedonic housing prices and the demand for clean air, J. Environ. Econ. Manag. 5 (1978) 81–102.

[31] T. Næs, P. Brockhoff, O. Tomic, Statistics for sensory and consumer science, 1st Edition, John Wiley & Sons Ltd, Chichester, United Kingdom, 2010.

[32] T. Næs, I. Berget, K. Liland, G. Ares, P. Varela, Estimating and interpreting more than two consensus components in projective mapping: INDSCAL vs. Multiple Factor Analysis (MFA), Food Qual. Prefer. 58 (2017) 45–60.

[33] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.

[34] E. Lock, K. Hoadley, J. Marron, A. Nobel, Joint and Individual Variation Explained (JIVE) for intergrated analysis of multiple data types, Ann. Appl. Stat. 7 (2013) 523–542.

[35] L. Humphreys, R. Montanelli, An investigation of the parallel analysis criterion for determining the number of common factors, Multivar. Behav. Res. 10 (1975) 193–206.

[36] W. Zwick, W. Velicer, Comparison of five rules for determining the number of components to retain, Psychol.

Bull. 99 (1986) 432–442.

[37] L. Glorfeld, An improvement on horn's parallel analysis methodology for selecting the correct number of factors to retain, Educ. Psychol. Meas. 55 (1995) 377–393.

[38] B. Thompson, L. Daniel, Factor analytic evidence for the construct validity of scores: a historical overview and some guidelines, Educ. Psychol. Meas. 56 (1996) 197–208.

[39] R. Ledesma, P. Valero-Mora, Determining the number of factors to retain in EFA: an easy-to-use computer program for carrying out parallel analysis, PARE 12 (2) (2007) 1–11.

Appendix A. Horn's parallel analysis

Horn's parallel analysis [13] is a Monte Carlo-based approach, whose basic idea is to compare the eigenvalues of the covariance matrix resulting from the data array under study with their sampling distribution, obtained simulating uncorrelated variables. A factor or component is retained if its respective eigenvalue is larger than e.g. the 99^{th} percentile of its sampling distribution. Since the 70s, Horn's parallel analysis has been often considered the best available option for PCA component selection in psychometrics [35–39].

Appendix B. Dray's method

In its most efficient form [14], Dray's method encompasses the following 9 algorithmic steps grouped in three consecutive phases:

- Phase I - Singular Value Decomposition of $\mathbf{X}$:
  1. Perform Singular Value Decomposition (SVD) on $\mathbf{X}$:

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} = \mathbf{TP}^{\mathrm{T}} \tag{B.1}$$

     where $\mathbf{U}$ ($N \times N$) and $\mathbf{V}$ ($J \times J$) contain the left and right singular vectors of $\mathbf{X}$, respectively, and $\mathbf{S}$ ($N \times J$) is a rectangular diagonal array whose non-zero diagonal elements are its singular values ($\sqrt{\lambda_1}$, $\sqrt{\lambda_2}, \ldots,$ $\sqrt{\lambda_Q}$);
  2. Compute for each a-th calculated component the so-called $RVDIM$ statistic:

$$RVDIM_a = \frac{\lambda_a}{\sqrt{\sum_{q=a}^{Q} \lambda_q^2}} \tag{B.2}$$

where λ_a corresponds to the a-th eigenvalue obtained after the decomposition of $\mathbf{X}$. $RVDIM_a$ is used for testing the statistical significance of the single factors.

- Phase II - Test for the first component:

3. For $a = 1$, randomly and independently permute the order of the entries within every column of $\mathbf{X}$ constructing a new matrix $\mathbf{X}_{perm}$, featuring uncorrelated variables;
4. Apply SVD to $\mathbf{X}_{perm}$ and calculate the $RVDIM$ index for the first extracted component:

$$RVDIM_{1,perm} = \frac{\lambda_{1,perm}}{\sqrt{\sum_{q=1}^{Q} \lambda_{q,perm}^2}} \tag{B.3}$$

where $\lambda_{1,perm}$ denotes the first eigenvalue obtained after the decomposition of $\mathbf{X}_{perm}$;

5. Iterate step 3 and 4 to generate a *null*-distribution for $RVDIM_{1,perm}$. If $RVDIM_1$ is found to be higher than its $(1-\alpha)\times 100^{\text{th}}$ percentile, the first component is considered statistically significant.

- Phase III - Test for the a-th component ($a > 1$):

6. For $a > 1$, calculate the residual matrix:

$$\mathbf{E}_a = \mathbf{X} - \sum_{q=1}^{a-1} \mathbf{u}_q \sqrt{\lambda_q} \mathbf{v}_q^{\mathrm{T}} = \mathbf{X} - \sum_{q=1}^{a-1} \mathbf{t}_q \mathbf{p}_q^{\mathrm{T}} \tag{B.4}$$

where $\mathbf{u}_q$, $\mathbf{v}_q$, $\mathbf{t}_q$ and $\mathbf{p}_q$ are the q-th column vectors of $\mathbf{U}$, $\mathbf{V}$, $\mathbf{T}$ and $\mathbf{P}$ (see Eq. B.1), respectively;

7. Randomly and independently permute the order of the entries within each column of $\mathbf{E}_a$ constructing a new matrix $\mathbf{E}_{a,perm}$. As specified in Section 2, unlike $\mathbf{E}_a$, $\mathbf{E}_{a,perm}$ has rank Q;
8. Perform SVD on $\mathbf{E}_{a,perm}$ and retain:

$$RVDIM_{a,perm} = \frac{\lambda_{1,perm}}{\sqrt{\sum_{q=1}^{Q} \lambda_{q,perm}^2}} \tag{B.5}$$

where $\lambda_{1,perm}$ is the first eigenvalue obtained after the decomposition of $\mathbf{E}_{a,perm}$;

9. Iterate step 7 and 8 to generate a *null*-distribution for $RVDIM_{a,perm}$. If $RVDIM_a$ is found to be higher than its $(1-\alpha)\times 100^{\text{th}}$ percentile, the a-th component is considered statistically significant.

The original procedure additionally includes a sequential Bonferroni correction for multiple testing to limit the increase of the Type I error and automatically stops as soon as the first non-significant factor is detected.

It is worth noting that, as detailed in [14], $RVDIM_a$ measures the similarity between the original data reconstruction $\hat{\mathbf{X}}_a = \mathbf{u}_a \sqrt{\lambda_a}\mathbf{v}_a^{\mathrm{T}}$ (where $\mathbf{u}_a/\mathbf{v}_a$ represents the a-th left/right singular vectors of $\mathbf{X}$) and $\mathbf{E}_a$. The higher this similarity, the higher the content of relevant information that the a-th component carries.

For the properties of the $RVDIM$ statistic, Dray's method is generally prone to recognise less significant components than expected. In fact, $RVDIM_a$ and $RVDIM_{a,perm}$ (for $a = 1,\ldots,Q$) are inversely proportional to the terms $\sqrt{\sum_{q=a}^{Q}\lambda_q^2}$ and $\sqrt{\sum_{q=1}^{Q}\lambda_{q,perm}^2}$, respectively, where λ_q corresponds to the q-th eigenvalue obtained after the decomposition of $\mathbf{X}$, and $\lambda_{q,perm}$ denotes the q-th eigenvalue obtained after the decomposition of $\mathbf{X}_{perm}$ (if $a = 1$) or $\mathbf{E}_{a,perm}$ (if $a > 1$). For each a, $\sum_{q=a}^{Q}\lambda_q$ and $\sum_{q=1}^{Q}\lambda_{q,perm}$ are identical, but that is not the case for $\sum_{q=a}^{Q}\lambda_q^2$ and $\sum_{q=1}^{Q}\lambda_{q,perm}^2$ owing to the redistribution of the total variation of $\mathbf{X}$ (if $a = 1$) or $\mathbf{E}_a$ (if $a > 1$) induced by the permutation of their elements, which modifies the single values of $\lambda_{q,perm}$ with respect to those of λ_q. On average, $\sqrt{\sum_{q=1}^{Q}\lambda_{q,perm}^2}$ is lower than $\sqrt{\sum_{q=a}^{Q}\lambda_q^2}$ when components accounting for relatively large amounts of data variation are still to be deflated. Consequently, for small a, the values of $RVDIM_{a,perm}$ may be overestimated, giving rise to a too conservative selection.

Selecting the number of factors in Principal Component Analysis by permutation testing - Numerical and practical aspects

Supporting Material

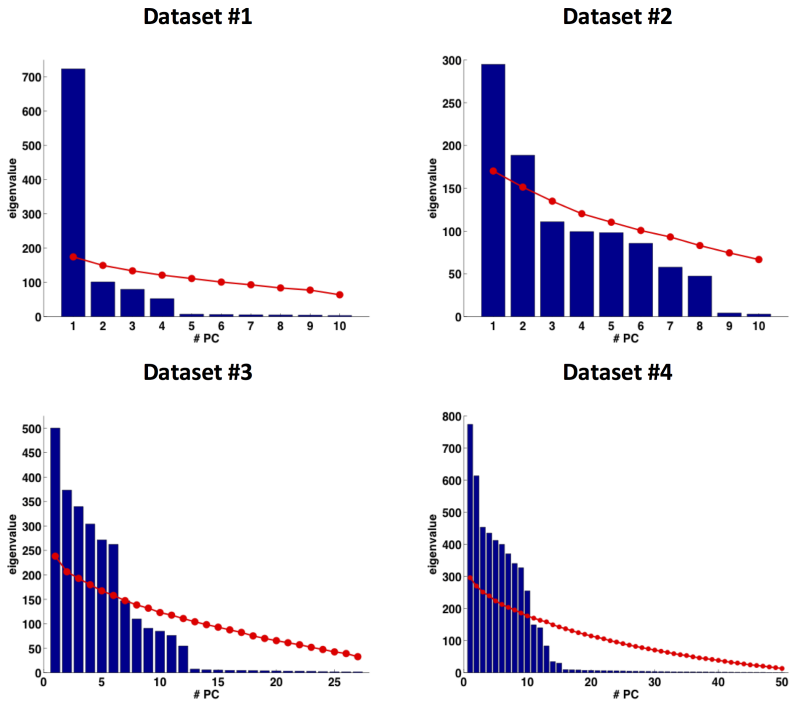

Figure SM1 - Results of the application of Horn's parallel analysis to the 4 synthetic datasets (noise level: 5%). The blue bars indicate the eigenvalues of the covariance matrices associated to the arrays under study, while the red dots correspond to the 99^{th} percentiles of their respective *null*-distributions generated after 300 permutations

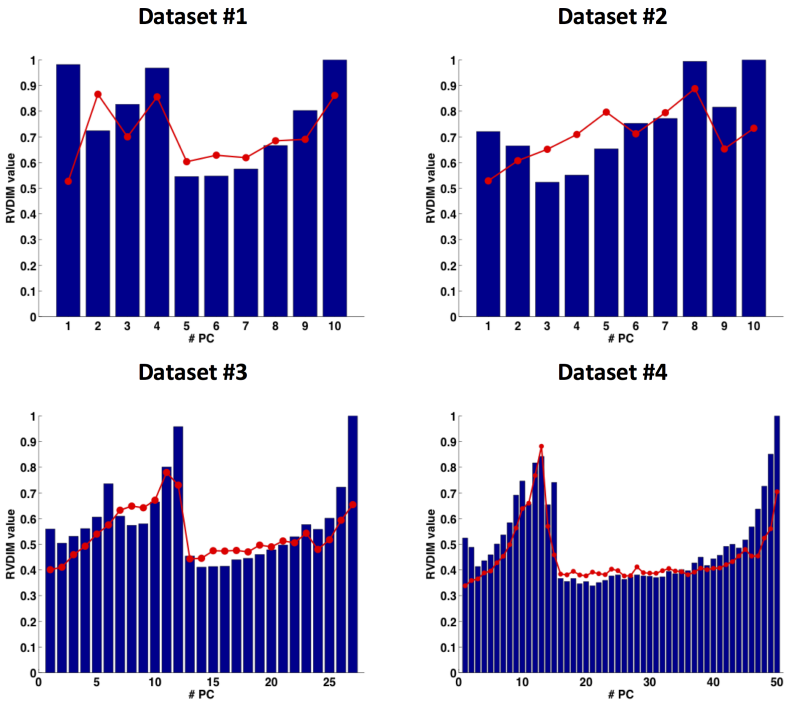

Figure SM2 - Results of the application of Dray's method to the 4 synthetic datasets (noise level: 5%). The blue bars indicate the $RVDIM_a$ values used for the testing procedure (see Appendix B for further details) and associated to the single components of the original matrices under study, while the red dots correspond to the 99^{th} percentiles of their respective *null*-distributions generated after 300 permutations

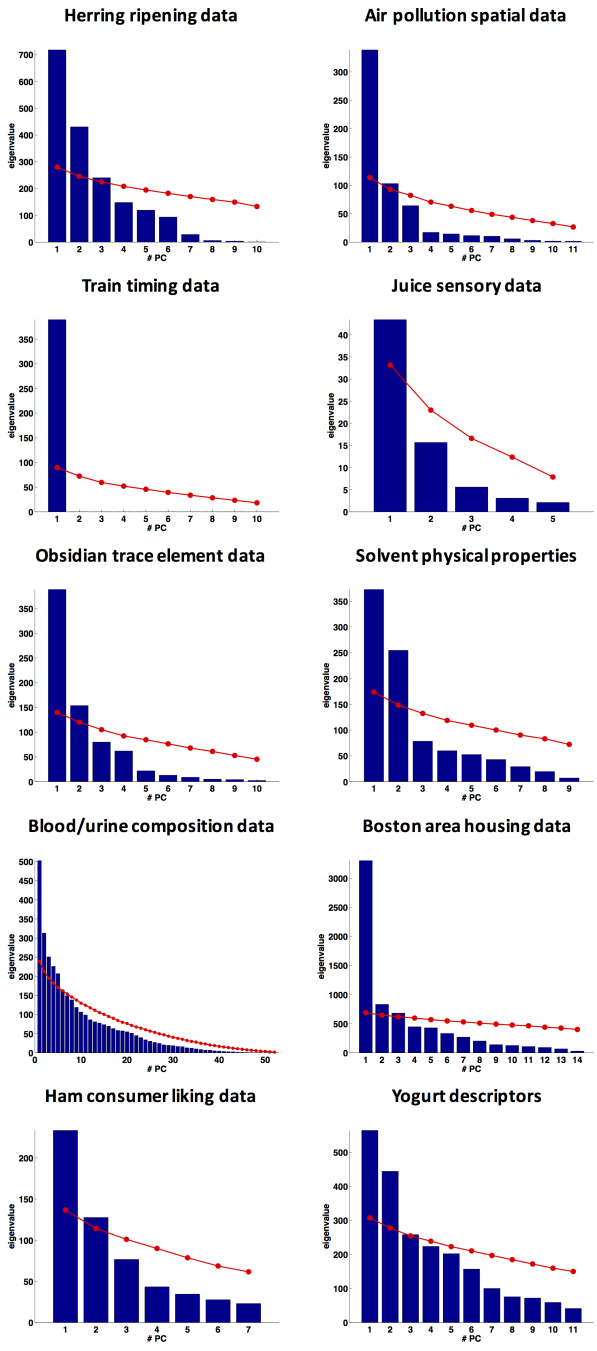

Figure SM3 - Results of the application of Horn's parallel analysis to the 10 real datasets. The blue bars indicate the eigenvalues of the covariance matrices associated to the arrays under study, while the red dots correspond to the 99$^{\text{th}}$ percentiles of their respective *null*-distributions generated after 300 permutations

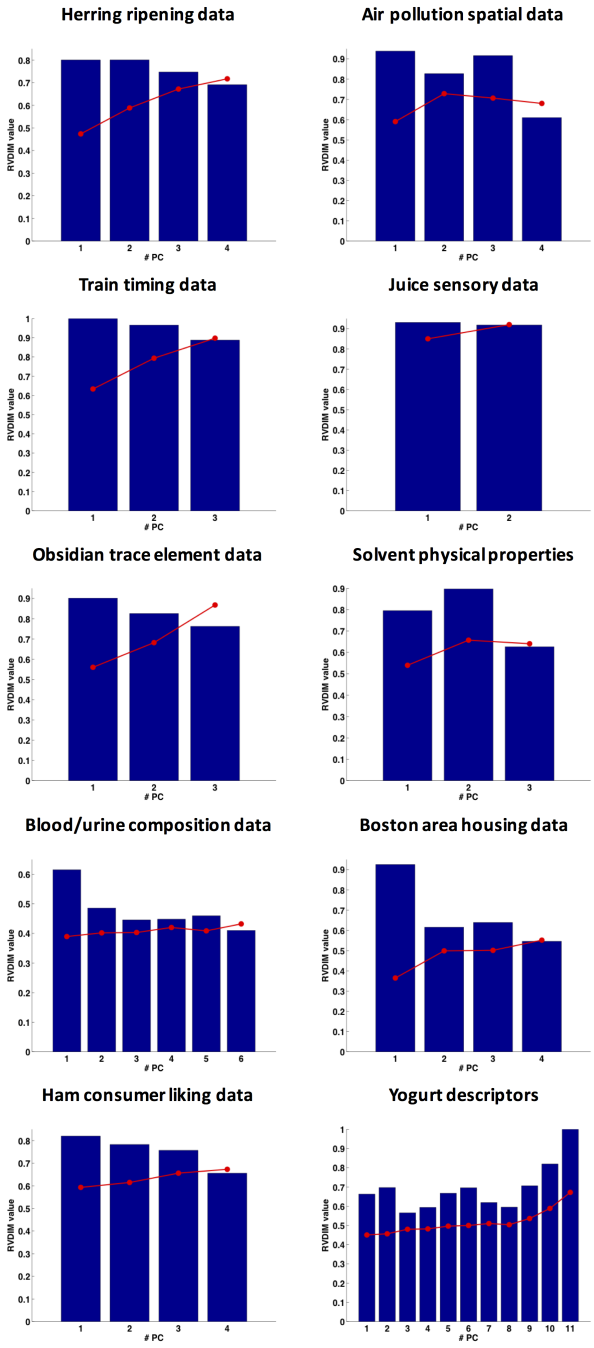

Figure SM4 - Results of the application of Dray's method to the 10 real datasets. The blue bars indicate the $RVDIM_a$ values used for the testing procedure (see Appendix B for further details) and associated to the single components of the original matrices under study, while the red dots correspond to the 99^{th} percentiles of their respective *null*-distributions generated after 300 permutations