

# Assessing the shared variation among high-dimensional data matrices: a modified version of the Procrustean correlation coefficient

E. Coissac<sup>1,\*</sup>, Co-Author<sup>2</sup> and C. Gonindard-Melodelima<sup>1</sup>

<sup>1</sup>Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, F-38000, France

<sup>2</sup>Department, Institution, City, Post Code, Country.

## Abstract

**Motivation:** Molecular biology and ecology are producing many high dimension data. Estimating correlation and shared variation between such data sets is an important step to distangle relationships among differents elements of a biological system. Unfortunately when using classical measure, because of the high dimension of the data, high correlation can be falsly infered.

**Results:** Here we propose a corrected version of the Procrustean correlation coefficient that is not sensible to the high dimension of the data. This allows for a correct estimation of the shared variation between two data sets and of the partial correlation coefficient between a set of matrix data.

**Availability:** The proposed corrected coefficients are implemented in the ProcMod R package available on <https://git.metabarcoding.org/lecasofts/ProcMod>

**Contact:** eric.coissac@metabarcoding.org

## 1 Introduction

Multidimensional data and even high-dimensional data, where the number of variables describing each sample is far larger than the sample count, are now regularly produced in functional genomics (*e.g.* transcriptomics, proteomics or metabolomics) and molecular ecology (*e.g.* DNA metabarcoding). Using various techniques, the same sample set can be described by several multidimensional data sets, each of them describing a different facet of the samples. This invites using data analysis methods able to evaluate mutual information shared by these different descriptions. Correlative approaches can be a first and simple way to decipher pairwise relationships of those data sets.

Since a long time ago, several coefficients have been proposed to measure correlation between two matrices (for a comprehensive review see

14 Ramsay *et al.*, 1984). But when applied to high-dimensional data, they  
 15 suffer from the over-fitting effect leading them to estimate a high corre-  
 16 lation even for unrelated data sets. Modified versions of some of these  
 17 matrix correlation coefficients have been already proposed to tackle this  
 18 problem. The  $RV_2$  coefficient (Smilde *et al.*, 2009) is correcting the origi-  
 19 nal RV coefficient (Escoufier, 1973) for over-fitting. Similarly, a modified  
 20 version of the distance correlation coefficient dCor (Székely *et al.*, 2007)  
 21 has been proposed by Székely and Rizzo (2013). dCor has the advantage  
 22 over the other correlation factors for not considering only linear relation-  
 23 ships. Here we will focus on the Procrustes correlation coefficient RLs  
 24 proposed by Lingoes and Schönemann (1974) and by Gower (1971). Let  
 25 define *Trace*, a function summing the diagonal elements of a matrix. For  
 26 a  $n \times p$  real matrix  $\mathbf{X}$  and a second  $n \times q$  real matrix  $\mathbf{Y}$  defining  
 27 respectively two sets of  $p$  and  $q$  centered variables characterizing  $n$  individ-  
 28 uals, we define  $\text{CovLs}(\mathbf{X}, \mathbf{Y})$  an analog of covariance applicable to vectorial  
 29 data following Equation (1)

$$\text{CovLs}(\mathbf{X}, \mathbf{Y}) = \frac{\text{Trace}((\mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{Y}')^{1/2})}{n-1} \quad (1)$$

30 and  $\text{VarLs}(\mathbf{X})$  as  $\text{CovLs}(\mathbf{X}, \mathbf{X})$ . RLs can then be expressed as follow  
 31 in Equation (2).

$$\text{RLs}(\mathbf{X}, \mathbf{Y}) = \frac{\text{CovLs}(\mathbf{X}, \mathbf{Y})}{\sqrt{\text{VarLs}(\mathbf{X}) \text{VarLs}(\mathbf{Y})}} \quad (2)$$

32 Among the advantages of RLs, its similarity with the Pearson's cor-  
 33 relation coefficient R (Bravais, 1844) has to be noticed. Considering  
 34  $\text{CovLs}(\mathbf{X}, \mathbf{Y})$  and  $\text{VarLs}(\mathbf{X})$ , respectively corresponding to the covari-  
 35 ance of two matrices and the variance of a matrix, Equation (2) high-  
 36 light the analogy between both the correlation coefficients. Besides, when  
 37  $p = 1$  and  $q = 1$ ,  $\text{RLs} = |\text{R}|$ . When squared RLs is an estimate, like  
 38 the squared Pearson's R, of the amount of variation shared between the  
 39 two datasets. This property allows for developing variance analyzing of  
 40 matrix data sets.

41 Moreover, Procrustean analyses have been proposed as a good alter-  
 42 native to Mantel's statistics for analyzing ecological data summarized by  
 43 distance matrices (Peres-Neto and Jackson, 2001). In such analyze dis-  
 44 tance matrices are projected into an orthogonal space using metric or  
 45 non metric multidimensional scaling according to the geometrical proper-  
 46 ties of the used distances. Correlations are then estimated between these  
 47 projections.

## 48 2 Approach

49 RLs is part of the procruste framework that aims to superimpose a set  
 50 of points with respect to another through three operations: a translation,  
 51 a rotation and a scaling. The optimal transfert matrix  $\text{Rot}_{X \rightarrow Y}$  can be  
 52 estimated from the singular value decomposition (SVD) of the covariance

53 matrix  $X'Y$ . SVD factorize any matrix as the product of three matrices  
 54 (Equation 3).

$$\mathbf{X}'\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \quad (3)$$

55  $\mathbf{U}$  and  $\mathbf{V}$  are two rotation matrices allowing to compute the transfer  
 56 matrix to superimpose  $\mathbf{X}$  on  $\mathbf{Y}$  or reciprocely  $\mathbf{Y}$  on  $\mathbf{X}$ . All the elements  
 57 of  $\mathbf{\Sigma}$  except its diagonal are equal to zero. The diagonal elements are the  
 58 singular values. Singular values are the extension of the eigenvalues to  
 59 non square matrices.  $\text{CovLs}(\mathbf{X}, \mathbf{Y})$  can also be computed from singular  
 60 values (Equation 4).

$$\text{CovLs}(\mathbf{X}, \mathbf{Y}) = \frac{\text{Trace}(\mathbf{\Sigma})}{n - 1} \quad (4)$$

61 This expression illustrates that actually  $\text{CovLs}(\mathbf{X}, \mathbf{Y})$  is the variance  
 62 of the projections of  $\mathbf{X}$  on  $\mathbf{Y}$  or of the reciproque projection. Therefore  
 63  $\text{CovLs}(\mathbf{X}, \mathbf{Y})$  and  $\text{RLs}(\mathbf{X}, \mathbf{Y})$  are always positive and rotation independante.  
 64 Here we propose to partitionate this variance in two components.  
 65 A first one corresponding to the actual shared information between  $\mathbf{X}$  and  
 66  $\mathbf{Y}$ , and a second part that corresponds to what two random matrices of  
 67 same structure than  $\mathbf{X}$  and  $\mathbf{Y}$  are sharing. This second part is estimated  
 68 as  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$  the mean of such random correlation.  $\text{ICovLs}(\mathbf{X}, \mathbf{Y})$ ,  
 69 the informative part of  $\text{CovLs}(\mathbf{X}, \mathbf{Y})$ , is computed using Equation (5).

$$\text{ICovLs}(\mathbf{X}, \mathbf{Y}) = \text{Max} \begin{cases} \text{CovLs}(\mathbf{X}, \mathbf{Y}) - \overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})} \\ 0 \end{cases} \quad (5)$$

70 Similarly the informative counter-part of  $\text{VarLs}(\mathbf{X})$  is defined as  
 71  $\text{IVarLs}(\mathbf{X}) = \text{ICovLs}(\mathbf{X}, \mathbf{X})$ , and  $\text{IRLs}(\mathbf{X}, \mathbf{Y})$  the informative Procruste  
 72 correlation coefficient as follow.

$$\text{IRLs}(\mathbf{X}, \mathbf{Y}) = \frac{\text{ICovLs}(\mathbf{X}, \mathbf{Y})}{\sqrt{\text{IVarLs}(\mathbf{X}) \text{IVarLs}(\mathbf{Y})}} \quad (6)$$

73 Like  $\text{RLs}(\mathbf{X}, \mathbf{Y})$ ,  $\text{IRLs}(\mathbf{X}, \mathbf{Y}) \in [0; 1]$  with the 0 value corresponding  
 74 to no correlation and the maximum value 1 reached for two strictly ho-  
 75 mothetic data sets.

76 The corollary of  $\text{ICovLs}(\mathbf{X}, \mathbf{Y})$  and  $\text{IVarLs}(\mathbf{X})$  definitions is that  
 77  $\text{ICovLs}(\mathbf{X}, \mathbf{Y}) \geq 0$  and  $\text{IVarLs}(\mathbf{X}) > 0$ . Therefore for  $M =$   
 78  $\{\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_k\}$  a set of  $k$  matrices with the same number of row, the  
 79 informative covariance matrix  $\mathbf{C}$  defined as  $\mathbf{C}_{i,j} = \text{ICovLs}(\mathbf{M}_i, \mathbf{M}_j)$  is  
 80 definite positive and symmetrical. This allows for defining the precision  
 81 matrix  $\mathbf{P} = \mathbf{C}^{-1}$  and the related partial correlation coefficient matrix  
 82  $\text{IRLs}_{\text{partial}}$  using Equation (7)

$$\text{IRLs}_{\text{partial}}(\mathbf{M}_i, \mathbf{M}_j) = \frac{\mathbf{P}_{i,j}}{\sqrt{\mathbf{P}_{i,i}\mathbf{P}_{j,j}}} \quad (7)$$

### 3 Methods

#### 3.1 Monte-Carlo estimation of $\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{Y})$

For every values of  $p$  and  $q$  including 1,  $\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{Y})$  can be estimated using a serie of  $k$  random matrices  $RX = \{\mathbf{R}\mathbf{X}_1, \mathbf{R}\mathbf{X}_2, \dots, \mathbf{R}\mathbf{X}_k\}$  and  $RY = \{\mathbf{R}\mathbf{Y}_1, \mathbf{R}\mathbf{Y}_2, \dots, \mathbf{R}\mathbf{Y}_k\}$  where each  $\mathbf{R}\mathbf{X}_i$  and  $\mathbf{R}\mathbf{Y}_i$  have the same structure respectively than  $\mathbf{X}$  and  $\mathbf{Y}$  in term of number of columns and of covariance matrix of these columns.

$$\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^k \text{CovLs}(\mathbf{R}\mathbf{X}_i, \mathbf{R}\mathbf{Y}_i)}{k} \quad (8)$$

Even when  $\mathbf{X} = \mathbf{Y}$  to estimate  $\text{IVarLs}(\mathbf{X})$ ,  $\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{X})$  is estimated with two independent sets of random matrix  $RX$  and  $RY$ , both having the same structure than  $\mathbf{X}$ .

#### Empirical assessment of $\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{Y})$

For two random vectors  $\mathbf{x}$  and  $\mathbf{y}$  of length  $n$ , the average coefficient of determination is  $\overline{R^2} = 1/(n - 1)$ . This value is independent of the distribution of the  $\mathbf{x}$  and  $\mathbf{y}$  values, but what about the independence of  $\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{Y})$  to the distributions of  $\mathbf{X}$  and  $\mathbf{Y}$ . To test this independence and to assess the reasonable randomization effort needed to estimate  $\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{Y})$ , this value is estimated for four matrices  $\mathbf{K}, \mathbf{L}, \mathbf{M}, \mathbf{N}$  of  $n = 20$  rows and respectively 10, 20, 50, 100 columns. Values of the four matrices are drawn from a normal or an exponential distribution, and  $k \in \{10, 100, 1000\}$  randomizations are tested to estimate  $\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{Y})$  and the respective standard deviation  $\sigma(\overline{\text{RCovLs}}(\mathbf{X}, \mathbf{Y}))$ . The VarLs of the generated matrices is equal to 1 therefore the estimated CovLs are equals to RLs.

#### 3.2 Simulating data for testing sensibility to overfitting

To test sensibility to overfitting correlations were mesured between two random matrices of same dimensions. Each matrix is  $n \times p$  with  $n = 20$  and  $p \in [2, 50]$ . Each  $p$  variables are drawn from a centered and reduced normal distribution  $\mathcal{N}(0, 1)$ . Eight correlation coefficients have been tested: RLs the original procruste coefficient, IRLs this work, RV the original R for vector data (Robert and Escoufier, 1976), *RV adjMaye*, *RV 2* and *RV adjGhaziri* three modified versions of RV (El Ghaziri and Qannari, 2015; Mayer *et al.*, 2011; Smilde *et al.*, 2009), *dCor* the original distance correlation coefficient (Székely *et al.*, 2007) and *dCor\_ttest* a modified version of *dCor* not sensible to overfitting (Székely and Rizzo, 2013). For each  $p$  value, 100 simulations were run. Computation of IRLs is estimated with 100 randomizations.

For  $p = 1$  random vectors with  $n \in [3, 25]$  are generated. As above data are drawn from  $\mathcal{N}(0, 1)$  and  $k = 100$  simulations are run for each  $n$ . The original Pearson correlation coefficient  $R$  and the modified version  $\text{IR}$  are used to estimate correlation between both vectors.

124 **3.3 Empirical assessment of the coefficient of de-**  
125 **termination**

126 The coefficient of determination ( $R^2$ ) represents the part of shared varia-  
127 tion between two variables.  $RLs^2$  keeps the same meaning when applied  
128 to two matrices. But because of over-fitting RLs and therefore  $RLs^2$  are  
129 over-estimated.

130 **between two matrices**

131 To test how the IRLs version of the coefficient of determination  $IRLs^2$  can  
132 perform to evaluate the shared variation, pairs of random matrices were  
133 produced for two values of  $p \in \{10, 100\}$  and  $n \in \{10, 25\}$ , and for several  
134 levels of shared variations ranging between 0.1 and 1 by step of 0.1. For  
135 each combination of parameters,  $k = 100$  simulations are run, and both  
136  $RLs^2$  and  $IRLs^2$  are estimated using 100 randomizations.

137 **between two vectors**

138 Coefficient of determination between two vectors also suffers from over-  
139 estimation when  $n$  the number of considered points is small. On average  
140 for two random vectors of size  $n$ ,  $\overline{R^2} = 1/(n - 1)$ . This random part of  
141 the shared variation inflates the observed shared variation even for not  
142 random vectors. In the context of multiple linear regression, Theil *et al.*  
143 (1958) proposed an adjusted version of the coefficient of determination  
144 (Equation 9) correcting for both the effect of the number of vector con-  
145 sidered ( $p$ ) and of the vector size ( $n$ ).

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1} \quad (9)$$

146 To evaluate the strength of that over-estimation and the relative effect  
147 of the correction proposed by Theil *et al.* (1958) and by IRLs, pairs of  
148 random vectors were produced for  $n \in \{10, 25\}$ , and for several levels  
149 of shared variations ranging between 0.1 and 1 by step of 0.1. For each  
150 combination of parameters,  $k = 100$  simulations are run, and  $R^2$ ,  $R_{adj}^2$   
151 and  $IRLs^2$  are estimated using 100 randomizations.

152 **partial determination coefficients**

153 To evaluate capacity of partial determination coefficient  $IRLs_{partial}^2$  to  
154 distangle nested correlations, four matrices **A**, **B**, **C**, **D** of size  $n \times p = 20 \times$   
155  $200$  are generated according to the schema: **A** shares 80% of variation with  
156 **B**, that shares 40% of variation with **C**, sharing 20% of variation with **D**.  
157 These direct correlations induce indirect ones spreading the total variation  
158 among each pair of matrices according to Figure 1. The simulation is  
159 repeated 100 times, for every simulation  $IRLs_{partial}^2$  and  $RLs_{partial}^2$  are  
160 estimated for each pair of matrices.

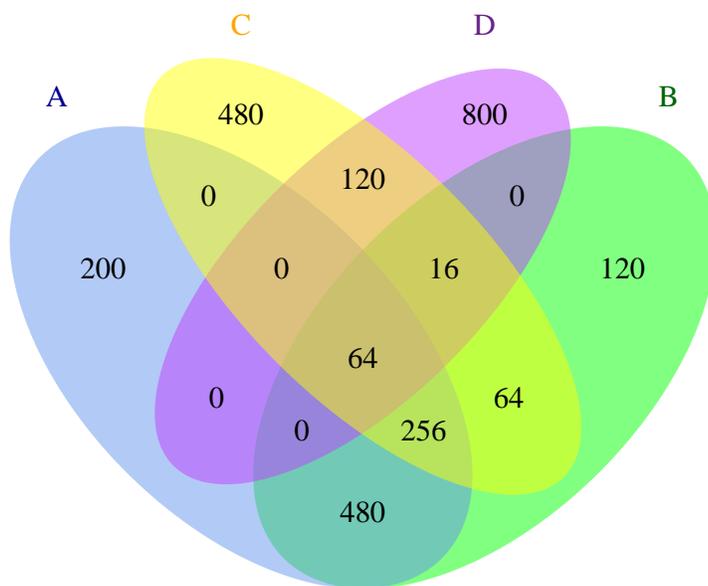


Figure 1: Theoretical distribution of the shared variation between the four matrices **A**, **B**, **C**, **D**, expressed in permille.

### 3.4 Testing significance of IRLs( $\mathbf{X}$ , $\mathbf{Y}$ )

Significance of IRLs( $\mathbf{X}$ ,  $\mathbf{Y}$ ) can be tested using permutation test as defined in Jackson (1995) or Peres-Neto and Jackson (2001) and implemented respectively in the `protest` method of the `vegan` R package (Dixon, 2003) or the `procuste.rtest` method of the `ADE4` R package Dray and Dufour (2007).

It is also possible to take advantage of the Monte-Carlo estimation of  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$  to test that  $\overline{\text{ICovLs}(\mathbf{X}, \mathbf{Y})}$  and therefore IRLs( $\mathbf{X}$ ,  $\mathbf{Y}$ ) are greater than expected under random hypothesis. Let counting over the  $k$  randomization when  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}_k$  greater than  $\overline{\text{CovLs}(\mathbf{X}, \mathbf{Y})}$  name this counts  $N_{>\text{CovLs}}$ .  $P_{\text{value}}$  of the test can be estimated following Equation (10).

$$P_{\text{value}} = \frac{N_{>\text{CovLs}}}{k} \quad (10)$$

#### Empirical assessment of $\alpha$ -risk for the *CovLs* test

To assess empirically the  $\alpha$ -risk of the `procuste` test based on the randomizations realized during the estimation of  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$ , distribution of  $P_{\text{value}}$  under the  $H_0$  is compared to a uniform distribution between 0 and 1 ( $\mathcal{U}(0, 1)$ ). To estimate such empirical distribution,  $k = 1000$  pairs of  $n \times p$  random matrices with  $n = 20$  and  $p \in \{10, 20, 50\}$  are simulated under the null hypothesis of independancy. Procruste correlation between whose matrices is tested based on three tests. Our proposed test (*CovLs.test*), the `protest` method of the `vegan` R package and the `procuste.rtest` method of the `ADE4` R package. Conformance of the distribution of each set of  $k$   $P_{\text{value}}$  to  $\mathcal{U}(0, 1)$  is assessed using the Cramer-Von Mises test (Csörgő and Faraway, 1996) implemented in the `cvm.test` function of the R package `gofstest`.

#### Empirical power assessment for the *CovLs* test

To evaluate relative power of the three considered tests, pairs of two random matrices were produced for various  $p \in \{10, 20, 50, 100\}$ ,  $n \in \{10, 15, 20, 25\}$  and two levels of shared variations  $R^2 \in \{0.05, 0.1\}$ . For each combination of parameters,  $k = 1000$  simulations are run. Each test are estimated based on 1000 randomizations for the *CovLs* test, or permutations for `protest` and `procuste.rtest`.

## 4 Results

### 4.1 Empirical assessment of $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$

Two main parameters can influence the Monte Carlo estimation of  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$ : the distribution used to generate the random matrices and  $k$  the number of random matrix pair. Two very different distribution are tested to regenerate the random matrices, the normal and the exponential distributions. The first one is symmetric where the second is not

Table 1: Estimation of  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$  according to the number of random matrices (k) aligned.

p	k	normal		exponential	
		mean	sd	mean	sd
10	10	0.5746	$1.3687 \times 10^{-2}$	0.5705	$1.1714 \times 10^{-2}$
	100	0.5824	$3.6425 \times 10^{-3}$	0.5794	$3.6452 \times 10^{-3}$
	1000	0.5806	$1.1564 \times 10^{-3}$	0.5801	$1.1515 \times 10^{-3}$
20	10	0.7682	$4.5568 \times 10^{-3}$	0.7729	$7.4071 \times 10^{-3}$
	100	0.7645	$2.1681 \times 10^{-3}$	0.7695	$2.0006 \times 10^{-3}$
	1000	0.7658	$6.8923 \times 10^{-4}$	0.7671	$6.6033 \times 10^{-4}$
50	10	0.9131	$3.4659 \times 10^{-3}$	0.9102	$2.6647 \times 10^{-3}$
	100	0.9088	$8.9318 \times 10^{-4}$	0.9086	$7.7424 \times 10^{-4}$
	1000	0.9086	$2.8590 \times 10^{-4}$	0.9091	$2.8450 \times 10^{-4}$
100	10	0.9557	$1.2939 \times 10^{-3}$	0.9532	$1.7316 \times 10^{-3}$
	100	0.9542	$4.7043 \times 10^{-4}$	0.9547	$4.6869 \times 10^{-4}$
	1000	0.9544	$1.4547 \times 10^{-4}$	0.9546	$1.4249 \times 10^{-4}$

with a high probability for small values and a long tail of large ones. Despite the use of these contrasted distributions, estimates of  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$  and of  $\sigma(\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})})$  are identical if we assume the normal distribution of the  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$  estimator and a 0.95 confidence interval of  $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})} \pm 2\sigma(\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})})$  (Table 1).

## 4.2 Relative sensibility of $IRLs(X, Y)$ to overfitting

$RLs$ , like  $RV$  and  $dCor$ , is sensible to overfitting which increase when  $n$  decrease, and  $p$  or  $q$  increase. Because  $RV$  is more comparable to  $R^2$  when  $RLs$  and  $dCor$  are more comparable to  $R$ ,  $RV$  values increase more slowly than  $RLs$  and  $dCor$  values with  $p$  (Figure 2A). Because of its definition  $IRLs$  values for non-correlated matrices are close to 0 whatever  $p$  (Figure 2A).

## 4.3 Evaluating the shared variation between two matrices

$RLs$  can be considered for matrices as a strict equivalent of Pearson's  $R$  for vectors. Therefore its squared value is an estimator of the shared variation between two matrices. But because of over-fitting the estimation is over-estimated. The proposed corrected version ( $IRLs$ ) of that coefficient is able to provide a good estimate of the shared variation and is perfectly robust to the over-fitting phenomenon (Figure 3). Only a small over evaluation is observable for the low values of simulated shared variation.

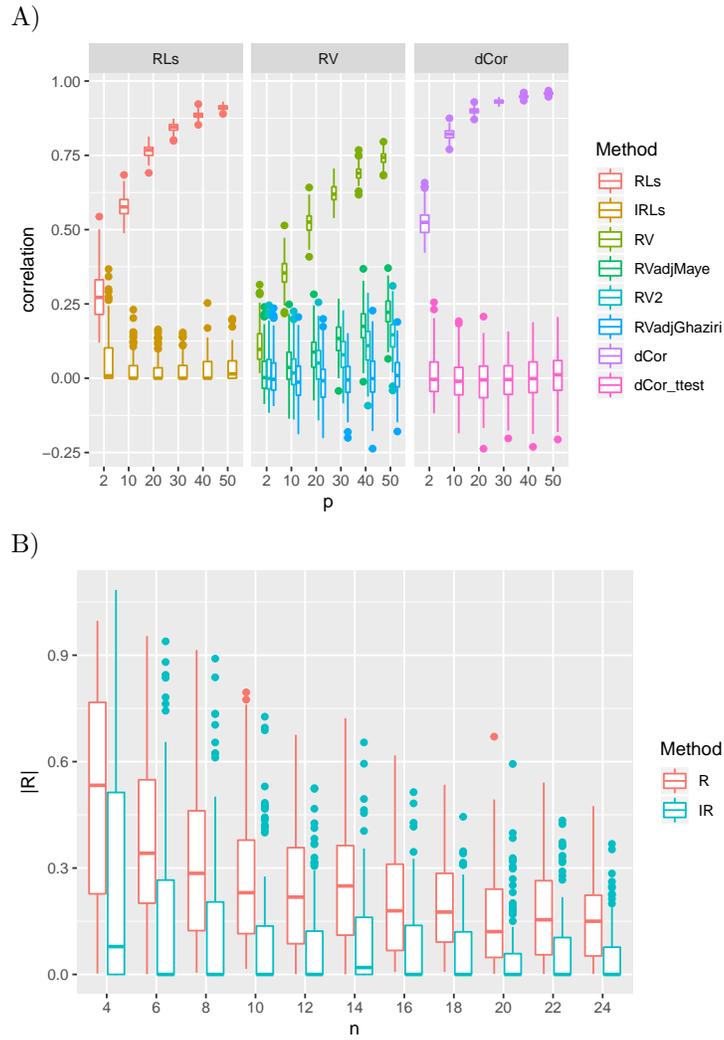


Figure 2: A) Sensibility to overfitting for various correlation coefficients. (A) Both simulated data sets are matrices of size  $(n \times p)$  with  $p > 1$ . B) Correlated data sets are vectors ( $p = 1$ ) with a various number of individuals  $n$  (vector length). A & B) 100 simulations are run for each combination of parameters

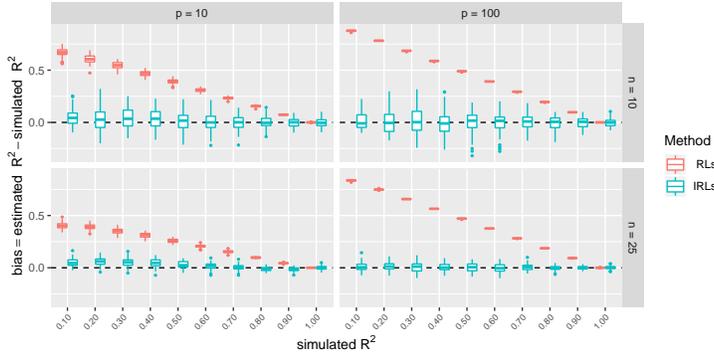


Figure 3: Shared variation ( $R^2$ ) between two matrices is measured using both the corrected (IRLs) and the original (RLs) versions of the procrustean correlation coefficient. A gradient of  $R^2$  is simulated for two population sizes ( $n \in \{10, 25\}$ ) and two numbers of descriptive variables ( $p \in \{10, 100\}$ ). The distribution of differences between the observed and the simulated shared variation is plotted for each condition. The black dashed line corresponds to a perfect match where measured  $R^2$  equals the simulated one.

222

### between two vectors

223

224

225

226

227

228

229

230

231

232

233

234

Vectors can be considered as a single column matrix, and the efficiency of IRLs<sup>2</sup> to estimate shared variation between matrices can also be used to estimate shared variation between two vectors. Other formulas have been already proposed to better estimate shared variation between vectors in the context of linear models. Among them the one presented in Equation 9, is the most often used and is the one implemented in R linear model summary function. On simulated data, IRLs<sup>2</sup> performs better than the simple  $R^2$  and its modified version  $R_{adj}^2$  commonly used (Figure 4). Whatever the estimator the bias decrease with the simulated shared variation. Nevertheless for every tested cases the median of the bias observed is smaller than with both other estimators, even if classical estimators well perform for large values of shared variation.

235

### partial coefficient of determination

236

237

238

239

240

241

242

243

244

245

The simulated correlation network between the four matrices **A**, **B**, **C**, **D** induced moreover the direct simulated correlation a network of indirect correlation and therefore shared variances (Figure 1). In such system, the interest of partial correlation coefficients and their associated partial determination coefficients is to measure correlation between a pair of variable without accounting for the part of that correlation which is explained by other variables, hence extracting the pure correlation between these two matrices. From Figure 1, the expected partial shared variation between **A** and **B** is  $480/(200+480) = 0.706$ ; between **B** and **C**,  $64/(480+120) = 0.107$ ; and between **C** and **D**  $120/800 = 0.150$ . All other

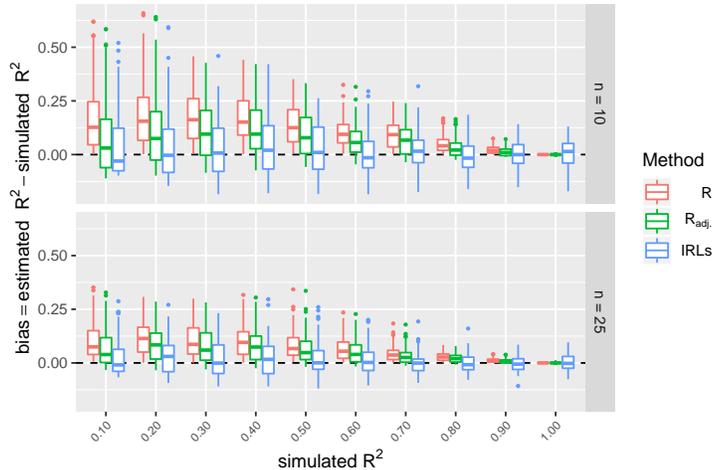


Figure 4: Shared variation between two vectors is measured using the classical  $R^2$ , its adjusted version  $R_{adj}^2$  and (IRLS $^2$ ). A gradient of shared variation is simulated for two vector sizes ( $n \in \{10, 25\}$ ). The black dashed line corresponds to a perfect match where measured  $R^2$  equals the simulated one.

Table 2:  $P_{values}$  of the Cramer-Von Mises test of conformity of the distribution of  $P_{values}$  correlation test to  $\mathcal{U}(0, 1)$  under the null hypothesis.

p	Cramer-Von Mises p.value		
	CovLs test	protest	procuste.rtest
10	0.323	0.395	0.348
20	0.861	0.769	0.706
50	0.628	0.783	0.680

246 partial coefficient are expected to be equal to 0. The effect of the correc-  
 247 tion introduced in IRLs is clearly weaker and on the partial coefficient  
 248 of determination (Figure 5) than on the full coefficient of determination  
 249 (Figure 3). The spurious random correlations, constituting the over-fitting  
 250 effect, is distributed over all the pair of matrices **A**, **B**, **C**, **D**.

#### 251 4.4 $p_{value}$ distribution under null hypothesis

252 As expected,  $P_{values}$  of the *CovLs* test based on the estimation of  
 253  $RCovLs(X, Y)$  are uniformly distributed under  $H_0$ . whatever the  $p$   
 254 tested (Table 2). This ensure that the probability of a  $P_{value} \leq \alpha$ -risk  
 255 is equal to  $\alpha$ -risk. Moreover  $P_{values}$  of the *CovLs* test are strongly  
 256 linearly correlated with those of both the other tests ( $R^2 = 0.996$  and  
 257  $R^2 = 0.996$  respectively for the correlation with `vegan::protest` and  
 258 `ade4::procuste.rtest`  $P_{values}$ ). The slopes of the corresponding linear  
 259 models are respectively 0.998 and 0.999.

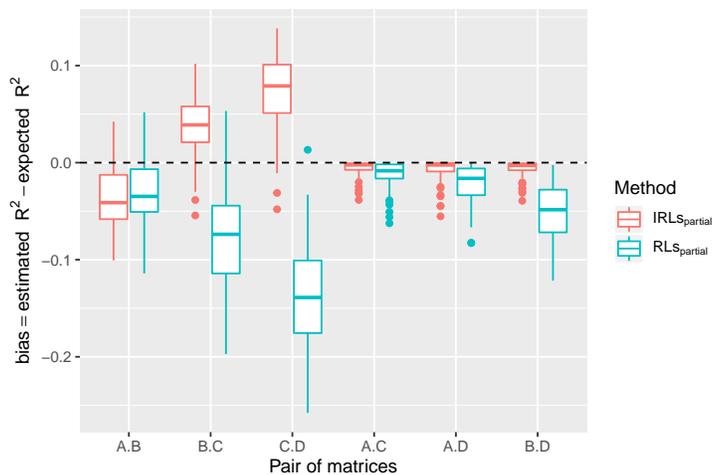


Figure 5: Estimation error on the partial determination coefficient. Error is defined as the absolute value of the difference between the expected and the estimated partial  $R^2$  using the corrected  $IRLs_{partial}$  and not corrected  $RLs_{partial}$  procruste correlation coefficient.

260  
261  
262  
263  
264  
265  
266  
267  
268

#### 4.5 Power of the test based on randomisation

Power of the *CovLs* test based on the estimation of  $\overline{RCovLs}(X, Y)$  is equivalent of the power estimated for both `vegan::protest` and `ade4::procuste.rtest` tests (Table 3). As for the two other tests, power decreases when the number of variable ( $p$  or  $q$ ) increases, and increase with the number of individuals and the shared variation. The advantage of the test based on the Monte-Carlo estimation of  $\overline{RCovLs}(X, Y)$  is to remove the need of running a supplementary set of permutations when IRLs is computed.

269

## 5 Discussion

270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281

Correcting the over-adjustment effect on metrics assessing the relationship between high dimension datasets is a constant effort over the past decade. Therefore, IRLs can be considered as a continuation of the extension of the toolbox available to biologists for analyzing their omics data. The effect of the proposed correction on the classical RLS coefficient is as strong as the other ones previously proposed for other correlation coefficients measuring relationship between vector data (see Figure 3, e.g. Smilde *et al.*, 2009; Székely and Rizzo, 2013). When applied to univariate data, RLS is equal to the absolute value of the Pearson correlation coefficient, hence, and despite it is not the initial aim of that coefficient, IRLs can also be used to evaluate correlation between two univariate datasets. Using IRLs for such data sets is correcting for spurious correlations when the number of

Table 3: Power estimation of the procruste tests for two low level of shared variations 5% and 10%.

	$R^2$	5%				10%			
	p	10	20	50	100	10	20	50	100
	n	$power = 1 - \beta\text{-risk}$							
Covls.test	10	0.49	0.45	0.40	0.45	0.76	0.68	0.70	0.68
	15	0.88	0.80	0.75	0.75	0.99	0.98	0.96	0.95
	20	0.99	0.96	0.94	0.93	1.00	1.00	1.00	1.00
	25	1.00	1.00	0.99	0.98	1.00	1.00	1.00	1.00
protest	10	0.50	0.45	0.40	0.45	0.77	0.70	0.70	0.68
	15	0.88	0.80	0.75	0.75	0.99	0.98	0.96	0.95
	20	0.99	0.96	0.94	0.93	1.00	1.00	1.00	1.00
	25	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
procruste.rtest	10	0.50	0.45	0.41	0.45	0.76	0.69	0.70	0.68
	15	0.88	0.80	0.74	0.75	0.99	0.98	0.96	0.96
	20	0.99	0.96	0.94	0.93	1.00	1.00	1.00	1.00
	25	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00

282 individual is small more efficiently than classical correction (see Figure 4,  
 283 Theil *et al.*, 1958).

284 The main advantage of IRLs over other matrix correlation coefficients  
 285 is that it allows for estimating shared variation between two matrices  
 286 according to the classical definition of variance partitioning used with  
 287 linear models. This opens the opportunity to develop linear models to  
 288 explain the variation of a high dimension dataset by a set of other high  
 289 dimension data matrices.

290 The second advantage of IRLs is that its definition implies that the  
 291 variance/co-variance matrix of a set of matrices is positive-definite. That  
 292 allows for estimating partial correlation coefficients matrix by inverting  
 293 the variance/co-variance matrix. The effect of the correction is less strong  
 294 on such partial coefficients than on full correlation, but the partial coef-  
 295 ficients that should theoretically be estimated to zero seem to be better  
 296 identified after the correction.

## 297 6 Conclusion

298 A common approach to estimate strength of the relationship between two  
 299 variables is to estimate the part of shared variation. This single value  
 300 ranging from zero to one is easy to interpret. Such value can also be  
 301 computed between two sets of variable, but the estimation is more than  
 302 for simple vector data subject to over estimation because the over-fitting  
 303 phenomena which is amplified for high dimensional data. With IRLs and  
 304 its squared value, we propose an easy to compute correlation and determi-  
 305 nation coefficient far less biased than the original Procrustean correlation  
 306 coefficient. Every needed function to estimate the proposed modified ver-  
 307 sion of these coefficients are included in a R package ProcMod available

308 for download from the Comprehensive R Archive Network (CRAN).

## 309 Acknowledgements

310 Text Text Text Text Text Text Text Text. Text Text Text Text Text Text  
311 Text Text. Text Text Text Text Text Text Text Text.

## 312 Funding

313 This work has been supported by the... Text Text Text Text.

## 314 References

315 Bravais, A. (1844). *Analyse mathématique sur les probabilités des erreurs*  
316 *de situation d'un point*. Impr. Royale.

317 Csörgő, S. and Faraway, J. J. (1996). The exact and asymptotic distribu-  
318 tions of Cramér-Von mises statistics.

319 Dixon, P. (2003). VEGAN, a package of R functions for community ecol-  
320 ogy. *J. Veg. Sci.*, **14**(6), 927–930.

321 Dray, S. and Dufour, A.-B. (2007). The ade4 package: Implementing the  
322 duality diagram for ecologists. *Journal of Statistical Software, Articles*,  
323 **22**(4), 1–20.

324 El Ghaziri, A. and Qannari, E. M. (2015). Measures of association between  
325 two datasets; application to sensory data. *Food Qual. Prefer.*, **40**, 116–  
326 124.

327 Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*,  
328 pages 751–760.

329 Gower, J. C. (1971). Statistical methods of comparing different multivari-  
330 ate analyses of the same data. *Mathematics in the archaeological and*  
331 *historical sciences*, pages 138–149.

332 Jackson, D. A. (1995). PROTEST: A PROcrustean randomization TEST  
333 of community environment concordance. *Écoscience*, **2**(3), 297–303.

334 Lingoes, J. C. and Schönemann, P. H. (1974). Alternative measures of  
335 fit for the schönemann-carroll matrix fitting algorithm. *Psychometrika*,  
336 **39**(4), 423–427.

337 Mayer, C.-D., Lorent, J., and Horgan, G. W. (2011). Exploratory analysis  
338 of multiple omics datasets using the adjusted RV coefficient. *Stat. Appl.*  
339 *Genet. Mol. Biol.*, **10**, Article 14.

340 Peres-Neto, P. R. and Jackson, D. A. (2001). How well do multivari-  
341 ate data sets match? the advantages of a procrustean superimposition  
342 approach over the mantel test. *Oecologia*, **129**(2), 169–178.

343 Ramsay, J. O., ten Berge, J., and Styan, G. P. H. (1984). Matrix correla-  
344 tion. *Psychometrika*, **49**(3), 403–423.

345 Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate  
346 statistical methods: The RV- coefficient. *J. R. Stat. Soc. Ser. C Appl.*  
347 *Stat.*, **25**(3), 257–265.

348 Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., and van  
349 Erk, M. J. (2009). Matrix correlations for high-dimensional data: the  
350 modified RV-coefficient. *Bioinformatics*, **25**(3), 401–405.

351 Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of  
352 independence in high dimension. *J. Multivar. Anal.*, **117**, 193–213.

353 Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and  
354 testing dependence by correlation of distances. *Ann. Stat.*, **35**(6), 2769–  
355 2794.

356 Theil, H., Cramer, J. S., Moerman, H., and Russchen, A. (1958). *Eco-*  
357 *nomic forecasts and policy*. North-Holland Publishing Company, Am-  
358 sterdam.

## 359 Appendix

### 360 A Notations

$\mathbf{x}$ (vector)	bold lowercase.
$\mathbf{X}$ (matrix)	bold uppercase.
$i = 1, \dots, n$	object index.
$j = 1, \dots, p$	variable index.
$k$	iteration index.
$\mathbf{X}'$	The transpose of $\mathbf{X}$ .
$\mathbf{XY}$	Matrix multiplication of $\mathbf{X}$ and $\mathbf{Y}$ .
$\text{Diag}(\mathbf{X})$	A column matrix composed of the diagonal elements of $\mathbf{X}$ .
$\text{Trace}(\mathbf{X})$	The trace of $\mathbf{X}$ .