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

E. Coissac [1,*], Co-Author [2] and C. Gonindard-Melodelima [1]

[1]Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, F-38000, France
[2]Department, Institution, City, Post Code, Country.

Abstract

Motivation: Molecular biology and ecology produce 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. Unfortunatly when using classical measures, 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 coeficient 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 coeficients 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.

For a long time, several coefficients have been proposed to measure correlation between two matrices (for a comprehensive review see Ramsay *et al.*, 1984). But when applied to high-dimensional data, they suffer from the over-fitting effect leading them to estimate a high correlation even for unrelated data sets. Modified versions of some of these matrix correlation coefficients have been already proposed to tackle this problem. The RV_2 coefficient (Smilde *et al.*, 2009) is correcting the original RV coefficient (Escoufier, 1973) for over-fitting. Similarly, a modified version of the distance correlation coefficient dCor (Székely *et al.*, 2007) has been proposed by SzéKely and Rizzo (2013). dCor has the advantage over the other correlation factors for not considering only linear relationships. Here we will focus on the Procrustes correlation coefficient RLs proposed by Lingoes and Schönemann (1974) and by Gower (1971). Define $Trace$, a function summing the diagonal elements of a matrix. For an $n \times p$ real matrix $\mathbf{X}$ and a second $n \times q$ real matrix $\mathbf{Y}$ defining respectively two sets of p and q centered variables caracterizing n individuals, we define $\mathrm{CovLs}(\mathbf{X},\mathbf{Y})$ an analog of covariance applicable to vectorial data following Equation (1)

$$\mathrm{CovLs}(\mathbf{X},\mathbf{Y}) = \frac{\mathrm{Trace}((\mathbf{XX'YY'})^{1/2})}{n-1} \tag{1}$$

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

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

Among the advantages of RLs, its similarity with Pearson's correlation coefficient R (Bravais, 1844) has to be noticed. Considering $\mathrm{CovLs}(\mathbf{X},\mathbf{Y})$ and $\mathrm{VarLs}(\mathbf{X})$, respectively corresponding to the covariance of two matrices and the variance of a matrix, Equation (2) highlight the analogy between both the correlation coefficients. Besides, when $p = 1$ and $q = 1$, $\mathrm{RLs} = |\mathrm{R}|$. When squared RLs is an estimate, like the squared Pearson's R, of the amount of variation shared between the two datasets. This property allows for developing variance analyzing of matrix data sets.

Moreover, Procrustean analyses have been proposed as a good alternative to Mantel's statistics for analyzing ecological data summarized by distance matrices (Peres-Neto and Jackson, 2001). In such analyze distance matrices are projected into an orthogonal space using metric or non metric multidimensional scaling according to the geometrical properties of the used distances. Correlations are then estimated between these projections.

2 Approach

RLs is part of the procruste framework that aims to superimpose a set of points with respect to another through three operations: a translation, a rotation and a scaling. The optimal transfert matrix $Rot_{X\to Y}$ can be estimated from the singular value decomposition (SVD) of the covariance matrix $X'Y$. SVD factorize any matrix as the product of three matrices

(Equation 3).

$$\mathbf{X}'\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{3}$$

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

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

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

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

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

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

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

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

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

3 Methods

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

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

$$\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})} = \frac{\Sigma_{i=1}^{k} \mathrm{CovLs}(\mathbf{RX}_i, \mathbf{RY}_i)}{k} \qquad (8)$$

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

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

For two random vectors $\mathbf{x}$ and $\mathbf{y}$ of length n, the average coefficient of determination is $\overline{\mathrm{R}^2} = 1/(n-1)$. This value is independent of the distribution of the $\mathbf{x}$ and $\mathbf{y}$ values, but what about the independence of $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$ to the distributions of $\mathbf{X}$ and $\mathbf{Y}$. To test this independance and to assess the reasonnable randomization effort needed to estimate $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$, 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 $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$ and the respective standard deviation $\sigma(\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})})$. 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 IR are used to estimate correlation between both vectors.

3.3 Empirical assessment of the coefficient of determination

The coefficient of determination (R^2) represente the part of shared variation between two variables. RLs^2 keeps the same meaning when applied to two matrices. But because of over-fitting RLs and therefore RLs^2 are over-estimated.

Between two matrices

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

Between two vectors

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

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

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

Partial determination coefficients

To evaluate the capacity of partial determination coefficient IRLs$^2_{partial}$ to distangle nested correlations, four matrices **A**, **B**, **C**, **D** of size $n \times p = 20 \times 200$ are generated according to the schema: **A** shares 80% of variation with **B**, that shares 40% of variation with **C**, sharing 20% of variation with **D**. These direct correlations induce indirect ones spreading the total variation among each pair of matricies according to Figure 1. The simulation is repeadted 100 times, for every simutation IRLs$^2_{partial}$ and RLs$^2_{partial}$ are estimated for each pair of matrices.

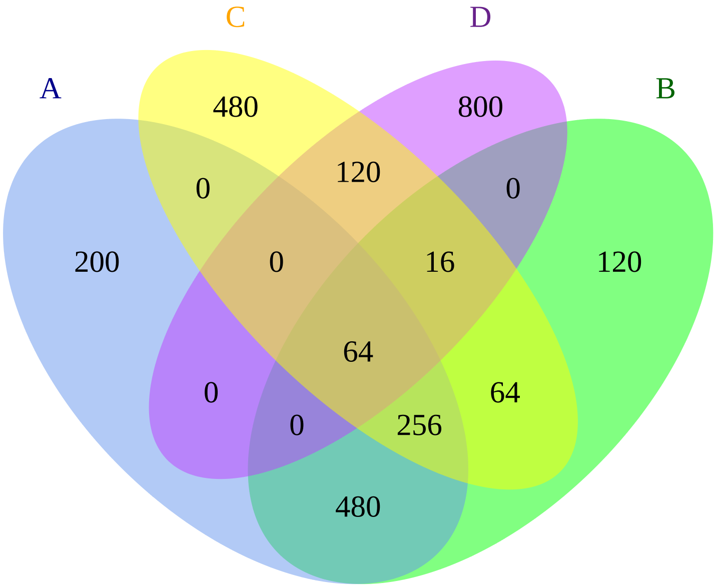

Figure 1: Theoritical distribution of the shared variation between the four matrices **A**, **B**, **C**, **D**, expressed in permille.

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

The significance of $\mathrm{IRLs}(\mathbf{X},\mathbf{Y})$ 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 $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$ to test that $\mathrm{ICovLs}(\mathbf{X},\mathbf{Y})$ and therefore $\mathrm{IRLs}(\mathbf{X},\mathbf{Y})$ are greater than expected under random hypothesis. Let us count over the k randomization when $\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})_k$ greater than $\mathrm{CovLs}(\mathbf{X},\mathbf{Y})$ and name this counts $N_{>\mathrm{CovLs}}$. The $\mathrm{P}_{\mathrm{value}}$ of the test can be estimated following Equation (10).

$$\mathrm{P}_{\mathrm{value}} = \frac{N_{>\mathrm{CovLs}}}{k} \tag{10}$$

Empirical assessment of α-risk for the CovLs test

To assess empirically the α-risk of the procruste test based on the randomisations realized during the estimation of $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$, the distribution of P_{value} under H_0 is compared to a uniform distribution between 0 and 1 ($\mathcal{U}(0,1)$). To estimate such an 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 independence. Procruste correlation between those 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_{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 `goftest`.

Empirical power assessment for the *CovLs* test

To evaluate relative the power of the three considered tests, pairs of to random matrices were produced for various $p \in \{10, 20, 50, 100\}$, $n \in \{10, 15, 20, 25\}$ and two levels of shared variations $\mathrm{R}^2 \in \{0.05, 0.1\}$. 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 $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$

Two main parameters can influence the Monte Carlo estimation of $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$: 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 $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$ according to the number of random matrices (k) aligned.

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

with a high probability for small values and a long tail of large ones. Despite the use of these contrasted distributions, estimates of $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$ and of $\sigma(\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})})$ are identical if we assume the normal distribution of the $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$ estimator and a 0.95 confidence interval of $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})} \pm 2\,\sigma(\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})})$ (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 overestimated. The proposed corrected vection (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 evalution is observable for the low values of simulated shared variation.

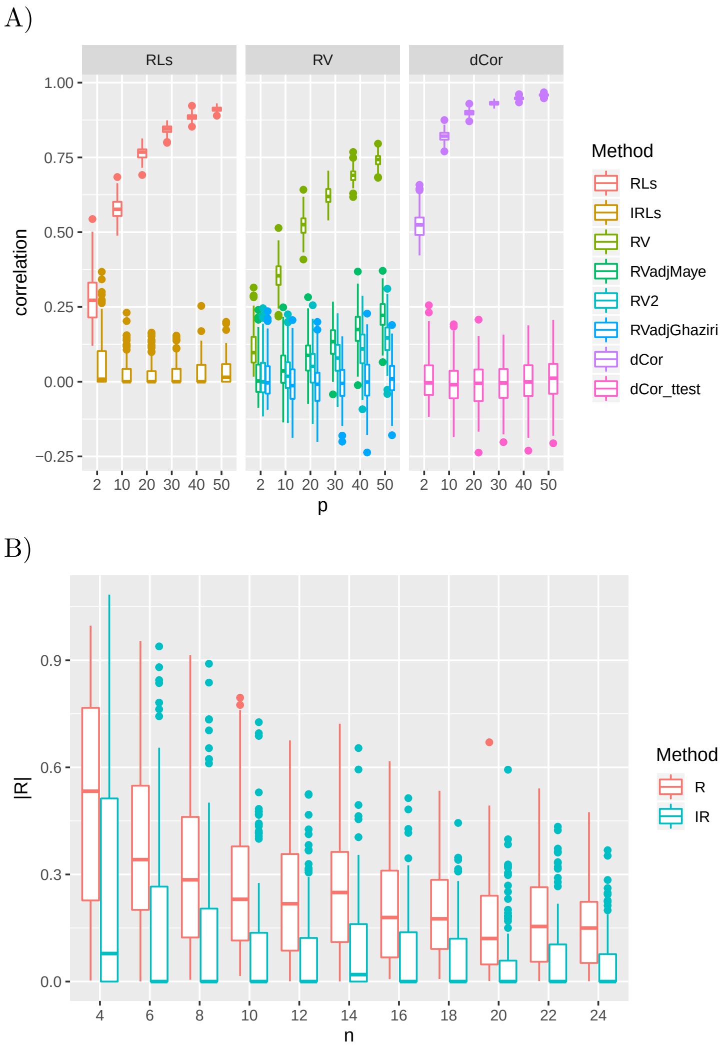

Figure 2: A) Sensibility to overfitting for various correraltion coefficients. (A) Both simulated data sets are matrices of size $(n \times p)$ with $p > 1$. B) Correlated data sets are vectors $(p = 1)$ with a various number of individuals n (vector length). A & B) 100 simulations are run for each combination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

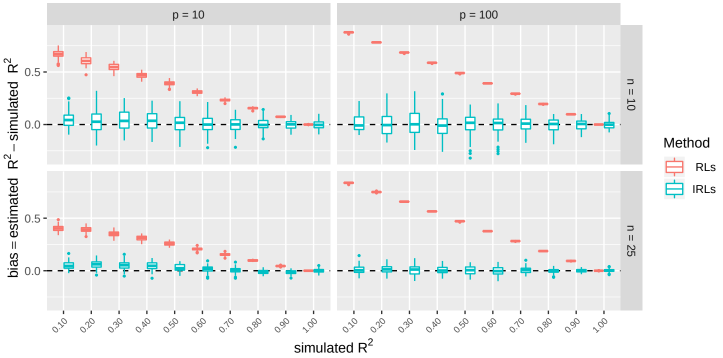

Figure 3: Shared variation (R^2) between two matrices is mesured using both the corrected (IRLs) and the original (RLs) versions of the procrustean correlation coefficient. A gradiant of R^2 is simulated for two population sizes ($n \in \{10, 25\}$) and two numbers of descriptive variables ($p \in \{10,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

Between two vectors

Vectors can be considered as a single column matrix, and the efficiency of IRLs^2 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^2 performs better than the simple R^2 and its modified version R^2_{adj} 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 perfom for large values of shared variation.

Partial coefficient of determination

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

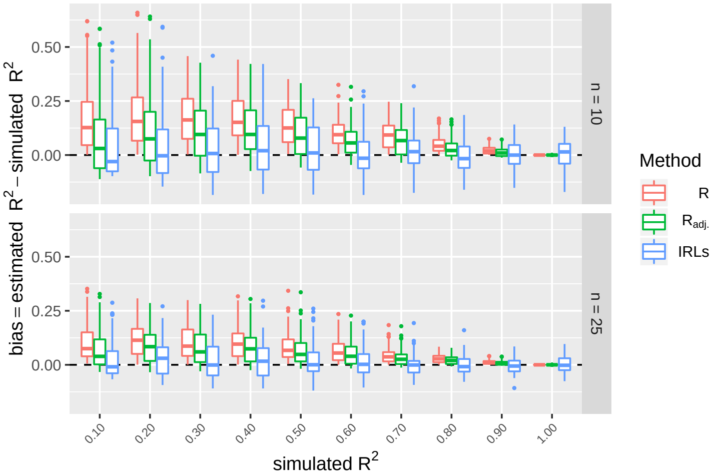

Figure 4: Shared variation between two vectors is mesured using the classical R^2, its ajusted version R^2_{adj} and (IRLs^2). A gradiant of shared variation is simulated for two vector sizes ($n \in \{10, 25\}$). The black dashed line corresponds to a perfect match where measured R^2 equals the simulated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

	Cramer-Von Mises p.value		
p	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

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

4.4 p_{value} distribution under null hyothesis

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

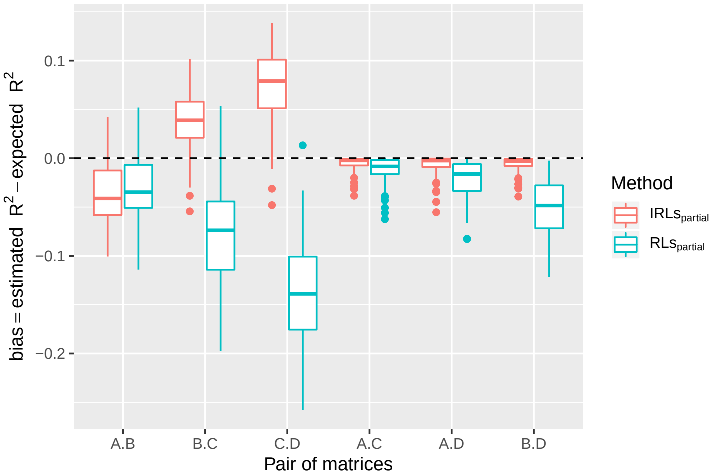

Figure 5: Estimation error on the partial determination coefficient. Error is defined as the absolute value of the difference between the expected and the estimated partial R^2 using the corrected IRLs$_{partial}$ and not corrected RLs$_{partial}$ procruste correlation coefficient.

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.

5 Discussion

Correcting the over-adjustment effect on metrics assessing the relationship between high dimension datasets has been a constant effort over the past decades. 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

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* $= 1 - \beta$-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
procuste.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

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

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

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

6 Conclusion

A common approach to estimate strengh of the relationship between two variables is to estimate the part of shared variation. This single value ranging from zero to one is easy to interpret. Such value can also be computed between two sets of variable, but the estimation is more than for simple vector data subject to over estimation because the over-fitting phenomena which is amplified for high dimensional data. With IRLs and its squared value, we propose an easy to compute correlation and determination coefficient far less biased than the original Procrustean correlation coefficient. Every needed function to estimate the proposed modified version of these coefficients are included in a R package ProcMod available for download from the Comprehensive R Archive Network (CRAN).

Acknowledgements

Text Text Text Text Text Text Text Text. Text Text Text Text Text Text Text Text. Text Text Text Text Text Text Text Text.

Funding

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

References

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

Csörgő, S. and Faraway, J. J. (1996). The exact and asymptotic distributions of Cramér-Von mises statistics.

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

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

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

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

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

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

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

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

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

Ramsay, J. O., ten Berge, J., and Styan, G. P. H. (1984). Matrix correlation. *Psychometrika*, **49**(3), 403–423.

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

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

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

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

Theil, H., Cramer, J. S., Moerman, H., and Russchen, A. (1958). *Economic forecasts and policy*. North-Holland Publishing Company, Amsterdam.

Appendix

A Notations

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