Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Data and text mining

A modified version of the Procruste correlation coefficient for high-dimensional data

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

[1] Department, Institution, City, Post Code, Country and
[2] Department, Institution, City, Post Code, Country.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.
Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Contact: eric.coissac@metabarcoding.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 one describing a different aspect 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 **?**). 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 (**?**) is correcting the original RV coefficient (**?**) for over-fitting. Similarly, a modified version of the distance correlation coefficient dCor (**?**) has been proposed by **?**. dCor has the advantage over the other correlation factors for not considering only linear relationships. Here we will focus on the Procrustes correlation coeficient RLs proposed by **?** and by **?**. Let the define $Trace$, a function summing the diagonal elements of a matrix.

For a $n \times p$ real matrix $\mathbf{X}$ and anothet a $n \times q$ real matrix $\mathbf{Y}$ defining respectively two sets of p and q centered variables caracterizing n individuals, we define $\mathrm{CovLs}(\mathbf{X},\mathbf{X})$ 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).

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

Procrustean analyses have been proposed as a good alternative to Mantel's statistics for analyzing ecological data, and more generally for every high-dimensional data sets (**?**). Among the advantages of Rls, its similarity with the Pearson correlation coefficient R (**?**) 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. Moreover, when $p = 1$ and $q = 1$, $\mathrm{RLs} = |\mathrm{R}|$.

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 reciproquly $\mathbf{Y}$ 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. $\mathrm{CovLs}(\mathbf{X},\mathbf{Y})$ can also be computed from singular values (Equation 4).

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

This expression illustrates that actually $\mathrm{CovLs}(\mathbf{X},\mathbf{Y})$ is the variance of the projections of $\mathbf{X}$ on $\mathbf{Y}$ or of the reciproque projection. Therefore $\mathrm{CovLs}(\mathbf{X},\mathbf{Y})$ and $\mathrm{RLs}(\mathbf{X},\mathbf{Y})$ 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 $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$ the mean of such random correlation. $\mathrm{ICovLs}(\mathbf{X},\mathbf{Y})$ the informative part of $\mathrm{CovLs}(\mathbf{X},\mathbf{Y})$ is computed using Equation (5).

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

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

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

Like $\mathrm{RLs}(\mathbf{X},\mathbf{Y})$ $\mathrm{IRLs}(\mathbf{X},\mathbf{Y}) \in [0;1]$ with the 0 value corresponding to not correlation and the maximum value 1 reached for two strictly homothetic data sets.

The corollary of $\mathrm{ICovLs}(\mathbf{X},\mathbf{Y})$ and $\mathrm{IVarLs}(\mathbf{X})$ definitions is that $\mathrm{ICovLs}(\mathbf{X},\mathbf{Y}) \geqslant 0$ and $\mathrm{IVarLs}(\mathbf{X}) > 0$. 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 $\mathbf{C}_{i,j} = \mathrm{ICovLs}(\mathbf{M}_i, \mathbf{M}_j)$ for is definite positive and symmetrical. This allows for defining the precision matrix $\mathbf{P} = \mathbf{C}^{-1}$ and the related partial correlation coeficent matrix $\mathrm{IRLs}_{partial}\ using Equation\ (7)$

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

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 $RX = \{\mathbf{RX}_1, \mathbf{RX}_2, ..., RX_k\}$ 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 standard deviation of these columns.

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

Even when $\mathbf{X} = \mathbf{Y}$ to estimate $\mathrm{VarLs}(\mathbf{X})$, $\overline{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$ 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 $\sigma_{\mathrm{RCovLs}(\mathbf{X},\mathbf{Y})}$.

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 (**?**), RVadjMaye, RV 2 and $adjGhaziri$ three modified versions of RV (**???**), dCor the original distance correlation coefficient (**?**) and *dCor_ttest* a modified version of dCor not sensible to overfitting (**?**). 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 two 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 0.05 and 1 by step of 0.05. 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

partial determination coefficients

To evaluate capacity of partial determination coefficient $\mathrm{IRLs}^2_{partial}$ to distangle nested correlations, four matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ of size $n \times p = 20 \times 200$ are generated according to the schema: $\mathbf{A}$ shares 80% of variation with $\mathbf{B}$, that shares 40% of variation with $\mathbf{C}$, sharing 20% of variation with $\mathbf{D}$. 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 $\mathrm{IRLs}^2_{partial}$ and $\mathrm{RLs}^2_{partial}$ are estimated for each pair of matrices.

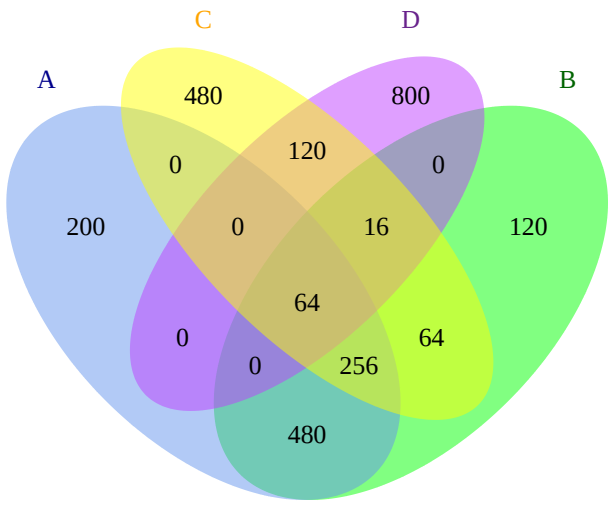

Fig. 1. Theoritical distribution of the shared variation between the four matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, expressed in permille.

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

Significance of $\mathrm{IRLs}(\mathbf{X}, \mathbf{Y})$ can be tested using permutation test as defined in **?** or **?** and implemented respectively in the `protest` method of the vegan R package (**?**) or the `procuste.rtest` method of the ADE4 R package **?**.

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 counting over the k randomization when $\mathrm{RCovLs}(\mathbf{X}, \mathbf{Y})_k$ greater than $\mathrm{CovLs}(\mathbf{X}, \mathbf{Y})$ name this counts $N_{>\mathrm{CovLs}}$. $\mathrm{P}_{\mathrm{value}}$ of the test can be estimated following Equation (9).

$$\mathrm{P}_{\mathrm{value}} = \frac{N_{>\mathrm{CovLs}}}{k} \tag{9}$$

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 $\overline{\mathrm{RCovLs}(\mathbf{X}, \mathbf{Y})}$, distribution of P_{value} under the H_0 is comparaed 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_{value} to $\mathcal{U}(0, 1)$ is assessed using the Cramer-Von Mises test (**?**) implemented in the `cvm.test` function of the R package `goftest`.

Empirical power assessment for the $CovLs$ test

To evaluate relative 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`.

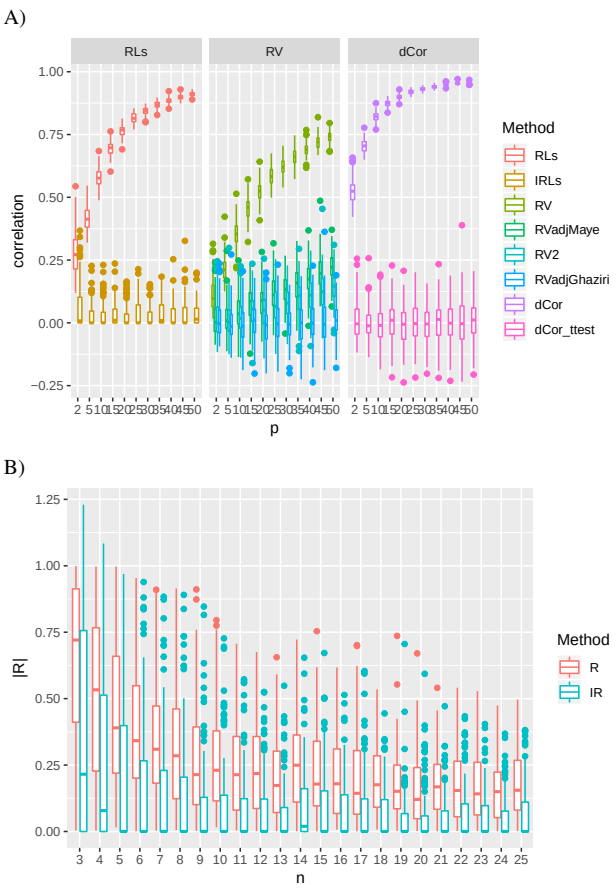

Fig. 2. A) Sensibility to overfitting for various correraltion coefficients. (A) Both simulated data sets are matrices of size $(n \times p)$ with $p > 1$. B) Correlated data sets are vectors $(p = 1)$ with a various number of individuals n (vector length). A & B) 100 simulations are run for each combination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

4 Results

4.1 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).

4.2 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 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.

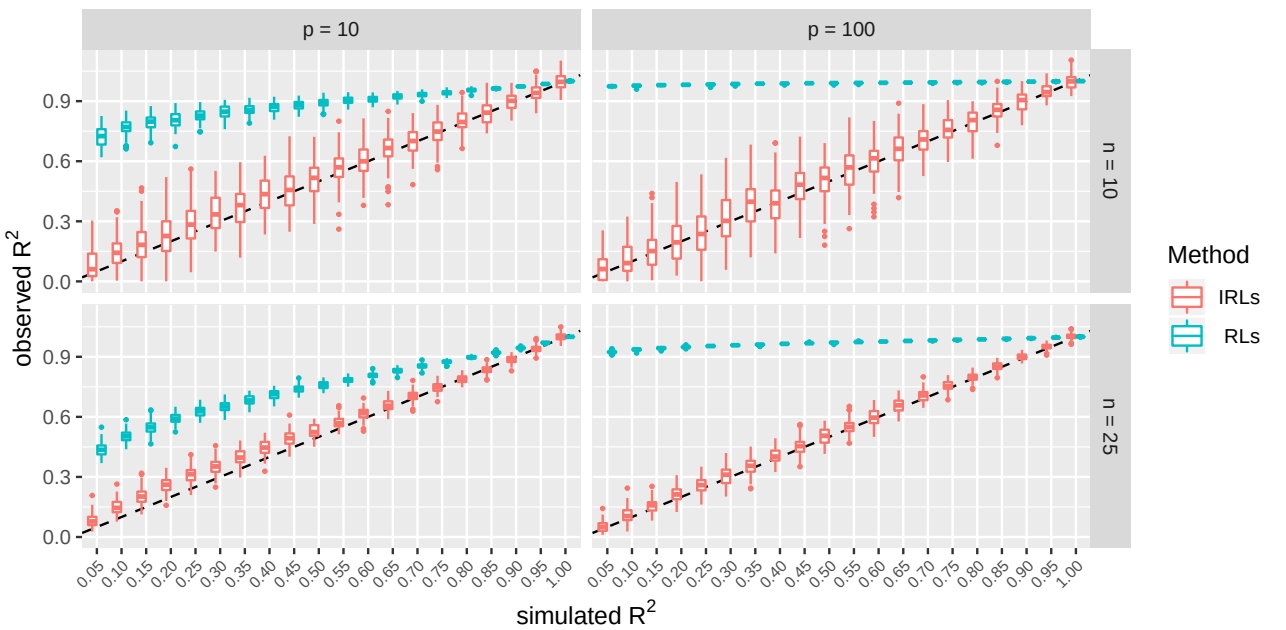

Fig. 3. Shared variation (R^2) between two matrices is mesured with both the corrected (IRLs) and the original (RLs) versions of the procrustean correlation coefficient. A gradiant of R^2 is simulated for two population sizes ($n \in \{10, 24\}$) and two numbers of descriptive variables ($p \in \{10, 100\}$). The black dashed line corresponds to a perfect match where measured R^2 equals the simulated one.

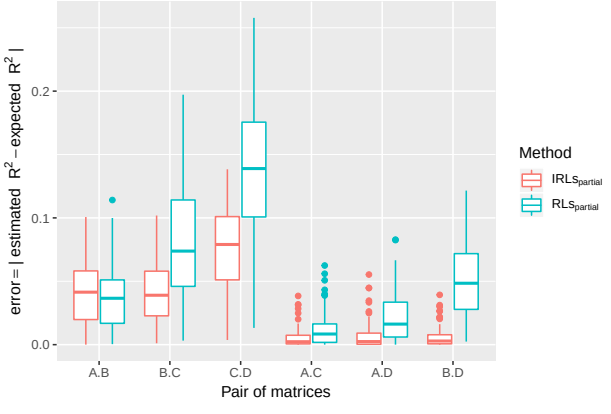

Fig. 4. Estimation error on the partial determination coefficient. Error is defined as the absolute value of the difference between the expected and the estimated partial R^2 using the corrected $\mathrm{IRLs}_{partial}$ and not corrected $\mathrm{RLs}_{partial}$ procruste correlation coefficient.

between two vectors

partial determination coefficients

The simulated correlation network between the four matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ 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 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 $\mathbf{A}$ and $\mathbf{B}$ is $480/(200 + 480) = 0.706$; between $\mathbf{B}$ and $\mathbf{C}$, $64/(480 + 120) = 0.107$; and between $\mathbf{C}$ and $\mathbf{D}$ $120/800 = 0.150$. All other partial coeficient are expected to be equal to 0.

Table 1. 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

4.3 p_{value} distribution under null hyothesis

As expected, P_{values} of the $CovLs$ test based on the estimation of $\overline{RCovLs(X, Y)}$ are uniformely distributed under H_0. whatever the p tested (Table 1). 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.

4.4 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 2). 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 $iscomputed$.

Table 2. 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

5 Discussion

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

6 Conclusion

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

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.

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 $\mathbf{Y}$.
$\mathbf{X} \circ \mathbf{Y}$	Hadamard product of $\mathbf{X}$ and $\mathbf{Y}$.
$\mathbf{X}^{\circ y}$	Hadamard power of $\mathbf{X}$.
$\mathrm{Diag}(\mathbf{X})$	A column matrix composed of the diagonal elements of $\mathbf{X}$.
$\mathrm{Trace}(\mathbf{X}$	The trace of $\mathbf{X}$.