

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

Eric Coissac^{a,*}, Christelle Gonindard-Melodelima^a

^a*Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, F-38000, France*

Motivation: Ecology studies can produce high dimension data. Estimating correlations and shared variation between such data sets are an important step in disentangling the relationships between different elements of a biological system. Unfortunately, if classical approaches are able to detect correlation estimations of shared variation between such dataset is always overestimated.

Results: Here we propose a corrected version of the Procrustean correlation coefficient that is robust to high dimensional data. This allows for a correct estimation of the shared variation between two data sets.

Availability: The proposed corrected coefficient is implemented in the ProcMod R package available on CRAN. The git repository is hosted at <https://git.metabarcoding.org/lecasofts/ProcMod>

Keywords: High dimension data, Correlation, Shared variation, Procruste

2000 MSC: 62H20

1. Introduction

Ecologists are accustomed to manipulating multidimensional datasets describing the environment in multiple terms: climate, chemistry, species diversity, genetic variation. Therefore, they use appropriate statistics, such as ordination, to visualize this type of data and

*Corresponding author

Email addresses: eric.coissac@metabarcoding.org (Eric Coissac),
christelle.gonindard@univ-grenoble-alpes.fr (Christelle Gonindard-Melodelima)

5 related tests to assess the significance of the observed patterns. To test the correlation be-
6 tween two multidimensional datasets, Mantel-test and Procrustes Multidimensional data and
7 even high-dimensional data, where the number of variables describing each sample is far
8 larger than the sample count, is now routinely produced in functional genomics (*e.g.* tran-
9 scriptomics, proteomics or metabolomics) and molecular ecology (*e.g.* DNA metabarcoding
10 and metagenomics). Using a range of techniques, the same sample set can be described by
11 several multidimensional data sets, each of them describing a different facet of the samples.
12 This enables data analysis methods to evaluate mutual information shared by these different
13 descriptions.

14 Correlative approaches are one of the simplest approaches to decipher pairwise relation-
15 ships between multiple datasets. For a long time, several coefficients have been proposed to
16 measure correlations between two matrices (for a comprehensive review see [Ramsay et al.,](#)
17 [1984](#)). However, when applied to high-dimensional data, these approaches suffer from over-
18 fitting, resulting in high estimated correlations even for unrelated data sets. The creation of
19 incorrect correlations from over-fitting consequently affects the biological interpretation of
20 the analysis ([Chariton et al., 2010](#)) can have downstream effects on the biological interpre-
21 tation of a study. A number of modified matrix correlation coefficients have been proposed
22 to address this issue. For example, the RV2 coefficient ([Smilde et al., 2009](#)) corrects for
23 overfitting of the original RV coefficient ([Escoufier, 1973](#)). Similarly, a modified version of
24 the distance correlation coefficient dCor ([Székely et al., 2007](#)) proposed by [Székely and Rizzo](#)
25 [\(2013\)](#) dCor has the advantage over the other correlation factors by considering by not being
26 restricted to linear relationships.

27 Here we focus on the Procrustes correlation coefficient (RLs) proposed by [Lingoes and](#)
28 [Schönemann \(1974\)](#) and by [Gower \(1971\)](#). Define *Trace*, a function summing the diagonal
29 elements of a matrix. For an $n \times p$ real matrix \mathbf{X} and a second $n \times q$ real matrix \mathbf{Y}
30 defining respectively two sets of p and q centered variables characterizing n individuals, we

31 define $\text{CovLs}(\mathbf{X}, \mathbf{Y})$ following Equation (1)

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

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

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

33 Considering $\text{CovLs}(\mathbf{X}, \mathbf{Y})$ and $\text{VarLs}(\mathbf{X})$, respectively corresponding to the covariance of
34 two matrices and the variance of a matrix, Equation (2) highlighting the analogy between
35 RLs and Pearson's correlation coefficient (R) (Bravais, 1844). When $p = 1$ and $q = 1$, RLs =
36 $|\text{R}|$. Like the squared Pearson's R, the squared RLs is an estimate of the amount of variation
37 shared between the two datasets.

38 Procrustean analyses have been proposed as a good alternative to Mantel's statistics
39 for analyzing ecological data summarized by distance matrices (Peres-Neto and Jackson,
40 2001). In Procrustean analyze, distance matrices are projected into an orthogonal space
41 using metric or non metric multidimensional scaling according to the geometrical properties
42 of the used distances. Correlations can then be estimated on these projections.

43 2. Approach

44 RLs is part of the Procrustes framework that aims to superimpose a set of points with
45 respect to another through three operations: a translation, a rotation, and a scaling. The
46 optimal transfer matrix $\text{Rot}_{X \rightarrow Y}$ can be estimated from the singular value decomposition
47 (SVD) of the covariance matrix $X'Y$. SVDs factorize any matrix as the product of three
48 matrices (Equation 3).

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

49 \mathbf{U} and \mathbf{V} are two rotation matrices which computes the transfer matrix to superimpose
50 \mathbf{X} on \mathbf{Y} or reciprocally \mathbf{Y} on \mathbf{X} . All the elements of $\mathbf{\Sigma}$ except its diagonal are equal to
51 zero. The diagonal elements are the singular values. Singular values are the extension of

52 the eigenvalues to non-square matrices. $\text{CovLs}(\mathbf{X}, \mathbf{Y})$ can also be computed from singular
 53 values (Equation [4](#)).

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

54 This expression actually illustrates that $\text{CovLs}(\mathbf{X}, \mathbf{Y})$ is the variance of the projections of
 55 \mathbf{X} on \mathbf{Y} or of the reciprocal projections. Therefore $\text{CovLs}(\mathbf{X}, \mathbf{Y})$ and $\text{RLs}(\mathbf{X}, \mathbf{Y})$ are always
 56 positive and rotation independent. Here we propose to partitionate $\text{Trace}(\boldsymbol{\Sigma})$, the variation
 57 amount corresponding to $\text{CovLs}(\mathbf{X}, \mathbf{Y})$, in two components. The first corresponds to the
 58 actual shared information between \mathbf{X} and \mathbf{Y} . The second, corresponds to the over-fitting
 59 effect. It that can be estimated as the average variation shared by two random matrices of
 60 same structure as \mathbf{X} and \mathbf{Y} noted $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$. $\text{ICovLs}(\mathbf{X}, \mathbf{Y})$, the informative part of
 61 $\text{CovLs}(\mathbf{X}, \mathbf{Y})$, is computed using Equation [5](#).

$$\text{ICovLs}(\mathbf{X}, \mathbf{Y}) = \text{CovLs}(\mathbf{X}, \mathbf{Y}) - \overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})} \quad (5)$$

62 Similarly the informative counter-part of $\text{VarLs}(\mathbf{X})$ is defined as $\text{IVarLs}(\mathbf{X}) =$
 63 $\text{ICovLs}(\mathbf{X}, \mathbf{X})$, and $\text{IRLs}(\mathbf{X}, \mathbf{Y})$ the informative Procrustes correlation coefficient as defined
 64 in Equation [6](#).

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

65 As in the case of $\text{RLs}(\mathbf{X}, \mathbf{Y})$, $\text{IRLs}(\mathbf{X}, \mathbf{Y}) \in [0; 1]$, the 0 value corresponding to no
 66 correlation and the maximum value 1 reflects two strictly homothetic data sets.

67 3. Methods

68 3.1. Reduction of data dimensions

69 The numerical algorithms used to compute IRLs , (*i.e.* SVD decomposition, generation of
 70 multivariate normal random numbers) present complexities that may hinder their application
 71 to high dimensional matrices. To solve this problem, the dimensions of \mathbf{X} , \mathbf{Y} matrices
 72 are first reduced using principal component analysis. Such a transformation loses no data

73 variation, but reduces the size of a $n \times p$ matrix to $n \times \min(n, p)$. The use of a multi-scale
74 method to reduce the size of the original data can also be applied to allow the use of distance
75 matrices as input data. In this case, the coordinates of individuals are estimated from their
76 distances using or principal coordinate analyze for metrics or a non-metric multi-dimensional
77 scaling for other distances. Here by \mathbf{X} , \mathbf{Y} will design multidimensional scaling projections
78 of the original data.

79 3.2. Monte-Carlo estimation of $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$

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

$$\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})} = \frac{\sum_{i=1}^k \text{CovLs}(\mathbf{RX}_i, \mathbf{RY}_i)}{k} \quad (7)$$

84 To estimate $\text{IVarLs}(\mathbf{X})$, which is equal to $\text{ICovLs}(\mathbf{X}, \mathbf{X})$, $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{X})}$ is estimated
85 with two independent sets of random matrix RX and RY , both having the same structure
86 than \mathbf{X} .

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

88 For two random vectors \mathbf{x} and \mathbf{y} of length n , the average coefficient of determination is
89 $\overline{R^2} = 1/(n - 1)$. This value is independent of the distribution of the \mathbf{x} and \mathbf{y} values, but
90 what about the independence of $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$ to the distributions of \mathbf{X} and \mathbf{Y} ? To test
91 this independence and to assess the reasonable randomization effort needed to estimate
92 $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$, this value is estimated for four matrices \mathbf{K} , \mathbf{L} , \mathbf{M} , \mathbf{N} of $n = 20$ rows and
93 10, 20, 50, 100 columns, respectively. Values of the four matrices are drawn from a normal
94 or an exponential distribution, with $k \in \{10, 100, 1000\}$ randomizations tested to estimate
95 $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$ and the respective standard deviation $\sigma(\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})})$. The VarLs of the
96 generated matrices is equal to 1, therefore the estimated CovLs are equals to the RLs .

97 3.3. Simulating data for testing sensibility to overfitting

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

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

111 3.4. Empirical assessment of the coefficient of determination

112 As in the case of the coefficient of determination (R^2), RLs^2 represents the part of shared
113 variation between two matrices. Because of over-fitting in high-dimension data, RLs and
114 therefore RLs^2 are over-estimated.

115 *Between two matrices*

116 To test how the IRLs version of the coefficient of determination $IRLs^2$ can perform to
117 evaluate the shared variation, pairs of random matrices were produced for two values of
118 $p \in \{10, 100\}$ and $n \in \{10, 25\}$, and for several levels of shared variations ranging between
119 0.1 and 1 using 0.1 increments. For each combination of parameters, $k = 100$ simulations
120 were run, and both RLs^2 and $IRLs^2$ were estimated using 100 randomizations. To generate
121 two random matrices \mathbf{A} , \mathbf{B} , sharing $w \in [0, 1]$ part of variation, two independent random
122 matrices \mathbf{A} and $\mathbf{\Delta}$ were generated such as $VarLs(\mathbf{A}) = 1$ and $VarLs(\mathbf{\Delta}) = 1$. The $\mathbf{\Delta}_{rot}$
123 matrix was computed as the alignment of $\mathbf{\Delta}$ on \mathbf{A} using the optimal Procrustes rotation.

124 Then \mathbf{B} is computed using equation 8.

$$\mathbf{B} = \mathbf{A} \times \sqrt{w} + \mathbf{\Delta}_{rot} \times \sqrt{1-w}. \quad (8)$$

125 *Between two vectors*

126 Coefficient of determination between two vectors also suffers from over-estimation when
127 n , the number of considered points, is small. On average, for two random vectors of size n ,
128 $\overline{R^2} = 1/(n-1)$. This random part of the shared variation inflates the observed shared vari-
129 ation even for non-random vectors. In the context of multiple linear regression, Theil (1958)
130 proposed an adjusted version of the coefficient of determination (Equation 9), correcting for
131 both the effect of the number of vectors (p) and the vector size (n).

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

132 To evaluate the strength of that over-estimation and the relative effect of the correction
133 proposed by Theil (1958) and by IRLs, pairs of random vectors were produced for $n \in$
134 $\{10, 25\}$, and for several levels of shared variations ranging between 0.1 and 1 using 0.1
135 increments. For each combination of parameters, $k = 100$ simulations were run, and R^2 ,
136 R_{adj}^2 and $IRLs^2$ were estimated using 100 randomizations.

137 3.5. Testing the significance of $IRLs(\mathbf{X}, \mathbf{Y})$

138 The significance of $IRLs(\mathbf{X}, \mathbf{Y})$ can be tested using permutation test as defined in Jackson
139 (1995) or Peres-Neto and Jackson (2001) and implemented respectively in the `protest`
140 method of the `vegan` R package (Dixon, 2003) or the `procuste.rtest` method of the `ADE4`
141 R package (Dray and Dufour (2007)).

142 It is also possible to take advantage of the Monte-Carlo estimation of $\overline{RCovLs(\mathbf{X}, \mathbf{Y})}$
143 to test that $ICovLs(\mathbf{X}, \mathbf{Y})$ and therefore $IRLs(\mathbf{X}, \mathbf{Y})$ are greater than expected under
144 random hypothesis. Over the k randomizations, $N_{>CovLs}$ is estimated by counting when
145 $RCovLs(\mathbf{X}, \mathbf{Y})_k > CovLs(\mathbf{X}, \mathbf{Y})$. The P_{value} of the test then can be estimated following

146 Equation (10).

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

147 *Empirical assessment of α -risk for the CovLs test*

148 To empirically assess the α -risk of the Procrustes test based on the randomisations
149 realized during the estimation of $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$, the distribution of P_{value} under H_0 was
150 compared to a uniform distribution between 0 and 1 ($\mathcal{U}(0, 1)$). To estimate the empirical
151 distribution, $k = 1000$ pairs of $n \times p$ random matrices with $n = 20$ and $p \in \{10, 20, 50\}$
152 were simulated under the null hypothesis of independence. Significance of the Procrustes
153 correlation between those matrices was tested using the three approaches: our proposed test
154 (*CovLs.test*) ; the `protest` method of the `vegan` R package ; the `procuste.rtest` method
155 of the `ADE4` R package. Conformance of the distribution of each set of k P_{value} to $\mathcal{U}(0, 1)$
156 was assessed using the Cramer-Von Mises test (Csörgő and Faraway, 1996) implemented in
157 the `cvm.test` function of the R package `goftest`.

158 *Empirical power assessment for the CovLs test*

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

164 4. Results

165 4.1. Empirical assessment of $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$

166 Two main parameters can influence the Monte Carlo estimation of $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$: the
167 distribution used to generate the random matrices, and k the number of random matrix
168 pair. Two very different distribution are tested to regenerate the random matrices, the
169 normal and the exponential distributions. The normal distribution is symmetric where the

p	k	normal		exponential	
		mean	sd	mean	sd
10	10	0.6048	4.3972×10^{-2}	0.5876	3.8187×10^{-2}
	100	0.5845	3.4226×10^{-2}	0.5795	3.6287×10^{-2}
	1000	0.5819	3.5359×10^{-2}	0.5803	3.6994×10^{-2}
20	10	0.7586	2.2819×10^{-2}	0.7596	2.3101×10^{-2}
	100	0.7683	2.1031×10^{-2}	0.7636	2.1718×10^{-2}
	1000	0.7657	2.0879×10^{-2}	0.7663	2.1731×10^{-2}
50	10	0.9090	1.0806×10^{-2}	0.9070	8.8522×10^{-3}
	100	0.9078	9.2631×10^{-3}	0.9086	9.4615×10^{-3}
	1000	0.9080	9.1205×10^{-3}	0.9084	9.4858×10^{-3}
100	10	0.9541	3.5242×10^{-3}	0.9532	6.9991×10^{-3}
	100	0.9550	4.3143×10^{-3}	0.9538	4.9438×10^{-3}
	1000	0.9548	4.6308×10^{-3}	0.9545	4.8369×10^{-3}

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

170 exponential is unsymmetrical with a high probability for small values and a long tail of large
171 ones. Despite the use of these contrasted distributions, estimates of $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$ and of
172 $\sigma(\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})})$ were identical if we assumed the normal distribution of the $\overline{\text{RCovLs}(\mathbf{X}, \mathbf{Y})}$
173 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).

174 4.2. Relative susceptibility of $IRLs(X, Y)$ to overfitting

175 RLs , like RV and $dCor$, is susceptible to overfitting which increases when n decreases,
176 and p or q increase. Because RV is more comparable to R^2 , when RLs and $dCor$ are
177 more comparable to R , RV values increase more slowly than RLs and $dCor$ values with p
178 (Figure 1A). As expected $IRLs$ values for non-correlated matrices are close to 0 regardless
179 of p (Figure 1A). The same correction of the overfitting effect can be observed for vectors
180 (Figure 1B)

181 4.3. Evaluating the shared variation

182 Between two matrices

183 For two matrices, our proposed corrected version of RLs^2 ($IRLs^2$) provided a good esti-
184 mate of the shared variation and was robust to the phenomenon of over-fitting (Figure 2).
185 Only a small over estimation was observed when $p = 10$ for the lowest values of the simulated
186 shared variations (≤ 0.4).

187 Between two vectors

188 Vectors can be considered as a single column matrix, and the efficiency of $IRLs^2$ to
189 estimate shared variation between matrices can also be used to estimate shared variation
190 between two vectors. Other formulas have been already proposed to better estimate shared
191 variation between vectors in the context of linear models. Among them the one presented in
192 Equation 9, is the most often used and is the one implemented in R linear model summary
193 function. On simulated data, $IRLs^2$ performs better than the simple R^2 and its modified
194 version R_{adj}^2 commonly used (Figure 3). Whatever the estimator the bias decrease with
195 the simulated shared variation. Nevertheless for every tested cases the median of the bias
196 observed is smaller than with both other estimators, even if classical estimators well perform
197 for large values of shared variation.

198 4.4. p_{value} distribution under null hypothesis

199 As expected, P_{values} of the $CovLs$ test based on the estimation of $\overline{RCovLs(X, Y)}$ are
200 uniformly distributed under H_0 . whatever the p tested (Table 2). This ensure that the
201 probability of a $P_{value} \leq \alpha$ -risk is equal to α -risk. Moreover P_{values} of the $CovLs$ test are
202 strongly linearly correlated with those of both the other tests ($R^2 = 0.996$ and $R^2 = 0.996$
203 respectively for the correlation with `vegan::protest` and `ade4::procuste.rtest` P_{values}).
204 The slopes of the corresponding linear models are respectively 0.997 and 0.996.

205 4.5. Power of the test based on randomisation

206 Power of the $CovLs$ test based on the estimation of $\overline{RCovLs(X, Y)}$ is equivalent of the
207 power estimated for both `vegan::protest` and `ade4::procuste.rtest` tests (Table 3). As

p	Cramer-Von Mises p.value		
	CovLs test	protest	procuste.rtest
10	0.132	0.126	0.111
20	0.455	0.510	0.474
50	0.814	0.764	0.833

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.

208 for the two other tests, power decreases when the number of variable (p or q) increases, and
 209 increase with the number of individuals and the shared variation. The advantage of the test
 210 based on the Monte-Carlo estimation of $\overline{RCovLs(X, Y)}$ is to remove the need of running a
 211 supplementary set of permutations when IRLs is computed.

212 5. Discussion

213 Correcting the over-adjustment effect on metrics assessing the relationship between high
 214 dimension datasets has been a constant effort over the past decades. Therefore, IRLs can
 215 be considered as a continuation of the extension of the toolbox available to biologists for
 216 analyzing their omics data. The effect of the proposed correction on the classical RLs
 217 coefficient is as strong as the other ones previously proposed for other correlation coefficients
 218 measuring relationship between vector data (see Figure 2, e.g. Smilde et al., 2009; Székely
 219 and Rizzo, 2013). When applied to univariate data, RLs is equal to the absolute value of the
 220 Pearson correlation coefficient, hence, and despite it is not the initial aim of that coefficient,
 221 IRLs can also be used to evaluate correlation between two univariate datasets. Using IRLs
 222 for such data sets is correcting for spurious correlations when the number of individual is
 223 small more efficiently than classical correction (see Figure 3, Theil, 1958).

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

	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	0.69	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

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

228 data matrices.

229 6. Conclusion

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

239 **Acknowledgements**

240 The authors would like to thank Dr Anthony Chariton & Dr Eric Marcon for their helpful
241 discussions and suggestions during the writing of the manuscript.

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

244 Chariton, A. A., Roach, A. C., Simpson, S. L., Batley, G. E., 2010. Influence of the choice of physical
245 and chemistry variables on interpreting patterns of sediment contaminants and their relationships with
246 estuarine macrobenthic communities. *Marine and Freshwater Research* 61 (10), 1109.

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

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

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

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

253 Escoufier, Y., 1973. Le traitement des variables vectorielles. *Biometrics*, 751–760.

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

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

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

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

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

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

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

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

269 Székely, G. J., Rizzo, M. L., May 2013. The distance correlation t-test of independence in high dimension.
270 *J. Multivar. Anal.* 117, 193–213.

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

273 Theil, H., 1958. Economic forecasts and policy., North-Holland Pub. Co., Amsterdam.

274 **Appendix**

275 *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} .

276 \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} .

$\mathbf{X}^{1/2}$ Matrix square root of \mathbf{X} .

$\text{Trace}(\mathbf{X})$ The trace of \mathbf{X} .

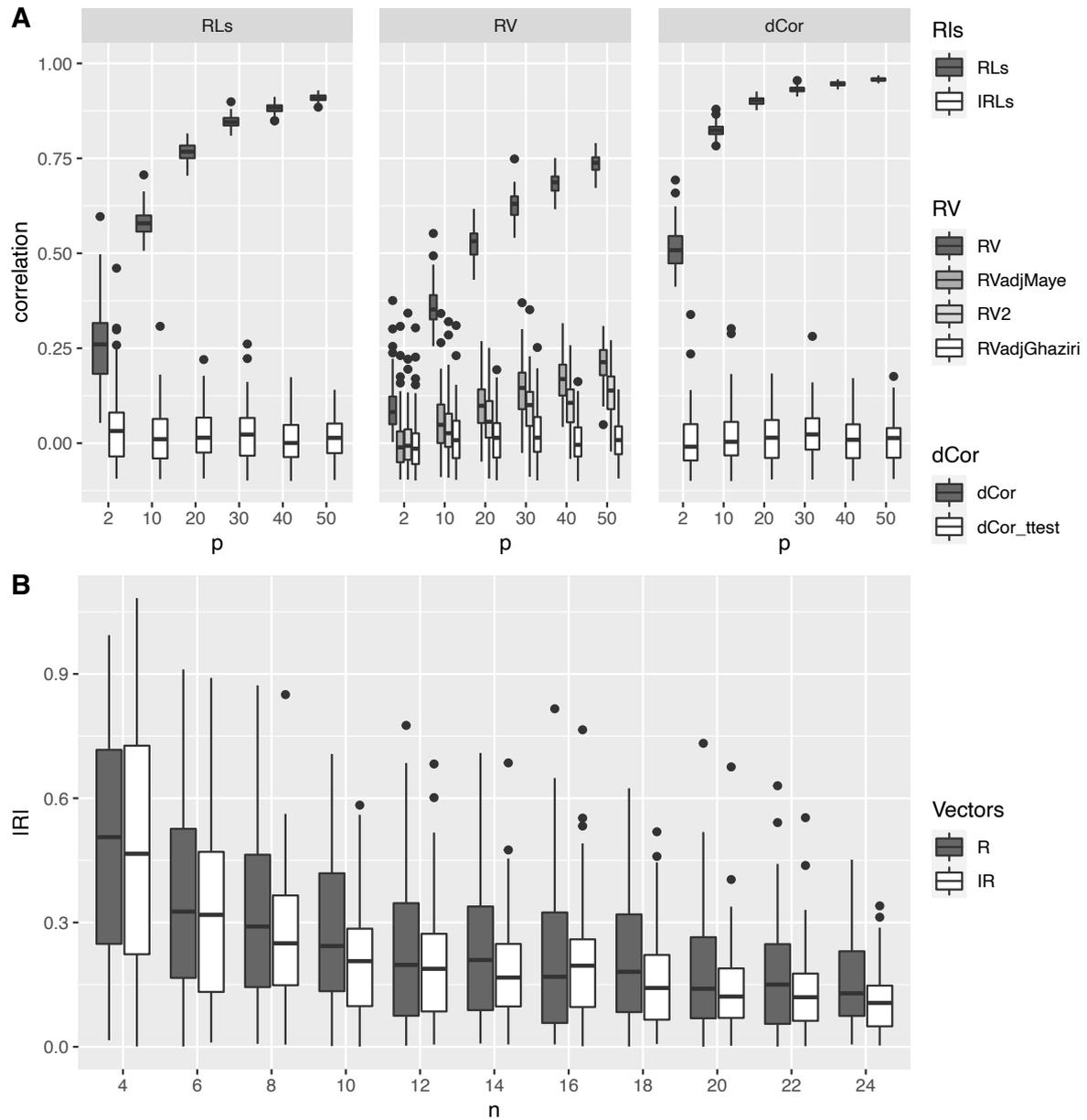


Figure 1: Susceptibility to overfitting for various correlation coefficients. A) Both simulated data sets are matrices of size $(n \times p)$ with $p > 1$. B) Correlated data sets are vectors ($p = 1$) with a various number of individuals n (vector length). For both A & B, 100 simulations were run for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

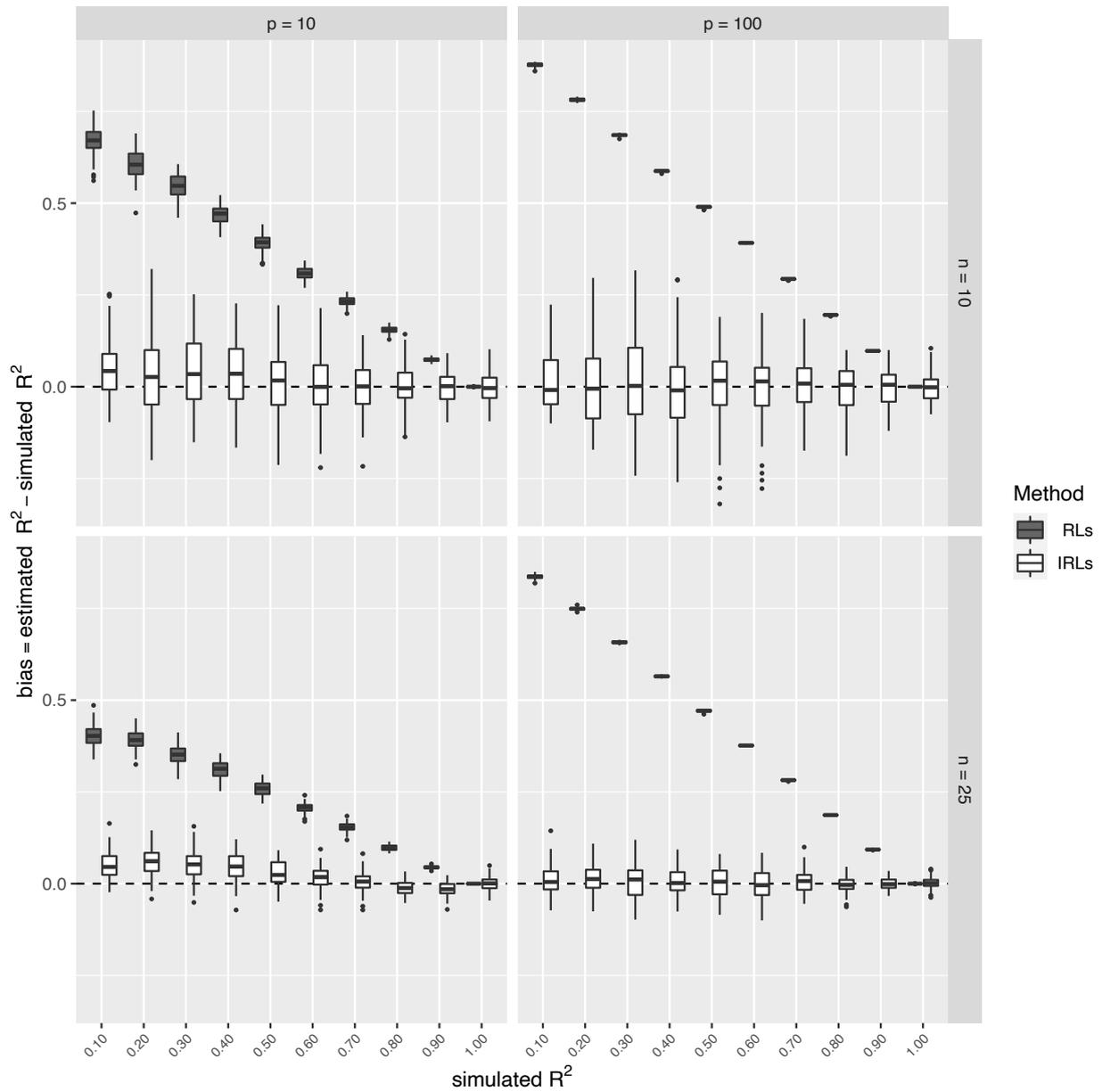


Figure 2: Shared variation (R^2) between two matrices has measured using both the corrected (IRLs) and the original (RLs) versions of the Procrustean correlation coefficient. A gradient of R^2 was simulated for two population sizes ($n \in \{10, 25\}$) and two numbers of descriptive variables ($p \in \{10, 100\}$). 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.

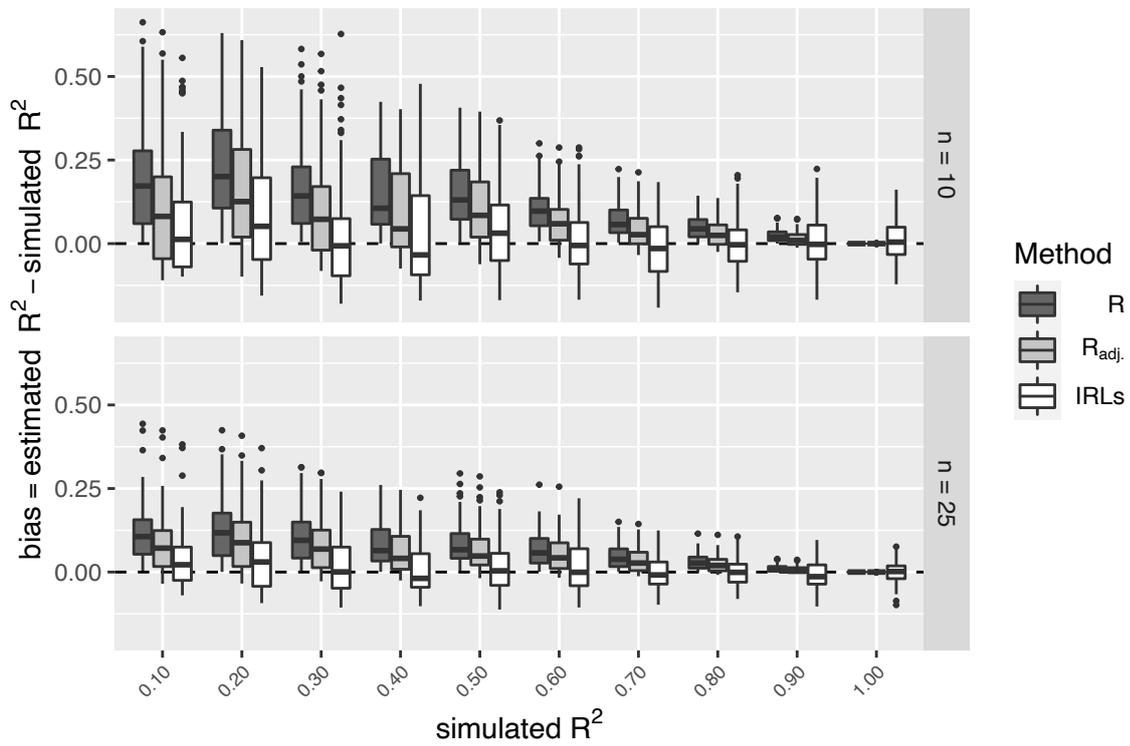


Figure 3: Shared variation between two vectors is measured using the classical R^2 , its adjusted version R_{adj}^2 and $(IRLs^2)$. A gradient of shared variation is simulated for two vector sizes ($n \in \{10, 25\}$). The black dashed line corresponds to a perfect match where measured R^2 equals the simulated one.