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.
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
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
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 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 coeficient Rls proposed by Lingoes and Schönemann (1974) and by Gower (1971). Let the define $Trace$, a function summing the diagonal elements of a matrix. For a $n \times p$ real matrix X and anothet a $n \times q$ real matrix Y defining respectively two sets of p and q centered variables caracterizing n individuals, we define $CovLs(X,Y)$ an analog of covariance applicable to vectorial data following Equation (1)

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

and $VarLs(X)$ as $CovLs(X,X)$. Rls can then be expressed as follow in Equation (2).

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

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 (Peres-Neto and Jackson, 2001). Among the advantages of Rls, its similarity with the Pearson correlation coefficient R (Bravais, 1844) has to be noticed. Considering $CovLs(X,Y)$ and $VarLs(X)$ 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 $q=1$, $Rls = |R|$.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

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).

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

U and V are two rotation matrices allowing to compute the transfer matrix to superimpose X on Y or reciproquly Y on 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. $CovLs(X,Y)$ can also be computed from singular values (Equation 4).

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

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

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

Similarly the informative counter-part of $VarLs(X)$ is defined as $IVarLs(X) = ICovLs(X,X)$.

3 Methods

Two methods are proposed to estimate $\overline{RCovLs(X,Y)}$. The first one is formal and applicable only when $p = 1$ and $q = 1$ the second one is based on a Monte-Carlo evaluation and is applicable for every p and q.

3.1 Formal estimation of $\overline{RCovLs(X,Y)}$

For two random real vectors of length n, x and y the mean of $R(x,y)^2$, $\overline{R(x,y)^2} = 1/(n-1)$. That equality is independent of the distribution of x and y. Let σ_x and σ_y being respectively the standard deviations of x and y, we can estimate $\overline{RCovLs(x,y)}$ by Equation (6).

$$\overline{RCovLs(x,y)} = \sigma_x\ \sigma_y\sqrt{\frac{1}{n-1}} \tag{6}$$

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

For every values of p and q including 1, $\overline{RCovLs(X,Y)}$ can be estimated using a serie of k random matrices $RX = \{RX_1, RX_2, ..., RX_k\}$ and $RY = \{RY_1, RY_2, ..., RY_k\}$ where each RX_i and RY_i have the same structure respectively than X and Y in term of number of columns and of standard deviation of these columns.

$$\overline{RCovLs(X,Y)} = \frac{\Sigma_{i=1}^{k} CovLs(RX_i, RY_i)}{k} \tag{7}$$

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

3.3 Estimation of $IRLs(X,Y)$

We proposed to define $IRLs(X,Y)$ the informative Procruste correlation coefficient as follow.

$$IRLs(X,Y) = \frac{ICovLs(X,Y)}{IVarLs(X)\ IVarLs(Y)} \tag{8}$$

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

3.4 Testing significance of $IRLs(X,Y)$

Significance of $IRLs(X,Y)$ 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 $\overline{RCovLs(X,Y)}$ to test that $ICovLs(X,Y)$ and therefore $IRLs(X,Y)$ are greater than expected under random hypothesis. Let counting over the k rendomization when $RCovLs(X,Y)_k$ greater than $CovLs(X,Y)$ name this counts $N_{>CovLs}$. P_{value} of the test can be estimated following Equation (9).

$$P_{value} = \frac{N_{>CovLs}}{k} \tag{9}$$

3.5 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), RVadjMaye, *RV2* and *RVadjGhaziri* 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.

3.6 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 $\overline{RCovLs(X,Y)}$, 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

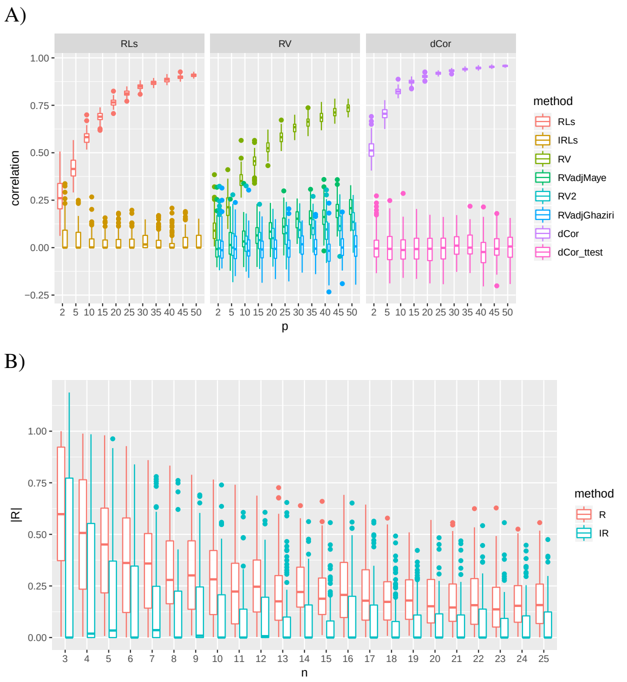

Fig. 1. A) Sensibility to overfitting for various correraltion coefficients. (A) Both simulated data sets are matrices of size $(n \times p)$ with $p > 1$. B) Correlated data sets are vectors $(p = 1)$ with a various number of individuals n (vector length). A & B) 100 simulations are run for each combination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

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`.

3.7 Empirical power assessment for the $CovLs$ test

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 1A). Because of its definition $IRLs$ values for non-correlated matrices are close to 0 whatever p (Figure 1A).

4.2 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 linearly 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.

4.3 Power of the test based on randomisation

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

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	`vegan::protest`	`ade4::procuste.rtest`
10	0.323	0.395	0.348
20	0.861	0.769	0.706
50	0.628	0.783	0.680

Fig. 2. Caption, caption.

4.4 Test1

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

5 Discussion

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text.

Table **??** shows that Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text.

6 Conclusion

(Table **??**) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method 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. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about 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. Figure 2 shows that the above method Text Text Text Text

Acknowledgements

Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

Funding

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

References

Bauer, M., Klau, G. W., and Reinert, K. (2007). Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics*, **8**, 271.

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.