Collinearity does not affect Procrustes analysis outputs: directions for plant and soil ecologists

Francy Lisboa [Corresp., 1], **Ruth Mitchell** [2], **Stephen J Chapman** [2], **Jacqueline Potts** [3], **Ricardo Berbara** [4]

[1] Statistics Division (ESS), Food and Agricultural Organization of the United Nations, Rome, Italy
[2] Ecological Sciences, James Hutton Institute, Aberdeen, Scotland - UK
[3] Biomathematics & Statistics Scotland, Aberdeen, Scotland - UK
[4] Soil Science, Universidade Federal Rural do Rio de Janeiro, Seropedica, Brazil

Corresponding Author: Francy Lisboa
Email address: Francy.Lisboa@fao.org

Background. The Procrustean residual vector (or PAM, an acronym for the alternative equivalent term Procrustean association metric) derived from Procrustes analysis can be seen as the univariate form of relationship between two or more data tables, which provides an interesting way for ecologists to place multivariate relationships as the central object of investigation in more familiar statistical approaches such as ANOVA and post hoc tests. However, many aspects need to be elucidated to make ecologists more confident in using Procrustes in their studies going beyond the simple comparisons. We attempted to address two questions: 1) How does the increasing number of correlated columns within an entire data table affect the Procrustes results? 2) Can the PAM be used for detecting how the correlation is partitioned across treatment levels within the original data table? **Methods.** Question 1) two data tables, **X** and **Y**, from a previous research were used to conduct the study. Four levels of correlation between variables (0.9, 0.7, 0.5, and 0.2) within the **X** data table were imposed to an increasing number of variables (6, 9, 12, and 15) to assess their effects on Procrustes relationship and its significance. Question 2) two simulated data tables covering four hypothetical categorical predictors (A, B, C, D) were created varying the relationship between them regarding the treatment A (0.2, 0.5, 0.7, 0.9) in order to assess the association between Procrustes and multiple mean comparisons method. **Results.** for the first question, we found that increasing the number of correlated variables across different imposed correlation levels (0.9, 0.7, 0.5, and 0.2) in the data table not subject to Procrustean linear transformation (translation and rotation), i.e. the **X** data table, had no effects either on the classical Procrustes outcomes related to the fit between data tables (*R* statistic and its *P* value), or on the significance of the ANOVA using the Procrustes association metric (PAM), which summarizes the multivariate correlation between two data tables, as the response variable. For the second question, increasing the between correlation levels between **X** and **Y** data tables for a specific set of rows in these

tables corresponding to a hypothetical treatment A resulted in PAMs that, when used in mean multiple comparisons, did show this treatment A as different from all others treatments B, C, and D from which X and Y were not related above (0.1). **Discussion**. Our results support that the Procrustes fit is only dependent on the information between data tables instead of within a data table. Finally, we showed that PAM, in fact, reflects the differences in multivariate correlation across data tables which can be useful for ecological questions addressing the partitioning of the multivariate correlation among different categorical levels (e.g. plots, time, land use type, etc.).

Francy J. G. Lisboa [1*], Ruth J. Mitchell [2], Stephen Chapman[2], Jacqueline Potts[3], Ricardo L. L. Berbara[4]

[1]Statitsics Division (ESS), Food and Agricultural Organization of the United Nations, Rome, Italy

[2] The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH United Kingdom

[3] Biomathematics & Statitsics Scotland, Craigiebuckler, Aberdeen, AB15 8QH United Kingdom.

[4] Soil Science Department, Agronomy Institute, Federal Rural University of Rio de Janeiro, Seropédica-RJ, 23890000, Brazil

[*]Corresponding author: Tel/Fax +39 0657056255; email address: Francy.Lisboa@fao.org

Abstract

Background. The Procrustean residual vector (or PAM, an acronym for the alternative equivalent term Procrustean association metric) derived from Procrustes analysis can be seen as the univariate form of relationship between two or more data tables, which provides an interesting way for ecologists to place multivariate relationships as the central object of investigation in more familiar statistical approaches such as ANOVA and post hoc tests. However, many aspects need to be elucidated to make ecologists more confident in using Procrustes in their studies going beyond the simple comparisons. We attempted to address two questions: 1) How does the increasing number of correlated columns within an entire data table affect the Procrustes results? 2) Can the PAM be used for detecting how the correlation is partitioned across treatment levels within the original data table?

Methods. Question 1) two data tables, **X** and **Y**, from a previous research were used to conduct the study. Four levels of correlation between variables (0.9, 0.7, 0.5, and 0.2) within the **X** data table were imposed to an increasing number of variables (6, 9, 12, and 15) to assess their effects on Procrustes relationship and its significance. Question 2) two simulated data tables covering four hypothetical categorical predictors (A, B, C, D) were created varying the relationship between them regarding the treatment A (0.2, 0.5, 0.7, 0.9) in order to assess the association between Procrustes and multiple comparison mean method.

Results. for the first question, we found that increasing the number of correlated variables across different imposed correlation levels (0.9, 0.7, 0.5, and 0.2) in the data table not subject to Procrustean linear transformation (translation and rotation), i.e. the **X** data table, had no effects either on the classical Procrustes outcomes related to the fit between data tables (*R* statistic and its *P* value), or on the significance of the ANOVA using the Procrustes association metric

(PAM), which summarizes the multivariate correlation between two data tables, as the response variable. For the second question, increasing the between correlation levels between **X** and **Y** data tables for a specific set of rows in these tables corresponding to a hypothetical treatment A resulted in PAMs that, when used in mean multiple comparisons, did show this treatment A as different from all others treatments B, C, and D from which X and Y were not related above (0.1).

Discussion. Our results support that the Procrustes fit is only dependent on the information between data tables instead of within a data table. Finally, we showed that PAM in fact reflects the differences in multivariate correlation across data tables which can be useful for ecological questions addressing the partitioning of the multivariate correlation among different categorical levels (e.g. plots, time, land use type, etc.).

Keywords: Procrustes association metric, multivariate data, correlation, ANOVA, ecology

Introduction

Analysis of Variance (ANOVA) is used as a tool to split the variability of a given outcome of interest into two basic components: 1) the variability explained by one or more categorical predictors; 2) the residual variation. In the ANOVA framework the response variables can vary in their nature, being classified as continuous or discrete, and univariate or multivariate. The simplest univariate context of ANOVA, that is, one response and one categorical predictor is obviously easier to analyze than the multivariate context; however for ecologists the univariate world rarely exists given that most ecological questions require one to handle multiple variables. Therefore the question arises: how can ecologists fit the natural

multivariate requirement of ecological research to the simplicity of the univariate ANOVA and post hoc test frameworks?

Lisboa et al. (2014b) showed how the results from Procrustes analysis (Gower, 1971), a multivariate statistical approach for correlating data tables representing sets of information coming from the same objects of study: plots, environmental gradient levels, experimental treatments, etc., could be used in downstream statistical analysis, including ANOVA and post hoc tests. Procrustes analysis has been shown to be statistically superior in some aspects (lower Type I error and higher power) than the traditional analogue approach, the Mantel test (Peres-Neto and Jackson, 2001) and one of the features that arise from Procrustes analysis is the possibility of providing the multivariate relationship among two, or more, data tables in a vector form made by residuals, the Procrustean residual vector, also named the Procrustean association metric (PAM) (Lisboa et al. 2014b). For example, assume that an ecologist wants to correlate two data tables, **X** and **Y**, the first one representing abiotic variables (climate, soil, elevation, etc.) and the second one representing a certain biological community (birds, bacteria, etc.). Moreover, assume that in **X** and **Y** all variables (columns) were measured from field plots, which represent the rows of the data tables. Procrustes analysis will find the "best" fit of homologous coordinates across the **X** and **Y** data tables by seeking to minimize of the sum of squares between corresponding coordinates in **X** and **Y**, i.e. the plots or rows of these tables. Given that the Procrustes fit is never perfect, the PAM stands for the residuals between corresponding coordinates (the plots or rows of **X** and **Y**) after the "best" fit among the tables has been found so that the lower the residual sizes in the PAM, the higher the correlations. The compilation of these residual differences between homologous rows in the **X** and **Y** data tables making up the Procrustean residual vector (PAM) represents a useful way to summarize information on the

92 relationship between the two matrices and makes it available for further statistical analysis, both
93 parametric and non-parametric (Lisboa et al., 2014b).

94 Despite the potential uses of this Procrustean feature in ecological research, the rigor of
95 this composite framework (made up of PAM – ANOVA – Post hoc tests) has not been assessed.
96 In particular we consider two questions: 1) How does the increasing number of correlated columns
97 within an entire data table affect the Procrustes results? 2) Can the PAM be used for detecting how
98 the correlation is partitioned across treatments levels within the original data table? For simplicity,
99 the Procrustes analysis results can be divided into two sets 1) *mainstream results*, which take
100 account of the correlation statistic (*R*) between multidimensional data tables and its significance:
101 (*P* value), i.e. the Procrustes fit; and 2) *downstream results*, which are related to statistics provided
102 by the analysis of the PAM in other statistical frameworks, like ANOVA, and multiple
103 comparisons of means using, for example, Tukey's HSD test.

104 With respect to the first question, Dray et al. (2003) argued that the Procrustes fit, that is
105 the mainstream results, is only influenced by the variation between matrices. Therefore, according
106 to these authors, the variation in the number of correlated columns within a data table should not
107 influence the mainstream results such as the *R* statistic and its *P* value. However, nothing is known
108 about the consequences of the number of correlated columns within a data table on the PAM. For
109 example, we do not know whether the increasing correlation within an entire data table is conveyed
110 to the PAM and whether it affects the outcome of using PAM as the response in an ANOVA
111 framework.

112 On the other hand, there are no papers exploring explicitly the following statement: "the
113 analysis of the PAM, by looking at consistencies of the residual size between homologous
114 coordinates across different treatments, could be useful for providing insights about differences in

terms of multivariate correlation". Such a statement is linked to the use of the PAM in downstream statistical analysis such as ANOVA and multiple comparisons of means. Thus we also investigated whether the PAM is able to detect differences in multivariate correlation among treatments when it is used in multiple comparison tests.

Material and Methods

First question: does the increasing number of correlated columns within an entire data table affect the Procrustes results?

To assess whether increasing the number of correlated columns within an entire data table affects the Procrustes results we used two data tables from a study by Lisboa et al. (2014a). In this study the authors used Procrustes analysis together with ANOVA and multiple comparisons of the means in order to assess how the strength of the "match" (correlation) between individual soil microbial community variables and individual soil fertility variables varied across four land use types (native forest, degraded pasture, improved crop, integrated crop-livestock-forest). The soil microbial community (PLFA profile) and the soil fertility data tables had the following dimensions: (n = 53, p = 20) and (n = 53, p = 15), respectively. Hereafter n and p stand for row and column numbers in the data tables, respectively.

Four correlation levels (0.2, 0.5, 0.7, and 0.9) were incorporated into different numbers of columns within the soil fertility data table (n = 53, p = 15), hereafter the **X** data table. The number of soil variables that were correlated was increased gradually (6, 9, 12, and 15), whereas the PLFA profile data table (n = 53, p = 20), **Y,** had its original correlation structure unaltered. After that, the **X** (soil fertility) and **Y** (PLFA profile) data tables were submitted to thirteen different pre-transformations. These 13 pre-Procrustes transformations were used to encompass the three

broadly different manners in which **X** and **Y** data tables can be used in the Procrustes analysis: raw data, dissimilarity/distance matrices or ordination axes (Fig. 1 c). Finally the Procrustes relationships between **X** (soil fertility) and **Y** (PLFA profile) were simulated one hundred times for each pre-transformation. As some of the dissimilarity metrics used are undefined for negative values, the **X** matrix was squared prior to analysis. From each set of 100 simulations/pre-transformations within a given number of correlated columns within **X** (6, 9, 12, 15) with a given correlation level (0.2, 0.5, 0.7, 0.9) we retained the following statistics: 1) the average of the Procrustean correlation statistic (*R* value); 2) the number of times that the *R* statistic was significant (*P* value); 3) the PAM average (a Procrustean residual vector from the average of the 100 simulations); 4) the residual size range within the PAM average (subtracting the highest and the lowest residual sizes linking the two data tables after Procrustes fit); 5) the number of times in which the ANOVA using the PAM average as response and the land use type as factor (4 levels) was considered significant ($P < 0.05$). Thus, we retained 13 sets of Procrustes information, which were then used for making graphs.

Correlation incorporated into soil fertility data table for assessing the first question

The process of incorporating distinct correlation levels into the soil fertility data table, the **X** data table, followed two basic steps:

1) Specific level - correlation matrix **M** generation (0.2, 0.5, 0.7 and 0.9);
2) Cholesky decomposition of **M** *into* $\mathbf{L}\mathbf{L}^{T}$, where **L** is a lower triangular matrix, and multiplication of $\mathbf{L}^{T}$ by the transpose of the soil fertility matrix **X**.

For generating the specific-level correlation table **M**, we used the R functions described by Hardin et al. (2013) which are intended for building correlation matrices with *noise* addition

(http://pages.pomona.edu/~jsh04747/research/simcor.r.). Here the *noise* added to the **M** entries was from -0.001 to 0.001. After obtaining **M**, its correlation structure levels were incorporated into the soil fertility data table by using the following R code:

```
fert.unc<-t((solve(t(chol(cov(fert.m)))))%*%t(fert.m)) # remove existing correlation
object <- t(chol(M))%*% t(fert.unc) # incorporates correlation structure levels
object.df <- t(object)      # creates the simulated soil fertility data frame
corrplot(cor(object.df)) # checks the correlation structure incorporated
```

Specifically, all correlation structure levels (0.2, 0.5, 0.7, and 0.9) were incorporated into the entire **X** data table ($n = 53$, $p = 15$) and from them the number of columns (variables) correlating was reduced gradually (15, 12, 9, and 6) within each correlation level. The remaining columns within the soil fertility were left without any correlation level imposed. For example, for evaluating the effects of correlation levels imposed into 6 columns of the entire **X** data table ($n = 53$, $p = 15$) the rest of the 9 columns within **X** were not correlated. The whole process from incorporating different correlation levels into an increasing number of columns within the **X** data table to the Procrustes analysis was repeated 100 times for each of the 13 pre-Procrustes transformations is described in Fig. 1 c. For each of the simulations the correlated columns were sampled at random without replacement from the soil fertility table. The R code for simulating the increasing number of correlated columns within the soil fertility table and then exploring its effects on Procrustes results can be found in the supplementary material S1.

Second question: can the PAM be used for detecting how the multivariate correlation is partitioned across different treatments?

For assessing whether the use of PAM in multiple comparisons of means is able to detect differences among treatments in terms of multivariate correlation, simulated data tables **X** and **Y**

were used. One can visualize this as if these data arose from a hypothetical scenario where **X** and **Y** are data tables derived from a study investigating how the multivariate correlation between the general plant community (**X** data table) and its functional traits (**Y** data table) is partitioned across an environmental gradient based on the time elapsed after an intense burning event. Also, one can consider that **X** and **Y** are encompassing four times elapsed after the burning event where plant community (**X** data table) and functional traits (**Y** data table) were measured at times A, B, C, and D.

Hereafter these four times will be referred as treatments. Four different correlation levels (0.2, 0.5, 0.7, 0.9) were only incorporated into the treatment A for both **X** and **Y** data tables, and this treatment A corresponds to the first ten rows of each of these data tables. For all other treatments (B, C, and D) the correlation between **X** and **Y** was never greater than 0.1 (Fig. 1 c). After correlation level incorporation, the 13 pre-Procrustes transformations were applied to both **X** and **Y** data tables before the Procrustes analysis (Fig. 1c). All steps from the **X** and **Y** data table generation to the Procrustes analysis were repeated 100 times. Also, these simulations were carried out varying the number of columns (variables, p) in relation to the number of rows (sites, n) so that **X** and **Y** were data tables with the follow dimensions: ($n = 40$, $p = 25$); ($n = 40$, $p = 45$) and ($n = 40$, $p = 80$). From each set of 100 simulations/pre-transformations within a given correlation level between **X** and **Y** data tables incorporated into the treatment A (0.2, 0.5, 0.7, 0.9), we retained the following information from the Procrustes results: 1) the number of times in which the treatment A came out as being different from all other treatment (A ≠ B, C, D) when using the average PAM in Tukey HSD (95%); 2) the average value of the Procrustean residual size in each treatment (A, B, C, D). Thus, for each correlation level between **X** and **Y** in the treatment A, we retained 13 sets of Procrustes information, which were used for making graphs.

Different correlation levels between **X** *and* **Y** *for a specific treatment*

For creating the simulated data tables with different between correlation levels for a specific categorical level in both **X** and **Y** tables (namely level A) we first created sets of three "big" tables: 1 (n = 10, p = 50), 2 (n = 10, p = 90), 3 (n = 10, p = 160). Four correlation structure levels were incorporated into each "big" table (0.2, 0.5, 0.7, 0.9), and this was carried out using the same procedure described for the soil fertility data table in the first part of this paper.

After the correlation structure was added to the "big" data tables, each one was broken down into two equal tables. For example, in the case of a "big" table (n = 10, p = 50) with a given correlation level of 0.2, it was divided into **X***corr*$_{0.2}$ (n = 10, p = 25) and **Y***corr*$_{0.2}$ (n = 10, p = 25) tables. Thus, each one of these "big" tables provided four pairs of **X** and **Y** data tables (n = 10, p = 25, 45, 80) representing different correlation levels between them for the treatment A, such that we have: (A**X***corr*$_{0.9}$, A**Y***corr*$_{0.9}$); (A**X***corr*$_{0.7}$, AY*corr*$_{0.7}$); (A**X***corr*$_{0.5}$, A**Y***corr*$_{0.5}$); (A**X***corr*$_{0.2}$, A**Y***corr*$_{0.2}$).

For taking account of other treatments (B, C, and D) we created three "big" tables: 1 (n = 30, p = 50), 2 (n = 30, p = 90), and 3 (n = 30, p = 160), with all columns p having the same correlation level (corr. < 0.1). These tables were broken down as in the same way as for treatment A. Thus, for each (n = 30, p = 25, 45, and 80) four pairs of **X** and **Y** tables were generated. The tables A**X***cov*$_i$ and A**Y***cov*$_i$ were then linked to BCD**X***corr*$_{<0.1}$ and BCD**Y***corr*$_{<0.1}$ tables, respectively, in order to build the entire **X** and **Y** tables (n = 40, p = 25, 45, 80) as shown in Fig. 1 b. The whole process of incorporating different correlation levels into **X** until the Procrustes analysis was simulated 100 times for each one of the 13 pre-Procrustes transformations described in Fig. 1 c. The R code used for simulating different correlation levels between **X** and **Y** data tables

for the treatment A and then exploring its effects on the Procrustes results can be found in the supplementary material S2.

Results

Imposed correlation effects on individual mainstream and downstream Procrustes results

Both Procrustes mainstream results, the Procrustean correlation statistic *R* and its significance (*P* value), remained constant irrespective of the increasing number of correlated variables within the **X** data table and the level of correlation incorporated into them (Fig. 2 a b). The constancy across the increasing number of correlated columns and their imposed correlation levels within the **X** data table was also true for Procrustes results involving the PAM, such as the measure of residual size variability across individual PAMs, the residual ranging size (maximum minus minimum residual sizes in the PAM linking the two data tables under analysis), and the number of significant ANOVA results using PAMs as response variable (Fig. 2 c - d). Such lack of effects of the increasing number of correlated columns/variables across different correlation levels were reinforced when the Procrustes results from the use of ordination axes and dissimilarity/distance matrices were considered separately (Fig. 1 and 2, supplementary material S3). Moreover, the use of PAMs in NMDS ordination (Euclidean distance) showed no clear grouping following both the number of correlated columns/variables and the level of correlation within the **X** data table (Fig. 3 a-b, supplementary material S3).

*Correlation between **X** and **Y** data tables for a specific treatment*

The mean percentage of significant ANOVAs using PAMs as the response variable increased as the correlation level between **X** and **Y** data tables for treatment A increased (Fig. 3 a). The mean number of times where treatment A was significantly different from all other

252 treatments (% A ≠ BCD) increased as the correlation level between **X** and **Y** for the treatment A
253 increased (Fig. 3 b).

254 For all dimensions of **X** and **Y**, the higher correlation levels between these two data tables
255 for the treatment A (0.7 and 0.9) were reflected by the mean Procrustes residual size for
256 treatment A being lower than others B, C, and D (Fig. 4 a - c). At the lower correlation levels
257 between the **X** and **Y** data tables for the treatment A (0.2 and 0.5) the mean Procrustes residual
258 size for treatment A was not different from the others, B, C, and D (Fig. 4 c).

259 **Discussion**

260 The use of the Procrustes residual vector (PAM) in ANOVA and multiple comparisons
261 is not widespread as an ecological routine (Lisboa et al., 2014b). The reasons for this are diverse,
262 including the lack of studies exploring the limitations of this composite framework. Here, we
263 have attempted to address two questions: 1) how does the increasing number of correlated
264 variables/columns within an entire data table affect the Procrustes results? 2) can the Procrustean
265 residual vector (i.e. the PAM) show differences among treatments in terms of multivariate
266 correlation when it is used in multiple comparisons of means?

267 *Correlation level within a data table does not clearly affect the Procrustes fit and PAM-ANOVA*
268 *results*

269 The most common use of Procrustes analysis in the ecological literature is for comparing
270 different methodologies. For example, in soil microbiology research Procrustes analysis has
271 typically been used for comparing ordination patterns from different methods of accessing the
272 soil microbial community (e.g. PLFA, T-RFLP, high throughput sequencing) (Vinten et al.,
273 2011). Others authors have used Procrustes to assess how sampling error levels could affect the
274 correlation between ordinations (Hirst and Jackson, 2007). All these examples used Procrustean

parameters, such as the Procrustean *R* correlation statistic and its significance for assessing the fit between methodologies, which stress that the Procrustes fit is usually the final aim of most studies using that approach.

The limited usage of Procrustes in the ecological literature, especially in plant and soil ecology (Lisboa et al., 2014b), results in no information on the consequences of the correlation within an entire data on Procrustes results and, in the sense of using the Procrustean residual vector (PAM) in other statistical frameworks such as ANOVA, there are no references. Here our results indicated no clear effects of the increasing number of correlated variables across different imposed correlation levels in the **X** data table. The Procrustes analysis used for carrying our simulation out was the so-called symmetric version in the vegan R package (Oksanen et al., 2013), which submits both **X** and **Y** data tables – multidimensional configurations to linear transformations (scaling, translation and rotation) in order to minimize the sum of squares of corresponding points (Peres-Neto and Jackson, 2001).

The low effect of the imposed correlation levels on Procrustes outcomes may have to do with the statement that the Procrustes fit only takes account the differences between **X** and **Y** configurations (Dray et al., 2003), which can be seen by:

$$fit_{XY} = \text{trace}((\mathbf{X} - \mathbf{Z})^T (\mathbf{X} - \mathbf{Z})) \quad \text{(Gower, 1971)},$$

where the **Z** matrix corresponds to the new set of coordinates arising from linear translation and rotation on the **Y** data table – multidimensional configurations. It can cast light on our results showing that the increasing number of correlated columns across different imposed correlation levels did not come up with differences in terms of both mainstream and downstream Procrustes results.

297 Since Procrustes takes "two to tango" by relating multidimensional configurations, which
298 in turn are affected by the kind of pre-transformation on the data tables (Legendre and Gallagher,
299 2001), it would be expected that both, **X** and **Y** data table – multidimensional configurations
300 (without scaling, translation and rotation movements) could have had some effect on Procrustes
301 outcomes. Nonetheless, our results suggest there was no effect of the **X** data table – pre-
302 transformations on their respective **X** data table – multidimensional configurations. It was due to
303 the high similarity between raw data, dissimilarity matrices and ordination axes in terms of
304 Procrustes results and PAMs, irrespective of the imposed correlation level. Thus, the existing
305 correlation within the non-translated and non-rotated configuration, in our case, the **X** data table,
306 may not be a hurdle for the Procrustes results, irrespective of using raw, dissimilarity matrices, or
307 ordination axes as entries.

308 *What does PAM tell us?*

309 An argument advocating the use of the Procrustes residual vector in a downstream
310 statistical approach using ANOVA and multiple comparisons is that the consistencies in the
311 Procrustean residual sizes, which are linking the two or more tables under investigation, could be
312 used to make inferences on the strength of the multivariate correlation across environmental
313 gradients. However, so far, no studies have explicitly explored such a statement. In fact the few
314 existing studies that used the PAM to make inference that goes beyond accessing the correlation
315 between data tables were based on that statement (Singh et al., 2008, Landeiro et al., 2011,
316 Siqueira et al. 2012, Lisboa et al., 2012, 2014a). Our results show that the correlation level
317 between **X** and **Y** data tables for the treatment A affected the ANOVA and multiple comparisons
318 of means using the PAM. We have found that the PAMs generated from the higher levels of
319 correlation (0.7 and 0.9) are more capable of discriminating the treatment (A) from the others (B,

C, and D). These results are supporting the statement in favor of using the PAM to assess the strength of the multivariate correlation across categorical levels.

One point that may create confusion is the interpretation of the PAM-multiple comparison of means results. We have used Tukey's HSD as it is a standard option in many studies, but we found that when only plotting the point estimate (mean) and the standard deviation of the residuals for each treatment the pattern of lower residual size in highest correlation levels was clearer (Fig. 4). The assumptions are that since the link between the two data tables is done through the use of residuals of the PAM after the best fit, then the lower the average residual size, the higher is the multivariate correlation for a specific treatment. Our results support this by showing that the mean residual size at treatment A is lower than the mean residual size in the other treatments B, C, and D when the correlation between **X** and **Y** for treatment A is high (0.7 and 0.9). Thus, the overall interpretation for PAM using multiple comparisons is that low mean residual size for a treatment indicates "strong" multivariate correlation.

Final considerations

We explored for the first time the effects of increasing the number of correlated columns across different imposed correlation levels within the **X** data table on the Procrustes results related to the fit (*R* statistic and its *P* value) and to the use of the Procrustean residual vector (PAM) in ANOVA. In addition, we also tried to show that the PAM when used in multiple comparisons can provide insights about differences among "treatments" in terms of multivariate relationships. We have only used data tables whose entries were quantitative, so only dissimilarities and transformations considered adequate for this kind of data were used in pre-transformation for getting the same dimension between **X** and **Y**. However, we do recognize that

343 binary data tables (presence/absence) are also important in ecology (Anderson et al., 2011) and
344 the evaluation of the correlation levels within binary data tables on Procrustes results must be an
345 objective of future investigations.

346 Procrustes analysis is a symmetric approach used to link two or more data tables
347 (Legendre and Legendre, 2012). This means that the data tables under analysis are evaluated on
348 an equal footing, that is, without setting which of them is response or predictor. Also Procrustes
349 does not have a regression step, which implies that the number of columns (variables, p) in a
350 matrix does not need to be lower than the number of rows (sites, n) as required by traditional
351 approaches to link data tables such as RDA (Redundancy analysis) and CCA (Canonical
352 Correspondence Analysis). In the present study we did not do a formal evaluation to test the
353 effects of $n > p$ on the Procrustes results as we focused on correlation effects. However, by
354 varying the dimensions of **X** and **Y** data tables in the second part of the paper ($n = 40$, $p = 25$, 45,
355 and 80) the results indicated that $n < p$ and $n > p$ may have similar effects on Procrustes results.

356 To our knowledge this is the first study exploring the correlation effects on the Procrustes
357 results and interpretation. Here we showed that both, the number of correlated variables and the
358 correlation levels within an entire data table, have no effects on the mainstream Procrustes
359 results related to the fit, such as R and its significance. In addition, the increasing correlation
360 level within a data table does not affect the results of ANOVA using PAM as the response.
361 Overall, our study supports the concept that the Procrustes fit only take into account the variation
362 between data tables. Finally, we were able to show that the PAM can reflect treatment
363 differences in terms of multivariate correlation when it is used in multiple comparisons of means.
364 It supports PAM – ANOVA – Multiple comparisons as an interesting composite approach for

getting additional information on how the strength of the multivariate correlation varies across categorical environmental levels.

Acknowledgements

The work of Ruth Mitchell, Stephen Chapman and Jacqueline Potts was funded by the Scottish Government Rural and Environment Science and Analytical Services Division (RESAS). Finally, the authors thanks to Mark Brewer and David Elston from BioSS for the helpful comments on earlier drafts of the manuscript.

Literature cited

Anderson, M. J, Crist, T. O, Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L, Sanders, N. J., Cornell, H, V., Comita, L. S., Davies, K. F., Harrison, S. P., Kraft, N. J., Stegen, J. C., Swenson, N. G. 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology Letters, 14: 19-28.

Dray, S., Chessel, D., Thioulouse, J. 2003. Procrustean Co-inertia analysis for the linking of multivariate datasets. Ecoscience **10**: 110-119.

Gower, J. C. 1971. Statistical methods of comparing different multivariate analyses on the same data. In: Hodson F, R., Kendall, D. G., Tautu, P., editors. Mathematics in the archeological and historical sciences. Edinburgh University Press, Edinburgh. pp. 138-149

Hardin, J., Garcia, S, R., Golan, D. 2013. A method for generating realistic correlation matrices. Annals of Applied Statistics **7**: 1733-1762

387 Hirst, C. N., Jackson, D. A. (2007) Reconstructing community relationships: the impact of
388 sampling error, ordination approach, and gradient length. Divers. Distrib 13: 361-371.
389 Landeiro V. L., Bini, L. M., Costa, F. R. C., Franklin, E., Nogueira, A., de Souza, J.L.P., Moraes,
390 J., Magnusson, W.E.2012. How far can we go in simplifying biomonitoring assessments?
391 An integrated analysis of taxonomy surrogacy, taxonomic sufficiency and numerical
392 resolution in a mega diverse region. Ecological Indicators 23: 366–373.
393 Legendre P & Gallagher ED. (2001) Ecologically meaningful transformations for ordination of
394 species data. Oecologia 129: 271–280
395 Legendre, P., Legendre, L. 2012. Numerical Ecology, 3rd English edn. Elsevier344 Science BV,
396 Amsterdam. 516 p. 483-495.
397 Lisboa, F. J. G., Chaer, G. M., Jesus, E. C., Gonçalves, F. S., Santos, F.M., de Faria, S. M.,
398 Castilho, A., Berbara, R. L. L. 2012. The influence of litter quality on the relationship
399 between vegetation and below-ground compartments: a Procrustean approach. Plant and
400 Soil **367**: 551-562.
401 Lisboa, F. J. G; Chaer, G. M; Fernandes, M. F; Berbara, R. L. L; Madari, B. E. 2014a. The match
402 between microbial community structure and soil properties is modulated by land use
403 types and sample origin within an integrated agroecosystem. Soil Biology and
404 Biochemistry 78: 97-108
405 Lisboa, F. J. G; Peres-Neto, P. R., Chaer, G. M., Jesus, E. C., Mitchell, R. J., Chapman, S. J.,
406 Berbara, R. L. L. 2014b. Much beyond Mantel: Bringing Procrustes Association Metric
407 to the Plant and Soil Ecologist's Toolbox. PLoS ONE **9**(6): e101238.
408 doi:10.1371/journal.pone.0101238

409 Oksanen, J.; Blanchet, F. G.; Kindt, R.; Legendre, P.; Minchin, P. R.; O'Hara, R. B.; Simpson, G.
410 L.; Solymos, P.; Stevens, M. H. H.; Wagner, H. vegan: Community Ecology Package, 2013
411 http://cran.r-project.org/web/packages/vegan/index.html.

412

413 Peres-Neto, P. R., Jackson, D. A. 2001. How well do multivariate data sets match? The
414 advantages of a Procrustean superimposition approach over the Mantel test. Oecologia
415 129: 169-178.

416 Schneider, J. W, Borlund, P. 2007. Matrix Comparison, Part 2: measuring the resemblance
417 between proximity measures or ordination results by use of the Mantel and Procrustes
418 statistics Journal of the American Society for Information Science and Technology
419 58:1596

420 Singh, B. K, Nunan, N., Ridgway, K. P., McNicol, J., Young, J. P. W, Daniell, T.J., Prosser, J.I.,
421 Millard P. 2008. Relationship between assemblages of mycorrhizal fungi and bacteria on
422 grass roots. Environmental Microbiology 10: 534–542.

423 Siqueira, T., Bini, L. M., Roque, F. O., Cottiene, K. 2012. A metacommunity framework for
424 enhancing the effectiveness of biological monitoring strategies. PLoS One 7: e43626.

425 Vinten, A. J. I., Artz, R. R. E., Thomas, N., Potts, J. M., Avery, L. M., Langan, S. J., Watson, H.,
426 Cook, Y., Taylor, C., Abel, C,. Reid E., Singh, B.K. 2011. Comparison of microbial
427 community assays for the assessment of stream biofilm ecology. Journal of
428 Microbiological Methods. 85:190-19

429

Figure 1(on next page)

Figure 1

General approach used in the study. **a**) Illustration of the first question addressed: the effects of increasing the number of variables correlating within **X** data table (soil fertility, n = 53, p = 15) on the Procrustes results. For each number of columns (6, 9, 12, 15) in the **X** data table, four correlation levels were imposed (0.2, 0.5, 0.7, 0.9). The **X** data table was related to the **Y** data table (PLFA profile, n =53, p = 20, none correlation structure imposed) by Procrustes analysis. **X** and **Y** data tables are from Lisboa et al. (2014b). **b**) Illustrates the second question: whether the correlation level between **X** and **Y** data tables (simulated data) incorporated into a specific treatment (treatment A) is reflected in the results of ANOVA analysis of the PAMs. The correlation between **X** and **Y** for all others treatments (B, C, and D) was not greater than 0.1. **c**) The different pre-Procrustes transformations in which **X** and **Y** were used in the Procrustes analysis (raw data, dissimilarity matrices, ordination axes). Each of these pre-transformations was simulated 100 times in order to get the results.

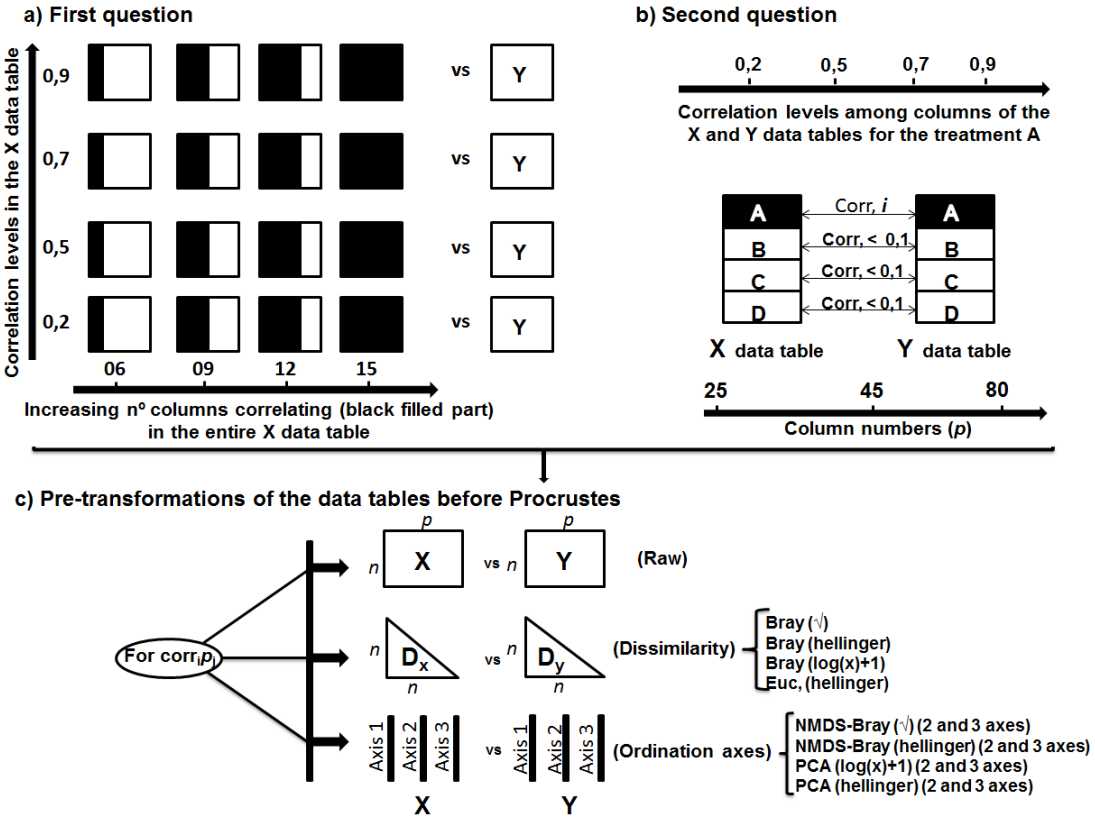

Figure 2(on next page)

Figure 2

Effects of increasing the number of variables correlated at different levels within an entire **X** data table on Procrustes results. **a)** Effect on Procrustes correlation statistic *R*; **b)** Effect on significance of Procrustean relationship (*P* value); **c)** Effect on residual size ranging within the vector of relationship (Procrustean association metric: PAM); **d)** Effect on ANOVA significance by using the PAMs as response and land use type (4 levels) as categorical predictor. The **X** (soil fertility) and the **Y** (PLFA profile) data tables are derived from Lisboa et al. (2014b). For each correlation levels (0.9, 0.7, 0.5, and 0.2) the number of variables correlating was increased from 6 to 15 (total variables). The correlation within **Y** data table was held fixed (original correlation structure). Means ± 1 SE of 13 pre-Procrustes transformations simulated 100 times are shown (Fig. 1c). d? L? :UZC

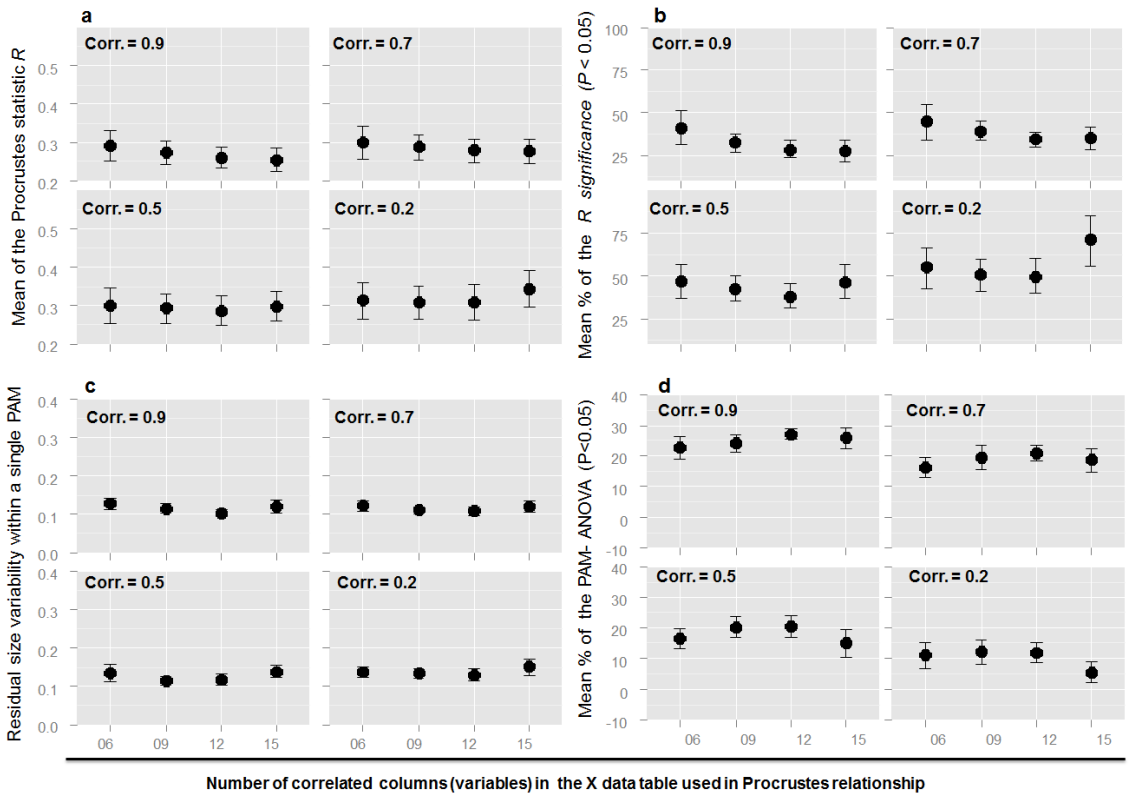

Figure 3(on next page)

Figure 3

Procrustes downstream results as affected by the correlation levels between **X** and **Y** data tables incorporated into a specific treatment A, while holding fixed the correlation level between **X** and **Y** for others treatments B, C, and D (Correlation < 0.1). **b**) Mean percentage of significant ANOVAs ($P < 0.05$) when the Procrustes residual vectors (PAMs) were used as response variable. **b**) Mean percentage of time when A treatment was significant different from all other treatments in multiple comparisons (Tukey, 95%, CI). Means ± 1 SE of 13 pre-Procrustes transformations simulated 100 times.

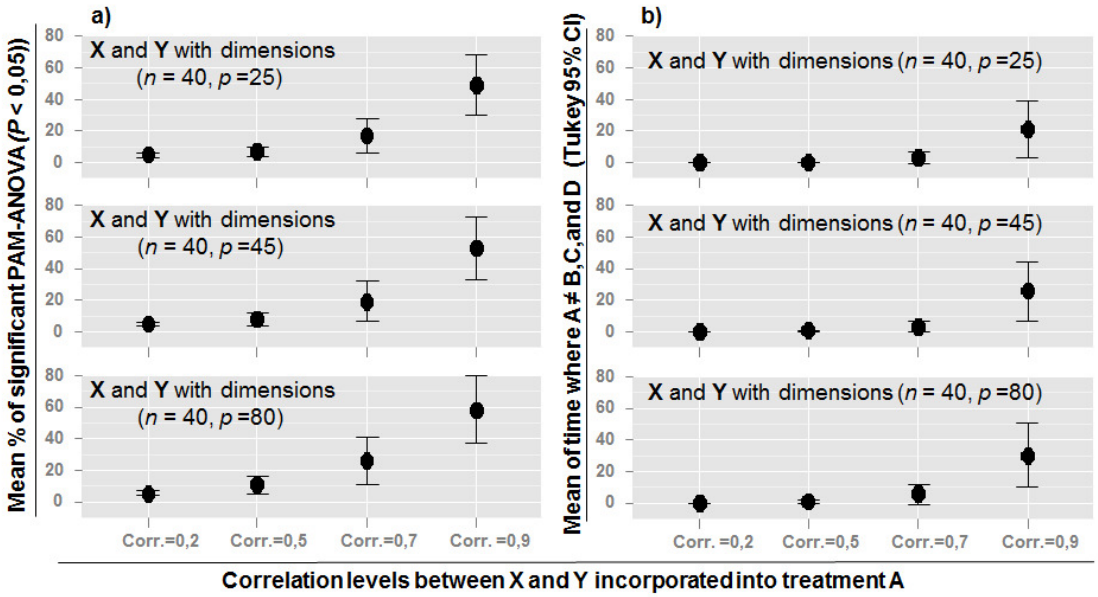
a)
b)
X and Y with dimensions (n = 40, p =25)
X and Y with dimensions (n = 40, p =45)
X and Y with dimensions (n = 40, p =80)
Mean % of significant PAM-ANOVA (P < 0,05))
Mean of time where A ≠ B,C,and D (Tukey 95% CI)
Corr.=0,2
Corr.=0,5
Corr.=0,7
Corr. =0,9
Correlation levels between X and Y incorporated into treatment A

Figure 4(on next page)

Figure 4

Assessing how multivariate correlation levels between two data tables incorporated into a specific treatment (treatment A: 0.9, 0.7, 0.5, and 0.2) are able to generate Procrustes residual vectors (PAMs) capable of differentiating this treatment from others (B, C, and D) in a multiple comparison. The correlation between **X** and **Y** for all other treatments (B, C, and D) was held fixed at < 0.1. **a**) PAMs from Procrustes relationships between **X** and **Y** data tables with dimensions (n = 40, p =25), where n = nº rows and p = nº columns. **b)** PAMs from Procrustes relationships between **X** and **Y** data tables with dimensions (n = 40, p =45). **c)** PAMs from Procrustes relationships between **X** and **Y** data tables with dimensions (n = 40, p =80). All PAMs used in the multiple comparisons were generated from simulations of 13 pre-Procrustes analysis transformations as described in (Fig. 1 c).