ECOSCIENCE 10 (1) : 110-119 (2003)

Procrustean co-inertia analysis for the linking of multivariate datasets[1]

Stéphane DRAY[2], Daniel CHESSEL & Jean THIOULOUSE, UMR CNRS 5558,
Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, 69622
Villeurbanne Cedex, France, e-mail: dray@biomserv.univ-lyon.fr

Abstract: Procrustes analysis is a method for fitting a set of points to another. These two sets of points are often defined by the measurements of two sets of variables for the same individuals (*e.g.*, measurements of species abundances and environmental variables at the same sites). We present a solution for graphical representation of the results of procrustes analysis when the number of variables in each of the two datasets exceeds two. This method is named procrustean co-inertia analysis because it is based on the joint use of procrustes analysis and co-inertia analysis, which is a coupling method for finding linear combinations of two sets of variables of maximal covariance. It provides better graphical representation of the concordance between the two datasets than classical co-inertia analysis. Moreover, distance matrices can be introduced in the analysis to improve its ecological meaning. Lastly, a randomization test equivalent to PROTEST is proposed as an alternative to the Mantel test. An ecological example is presented to illustrate the method.
Keywords: co-inertia, distance matrix, graphical representation, inter-battery analysis, matrix concordance, procrustes analysis, two-blocks partial least-squares.

Résumé : L'analyse procrustéenne est une méthode permettant d'ajuster un nuage de points sur un autre. Ces deux nuages de points sont souvent définis par deux ensembles de variables mesurées sur les mêmes individus (par exemple, les mesures d'abondances spécifiques et de variables environnementales issues de mêmes sites). Nous présentons une solution pour la représentation graphique des résultats d'une analyse Procruste quand le nombre de variables dans chacun des deux jeux de données est supérieur à deux. Cette méthode est appelée analyse de co-inertie procrustéenne car elle est basée sur l'utilisation conjointe de l'analyse de procruste et de l'analyse de co-inertie. Cette dernière est une méthode visant à trouver des combinaisons linéaires entre deux ensembles de variables de covariance maximale. L'analyse de co-inertie procrustéenne fournit de meilleures représentations graphiques de la concordance de deux jeux de données que l'analyse de co-inertie classique. De plus, des matrices de distances peuvent être introduites dans l'analyse afin d'en améliorer l'interprétation écologique. Enfin, un test de randomisation, équivalent à PROTEST, est proposé comme alternative au test de Mantel. Une illustration écologique est présentée.
Mots-clés : co-inertie, matrice de distance, représentation graphique, analyse inter-batterie, concordance de matrice, analyse Procruste, analyse deux blocs aux moindres carrés potentiels.

Nomenclature: Allardi & Keith, 1991.

Introduction

Procrustes was the leader of a band of brigands in Greek mythology. He was in the habit of putting his victims in a bed and stretching or cutting their limbs in such a way that they "fit" in the bed (Digby & Kempton, 1987). By analogy, procrustes analysis is a method based on rotation, reflection, translation, and dilation of a set of points in order to fit it to another fixed set of points (Gower, 1971). Ecologists often deal with the coupling of species data and environmental data measured over several sites, but the use of procrustes rotation is unusual for this task, and Jackson (1995) claims that "Procrustean methods are used infrequently in ecology. This lack of use likely reflects the previously limited availability of the procedure." Another reason for this lack of interest is probably the fact that the procrustes method provides a measurement of the concordance between the two datasets but produces a graphical representation of this concordance only when there are two variables in each dataset. If there are more than two variables in one dataset, an alternative may be to perform two separate principal component analyses (PCAs) to summarize the main patterns of variation of each dataset and then to study their concordance with procrustean analysis of the first two principal components (Peres-Neto & Jackson, 2001). Original variables may then be plotted by their correlations with principal components. This approach is interesting, but it requires that the main variation of the two datasets be described by neither more nor less than their first two principal components. Graphical representations of species, sites, and environmental variables through biplot or triplot (Gabriel, 1971) are actually the usual procedures used to interpret the results of a multivariate analysis (ter Braak, 1994). Hence, coupling multivariate analyses such as co-inertia analysis (CIA: Dolédec & Chessel, 1994), redundancy analysis (RDA: Rao, 1964), or canonical correspondence analysis (CCA: ter Braak, 1986), which provide graphical representation of the results, are preferred. However, procrustes analysis has demonstrated its usefulness for comparing the results of different ordination methods applied to the same dataset (Digby & Kempton, 1987; Jackson, 1993) or

[1]Rec. 2002-04-24; acc. 2002-09-05.
[2]Author for correspondence.

for studying the concordance of ecological tables (Fasham & Foxton, 1979; Kenkel & Bradfield, 1986; Paszkowski & Tonn, 2000; Olden, Jackson & Peres-Neto, 2001).

In this paper, we present a solution for graphical representation of the results of procrustes analysis. We named this approach procrustean co-inertia analysis (PCIA) because it is based on the principles of procrustes analysis and co-inertia analysis. We present the method and demonstrate its efficiency for the coupling of ecological data and distance matrices. PCIA enables the incorporation of more ecological meaning in statistical analyses than classical methods such as canonical correspondence analysis (Legendre & Anderson, 1999; Legendre & Gallagher, 2001). An ecological example is presented.

Procrustes analysis

To illustrate this method, we analyzed data concerning the cephalofacial growth of a monkey (*Macaca nemestrina*) studied at the ages of 0.9 and 5.77 years, using the spatial coordinates of 72 fixed points (Olshan, Siegel & Swindler, 1982). Data are represented in figure 1a,b. Data concerning the 5.77-year-old monkey have been rotated by 90 degrees for the purposes of this example.

In this paper, procrustes analysis is based on a least-square procrustean rotation, and other types of rotation are not considered. Procrustes analysis aims to transform a set of points to fit another set of points. Let **X** (*n* by *p*) and **Y** (*n* by *q*) be two matrices containing the values of respectively *p* and *q* variables for the same *n* individuals. The two configurations of points are given by matrices **X** and **Y** in an *r*-dimensional space. If the number of variables is the same in the two tables, then $r = p = q$. If the number of variables of **X** is not equal to the number of variables of **Y**, then $r = \max(p,q)$ and columns of zeros are added to the smaller table to match the size of the larger one. The fit between the two configurations of points is measured by

$$d^2(\mathbf{X},\mathbf{Y}) = \|\mathbf{X}-\mathbf{Y}\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{r}\left(x_{ij}-y_{ij}\right)^2 \qquad [1]$$

The fit is simply measured by the square Euclidean distances between the rows of **X** and those of **Y** (square Euclidean norm). Centring tables **X** and **Y** in order to make the centroid of **X** coincide with that of **Y** is important for the rotational process (Figure 1c,d). Taking **X** to be fixed, a rotation **R** is applied to the coordinates of **Y** so that the sum of squared distances,

$$\begin{aligned} m^2_{\mathbf{YX}} &= d^2(\mathbf{X},\hat{\mathbf{Y}}) = \left\|\mathbf{X}-\mathbf{Y}\mathbf{R}^t\right\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{r}\left(x_{ij}-\hat{y}_{ij}\right)^2 \\ &= trace\left[\left(\mathbf{X}-\mathbf{Y}\mathbf{R}^t\right)^t\left(\mathbf{X}-\mathbf{Y}\mathbf{R}^t\right)\right] \\ &= trace\left(\mathbf{X}^t\mathbf{X}\right)+trace\left(\mathbf{Y}^t\mathbf{Y}\right)-2trace\left(\mathbf{R}\mathbf{Y}^t\mathbf{X}\right) \end{aligned} \qquad [2]$$

is minimum. **R** (*r* by *r*) is an orthogonal matrix satisfying $\mathbf{R}^t\mathbf{R}=\mathbf{I}_r=\mathbf{R}\mathbf{R}^t$. If we consider the singular value decomposition of $\mathbf{Y}^t\mathbf{X}=\mathbf{V}\Theta\mathbf{U}^t$, where Θ is a diagonal matrix with the singular values θ_{ii}, then the solution of procrustes analysis is simply expressed as $\mathbf{R}=\mathbf{U}\mathbf{V}^t$ (Figure 1e,f).

In ecological studies, a preliminary rescaling of **X** and **Y** may be necessary if the two tables contain different types of variables. Each column can be standardized to a variance equal to one if the different variables of one table are in different units (*e.g.*, temperature, slope). If all the variables of one table are in the same units (*e.g.*, abundances of species), one can choose to keep the relative variance between columns with a global rescaling. This rescaling can be

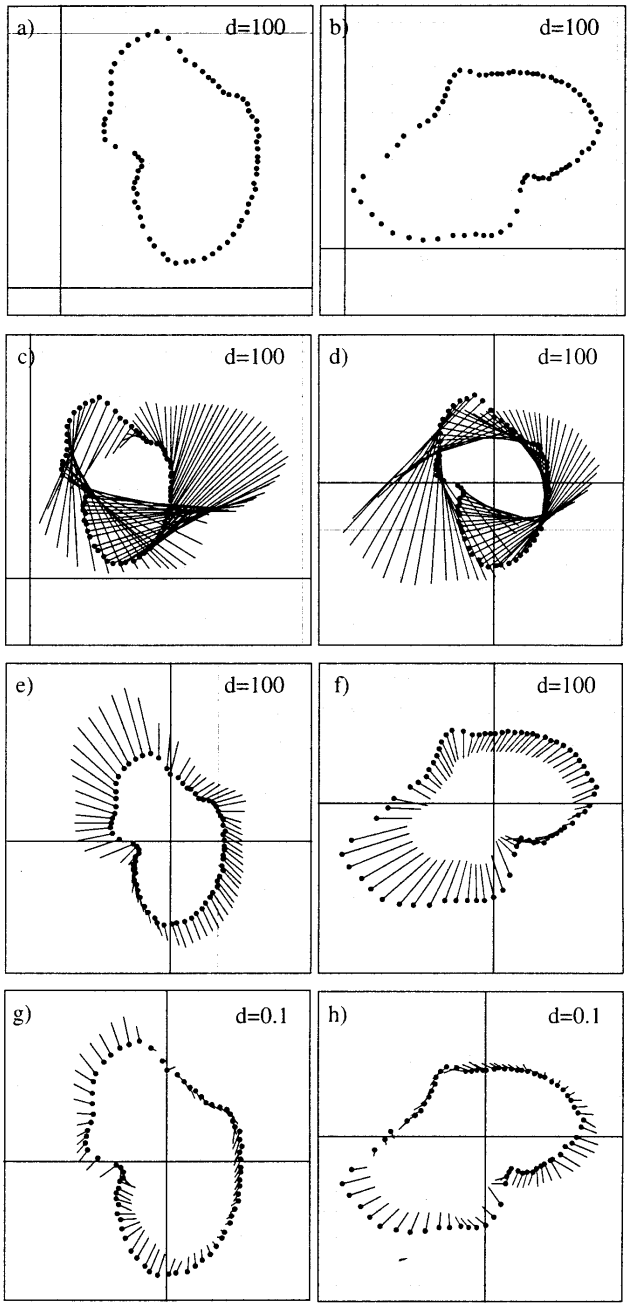

FIGURE 1. Procrustes analysis of monkey data. Original data for 0.9-year-old a) and 5.77-year-old monkey b). Graphical representation of the concordance of the two tables for original data c) and centred data d). Line segments are residuals between scores of 0.9-year-old and 5.77-year-old monkey. Results from procrustes analysis fitting data of 5.77-year-old monkey to data of 0.9-year-old monkey e) and fitting data of 0.9-year-old monkey to data of 5.77-year-old monkey f). Results from rescaled procrustes analysis fitting data of 5.77-year-old monkey to data of 0.9-year-old monkey g) and fitting data of 0.9-year-old monkey to data of 5.77-year-old monkey h). The value of d indicates the size of squares of the grid.

asymmetric if the modifications are defined by only one set of points (Schönemann & Carroll, 1970). In the rest of the paper we have adopted a symmetric rescaling. For this task, Gower (1971) proposes the use of a common scale, transforming the matrices by

$$\frac{\mathbf{X}}{\sqrt{trace\left(\mathbf{X}^t\mathbf{X}\right)}} \text{ and } \frac{\mathbf{Y}}{\sqrt{trace\left(\mathbf{Y}^t\mathbf{Y}\right)}}$$

to have unit sum of squares. This rescaling implies that the procrustes analysis focuses only on variations of shape and removes the variations in size (Figure 1g,h). In order to simplify notations, we name the centred and scaled tables **X** and **Y** in the rest of the paper. $\mathbf{X}_{rot}=\mathbf{XUV}^t$ is the configuration of the points from **X** that fits on **Y** and $\mathbf{Y}_{rot}=\mathbf{YVU}^t$ is the configuration of the points from **Y** that fits on **X**. The bi-representation of the points is easy when $r = 2$ because the two sets of points (**X** and $\mathbf{Y}_{rot}$ or **Y** and $\mathbf{X}_{rot}$) are contained in a plane. When $r > 2$, Peres-Neto and Jackson (2001) proposed to first modify the data (the first two principal components are used in the place of the original data) and then perform the procrustean rotation between the first two principal components of each table.

In the case where $r > 2$ and no preliminary analyses are performed, it is necessary to project the rows of **X** and $\mathbf{Y}_{rot}$ or those of **Y** and $\mathbf{X}_{rot}$ on a plane to visualize the fitting between the two datasets. Several possibilities have been proposed (Mouttet, 1981), such as the principal component analysis (PCA) of the bound tables

$$\left(e.g.\begin{bmatrix}\mathbf{X}_{rot}\\ \mathbf{Y}\end{bmatrix}\right)$$

or the PCA of the average table

$$\left(e.g.\ \frac{1}{2}\left(\mathbf{X}_{rot}+\mathbf{Y}\right)\right)$$

The problem with this approach is that the graphical representations obtained by the PCA of

$$\begin{bmatrix}\mathbf{X}_{rot}\\ \mathbf{Y}\end{bmatrix} \text{ or } \frac{1}{2}\left(\mathbf{X}_{rot}+\mathbf{Y}\right)$$

is different from those obtained by the PCA of

$$\begin{bmatrix}\mathbf{Y}_{rot}\\ \mathbf{X}\end{bmatrix} \text{ or } \frac{1}{2}\left(\mathbf{Y}_{rot}+\mathbf{X}\right)$$

All these approaches provide quite similar but different representations. Theoretically, as rotations do not change the shape of the configurations of $2n$ points (n reference points and n rotated points), the representation in a low-dimensional space must be unique for the two analyses (*i.e.*, **X** and $\mathbf{Y}_{rot}$ or **Y** and $\mathbf{X}_{rot}$).

Co-inertia analysis

Consider now that **X** contains the measurements of p environmental variables and **Y** the abundances of q species in n sites. There is a cloud of the sites in the species hyperspace and another one in the environmental hyperspace. CIA is based on the diagonalization of

$$\mathbf{R}^{1/2}\mathbf{Y}^t\mathbf{DXQX}^t\mathbf{DYR}^{1/2},$$

where **D** is the diagonal matrix of row weights, and **Q** and **R** are respectively the diagonal matrices of column weights of **X** and **Y**. Mathematical details are not described in this paper and can be found in Dolédec and Chessel (1994). The diagonalization of CIA results in a set of site scores in the species hyperspace (*i.e.*, linear combination of species) and a set of site scores in the environmental hyperspace (*i.e.*, linear combination of environmental variables) with maximal square covariance. This analysis, by maximizing the square covariance, maximizes simultaneously the variance of the sites in the species viewpoint, the variance of the sites in the environmental viewpoint, and the square correlation:

$$\text{cov}^2(\mathbf{Xu}, \mathbf{Yv}) = \text{corr}^2(\mathbf{Xu}, \mathbf{Yv})\text{var}(\mathbf{Xu})\text{var}(\mathbf{Yv}) \qquad [3]$$

To represent the concordance between the two datasets, the two sets of site scores can be normalized, and their correlations can then be plotted. But because the scores are normalized after the analysis, this representation is not the best, so it is the correlations that are plotted, while the analysis maximizes the covariances. Torre and Chessel (1995) present an extension of CIA in the case of fully matched tables (*i.e.*, same individuals and same variables). This analysis can be used, for example, to study the temporal stability between two tables containing measurements of the same variables at the same sites at two dates. For this kind of table, the sites are represented twice in the same hyperspace, so CIA finds only one co-inertia axis instead of the pair of co-inertia axes (one in each hyperspace) found in the usual case. The co-inertia axis maximizes the covariance of the projected coordinates of the two multidimensional clouds in the same hyperspace. If we consider two totally matched tables **A** and **B**, the co-inertia axes are the eigenvectors of

$$\frac{1}{2}\left(\mathbf{A}^t\mathbf{B}+\mathbf{B}^t\mathbf{A}\right).$$

Procrustean co-inertia analysis

The CIA of fully matched tables can be applied to **X** and $\mathbf{Y}_{rot}$ because these two datasets are contained in the same hyperspace. It results in finding the eigenvectors of

$$\frac{1}{2}\left(\mathbf{Y}_{rot}^t\mathbf{X}+\mathbf{X}^t\mathbf{Y}_{rot}\right)=\frac{1}{2}\left(\mathbf{UV}^t\mathbf{Y}^t\mathbf{X}+\mathbf{X}^t\mathbf{YVU}^t\right)$$

$$=\frac{1}{2}\left(\mathbf{UV}^t\mathbf{V}\Theta\mathbf{U}^t+\mathbf{U}\Theta\mathbf{V}^t\mathbf{VU}^t\right)=\mathbf{U}\Theta\mathbf{U}^t$$

In the same way, if **Y** is fixed, the CIA of totally matched tables **Y** and $\mathbf{X}_{rot}$ finds the eigenvectors of $\mathbf{V}\Theta\mathbf{V}^t$. The environmental rows score data are obtained by projecting the rows of **X** on **U** or those of $\mathbf{X}_{rot}$ on **V** and are contained in

$$\mathbf{S}_\mathbf{X} = \mathbf{X}_{rot}\mathbf{V} = \mathbf{XUV}^t\mathbf{V} = \mathbf{XU}.$$

In the same way, the species rows score date are obtained by projecting the rows of **Y** on **V** or those of $\mathbf{Y}_{rot}$ on **U** and are contained in

$$\mathbf{S}_\mathbf{Y} = \mathbf{Y}_{rot}\mathbf{U} = \mathbf{YVU}^t\mathbf{U} = \mathbf{YV}.$$

Variables can be represented by the coefficients contained in **U** (environmental variables) and **V** (species). Hence, these two analyses (**X** and $\mathbf{Y}_{rot}$ or **Y** and $\mathbf{X}_{rot}$) provide the same representation and the same quality of representation. So, PCIA provides a common solution to the problem of

representation in procrustes analysis. PCIA finds co-inertia axes maximizing the covariance between linear combinations of **X** and $\mathbf{Y}_{rot}$ (or **Y** and $\mathbf{X}_{rot}$). The appendix presents a complete example of computation of PCIA.

RANDOMIZATION TEST

PROTEST (Jackson, 1995) is a permutation test available for procrustes analysis. If the two tables have been rescaled,

$$m^2_{\mathbf{YX}} = m^2_{\mathbf{XY}} = m^2 = 2\left(1 - \sum_{k=1}^{r} \theta_k\right)$$

is a goodness-of-fit statistic (θ_k are the singular values of $\mathbf{Y}^t\mathbf{X}$). PROTEST is based on the statistic

$$m_{12} = 1 - \left(\sum_{k=1}^{r} \theta_k\right)^2$$

(*i.e.*, based on sum of singular values of $\mathbf{Y}^t\mathbf{X}$) and consists of computing new values of m_{12} after permuting entire rows of one table. The observed value is then compared to the set of values obtained by permutation. The hypothesis that there is no link between the two tables is tested against the hypothesis that there is a significant common structure. A permutation test of the same hypothesis is also available in CIA, based on the total co-inertia, which is simply, except for a constant,

$$\alpha^2 = \sum_{k=1}^{r} {\theta_k}^2$$

(*i.e.*, sum of eigenvalues of $\mathbf{Y}^t\mathbf{X}\mathbf{X}^t\mathbf{Y}$). The co-inertia test is based on

$$\sum_{k=1}^{r} \mathrm{cov}^2\left(\mathbf{X}\mathbf{u}_k, \mathbf{Y}\mathbf{v}_k\right),$$

while PROTEST is based on

$$\sum_{k=1}^{r} \mathrm{cov}\left(\mathbf{X}\mathbf{u}_k, \mathbf{Y}\mathbf{v}_k\right).$$

The test of CIA is based on the computation of *trace*($\mathbf{Y}^t\mathbf{X}\mathbf{X}^t\mathbf{Y}$) and is strictly equivalent to a test based on the RV coefficient (Heo & Gabriel, 1997). PROTEST is based on the residuals of the analysis, so a low value indicates a link, whereas the RV-test measures the co-structure of the two matrices, so a high value indicates a link. In the case of two distance matrices, PROTEST is as powerful or more powerful than the usual Mantel test (Peres-Neto & Jackson, 2001). Empirical experiences based on many numerical examples show that the results from PROTEST and RV-tests are very similar.

INTRODUCTION OF ECOLOGICAL DISTANCES

The Euclidean distance used in PCA and RDA and the Chi-square distance of CA and CCA are not always appropriate for the analysis of ecological data. In the same way, PCIA on raw data (abundance of species) lacks of ecological consideration for the analysis of the relationships between sites and species. The introduction of distances that are considered better for ecological data, such as Bray-Curtis distance, into ordination methods is therefore a significant improvement. Distance-based redundancy analysis (Legendre & Anderson, 1999; McArdle & Anderson, 2001) allows one to pair a distance matrix based on species data to a table of environmental or experimental data. This approach consists of computing a matrix of distances among sites from table **Y**, and the principal coordinates of this distance matrix are computed with principal coordinate analysis (PCoA), possibly including a correction for negative eigenvalues. The data describing the experiment are then linked to the principal coordinates through RDA. This approach allows the introduction of ecological distances into ordination methods, but the species cannot be plotted on the graphical representation because table **Y** has been replaced by its principal coordinates. Legendre and Gallagher (2001) propose that species score be plotted through the use of weighted correlation with principal coordinates. Another alternative is available to avoid this problem: table **Y** can be transformed so that the Euclidean distance among sites computed with the transformed table **Y** is ecologically meaningful (Legendre & Gallagher, 2001). For example, if we consider the transformation,

$$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{q} y_{ij}^2}}$$

then Euclidean distances measured between sites with **Y'** correspond to Chord distances measured with **Y**. Biplot and triplot are then available with ecological distances. However, some interesting distances that are not Euclidean based (*e.g.*, Bray-Curtis: Legendre & Legendre, 1998) cannot be obtained using a simple transformation of **Y**. One of the advantages of PCIA is that it can also be used to pair transformed data or to pair any distance matrix to a set of environmental variables. Moreover, the robustness of CIA concerning the number of variables relative to the number of individuals allows the coupling of two distance matrices. In this case, the original data are used to compute distances between sites from table **X** and from table **Y**. Two separate PCoAs are then applied to the two distance matrices, and two sets of principal coordinates are then obtained. The PCoA of an *n* by *n* distance matrix often produces *n-1* principal coordinates, so the two tables of principal coordinates have *n* rows and *n-1* columns. The two sets of principal coordinates can be linked by PCIA. Indeed, CIA is very robust concerning the number of variables, unlike redundancy analysis, which requires that there be few explanatory variables compared to the number of individuals and is therefore not appropriate in this case. So, PCIA is the only alternative for the coupling of two distance matrices.

Ecological illustration

The following example (Verneaux, 1973) is proposed to illustrate the PCIA approach. It concerns the relationships between 11 environmental variables and the distributions of 27 fish species in the Doubs river (France). Environmental variables were measured to describe the morphological aspects of the river (distance to the source, altitude, slope, flow) and the water quality (pH, calcium, phosphate, nitrate, ammonium, oxygen, biological demand in oxygen) at 29 sites. This dataset has previously been analyzed by CCA (Chessel, Lebreton & Yoccoz, 1987). Data were centred by column, and environmental variables were scaled to unit variance. A barplot of singular values indicated that PCIA based on original data identifies a two-axes structure (Figure 2a). The first axis is related with the upstream-

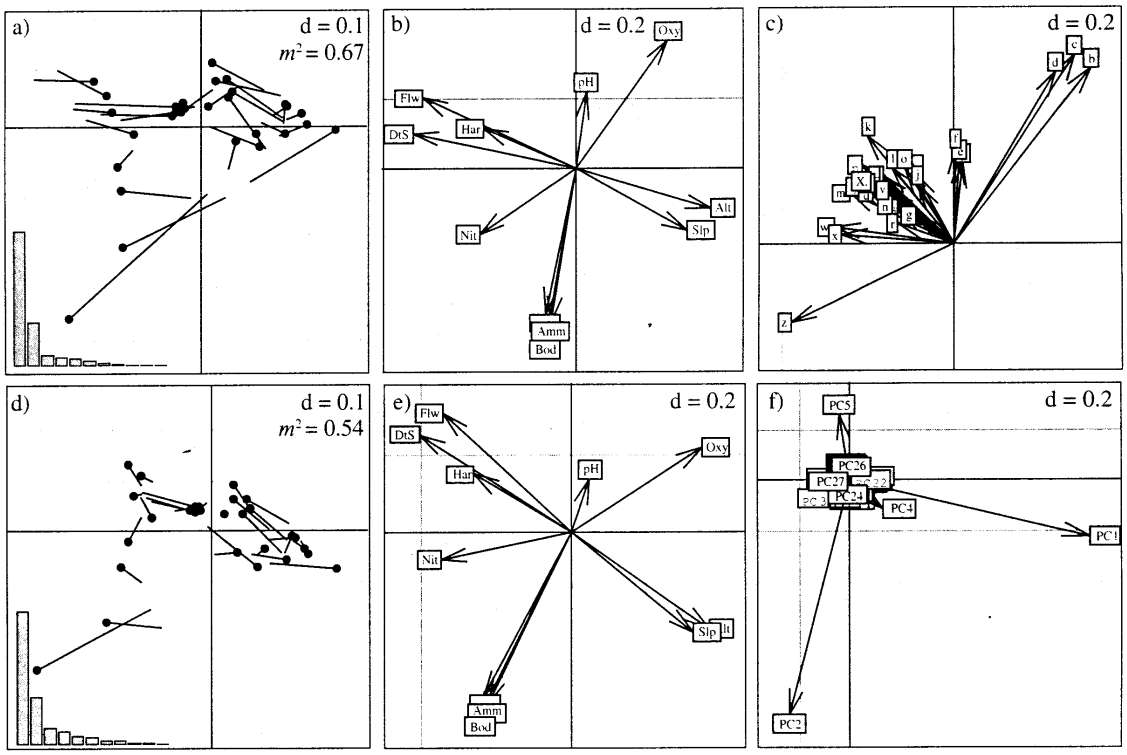

FIGURE 2. Procrustean co-inertia analysis of fish data. Coupling of original data (a-c) and coupling of Bray-Curtis distance matrix based on species data with environmental variables (d-f). (a,d) concordance between sites. Line segments are residuals between scores obtained from the environmental matrix and the fish species matrix. Barplot of singular values and m^2 statistic are indicated. (c,f) coefficients of fish species or coefficients of their principal coordinates. (b,e) coefficients of environmental variables. The value of d indicates the size of squares of the grid.

downstream structure of the river (altitude and slope decreasing, flow increasing with the distance to the source). The second axis is a pollution factor (ammonium, phosphate, nitrate), which produces a decrease of the species richness in several sites (Figure 2b). Species are separated essentially along the upstream-downstream structure (trout, minnow, and loach versus the others), and bleak seems to be unaffected by pollution (Figure 2c). The fit between the two tables is good (m^2 = 0.67), and high residuals essentially concern polluted sites. We constructed a Bray-Curtis distance matrix from the fish data and linked it to environmental variables. Use of the Bray-Curtis distance (Figure 2d,f) improves the fit of the analysis (decrease of the m^2 statistic from 0.67 to 0.54), and the two axes corresponding to the upstream-downstream structure and to the pollution factor are also identified. The improvement of the fit is especially noticeable for the polluted sites, because the Bray-Curtis distance is more sensitive to variations in relative composition and less sensitive to variations in absolute abundance than the Euclidean distance. Hence, the effects of pollution on species richness have low importance in this analysis. Note that the use of a distance matrix implies that species are replaced by their principal coordinates on the plot (Figure 2f). Lastly, we computed the Bray-Curtis distance matrix based on species data and the Euclidean distance matrix with environmental data. PCIA was then performed between the two sets of principal coordinates of these two distance matrices. The results are obviously the same as those obtained when coupling environmental variables to the Bray-Curtis distance, but environmental variables are replaced by their principal coordinates on the plot (Figure 3c). This example is only given to illustrate the use of the different permutation tests, because the Mantel test can only be applied in the case of two distance matrices. Permutation tests have been performed for this last case (Figure 3d,f). The three tests were significant (p-value < 0.0001), but the position of the observed values compared to the distribution of the simulated values indicates that the Mantel test seems to have less power than the two others.

Discussion

PCIA is demonstrably efficient for linking multivariate datasets. The method accepts various kinds of data, such as raw data, modified data, or distance matrices (Figure 4). This is very promising. Like the approach introduced by distance-based RDA, it will enable the inclusion of greater ecological meaning in ordination methods. PCIA provides a convenient solution for graphical representation of results of procrustes analysis when there are more than two variables in one dataset. Representation of the results obtained by the approach of Peres-Neto and Jackson (2001) is similar to that obtained by PCIA when the structure of each table can be well summarized by the first two principal components. If the structure of one table is more or less complex, our approach is better, because it takes into account the global structure of the data and not only the part of the structure that can be summarized on a plane. Moreover, representation of the concordance between the two datasets in PCIA is better than the representation used in classical CIA, because

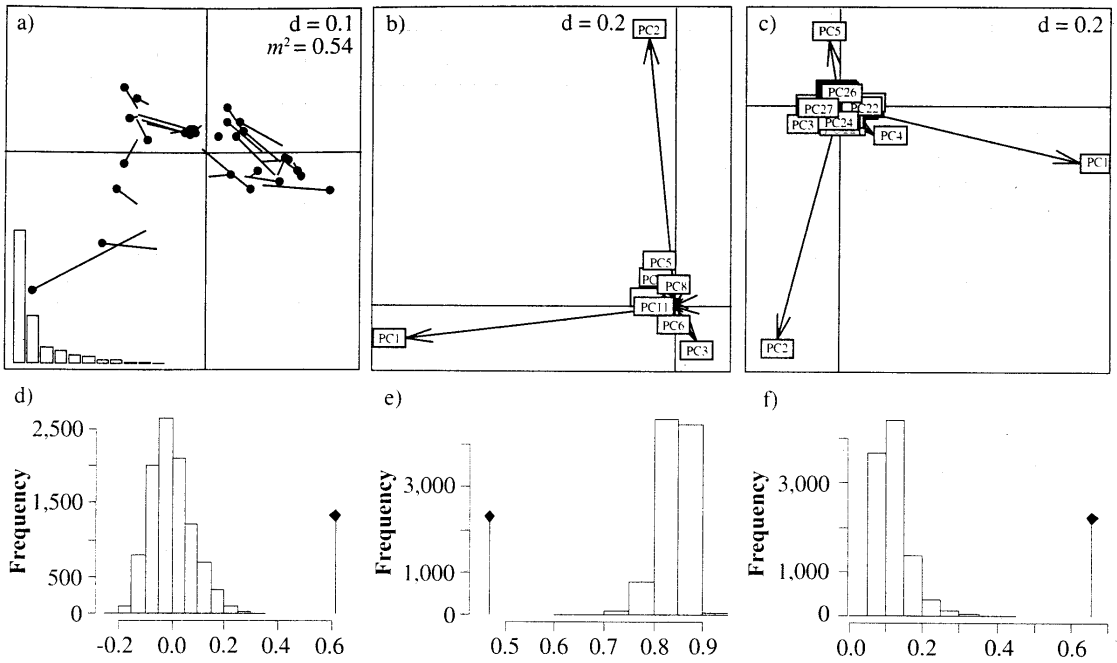

FIGURE 3. Procrustean co-inertia analysis between two distance matrices (Bray-Curtis distance for species data and Euclidean distance for environmental data). a) Concordance between sites. Line segments are residuals between scores obtained from the environmental matrix and the fish species matrix. Barplot of singular values and m^2 statistic are indicated. b) Coefficients of principal coordinates of the distance matrix based on species data. c) Coefficients of principal coordinates of the distance matrix based on environmental variables. (d-f) Results of randomization tests (observed value is indicated by the vertical line) with 9,999 permutations: d) Mantel test ($p < 0.0001$), e) PROTEST ($p < 0.0001$), and f) RV-test ($p < 0.0001$). The value of d indicates the size of squares of the grid.

the two systems of sites are in the same hyperspace and no rescaling is needed. Classical CIA is more general and can accept qualitative variables and various weighting of sites and of variables. PCIA is devoted only to the analysis of quantitative variables and cannot include weights, although the possibility of weighting rows and columns in PCIA is conceivable. In the case of quantitative environmental variables, PCIA appears to be a good alternative, whereas when table **X** contains the design of an experiment, methods based on RDA (*e.g.*, distance-based RDA) are more suitable, because the variables in table **X** are fixed by the experiment and it is then necessary to take into account this dissymmetry. Thus, classical CIA, distance-based RDA, and PCIA are three complementary tools.

PCIA aims to find linear combinations of maximal covariance. The computation of such axes, in the case of two sets of normalized variables, originated in the inter-battery analysis of Tucker (1958). CIA is a generalization of this method for various kinds of data, and it adds geometric interpretation, allowing graphical representation of the results. The two-block partial least-squares analysis (2B-PLS: Rohlf & Corti, 2000), which is similar to Tucker's approach, also aims to find combinations of variables of maximal covariance. PCIA and all these other methods perform exactly the same computation (diagonalization of the same matrix), but the major advantage of PCIA is that it produces graphical representations of the concordance of the two datasets. While other methods (*e.g.*, 2B-PLS) produce one plot for each dataset, PCIA enables the two datasets to be plotted on one graphic (Figure 2a,d). This is very helpful for interpreting the results, because the concordance for each observation is represented.

In the context of unimodal response species curves, canonical correspondence analysis has proved its efficiency for separating species niche centroids. PCIA used with the Chi-square distance can also be used in the unimodal context as an alternative to CCA, avoiding CCA's much-debated step of weighting sites by their species richness (Dolédec, Chessel & Gimaret-Carpentier, 2000). Moreover, CCA is more restricted concerning the number of variables relative to the number of sites. CCA is based on multivariate regression and requires a low number of explanatory variables. In addition, PCIA is more general than CCA and can be used in many contexts other than the study of species-environment relationships. Concerning randomization procedures, the RV-test of PCIA provides results similar to those of PROTEST, but is less time-consuming, because PROTEST performs a singular value decomposition for each step, whereas the RV-test is based only on computation of the trace of a matrix.

Lastly, a weighted version of PCIA can be used to incorporate weights of individuals and variables, as in classical CIA. Moreover, CIA has been extended to the case of coupling k tables ($k > 2$) under the name of multiple co-inertia analysis (Chessel & Hanafi, 1996), to the analysis of the concordance of k tables with a reference table (Lafosse & Hanafi, 1997), and to the analysis of k pairs of tables (Simier *et al.*, 1999). These different approaches probably will help to provide graphical representation for generalized procrustes analysis (Gower, 1975) in the case of more than two tables.

All analyses and graphical representations were made with ADE-4 software (Thioulouse *et al.*, 1997), freely distributed at http://pbil.univ-lyon1.fr/ADE-4/ADE-4.html.

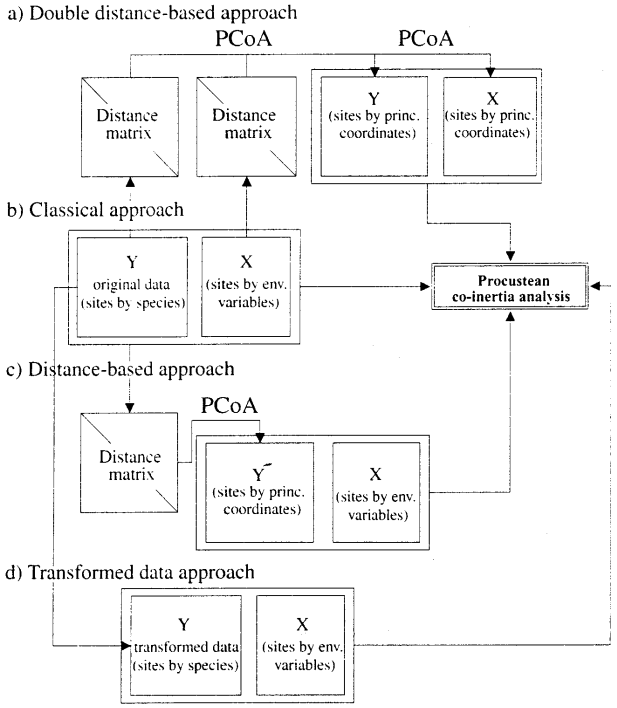

FIGURE 4. Different approaches available for procrustean co-inertia analysis. a) Double distance-based approach. The two original tables are used to create two between-sites distance matrices. Principal coordinate analyses are then performed, and two new tables (sites by principal coordinates) are created. These two tables can be linked by PCIA. b) Classical approach. The original data are directly linked by PCIA. c) Distance-based approach. Species matrix is used to construct a distance matrix. Principal coordinates obtained by PCoA are then linked to environmental matrix by PCIA. d) Transformed data approach. Species matrix is modified so that the Euclidean distances measured between sites of the modified table correspond to distances of ecological interest (*e.g.*, Chi-square). The transformed matrix is linked to environmental variables by PCIA.

A new version, developed in the form of an *R* package (Ihaka & Gentleman, 1996), is available.

Acknowledgements

We wish to thank M. R. T. Dale and two anonymous reviewers, whose suggestions and comments enabled us to improve the first version of this text.

Literature cited

Allardi, J. & P. Keith, 2001. Atlas des poissons d'eau douce de France. Muséum National d'Histoire Naturelle, Paris.

Chessel, D. & M. Hanafi, 1996. Analyse de la co-inertie de *K* nuages de points. Revue de Statistique Appliquée, 44: 35-60.

Chessel, D., J. D. Lebreton & N. Yoccoz, 1987. Propriétés de l'analyse canonique des correspondances. Une utilisation en hydrobiologie. Revue de Statistique Appliquée, 35: 55-72.

Digby, P. G. N. & R. A. Kempton, 1987. Multivariate Analysis of Ecological Communities. Chapman and Hall, London.

Dolédec, S. & D. Chessel, 1994. Co-inertia analysis: An alternative method for studying species-environment relationships. Freshwater Biology, 31: 277-294.

Dolédec, S., D. Chessel & C. Gimaret-Carpentier, 2000. Niche separation in community analysis: A new method. Ecology, 81: 2914-2927.

Fasham, M. J. R. & P. Foxton, 1979. Zonal distribution of pelagic decapods (Crustacea) in the eastern North Atlantic and its relation to the physical oceanography. Journal of Experimental Marine Biology and Ecology, 37: 225-253.

Gabriel, K. R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58: 453-467.

Gower, J. C., 1971. Statistical methods of comparing different multivariate analyses of the same data. Pages 138-149 *in* P. Tautu (ed.). Mathematics in the Archaeological and Historical Sciences. Edinburgh University Press, Edinburgh.

Gower, J. C., 1975. Generalized procrustes analysis. Psychometrika, 40: 33-51.

Heo, M. & K. R. Gabriel, 1997. A permutation test of association between configurations by means of the RV coefficient. Communications in Statistics, Simulation and Computation, 27: 843-856.

Ihaka, R. & R. Gentleman, 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 5: 299-314.

Jackson, D. A., 1993. Multivariate analysis of benthic invertebrate communities: The implication of choosing particular data standardizations, measures of association, and ordination methods. Hydrobiologia, 268: 9-26.

Jackson, D. A., 1995. PROTEST: A PROcrustean Randomization TEST of community environment concordance. Écoscience, 2: 297-303.

Kenkel, N. C. & C. E. Bradfield, 1986. Epiphytic vegetation on *Acer macrophyllum*: A multivariate study of species-habitat relationships. Vegetatio, 68: 43-53.

Lafosse, R. & M. Hanafi, 1997. Concordance d'un tableau avec *K* tableaux : définition de K + 1 uples synthétiques. Revue de Statistique Appliquée, 45: 111-126.

Legendre, P. & M. J. Anderson, 1999. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. Ecological Monographs, 69: 1-24.

Legendre, P. & E. D. Gallagher, 2001. Ecologically meaningful transformations for ordination of species data. Oecologia, 129: 271-280.

Legendre, P. & L. Legendre, 1998. Numerical Ecology, 2nd edition. Elsevier Science, Amsterdam.

McArdle, B. H. & M. J. Anderson, 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. Ecology, 82: 290-297.

Mouttet, F., 1981. Comparaison de tableaux par la méthode Procuste. Université Paris VI, Paris.

Olden, J. D., D. A. Jackson & P. R. Peres-Neto, 2001. Spatial isolation and fish communities in drainage lakes. Oecologia, 127: 572-585.

Olshan, A. F., A. F. Siegel & D. R. Swindler, 1982. Robust and least-squares orthogonal mapping: Methods for the study of cephalofacial form and growth. American Journal of Physical Anthropology, 59: 131-137.

Paszkowski, C. A. & W. M. Tonn, 2000. Community concordance between the fish and aquatic birds of lakes in northern Alberta, Canada: The relative importance of environmental and biotic factors. Freshwater Biology, 43: 421-437.

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

Rao, C. R., 1964. The use and interpretation of principal component analysis in applied research. Sankhya A, 26: 329-359.

Rohlf, F. J. & M. Corti, 2000. Use of two-block partial least-squares to study covariation in shape. Systematic Biology, 49: 740-753.

Schönemann, P. H. & R. M. Carroll, 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, 35: 245-256.

Simier, M., L. Blanc, F. Pellegrin & D. Nandris, 1999. Approche simultanée de *K* couples de tableaux : Application à l'étude des relations pathologie végétale-environment. Revue de Statistique Appliquée, 47: 31-46.

ter Braak, C. J. F., 1986. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. Ecology, 67: 1167-1179.

ter Braak, C. J. F., 1994. Canonical community ordination. Part I: Basic theory and linear methods. Écoscience, 1: 127-140.

Thioulouse, J., D. Chessel, S. Dolédec & J. M. Olivier, 1997. ADE-4: A multivariate analysis and graphical display software. Statistics and Computing, 7: 75-83.

Torre, F. & D. Chessel, 1995. Co-structure de deux tableaux totalement appariés. Revue de Statistique Appliquée, 43: 109-121.

Tucker, L. R., 1958. An inter-battery method of factor analysis. Psychometrika, 23: 111-136.

Verneaux, J., 1973. Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Université de Besançon, Besançon.

APPENDIX I. Species code for the 27 fish species

Code	English name	Scientific name
a	chub	*Cottus gobio*
b	trout	*Salmo trutta fario*
c	minnow	*Phoxinus phoxinus*
d	loach	*Nemacheilus barbatulus*
e	grayling	*Thymallus thymallus*
f	soufie	*Telestes soufia agassizi*
g	nase	*Chondrostoma nasus*
h	toxostome	*Chondrostoma toxostoma*
i	dace	*Leuciscus leuciscus*
j	chub	*Leuciscus cephalus cephalus*
k	barbel	*Barbus barbus*
l	spirlin	*Spirlinus bipunctatus*
m	gudgeon	*Gobio gobio*
n	pike	*Esox lucius*
o	perch	*Perca fluviatilis*
p	bitterling	*Rhodeus amarus*
q	pumpkinseed	*Lepomis gibbosus*
r	red-eye rudd	*Scardinius erythrophtalmus*
s	carp	*Cyprinus carpio*
t	tench	*Tinca tinca*
u	bream	*Abramis brama*
v	black bullhead	*Ictalurus melas*
w	ruffe	*Acerina cernua*
x	roach	*Rutilus rutilus*
y	silver bream	*Blicca bjoerkna*
z	bleak	*Alburnus alburnus*
+	eel	*Anguilla anguilla*

APPENDIX II. Environmental variables recorded at 29 sites and codes used as labels in the figures.

No.	Code	Environmental variable
1	DtS	Distance to the source (km × 10)
2	Alt	Altitude (m)
3	Slp	Slope (‰ × 10)
4	Flw	Minimum flow ($m^3\ s^{-1} \times 100$)
5	pH	pH (× 10)
6	Har	Total hardness (mg l^{-1} of calcium)
7	Pho	Phosphate (mg $l^{-1} \times 100$)
8	Nit	Nitrate (mg $l^{-1} \times 100$)
9	Amm	Ammonium (mg $l^{-1} \times 100$)
10	Oxy	Dissolved oxygen (mg $l^{-1} \times 10$)
11	Bod	5-days Biological Oxygen Demand (mg $l^{-1} \times 10$)

APPENDIX III. Numerical example of PCIA

Consider the two matrices:

$$\mathbf{X} = \begin{bmatrix} -0.02 & 0.24 \\ -0.74 & 1.84 \\ 0.35 & 1.99 \\ -0.27 & 2.21 \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} -0.90 & -0.01 & -0.90 \\ 0.09 & 0.23 & 0.12 \\ 1.48 & 0.46 & 1.74 \\ 1.22 & 0.70 & 1.66 \end{bmatrix}$$

The first step consists of centring and normalizing $\left(\frac{\mathbf{X}}{\sqrt{trace\left(\mathbf{X}^t\mathbf{X}\right)}} \right)$ the two matrices:

$$\mathbf{X} = \begin{bmatrix} 0.09 & -0.76 \\ -0.33 & 0.15 \\ 0.30 & 0.24 \\ -0.06 & 0.37 \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} -0.46 & -0.12 & -0.52 \\ -0.13 & -0.04 & -0.18 \\ 0.34 & 0.04 & 0.37 \\ 0.25 & 0.12 & 0.34 \end{bmatrix}$$

The singular value decomposition $\mathbf{Y}^t\mathbf{X}=\mathbf{V}\Theta\mathbf{U}^t$ is then performed and results in:

$$\Theta = \begin{bmatrix} 0.80 & 0 \\ 0 & 0.02 \end{bmatrix}, \mathbf{U} = \begin{bmatrix} -0,17 & -0.99 \\ -0.99 & 0.17 \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} -0.65 & -0.09 \\ -0.17 & 0.98 \\ -0.74 & -0.15 \end{bmatrix}$$

Rotated matrices are simply computed by $\mathbf{X}_{rot}=\mathbf{XUV}^t$ and $\mathbf{Y}_{rot}=\mathbf{YVU}^t$.

$$\mathbf{X}_{rot} = \begin{bmatrix} -0.46 & -0.34 & -0.51 \\ 0.03 & 0.36 & 0.02 \\ 0.21 & -0.20 & 0.25 \\ 0.22 & 0.18 & 0.24 \end{bmatrix} \text{ and } \mathbf{Y}_{rot} = \begin{bmatrix} -0.12 & -0.70 \\ -0.04 & -0.22 \\ 0.13 & 0.48 \\ 0.03 & 0.44 \end{bmatrix}$$

APPENDIX III. continued.

The concordance between the rows of the two tables can be represented with the two systems of scores $\mathbf{S_X}=\mathbf{XU}$ and $\mathbf{S_Y}=\mathbf{YV}$.

$$\mathbf{S_X} = \begin{bmatrix} 0.74 & -0.21 \\ -0.10 & 0.35 \\ -0.29 & -0.25 \\ -0.35 & 0.12 \end{bmatrix} \text{ and } \mathbf{S_Y} = \begin{bmatrix} 0.71 & 0.00 \\ 0.22 & 0.00 \\ -0.50 & -0.05 \\ -0.44 & 0.04 \end{bmatrix}$$

Variables of **X** and **Y** are represented respectively by the coefficients contained in **U** and **V**.