American Journal of Theoretical and Applied Statistics
2022; 11(1): 36-44
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20221101.16
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Multivariate Analysis of a Sequence of Paired Data Matrices: Succesive and Simultaneous Approaches

Rodnellin Onesime Malouata[1,*], Chedly Gélin Louzayadio[2], Bernédy Nel Messie Kodia Banzouzi[3]

[1]Department of Licence, Higher Institute of Management, Marien Ngouabi University, Brazzaville, Congo
[2]Department of Licence, Faculty of Economics, Marien Ngouabi University, Brazzaville, Congo
[3]Department of Mathematics, Faculty of Science and Technology, Marien Ngouabi University, Brazzaville, Congo

Email address:

onesimero@gmail.com (R. O. Malouata), gelinlouzayadio@gmail.com (C. G. Louzayadio), bernedy.kodia@umng.cg (B. N. M. K. Banzouzi)
*Corresponding author

To cite this article:

Rodnellin Onesime Malouata, Chedly Gélin Louzayadio, Bernédy Nel Messie Kodia Banzouzi. Multivariate Analysis of a Sequence of Paired Data Matrices: Succesive and Simultaneous Approaches. *American Journal of Theoretical and Applied Statistics*. Vol. 11, No. 1, 2022, pp. 36-44. doi: 10.11648/j.ajtas.20221101.16

Received: August 20, 2021; **Accepted:** September 7, 2021; **Published:** February 14, 2022

Abstract: The relationship between two data matrices has been studied in the interbattery factor analysis. When two data matrices are partitioned in rows, the relationship between two data matrices has been studied in the STATICO method. The main advantage of this method is the optimality of the compromise of co-structures. It is well known that the weighting coefficients of the compromise may be contrary sign in some cases and make it uninterpretable. Thus, many multivariate data analysis methods have been developed, particularly those designed to tackle the fundamental issue: the description of the relationships between two data matrices. This can be studied by successive modeling approaches as well as by a simultaneous modeling approach. These methods are based on co-inertia and can be reduced to finding the maximum, minimum, or other critical values of a ratio of quadratic forms. However, all these methods are successive. In this paper, we propose two algorithms. The first one called sDO-CCSWA (successive Double-Common Component and Specific Weight Analysis) maximizes the sum of squared covariances, by first finding the best pair-component solution, and repeating that process in the respective residual spaces. The sDO-CCSWA is a new monotonically convergent algorithm obtained by searching for a fixed point of the stationary equations. The second approach is a simultaneous algorithm (DO-CCSWA) which maximizes the sum of squared covariances.

Keywords: Interbattery Factor Analysis, STATICO, Common Component and Specific Weight Analysis

1. Introduction

Interbattery factor analysis [16] allows to investigate the relationships between two data sets X and Y of p and q variables observed on n individuals. Interbattery factor analysis consists of finding components $c_X = Xu$ and $c_Y = Yv$, such that their covariance is maximal. Thus, Interbattery factor analysis consists of maximizing:

$$f(u,v) = cov(Xu, Yv), \tag{1}$$

subject to the normalization constraints

$$||u|| = ||v|| = 1 \tag{2}$$

Another way to study the relationship between two data matrices consists of finding the component $c_X = Xu$ which characterizes the set covariances of the variables y_l $(l = 1, \ldots, q)$ of Y, and $c_Y = Yv$ characterizes the set covariances of the variables x_h $(h = 1, \ldots, p)$ of X. This way can be found in [11]. Thus, they maximize the following criterion

$$f(u,v) = \left[\sum_{h=1}^{p} cov^2(Yv, x_h)\right]\left[\sum_{l=1}^{q} cov^2(Xu, y_l)\right] \tag{3}$$

subject to the normalization constraints (2). This criterion is equivalent to maximize

$$f(u,v) = (u'Ku)(v'Hv) \tag{4}$$

where $K = V_{XY}V_{YX}$ and $H = V_{YX}V_{XY}$ are two positive semidefinite symmetric matrices [12]. This method can be further generalized by applying it to two different sets of matrices, centered column wise, which we shall refer to as $X = [X_1', \ldots, X_i', \ldots, X_M']'$ and $Y = [Y_1', \ldots, Y_i', \ldots, Y_M']'$, measured on different groups of individuals. $n_i \times p$ data matrix X_i and $n_i \times q$ data matrix Y_i are called a group. The number of individuals of each pair of groups can differ from one pair of groups to another. The main aim is to investigate the stability of the relationships between pairs of groups of variables. To study the stability of relationships between several pairs of matrices, [14] has proposed the STATICO method. STATICO is a Partial Triadic Analysis [15] on the series of cross product matrices obtained by crossing the two data matrices of a pair. It benefits from the three-steps computation scheme of STATIS-like methods (interstructure, compromise, intrastructure). It is well known that the weighting coefficients of the compromise may be contrary sign in some cases. Thus, alternative methods have been proposed which maximize the sum of covariances and the sum of squared covariances between components, with orthonormality constraints on the components. For instance, [9] has proposed sCIA3 (succesive co-inertia analysis 3), which maximizes

$$f(u,v) = \left[\sum_{i=1}^{M}\sum_{h=1}^{p} cov^2(x_{ih}, Y_i v)\right]\left[\sum_{i=1}^{M}\sum_{l=1}^{q} cov^2(X_i u, y_{il})\right] \tag{5}$$

Subject to the normalization constraints (2). This criterion is equivalent to maximize

$$f(u,v) = \sum_{i=1}^{M}(u' K_i u)\sum_{i=1}^{M}(v' H_i v) \tag{6}$$

Subject to the constraints $||u|| = ||v|| = 1$, with $K_i = V_{X_iY_i}V_{Y_iX_i}$ and $H_i = V_{Y_iX_i}V_{X_iY_i}$. Authors proposed an orthogonal approach to obtain partial components of the matrices X_i and Y_i. All these methods find loading vectors u and v which can be seen as common components as is the case with [5].

In addition, the criteria developed by [9] and [14] are successive approaches. That is, after finding an optimal pair of first components, these components are fixed and the second components are found in the respective residual spaces, orthogonal to the first components. This approach has its merits in finding the single most common component of all sets. However, when more than one component is desired, components other than the first one may be much easier to discern when a simultaneous approach is adopted. The situation is very unlike standard Principal Component Analysis, where the components are nested and the explained variance can be redistributed by rotation. These methods possess neither of these properties, which makes it worthwhile considering a simultaneous alternative [10]. In this paper, a simultaneous algorithm called DO-CCSWA (Double-Common Component and Specific Weight Analysis) is proposed. The DO-CCSWA criterion maximizes the product of two sums of squared covariances. A successive approach (sDO-CCSWA) which maximizes the product of two sums of squared covariances is also proposed. This approach is a sequence of the evolution of the criterion (6). It is important to note that the optimization criteria are considered as a double common component and specific weight analysis (CCSWA) proposed by [5]. It is also worth mentioning that CCSWA and HPCA (Hierarchical principal component analysis described in [17]) are two equivalent methods [4].

Finally, we will conclude this paper with a detailed analyses of a practical example. This paper is organized as follows: In sections 2 and 3, we will propose the sDO-CCSWA method and its simultaneous approach. In section 4, an overview of application of DO-CCSWA for two matrices is given.

2. Successive Approach of DO-CCSWA

In this section, we propose a method for analyzing the relationships between two sets of variables. The sDO-CCSWA method consists of optimizing the following criterion:

$$\text{Maximize} \quad f(u,v) = \left[\sum_{i=1}^{M}(u' K_i u)^2\right]\left[\sum_{i=1}^{M}(v' H_i v)^2\right] \tag{7}$$

Subject to the normalization constraints

$$||u|| = ||v|| = 1.$$

The objective of sDO-CCSWA is to find loading vectors u and v. These loading vectors must be the same for all matrices X_i and Y_i and allow to find specific weights. These specific weights can be considered as the projected variances associated with matrices X_i and Y_i. However, the loading vectors are obtained sequentially: the related group components are constrained to be orthogonal to the previous ones.

The following Lagrangian function related to optimization problem (7) is considered:

$$L(u,v,\lambda_1,\lambda_2) = \left[\sum_{i=1}^{M}(u' K_i u)^2\right]\left[\sum_{i}^{M}(v' H_i v)^2\right] + 2\lambda_1(1 - u' u) + 2\lambda_2(1 - v' v) \tag{8}$$

Where λ_1 and λ_2 are the Lagrange multipliers. The maximum of f follows from the requirement that the first order partial derivatives of L are simultaneously zero at the maximum of f and that the Hessian is negative. The following proposition specifies the role of the loading vectors in the criterion to be maximized.

Property 2.1 [Solution of order 1] The loading vectors u and v verify the stationary equations:

$$\sum_{i=1}^{M}(u^{'}K_iu)K_iu = r_uu \tag{9}$$

$$\sum_{i=1}^{M}(v^{'}H_iv)H_iv = r_vv \tag{10}$$

$$\lambda = r_ur_v \quad \text{where} \quad r_u = \sum_{i=1}^{M}(u^{'}K_iu)^2 \quad \text{and} \quad r_v = \sum_{i=1}^{M}(v^{'}H_iv)^2$$

Proof 2.1: Canceling the partial derivatives of the Lagrangian function with respect to u, v, λ_1 and λ_2 yields the following stationary equations:

$$\sum_{i=1}^{M}(v^{'}H_iv)^2\sum_{i=1}^{M}(u^{'}K_iu)K_iu = \lambda_1u \tag{11}$$

$$\sum_{i=1}^{M}(u^{'}K_iu)^2\sum_{i=1}^{M}(v^{'}H_iv)H_iv = \lambda_2v \tag{12}$$

with the normalization constraints

$$u^{'}u = 1 \quad \text{et} \quad v^{'}v = 1 \tag{13}$$

By combining (11), (12) and (13), we show that

$$\lambda = \lambda_1 = \lambda_2 = \sum_{i=1}^{M}(u^{'}K_iu)^2\sum_{i=1}^{M}(v^{'}H_iv)^2 = r_ur_v = f(u,v) \tag{14}$$

By replacing (14) into (11) and (12), it follows the stationary equations (9) and (10).

These stationary equations have no analytical solution, but they can be used to build a monotonically convergent algorithm for optimization problem (7). After having centered and standardized the matrices X_i and Y_i, we set $X_{i,0} = X_i$ and $Y_{i,0} = Y_i$. We use the following algorithm:

A. Initialization

A.1 Choose randomly u_0 and v_0 such that $||u_0|| = ||v_0|| = 1$ and ε $(e.g., 0.00001)$;

A.2 Compute $f(u_0, v_0) = \sum_{i=1}^{M}(u_0^{'}K_iu_0)^2\sum_{i=1}^{M}(v_0^{'}H_iv_0)^2$.

B. Computing of the updates
For $k = 1, 2, \ldots$

B.1 Compute and normalize the loading vectors

$$u_k = \sum_{i=1}^{M}(v_{k-1}^{'}H_iv_{k-1})^2\sum_{i=1}^{M}(u_{k-1}^{'}K_iu_{k-1})K_iu_{k-1}$$

B.2 Compute and normalize the loading vectors

$$v_k = \sum_{i=1}^{M}(u_k^{'}K_iu_k)^2\sum_{i=1}^{M}(v_{k-1}^{'}H_iv_{k-1})H_iv_{k-1}$$

C. Test

C.1 While $f(u_k, v_k) - f(u_{k-1}, v_{k-1}) \geq \varepsilon$, set $u_0 = u_k$ and $v_0 = v_k$ and go to B
Else we stop the algorithm and go to C.2.

C.2 For $i = 1, \cdots M$,
Compute $c_{X_{i,k}} = X_iu_k$ and $c_{Y_{i,k}} = Y_iv_k$;
Compute the specific weights $\alpha_{i,k} = u_k^{'}K_iu_k = \sum_{l=1}^{q} cov^2(X_iu_k, y_{il})$ and $\delta_{i,k} = v_k^{'}H_iv_k = \sum_{h=1}^{p} cov^2(Y_iv_k, x_{ih})$
End

End

The specific weights are positive and represent the proportions of the explained variances by u_k and v_k. It is important to note that the relative proportion can be found at the step k as follows:

$$q_{i,k} = \frac{(u_k^{'}K_iu_k)^2}{\sum_{i=1}^{M}(u_k^{'}K_iu_k)^2} = \frac{\alpha_{i,k}^2}{r_{u_k}} = \frac{\alpha_{i,k}^2r_{v_k}}{\lambda_k} \quad \text{and} \quad s_{i,k} = \frac{(v_k^{'}H_iv_k)^2}{\sum_{i=1}^{M}(v_k^{'}H_iv_k)^2} = \frac{\delta_{i,k}^2}{r_{v_k}} = \frac{\delta_{i,k}^2r_{u_k}}{\lambda_k}$$

We will now show that this algorithm for maximizing $f(u, v)$ subject to $||u|| = ||v|| = 1$ is monotonically convergent. Here, the algorithm is said to be monotonically convergent if and only if there is continuous and bounded function f such that que $f(u, v) < f(\bar{u}, \bar{v})$, with $\bar{u}$ and $\bar{v}$, the updates.

By setting $\alpha_i = u^{'} K_i u$, $\beta_{i,v} = r_v \alpha_i$ and $b_i = V_{Y_i X_i} u$ a vector of $\mathbb{R}^q$, the function (7) can be written

$$f(u,v) = u^{'} \sum_{i=1}^{M} r_v \alpha_i K_i u = u^{'} \sum_{i=1}^{M} \beta_{i,v} V_{X_i Y_i} b_i \qquad (15)$$

By fixing the loading vector v, we can set

$$\bar{u} = \frac{\sum_{i=1}^{M} \beta_{i,v} V_{X_i Y_i} b_i}{\|\sum_{i=1}^{M} \beta_{i,v} V_{X_i Y_i} b_i\|} \qquad (16)$$

We show that the passage in the loop of the iteration 1 modified $f(u)$ so that the function $\Delta f(u) = f(\bar{u}) - f(u)$ is nonnegative for any vector u. By setting $\theta = \|\sum_{i=1}^{M} \beta_{i,v} V_{X_i Y_i} b_i\|$, we obtain:

$$\Delta f(u) = f(\bar{u}) - f(u) = \sum_{i=1}^{M} \beta_{i,v} b_i^{'} V_{Y_i X_i} \bar{u} - u^{'} \sum_{i=1}^{M} \beta_{i,v} V_{X_i Y_i} b_i = \theta(1 - cos(u, \sum_{i=1}^{M} \beta_{i,v} V_{X_i Y_i} b_i)) \qquad (17)$$

Where $cos(u, \sum_{i=1}^{M} \beta_{i,v} V_{X_i Y_i} b_i)$ is the cosine between u and $\sum_{i=1}^{M} \beta_{i,v} V_{X_i Y_i} b_i$. This cosine is smaller or equal to 1. Which implies $f(u)$ increases at each iteration of the algorithm or $f(u) \leq f(\bar{u})$ in other words $f(u,v) \leq f(\bar{u},v)$.

By setting $\gamma_{i,\bar{u}} = (v' H_i v) r_{\bar{u}}$, $f(v) = f(\bar{u},v)$ and $c_i = V_{X_i Y_i} v$ a vector of $\mathbb{R}^p$, the function $f(\bar{u},v)$ can be written $f(\bar{u},v) = v^{'} \sum_{i=1}^{M} \gamma_{i,\bar{u}} V_{Y_i X_i} c_i$. By setting

$$\bar{v} = \frac{\sum_{i=1}^{M} \gamma_{i,\bar{u}} V_{Y_i X_i} c_i}{\|\sum_{i=1}^{M} \gamma_{i,\bar{u}} V_{Y_i X_i} c_i\|} \qquad (18)$$

and by showing as previously $f(\bar{u},v) \leq f(\bar{u},\bar{v})$, it follows that

$$f(u,v) \leq f(\bar{u},v) \leq f(\bar{u},\bar{v}) \qquad (19)$$

It can be concluded that the algorithm is monotone. The function $f(u,v)$ being bounded and continuous in the compact set define by the constraints, then it converges. By modifying the starting point of the algorithm, we do not find the same maximum because we can obtain a local maximum on the surface defined by the constraints.

At the order $s > 1$, the sDO-CCSWA criterion can be written:

$$f(u_s, v_s) = \sum_{i=1}^{M} (u_s^{'} K_i u_s)^2 \sum_{i=1}^{M} (v_s^{'} H_i v_s)^2 \qquad (20)$$

subject to the constraints $||u_s|| = ||v_s|| = 1$ and the orthonormality constraints $u_s^{'} u_t = v_s^{'} v_t = 0$ for $s = 1, \cdots, r$ and $t < s$, where $r \leq min(p,q)$ is the rank of the matrices $V_{X_i Y_i}$.

After having shown the convergence of the algorithm, the solutions of order $s > 1$ of optimization problem (7) are given in proposition 2.2 below.

Property 2.2 If we note $X_{i,0} = X_i$ and $Y_{i,0} = Y_i$, the loading vectors u_s and v_s ($s \geq 2$) are the solutions of order 1 of the sDO-CCSWA of the matrices $X_{i,s-1}$ and $Y_{i,s-1}$, with $X_{i,s-1} = X_{i,s-2} P^{\perp}_{u_{s-1}}$, $Y_{i,s-1} = Y_{i,s-2} P^{\perp}_{v_{s-1}}$. These loading vectors verify the following stationary equations:

$$\sum_{i=1}^{M} (u_s^{'} K_{i,s-1} u_s) K_{i,s-1} u_s = r_{u_s} u_s \qquad (21)$$

$$\sum_{i=1}^{M} (v_s^{'} H_{i,s-1} v_s) H_{i,s-1} v_s = r_{v_s} v_s \qquad (22)$$

$$\lambda_s = r_{u_s} r_{v_s} \quad \text{where} \quad r_{u_s} = \sum_{i=1}^{M} (u_s^{'} K_{i,s-1} u_s)^2 \quad \text{et} \quad r_{v_s} = \sum_{i=1}^{M} (v_s^{'} H_{i,s-1} v_s)^2.$$

To solve this system of equations, we apply the same algorithm.

A similar proof of this result can be found in [8].

sDO-CCSWA allows to find two sets of orthonormal loading vectors $(u_s)_s$ and $(v_s)_s$. Otherwise, sets of components are not orthogonal. To represent p variables of the data matrix X_i, we project rows of the matrix K_i on the loading vectors u_s. In the same way, to represent q variables of the data matrix Y_i, we project rows of the matrix H_i on the loading vectors v_s.

Loading vectors u_s and v_s and components $c_{X_{i,s}}$ and $c_{Y_{i,s}}$ may be calculated for the deflation. Undesirable properties of some methods, such as convergence problems or loss of data information due to deflation procedures, are studied by [17] and recently in [1].

The sDO-CCSWA criterion (7) is equivalent to the following maximization criterion:

$$f(u,v) = \sum_{i=1}^{M} (u^{'} K_i u)^2 + \sum_{i=1}^{M} (v^{'} H_i v)^2 \qquad (23)$$

Subject to the normalization constraints

$$||u|| = ||v|| = 1.$$

3. Simultaneous Approach of DO-CCSWA

In this section, we will extend the sequential model to a general model. As demonstrated by [3], the advantage of DO-CCSWA algorithm is that it is monotonically convergent, it reaches a stable solution faster and has a better performance in terms of convergence speed compared to sDO-CCSWA algorithm. By using the deflation procedure, some of the information in data matrices may be lost in the deflation step. Let $U = [u_1, \cdots, u_r]$ be the $p \times r$ loading matrix containing the loading vectors u_s and $V = [v_1, \cdots, v_r]$ be the $q \times r$ loading matrix containing the loading vectors v_s. The problem consists of finding two loading matrices U and V associated with two sets of variables. We define the DO-CCSWA method as the following maximization problem:

$$f(U,V) = \left[\sum_{i=1}^{M} \|diag(U^{'} K_i U)\|^2\right] \left[\sum_{i=1}^{M} \|diag(V^{'} H_i V)\|^2\right] \tag{24}$$

Subject to the constraints

$$U^{'}U = V^{'}V = I_r.$$

It is important to note that $||diag(A)||^2 = tr(Adiag(A)) = \sum_{s=1}^{r} a_{ss}^2$ where $A = (a_{sl})_{1 \leq s,l \leq r}$.

Furthermore, it is of interest to note that maximization of the above criterion can be written more explicitly as the maximization of criterion:

$$f(U,V) = \left[\sum_{i=1}^{M}\sum_{s=1}^{r} (u_s' K_i u_s)^2\right] \left[\sum_{i=1}^{M}\sum_{s=1}^{r} (v_s' H_i v_s)^2\right].$$

From this new description of DO-CCSWA, it is clear that the case $r = 1$ leads to sDO-CCSWA.

Many papers have dealt with the problem of finding general principles from which families of algorithms can be constructed. The approach presented in the present paper is almost the same as those by [10], [6] and recently by [7].

To maximize f subject to the constraints $U^{'}U = V^{'}V = I_r$, we present a monotonically convergent algorithm. This iterative algorithm that increases the function f monotonically subject to the constraints $U^{'}U = V^{'}V = I_r$ can be constructed by looking for two updates U^* and V^* of U and V, so that $f(U,V) \leq f(U^*, V^*)$. For reasons of notational convenience, let T_{UV} and H_{UV} be the current matrices:

$$T_{UV} = (\sum_{i=1}^{M} K_i U diag(U^{'} K_i U))(\sum_{i=1}^{M} tr(V^{'} H_i V diag(V^{'} H_i V))$$

$$\text{and} \quad K_{UV} = (\sum_{i=1}^{M} H_i V diag(V^{'} H_i V))(\sum_{i=1}^{M} tr(U^{'} K_i U diag(U^{'} K_i U))$$

In order to see how the DO-CCSWA approach can be applied in the present context, let us consider briefly (25):

$$f(U,V) = tr(U^{'} T_{UV}) = tr(V^{'} K_{UV}) \tag{25}$$

It is well known that, if PDQ' is the singular value decomposition of the (p,r) matrix A, with $P'P = Q'Q = QQ' = I_r$ and D is a diagonal matrix, nonnegative and with diagonal elements in weakly descending order, then the maximum of $tr(T'A)$ subject to $T'T = I_r$ is obtained when $Z = PQ'$ [2].

Before starting the iterations, we first centered and standardized the data matrices X_i and Y_i and we set $X_{i,0} = X_i$ and $Y_{i,0} = Y_i$. The algorithm derived above can be summarized as

A. Initialization

- A1. Choose randomly U_0 and V_0 such that $U_0^{'}U_0 = V_0^{'}V_0 = I_r$ and ε $(e.g., 0.00001)$;
- A2. Compute $f(U_0, V_0)$

B. Computing of the update of U

- B.1 consider the singular value decomposition: $T_{UV} = P\Delta L^{'}$, $P^{'}P = L^{'}L = LL' = I_r$
- B.2 Set $U^* = PL^{'}$

C. Computing of the update of V

- C.1 consider the singular value decomposition: $H_{UV} = F\Theta G^{'}$, $F^{'}F = G^{'}G = GG^{'} = I_r$
- C.2 Set $V^* = FG^{'}$

D. Test

- D.1 While $f(U^*, V^*) - f(U,V) \geq \varepsilon$, set $U = U^*$ and $V = V^*$ and go to B, else we stop the algorithm and go to D.2.
- D.2 For $i = 1, \cdots M$,
  compute the components $c_{X_{i,k}} = X_i u_k$ and $c_{Y_{i,k}} = Y_i v_k$.
  compute the distance between $diag(U^{'} K_i U)$ and $diag(V^{'} H_i V)$: $d_i = \|diag(U^{'} K_i U) - diag(V^{'} H_i V)\|$
  compute the weights $\alpha_i = ||diag(U' K_i U)|| =$

$\sqrt{\sum_{s=1}^{r}(u_s' K_i u_s)^2}$ and $\delta_i = ||diag(V'H_iV)|| = \sqrt{\sum_{s=1}^{r}(v_s' H_i v_s)^2}$
End

End

Specific weights α_i and δ_i are nonzero and positive. It is important to note that the relative proportions can be found as follows:

$$q_i = \frac{\alpha_i^2}{R_U} \quad \text{et} \quad s_i = \frac{\delta_i^2}{R_V}$$

with $R_U = \sum_{i=1}^{M} ||diag(U'K_iU)||^2$ and $R_V = \sum_{i=1}^{M} ||diag(V'H_iV)||^2$. It can be shown that at each step of the iteration the value of f increases. We only show the monotony related to U. That of V is demonstrated in the same way. Thus

$$f(U,V) \leq f(U^*,V) \leq f(U^*,V^*) \tag{26}$$

We consider the sth diagonal element of $U^{*'}T_{UV}$ written as

$$u_s^{*'}(\sum_{i=1}^{M} K_i u_s u_s' K_i) u_s (\sum_{i=1}^{M} \|diag(V' H_i V)\|^2) \tag{27}$$

where u_s, $s = 1, \cdots, r$, is a column of U. Let us set $G_s = \sum_{i=1}^{M} K_i u_s u_s' K_i$ a positive semidefinite symmetric matrix. Firstly, from $\|G_s^{\frac{1}{2}} u_s - G_s^{\frac{1}{2}} u_s^*\|^2 \geq 0$, we obtain

$$u_s' G_s u_s + u_s^{*'} G_s u_s^* \geq 2 u_s^{*'} G_s u_s \tag{28}$$

$\sum_{i=1}^{M} \|diag(V' H_i V)\|^2 \geq 0$ and by multiplying this number in the two members of (28), it follows

$$f(U,V) + (\sum_{i=1}^{M} \|diag(U^{*'} K_i U)\|^2)(\sum_{i=1}^{M} \|diag(V' H_i V)\|^2) \geq 2(\sum_{i=1}^{M} tr(U^{*'} K_i U diag(U' K_i U))(\sum_{i=1}^{M} \|diag(V' H_i V)\|^2) \tag{29}$$

According to [2], the update U^* of U verify the inequality

$$\sum_{i=1}^{M} tr(U^{*'} K_i U diag(U' K_i U))(\sum_{i=1}^{M} \|diag(V' H_i V)\|^2) \geq f(U,V) \tag{30}$$

Combining this with (29) shows that

$$(\sum_{i=1}^{M} \|diag(U^{*'} K_i U)\|^2)(\sum_{i=1}^{M} \|diag(V' H_i V)\|^2) \geq f(U,V) \tag{31}$$

On the other hand, from

$$\|(K_i)^{\frac{1}{2}} u_s u_s' (K_i)^{\frac{1}{2}} - (K_i)^{\frac{1}{2}} u_s^* u_s^{*'} (K_i)^{\frac{1}{2}}\|^2 \geq 0 \tag{32}$$

We obtain

$$u_s' K_i u_s (u_s' K_i u_s) + u_s^{*'} K_i u_s^* (u_s^{*'} K_i u_s^*) \geq 2 u_s^{*'} K_i u_s (u_s^{*'} K_i u_s) \tag{33}$$

Multiplying (33) by $\sum_{i=1}^{M} \|diag(V' H_i V)\|^2$ and summing over s and i, we obtain

$$f(U,V) + (\sum_{i=1}^{M} \|diag(U^{*'} K_i U^*)\|^2)(\sum_{i=1}^{M} \|diag(V' H_i V)\|^2) \geq 2(\sum_{i=1}^{M} \|diag(U^{*'} K_i U)\|^2)(\sum_{i=1}^{M} \|diag(V' H_i V)\|^2) \tag{34}$$

Finally, using inequality (31) we find

$$f(U^*,V) \geq f(U,V) \tag{35}$$

It can be concluded that the update U^* increases the function (24). After switching the roles of U and V , a parallel development can be given to prove that updating V by V^* and to show the monotony respect to V.

The function f being bounded, continuous and monotone in particular on the sets of loading vectors of U and V, then the algorithm converges to a local maximum of f. To obtain a global maximum, the algorithm is executed several times by starting on several initializations. The largest local maximum of the local maxima is retained to become the global maximum of the function. The particular choice of $U = U^*$ implies that $U^{*'}T_{UV} = LP'P\Delta L' = L\Delta L'$ is a positive semidefinite symmetric matrix and Δ contains the positive or null singular values.

The representation of the individuals of the DO-CCSWA simultaneous approach is similar to the sDO-CCSWA successive approach. It is enough to use their coordinates in the matrices X_iU^* and Y_iV^* for the two groups of variables.

The variables can be projected as supplementary elements to help interpret the results of the analysis. To represent p variables of the data matrix X_i, we project rows of the matrix K_i on the columns of the matrices U^*. In the same way, to represent q variables of the data matrix Y_i, we project rows of the matrix H_i on the columns of the matrices V^*.

The DO-CCSWA criterion (24) is equivalent to the following maximization criterion:

$$f(U,V) = \sum_{i=1}^{M} \|diag(U^{'} K_i U)\|^2 + \sum_{i=1}^{M} \|diag(V^{'} H_i V)\|^2 \tag{36}$$

Subject to the constraints

$$U^{'}U = V^{'}V = I_r.$$

4. Results

We illustrate the simultaneous DO-CCSWA method with an example of ecological datasets. We reanalyze the datasets that have been acquired by [13] and which serve as an illustration in STATICO [14]. Specifically, we reanalyze two data matrices: one data matrix X with 24 rows and 13 columns, containing the ephemeroptera species and one data matrix Y with 24 rows and 10 columns, containing the environmental variables.

The rows of both matrices correspond to 6 sampling sites ordered upstream-downstream along a small stream, the Méaudret. These 6 sites are sampled 4 times, in Spring, Summer, Autumn and Winter.

The 24×13 data matrix X is partitioned in four 6×13 data matices X_i. The 13 columns of the species data table correspond to 13 Ephemeroptera species, which are known to be highly sensitive to water pollution. These species are as follows: Eda=Ephemera, Bsp=Baetis sp, Brh=Baetis rhodani, Bni=Baetis niger, Bpu=baetis pumilus, Cen=centroptilum, Ecd=Ecdyonurus, Rhi=Rhihrogena, Hla=Habrophlebialauta, Hab=Habroletoides modesta, Par=Paraletophlebia, Cae=Caenis, Eig=Ephemerella ignita.

In addition, 24×10 data matrix Y is partitioned in four 6×10 data matices Y_i. The 10 environmental variables are physico-chemical measures: Temp=water temperature, flow, pH, Cond=conductivity, Oxyg=oxygen, BDO5=biological oxygen demand, Oxyd=oxidability, Ammo=ammonium, Nitr=nitrates and Phos=phosphates.

The problem is to investigate the stability relationships between Ephemeroptera species distribution and the quality of water in the site typology.

The distances of each season are given in Table 1. These distances allow to describe the evolution of the relationship species-environment. The distance between two configurations of points also measure the fit. These distances are squared Euclidian distances between the rows of $U'K_iU$ and those of $V'H_iV$.

A constancy of these distances allows to conclude the stability of the relationship. Autumn and Summer are the two most important seasons for comparing the data sets, while Winter and Spring are slightly less important. The variation of the distance between Autumn and Summer is equal to 754.14. This is not the case in Winter and Spring, where they differ from other seasons. We find exactly the same results with the STATICO method where the structures are the strongest in Autumn, both for environmental variables and for Ephemeroptera species.

The specific weights of the data matrices over two updates containing the loading vectors are given in Table 2. It would give a complete view of the variability of two data sets in each season. Thus, with the data matrices X and Y, the relative proportions of the specific weights show that the environmental variability and the diversity of the ephemeroptera species are the weakest in Winter and Spring, while Autumn has the highest relative proportion. The two data matrices X and Y have very similar results. That is, there is a common structure between two data matrices. Table 2 also shows that much of the information is contained in the data matrices X_3 and Y_3 (In Autumn).

Figure 1 on the left shows the position of the sites over two loading vectors. It appears that a general organization of the sites and species from one season to another. The sites S1 and S6 have perfectly similar behavior characterized more or less by a strong presence of the species (Bpu, Hla, Eda, Bsp, Eig), while site S2 has bad scores for these same variables. Overall, we note a size effect on axis 1 for the ephemeroptera species on the right. The first axis is a pollution factor (Bpu, Hla, Eda, Bsp, Eig), which produces a decrease of the environmental variables richness in several sites.

The second axis (horizontal) opposes site S2 to site S6 on the left. Figure 2 on the left shows the position of the sites over two loading vectors. It shows that flows are high in the sites S4 and S5 in Spring while the water temperature ("Temp") is strong in Winter. In Autumn, site S2 is characterized by phosphate ("Phosp"), Ammonium ("Ammo") and biological oxygen demand ("BDO5"). In the site S5, we find the importance of nitrates ("Nitr") in relation to temperature ("Temp"). The second axis is related with the upstream-downstream structure of the river.

Figures 1 and 2 on the right represent the projection of the columns of the matrices of two sets of variables. The species are obtained by projecting the rows of K_i on U^* and are contained XU^*. In the same way, the environmental variables are obtained by projecting the rows of H_i on V^* and are contained YV^*. This step is done for STATICO at the intrastructure.

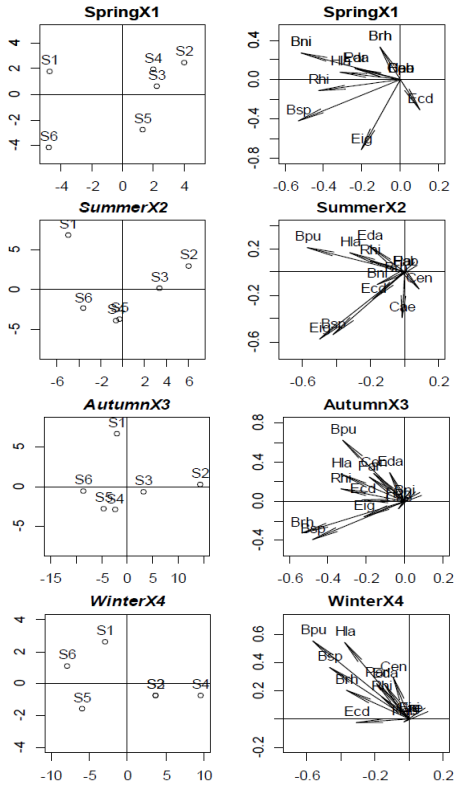

Figure 1. *Map of the sites on the left and map of the species at each season on the right.*

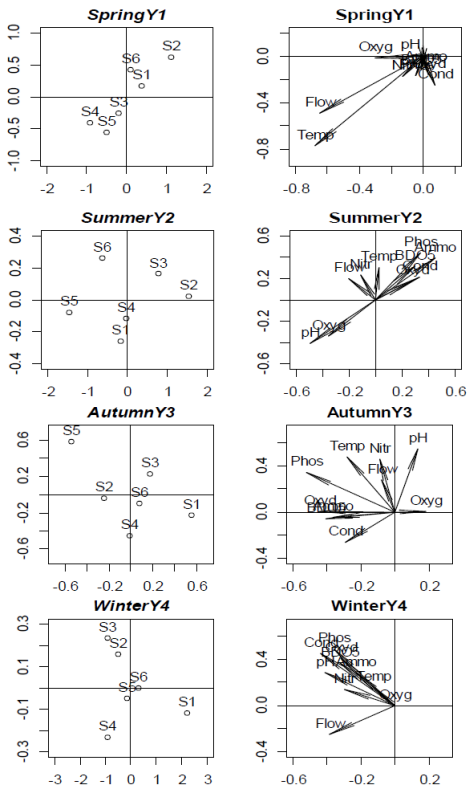

Figure 2. *Map of the sites on the left and map of the environmental variables at each season on the right.*

5. Conclusion

The DO-CCSWA method is algorithmic and uses matrix operators obtained by crossing doubly the two data matrices of a pair at each season. The main aim of DO-CCSWA is to investigate the relationships between two data sets structured in groups of variables. We have shown in this paper that a single simple and powerful algorithm can be used for investigate the relationships between two data sets. The DO-CCSWA algorithm converges monotonically to a stationary point, but convergence to the global optimum is not guaranted. This method is different to STATICO and s3CIA. These methods are sequential. STATICO is a Partial Triadic Analysis on the series of cross product matrices obtained by crossing the two data matrices of a pair. It benefits from the three-steps computation scheme of STATIS-like methods (interstructure, compromise, intrastructure). We wish to note that the simultaneous DO-CCSWA method proposed is more efficient than the s3CIA and STATICO methods in terms of determining the solution and the used criterion. The special case $r = 1$ of DO-CCSWA is a sequential method sDO-CCSWA which leads to the same results. We have illustrated DO-CCSWA for two matrices of data on an ecology testing example.

References

[1] Bougeard, S., Abdi, H., Saporta, G., & Niang, N. (2018). Clusterwise analysis for multiblock component methods. *Advances in Data Analysis and Classification, 12 (2), 285-313.*

[2] Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika 31, 33-42.*

[3] Hanafi, M., Dolce, P., & Hadri, Z. E. (2021). Generalized properties for Hanafi-Wold's procedure in partial least squares path modeling. *Computational Statistics, 36, 603-614.*

[4] Hanafi, M., Kohler, A., & Qannari, E. M. (2010). Shedding new light on hierarchical principal component analysis. *Journal of Chemometrics, 24, 703-709.*

[5] Hanafi, M., & Qannari, E. M. (2008). Nouvelles propriétés de l'analyse en composantes communes et poids spécifiques. *Journal de la Société Française de Statistique; tome 149 No 2, 75-97.*

[6] Hanafi, M., & Kiers, H. A. L. (2006). Analysis of K sets of data, with differential emphasis on agreement between and within sets. *Computational Statistics & Data Analysis, 51, 1491-1508.*

[7] Kiers, H. A. L., & Giordani, P. (2020). Candecomp/Parafac with zero constraints at arbitrary positions in a loading matrix. *Chemometrics and Intelligent Laboratory Systems, 207, 104145.*

[8] Kissita, G. (2003). Les analyses canoniques généralisées avec tableau de référence généralisé: éléents théoriques et appliqués. *PhD thesis, University of Paris Dauphine, France.*

[9] Kissita, G., Malouata, R. O., Mizère, D., & Makany, R. A. (2013). Proposition of analyses of links between two vertical multi-tables : methods (sVMA and sOVMA) and (sCIA3 et sOCIA3). *Applied mathematical Sciences, Vol. 7, no 131, 6503-6525.*

[10] Lafosse, R., & Ten Berge, J. M. F. (2006). A simultaneous CONCOR algorithm for the analysis of two partitioned matrices. *Computational Statistics & Data Analysis, 50, 2529-2535.*

[11] Lafosse, R., & Hanafi, M. (1997). Concordance d'un tableau avec K tableaux: définition de $K + 1$ uplés synthétiques. *Revue de Statistique Appliquée, XLV(4): 111-126.*

[12] Malouata, R. O. (2015). Proposition d'analyse de co-inertie d'une série de couples de tableaux: éléments théoriques et appliqués. *PhD thesis, Marien Ngouabi University, Congo.*

[13] Pegaz-Maucet, D. (1980). Impact d'une perturbation d'origine organique sur la dérive des macro-invertébrés bentiques d'un cours d'eau. Comparaison avec le benthos. *PhD thesis, University of Lyon I, France.*

[14] Simier, M., Blanc, L., Pellegrin, F., & Nandris, D. (1999). Approche simultanée de k couples de tableaux: Application à l'étude des relations pathologie végétale-environnement. *Revue de Statistique Appliquée 47 31-46.*

[15] Thioulouse, J. & Chessel, D. (1987). Les analyses multitableaux en ecologie factorielle. i : De la typologie d'etat à la typologie de fonctionnement par l'analyse triadique. *Acta Oecologica, Oecologia Generalis 8 463-480.*

[16] Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika, 23, 111-136.*

[17] Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics, 12, 301-321.*