Lakhesis: Consensus Seriation via Iterative Regression of Partial Rankings for Binary Data

Stephen A. Collins-Elliott

Department of Classics, University of Tennessee, Knoxville, USA

ARTICLE HISTORY
Compiled June 25, 2024

ABSTRACT
When it comes to seriating a matrix of binary data, dimensional reduction techniques like correspondence analysis and its derivatives often perform better than model-driven methods. Yet, there is a larger problem in that, first, seriation is data-dependent and not guaranteed for every binary matrix, and second, that high-dimensional matrices may yet produce plots that are difficult to interpret. The blame for such poor results may be cast onto the data themselves as being not conducive to seriation. Consequently, the onus is thus placed on the investigator to *a priori* have well-seriated data to begin with. The question then also arises, in the process of exploring the data and identifying multiple partial, well seriated sequences, how to harmonize them into a single ranking. The solution proposed here involves an iterative procedure called "Lakhesis," which uses an agglomerative process of regression of the partial sequences to resolve missing observations in producing a single consensus seriation. Per optimality measures, Lakhesis has the capacity to outperform current conventional approaches to seriation. The R `lakhesis` package provides an graphical interface for exploring binary data and selecting seriations, toward selecting well-seriated "strands," which are then "lakhesized" into a single consensus seriation.

KEYWORDS
Seriation, ordination, binary data, correspondence analysis, Spearman's rank correlation

1. Introduction

Seriation is a statistical problem in which one seeks to rearrange the rows and columns of a matrix in order to create an optimal order for the matrix values [26]. Also called sequencing or ordination, seriation represents a problem found in several fields, such as archaeology, ecology, and the biosciences. Given the high number of permutations of potential sequences to evaluate, a broad array of techniques have been cultivated to address the problem. Approaches to seriation have often treated count or frequency data, treating binary (0/1) data as a reductive case of the former. Yet, binary data can pose their own challenges. Quadratic ordination, i.e., using a generalized linear model to perform a logistic regression with a quadratic term, has long been the choice of model for frequency matrices. In a binary matrix, however, perfect separation of 0s/1s, which would be ideal with well-seriated data, would result in a failure of model

CONTACT Stephen A. Collins-Elliott. Email: sce@utk.edu

fit. Moreover, given a low count and sparse size, distributions might fail to converge to an optimally seriated order. This makes dimensional reduction techniques, especially correspondence analysis (CA), more suitable for binary seriation, but herein too there are challenges. Not all matrices will necessarily be well-seriated, making exploration a necessary part of seriation. Moreover, if one has determined multiple, partial seriations within their data, the need arises for a means to harmonize these partial sequences, to create what Hahsler et al. call a single "consensus seriation" [20]. This paper offers a practical solution to these challenges, with the development of (1) an interactive, graphical platform to perform exploratory seriations, (2) a new means to seriate correspondence analysis (CA) scores via Procrustes-fitting to a reference curve, (3) a means of producing a consensus seriation, and (4) defining critical measures. These tools are freely available in open source R `lakhesis` package.

1.1. Background

As mentioned above, there are a variety of approaches and tools available for seriation. Permutation-based and other combinatoric methods were advanced by [1, 8, 34, 44], which are discussed succinctly by Hahlser et al. [20]. In addition to non-metric multidimensional scaling (nMDS; [see 29–31]), correspondence analysis (CA) has been a popular method for seriation [3, 4, 18, 21]; Ihm [26] notes that the earliest form of CA can be found in [23]. Given that CA will display points along a curve if they follow a seriated pattern, detrended correspondence analysis (DCA) was pioneered by Hill and Gauch [22] to "unbend" the curve by creating partitions along the first component axis and shifting these partitions to produce a more linear form [see pp. 379-396 in 5, for more background]. Putting constraints on CA scores has been studied by [40, 51], while establishing bootstrap confidence intervals on CA-based seriations has been performed by [2, 33, 39]. DCA has been debated in ecology but remains popular [53, 54]. Model-based methods, namely quadratic ordination (QO), emerged in ecology around the same time as CA and have the benefit of being theoretically more sound, although at the time they were more numerically intensive [14, 15]. There is a significant literature on model-based ordination, especially with constraints, which have deep affinities with canonical correspondence analysis (CCA), factor analysis, and latent variable analysis [16, 25, 27, 43, 48–50, 57].

The aforementioned approaches are well served by numerous packages in R [42]. The `seriation` package [20] contains several functions for seriating and critically evaluating seriated matrices from a combinatorial approach. Dimensional reduction techniques like nMDS, PCA, and CA are available in `MASS` [52] and `ca` [37]. The `vegan` package [38] contains functions to perform DCA and CCA, among many other dimensional reduction techniques. `VGAM` [58] contains functions for constrained and unconstrained QO. Bayesian generalized linear models (GLMs) with and without constraints can be fit using the `boral` package [24]. The recent `ecoCopula` package [41] uses copulas for ordination.

1.2. Contribution

Before stipulating the contributions of this paper, it can be noted that this work came about in the attempt to adapt copulas for the case of seriating binary data [11, 41]. While copulas show promise for ordination of high-dimensional frequency data, binary data posed challenges. To start, the perfect separation of 0s and 1s in a sequence of row or column values (an optimal arrangement in seriation) would result in a failure of model fit in a quadratic regression. Furthermore, given the low and often sparse values, unconstrained model-fitting with high-dimensional binary data using simulation may also take an exceedingly long time to converge to a solution, or even fail. As such, dimensional reduction techniques like CA remain the most useful for this case. But, straightforward application of CA can just as often return results that are not clearly seriated or which form undesirable clusters. In archaeology, for example, creating a single incidence matrix (in which each row represents a context and each column represents a find-type) from multiple, contemporary sites will tend to elicit a plot which produces clusters of those sites [e.g., 32], rather than the expected horseshoe curve or arch effect, since the rows and columns have less overlap with those from other sites than they with those from the same site or locality. More generally, an exceedingly high dimensional matrix may lose too much variance (inertia) in the process of dimensional reduction.

Given that the choice of row and column elements plays such a decisive role in determining seriated sequences, the R package `lakhesis` provides an interactive platform to aid in the heuristic seriation of binary matrices and to evaluate their consensus seriation critically. Intended to complement functions already available in `seriation` and `vegan`, it does not reduplicate nor replace those packages. The dependencies of `lakhesis` include `ca` [37], `ggplot2` [55], `shiny` [7], `shinydashboard` [6], `bslib` [46], and `readr` [56], the latter of which was needed for properly importing Unicode characters from an initial `.csv` file. Figures for this paper were aided with `gridExtra` [36].

In the R `lakhesis` package, an interactive `shiny` app called the Lakhesis Calculator was developed to allow a user to perform CA in an exploratory fashion on a matrix of binary data. The investigator can select row and column scores and then re-run CA on that selection, in effect "exploring" lower dimensional pockets of space. To establish a seriated order, a new method of ranking was developed, in which CA scores are fit via a Procrustes method to the curve elicited by of a "reference matrix" of the same size, which contains an "ideal" seriation. The scores of the data are then projected onto the curve, which are used to determine their ordering. Two plots are shown in the Calculator, one of the Procrustes-fit CA scores plot and the other showing the distance of points from the reference curve-fit. The investigator may either log the displayed seriated sequence of rows and columns as a "strand" (a seriated matrix which contains a partial subset of the row and column elements of the full input matrix), and/or select row and column points on which to re-run CA.

User-selected strands can then merged into a single consensus seriation. This is accomplished using an iterative process of simple linear regression, performed separately on the row and column rankings. By using a measure of optimality called "concentration" (on which see Sec. 4.2) which conveys how well seriated a binary incidence matrix is, pairwise regression is first performed on all strands (with each strand serving as independent or dependent variate for every other strand). Then, the ranking of the joint row or column elements of the strand (as independent variate) are regressed onto the other (as dependent variate), with any discrepancy between the regression and actual value of the joint elements resolved by taking their mean. This will result in an "imputed" ranking for any elements missing in the dependent strand. The values are all then re-ranked. Performing this operation on all strands pairwise, the pair which elicits the most optimal concentration measure is chosen, and the new regression serves as the dependent variate for the next iteration, in which each remaining strand is regressed, and the one which elicits the lowest concentration score is chosen.

The end result is that all rankings will be regressed into a single, optimal "consensus seriation." Regression has the benefit in that sequences which have the same order but which may run in reverse are easily accommodated (their slope will merely be negative).

The critical evaluation of any one strand with respect to the consensus of all is performed on the basis of two coefficients, agreement and concentration. For agreement, the square of Spearman's rank correlation coefficient [47] is applied to row and column sequences separately, ρ_r^2 and ρ_c^2, after removing any pairs which have an `NA` value. The product of these two coefficients is then used as a measure of overall agreement between any two strands. Furthermore, the optimality of each strand can be assessed. Finally, a deviance test on the rows and columns of the resulting consensus seriation can be used to identify which row or column elements might not be conducive to obtaining a well-seriated sequence. These criteria can be used by the investigator to select or de-select strands and elements that may not be conducive to a seriated sequence.

The paper proceeds to outline the processes of Procrustes fitting CA scores in Section 2, which are used to derive a seriated sequence. In Section 3, the process of aligning the strands into a consensus via iterative regression is discussed. In Section 4, coefficients of agreement and concentration are defined. Section 5 then examines an application to a seriation of Early Iron Age tombs from southern Etruria, to assess the method's performance alongside other current approaches. The manual, "A Guide to Lakhesis," can be consulted for the functionality of the `lakhesis` package.

2. Exploratory CA with Procrustes Fitting

Let $\mathbf{Y}$ be the initial incidence matrix, of $u \times v$ size, in which each cell contains a 0 for absence and 1 for presence. Any matrix $\mathbf{X}$ is formed of selected rows and columns of $\mathbf{Y}$, provided that there is overlap in resulting columns and rows of $\mathbf{X}$, and each row sum and column sum of $\mathbf{X}$ is greater than 1. Columns or rows which attest only one incidence are omitted for expediency. To initialize, let $\mathbf{X} = \mathbf{Y}$.

2.1. Potential Matrix Transposition

Let the size of $\mathbf{X}$ be $n \times k$, where $n < k$, in which each cell contains a 0 for absence and 1 for presence. If it should be the case that $k < n$ for $\mathbf{X}$, the matrix is transposed, and the results for the row and column scores are then swapped at the end of the step. The reason for this condition, that the number of rows must be less than the number of columns, has to do with the production of a parabolic effect in the plot of the scores elicited by CA. The row and column scores are to be fitted to those of a reference matrix of the same size that contains a known "ideal" seriation, which is discussed next, wherein the parabola will only reliably be produced for row scores when $n < k$.

This reference matrix $\mathbf{R}$ is of the same size as $\mathbf{X}$, $n \times k$. Each ith row of $\mathbf{R}$ contains 1s across the columns, from the (i, i)th cell to the $(i, i + k - n)$th cell, and 0s in all other cells. That is, for an incidence matrix 5×7 in size, a corresponding reference

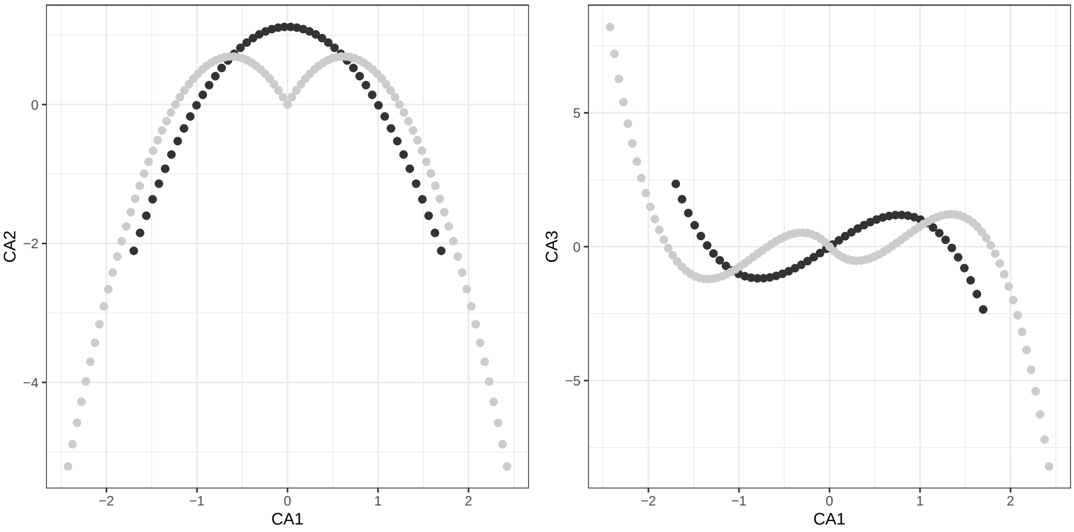

Figure 1. CA scores of a reference matrix $\mathbf{R}$ of size 50×100, showing the projections of the first and second principal axes on the left and the first and third axes on the right. Row scores are in black and column scores in gray. If $\mathbf{R}$ had been transposed (of size 100×50), the row scores would have taken the place of the column scores, and vice versa.

matrix $\mathbf{R}$ is

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

When performing CA, row scores of this reference matrix will follow a parabolic curve. If there had a larger number of rows than columns ($n > k$), then the row scores would resemble a double-parabola which appears truncated and symmetric around the second axis, and the column scores will instead follow a parabolic curve. This is a result of the "coiling" of the scores around the third principal axis, which is visible in three dimensions (Fig. 1). For the purpose of Procrustes fitting, if the number of rows is greater than that of the columns, the matrix $\mathbf{X}$ is transposed, and then the results for rows and columns swapped after fitting.

2.2. Correspondence Analysis

Perform correspondence analysis (CA) on $\mathbf{X}$ and $\mathbf{R}$ separately, the details of which may be found in the work of Greenacre and Nenadic [18, 19, 37]. CA involves centering the initial data matrix with respect to both row and column values, and then standardizing them with respect to both row and column values, essentially giving the chi-squared distance of each cell from the origin. This matrix of standardized residuals, $\mathbf{S}$, is then subjected to singular value decomposition (SVD),

$$\mathbf{S} = \mathbf{U\Sigma V}^\top.$$

The principal scores for each of the rows will be

$$\mathbf{G_n} = \mathbf{W_n}^{-\frac{1}{2}}\mathbf{U\Sigma}$$

and those of of the columns will be

$$\mathbf{G_k} = \mathbf{W_k}^{-\frac{1}{2}}\mathbf{V\Sigma}.$$

where $\mathbf{W_n}^{-\frac{1}{2}}$ and $\mathbf{W_k}^{-\frac{1}{2}}$ are diagonal matrices containing inverse square root of the row and column sums. The first two dimensions of the principal scores are then used for Procrustes fitting. Select the first two columns of $\mathbf{G_n}$ as $\mathbf{G_r}$, of size $n \times 2$, and the first two columns of $\mathbf{G_k}$ as $\mathbf{G_c}$, of size $k \times 2$. Let $\mathbf{L_r}$ and $\mathbf{L_c}$ be the first two columns of row and column scores of the reference matrix, respectively.

2.3. Procrustes Fitting

Procustes fitting [see 17] of $\mathbf{G_r}$ on $\mathbf{L_r}$ consists of the following three steps.

2.3.1. Centering

First, center the points of $\mathbf{G_r}$ and $\mathbf{L_r}$ around the origin. Let μ_j represent the mean of the jth column of $\mathbf{G_r}$, and $\boldsymbol{\mu_j}$ represent a column vector of size $n \times 1$ whose values are all μ_j. The centered matrix is then

$$\mathbf{G_{rc}} = \mathbf{G_r} - \begin{bmatrix} \boldsymbol{\mu_1} & \boldsymbol{\mu_2} \end{bmatrix}.$$

Similarly, to center the reference matrix row scores, where $\boldsymbol{\lambda_j}$ represents the mean of its jth column,

$$\mathbf{L_{rc}} = \mathbf{L_r} - \begin{bmatrix} \boldsymbol{\lambda_1} & \boldsymbol{\lambda_2} \end{bmatrix}.$$

2.3.2. Scaling

Second, scaling is performed using the Euclidean distance of the point furthest from the origin. Let $\mathbf{g_{rcj}}$ represent the jth column vector of $\mathbf{G_{rc}}$. Then,

$$\mathbf{d_g^2} = \mathbf{g_{rc1}}^\top \mathbf{g_{rc1}} + \mathbf{g_{rc2}}^\top \mathbf{g_{rc2}}$$
$$d_g = \sqrt{\max_i \left(\mathbf{d_g^2}\right)}$$

and similarly for the reference matrix, determine the magnitude of the point furthest from the origin:

$$\mathbf{d_l^2} = \mathbf{l_{rc1}}^\top \mathbf{l_{rc1}} + \mathbf{l_{rc2}}^\top \mathbf{l_{rc2}}$$
$$d_l = \sqrt{\max_i \left(\mathbf{d_g^2}\right)}.$$

Then the scaled points for the data and reference matrix will be

$$\mathbf{G_{rcs}} = \frac{1}{d_g}\mathbf{G_{rc}}$$

and

$$\mathbf{L_{rcs}} = \frac{1}{d_l}\mathbf{L_{rc}}.$$

2.3.3. Rotation

Third, the points of the data matrix are rotated around the origin to fit the centered and scaled points of the reference row scores. Typically rotation is accomplished by the identification of landmark points that are used to match one shape to another. In the absence of these landmark points, an iterative process is used to match each data point with the median reference point, that is the point at or nearest to $(0, f(0))$, where $f(x) = \beta_2 x^2 + \beta_0$ is the curve fitted to the centered and scaled reference row points. The Euclidean distance from each data point to any nearest reference point is then summed as residuals. The rotation which minimizes this squared residual sum is then selected.

Let $(\tilde{l}_1, \tilde{l}_2)$ be the median point contained in $\mathbf{L_{rcs}}$. Let $\tilde{\theta}_l = \arctan2(\tilde{l}_2, \tilde{l}_1)$.

Then, for each ith row in $\mathbf{G_{rcs}}$:

- Let g_{rcsi1} and g_{rcsi2} be the first and second axis coordinates of the point represented by the ith row of G_{rcs}.
- Compute $\theta_i^* = \arctan2(g_{rcs2}, g_{rcs1}) - \tilde{\theta}_l$.
- The rotation matrix will accordingly be

$$\mathbf{\Theta_i^*} = \begin{bmatrix} \cos\theta_i^* & \sin\theta_i^* \\ -\sin\theta_i^* & \cos\theta_i^* \end{bmatrix}$$

Then,

$$\mathbf{G_{rcs}^{(i*)}} = \mathbf{G_{rcs}}\mathbf{\Theta_i^*}$$

- Let $d_k^{(i*)}$ be the Eucldiean distance of the kth row of $\mathbf{G_{rcs}^{(i*)}}$ to whichever row of $\mathbf{L_{rcs}}$ is nearest to it.
- Then the residual sum of squares for the ith rotation will be

$$D_{i*} = \sum_{k=1}^{n} d_k^{(i*)}$$

Accordingly, each potential rotation is evaluated with regard to its goodness-of-fit, in terms of the Euclidean distances from a one-to-many mapping of the points of the scores of the data scores to those of the reference scores.

Finally, then, the rotation matrix $\mathbf{\Theta}$ is chosen as the $\mathbf{\Theta_i}$ for which D_{i*} is the smallest.

The final step of the Procrustes fitting then rotates the matrix $\mathbf{G_{rcs}}$ by multiplying it by $\mathbf{\Theta}$:

$$\mathbf{\Gamma_r} = \mathbf{G_{rcs}}\mathbf{\Theta}$$

Similarly, for the CA column points in the data matrix, $\mathbf{G_c}$, apply the same transformations as above using the centering, scaling, and rotation based on the row values (performing Procrustes fit using the means, scaling, and rotations derived from the column values would cause the points to be misaligned):

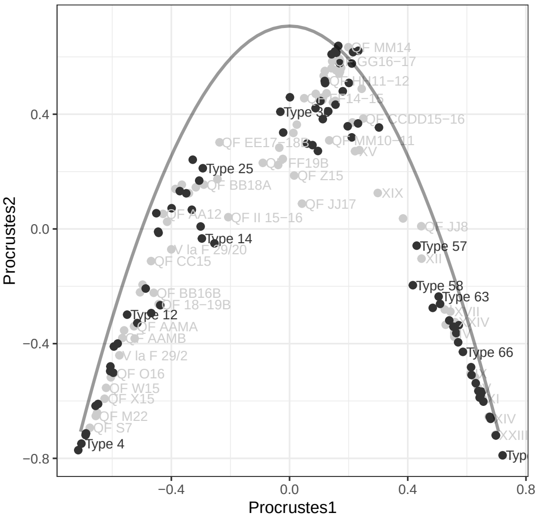

Figure 2. Procrustes-fitting CA scores of the `quattrofontanili` data included in the `lakhesis` package and their sequencing after projection onto the reference curve. Data scores are in black (rows/contexts) and gray (columns/find-types), with the curve of the reference scores shown as a line.

- Centering on the origin:

$$\mathbf{G_{cc}} = \mathbf{G_c} - \begin{bmatrix} \boldsymbol{\mu_1} & \boldsymbol{\mu_2} \end{bmatrix}$$

- Scaling:

$$\mathbf{G_{ccs}} = \frac{1}{d_x}\mathbf{G_{cc}}$$

- Rotation:

$$\mathbf{\Gamma_c} = \mathbf{G_{ccs}}\mathbf{\Theta}$$

Then, let $f(x) = \beta_2 x^2 + \beta_0$ be the "reference" curve fit to the points in $\mathbf{L_{rcs}}$. Let $y = f(x)$, and let (x_{ri}, y_{ri}) be the point on the line (x, y) closest to the ith row of $\mathbf{\Gamma_r}$, and let (x_{cj}, y_{cj}) be the point on (x, y) closest to the jth row of $\mathbf{\Gamma_c}$, in terms of their Euclidean distance. Let the vectors $\mathbf{x_r}$ and $\mathbf{x_c}$ represent all x_{ri} and x_{cj}, respectively. Note that these vectors are not simply a projection of the scores onto the first principal axis, since the best-fitting point on the reference curve is not orthogonal to the first principal axis. Figure 2 illustrates these steps from the initial CA plot of the data to the final fit of row and column scores.

2.4. Strand Selection

Let $\mathbf{s_r}$ then be the rankings of $\mathbf{x_r}$, and let $\mathbf{s_c}$ be the rankings of $\mathbf{x_c}$. The row and column rankings are swapped if the matrix $\mathbf{X}$ was transposed (as mentioned in Section 2.1 on page 4). The set $\{\mathbf{s_r}, \mathbf{s_c}\}$ determines a "strand" $\mathbf{H}$, a particular subset-seriation of the initial matrix $\mathbf{X}$ (this can also be conceived as the product of $\mathbf{X}$ and a permutation matrix).

The investigator selects any subset of rows and columns of $\mathbf{Y}$ to be $\mathbf{X}$, choosing specific rows and columns $n \leq u, k \leq v$, and selecting m number of row and column strands at their discretion. The `lakhesis` package allows the investigator to perform this task graphically using the CA plot, which will construct an $\mathbf{X}$ using selected row and column score points. Rows and columns which contain all 0s are automatically suppressed. At the end of the process of exploring and selecting seriations, there will be m number of strands, each with a find-type and a context ranking, $\mathbf{s_{r1}}, \mathbf{s_{c1}}, \ldots, \mathbf{s_{rm}}, \mathbf{s_{cm}}$.

3. Consensus Seriation: Lakhesis Analysis

In order to arrive at a consensus seriation, the rankings of the rows and columns are processed separately. Let $\tilde{u} \leq u$ be the number of rows which are contained in the strands, and $\tilde{v} \leq v$ be the number of columns (i.e., not every row or column in the original matrix $\mathbf{Y}$ needs to have been selected).

Given the missing values, a method of imputation is necessary in order to derive a consensus seriation using all of the available information from each strand. This problem is non-trivial, and finds analogues in ranked-voting or ranked-preference problems [e.g., 13, 45]. However, it differs from ranked-voting problems in that the objective is not to find a single winner or first-ranked element, nor it is to impute rankings on the basis of external variates such as a category or type as with preferences. Rather, it is to determine a single ranking on the basis of the information provided by a set of partial rankings from a larger data set, which is optimal for seriation according to a given principle.

Stringing together these ranked strands is here called "Lakhesis" analysis, named for the ancient Greek goddess who measured the thread of fate. Owing to the need to impute missing rankings, is an iterative process which works in an agglomerative fashion, building a consensus seriation from the bottom up. It uses the principle of concentration (see below, Section 4.2) in determining the order in which to regress strands.

(1) To initialize, perform pairwise linear regression on all strands' row and column elements separately. Each strand must have at least 4 joint elements with at least one other strand.
    (a) For any two strands, select one as the dependent (y), $\mathbf{H_y}$, and another as the independent (x) variate, $\mathbf{H_x}$.
(2) A regression is then performed in the following manner, to merge the two rankings. Regressions are performed separately for rows (i.e., s_{rx} on s_{ry}) and columns (s_{cx} on s_{cy}):
    (a) Remove all elements for which there is an `NA` value in either strand.
    (b) Use the regression line, $f(x) = \beta_1 x + \beta_0$, to project the ranks of the independent variate onto the dependent (thereby supplying rankings for any `NA` values in the dependent variate which are attested in the independent).
    (c) If $y \neq f(x)$, the mean of y and $f(x)$ is used for that element.
    (d) The values of the dependent and regressed independent variates are reranked, forming a new seriated incidence matrix, $\mathbf{H_{xy}}$.
(3) The concentration coefficient κ (Sec. 4.2) is computed for each pairwise regression's seriated incidence matrix, and the pair of strands is chosen which elicits the lowest concentration measure.
(4) The "merged" rankings from this seriation constitute the dependent variate for

the next regression.

(5) For all remaining strands, treat each as the independent variate, and the merged rankings from the previous regression as the dependent variate.

(6) Repeat Steps 2 – 5 until all strands have been regressed, resulting in a single consensus seriation.

Given that the order in which strands are regressed has a bearing on the resulting consensus seriation (i.e., a merged consensus seriation $\mathbf{H_{123}}$ may differ from $\mathbf{H_{213}}$), the use of the concentration principle to determine the selection of strands is necessary. Especially if the number of strands were to exceed 10, all possible permutations of the order of strands in their regression could not be evaluated in a reasonable amount of time. Similarly, PCA (which was initially explored as a method of harmonizing rankings) would only work as a method of ranking imputation if there were at least four elements jointly shared across all strands, a restriction which was undesirable given a potential need to perform consensus seriation on a large set of data in piecemeal fashion. Furthermore, as the rotation of axes in CA, as in PCA, can vary, it is possible for strands to have the same seriated sequence but in reverse order. Linear regression has a virtue in being able to accommodate identical sequences in reverse order, as their relationship will merely have a negative slope. Having at minimum four joint elements serves to mitigate any potential misalignment, but having a diagnostic criteria of agreement among strands with the resulting consensus will also enable the investigator to remove highly discrepant strands from the consensus (see Section 4.1).

Rather than slicing matrices in the R console and re-running commands, the process of seriating matrices and evaluating their consensus seriation is performed in the Lakhesis Calculator, which performs the above operations at the click of a button. The following section discusses the diagnostic criteria used in the Calculator to evaluate the quality of investigator-selected strands and their consensus seriation, via coefficients of agreement and optimality, as well as measures of goodness-of-fit for individual row and column elements, which can be used to improve the resultign seriation.

4. Critical Evaluation: Coefficients and Goodness-of-Fit

Two coefficients are introduced here, one of agreement, which is based on Spearman's rank correlation coefficient [47], ρ, and another of concentration, to address optimality for the particular case of binary data. Additional measures of optimality have been discussed by Hahsler et al. [20] (some of which assess the gradient of frequency seriations and so are not ideal for binary data). Additional measures that are applicable are implemented in Section 5.

4.1. Agreement

Agreement is computed on the basis of the square of Spearman's rank correlation coefficient [47], ρ, between one strand and the consensus seriation, omitting any elements with missing values. For example, agreement for one ranking `c(1,2,4,NA,3)` and another `c(NA, 4, 3, 2, 1)`, will involve first involve omitting to `NA` values to obtain `c(2,4,3)` and `c(4,3,1)`, whose squared correlation would come to $\rho^2 = 0.11$. For two strands, the squared Spearman correlation coefficients are obtained separately for rows, labeled ρ_r^2 and for columns, labeled ρ_c^2. The product of these two coefficients, $\rho_r^2 \rho_c^2$, represents how well two strands agree with one another in terms of their rankings.

Strands which have lower agreement can be deleted in order to see if the optimality of the overall consensus seriation will be improved.

4.2. Concentration

In this paper, an optimality measure is proposed by extending the Kendall-Doran concentration principle [12, 28], which is ideal for binary data, here called κ. The Kendall-Doran concentration principle is based on the notion that in a "perfect" seriation, one would see all attestations of any one column element in a contiguous order with no zeros interspersed, and so Kendall-Doran concentration measures the number of elements between the first and last attestation of a column, to be compared against the sum of the incidence matrix: if a seriation is most optimal, the concentration measure will be identical to the sum of the matrix (there will be no 0s between occurrences of a column).

The extension of the Kendall-Doran concentration principle is here extended here to apply not just to columns but also to rows, and also to weight it by the sum of the incidence matrix. This modified concentration measure is labeled κ, and is computed as follows. Let ν be the sum of all cells of the seriated incidence matrix (i.e., the strand) $\mathbf{H}$ ordered by the row and column rankings. Then let $j(\mathbf{H_{i\cdot}})$ be a function that represents the smallest column index of the ith row of $\mathbf{H}$ that contains a value of 1, and let $J(\mathbf{H_{i\cdot}})$ similarly be a function that represents the largest index that contains a value of 1. Similarly, let $i(\mathbf{H_{\cdot j}})$ and $I(\mathbf{H_{\cdot j}})$ be functions that return the smallest and largest row index of jth column of $\mathbf{H}$ which contains a value of 1. Then, the concentration coefficient κ is defined as:

$$\kappa = \frac{1}{2\nu}\left\{\sum_{i=1}^{n}\left[J(\mathbf{H_{i\cdot}}) - j(\mathbf{H_{i\cdot}})\right] + \sum_{j=1}^{k}\left[I(\mathbf{H_{\cdot j}}) - i(\mathbf{H_{\cdot j}})\right]\right\} \tag{1}$$

which will be bounded by 1, which indicates a perfect seriation, increasing the less optimal a seriation becomes (i.e., it is a "loss" measure).

4.3. Deviance Testing

Finally, a measure of goodness-of-fit can be achieved for each row and column element of $\mathbf{H}$ using a quadratic-logistic model, as has been conventional in ecology given its flexibility in representing the concentration [15, 49], along the lines of a chi-squared likelihood ratio test of deviance. That is, a row or column comprising a series of 0s and 1s, $\mathbf{h_{i\cdot}}$ or $\mathbf{h_{\cdot j}}$, represents the dependent variate, while the indices $\mathbf{i}$ or $\mathbf{j}$ represent the independent variate. Using a GLM, the model for the rows is

$$p(i) = \frac{1}{1 + \exp\left[-\beta_0 - \beta_1 i + \beta_2 i^2\right]} \tag{2}$$

and one likewise uses j and $p(j)$ for column model. It should be noted that the optimal condition for a seriation, as stipulated by the coefficient of concentration κ, would results in a failure of fit for the quadratic-logistic model, since there will be perfect separation of the 0/1 values along each column and row. Accordingly, this goodness-of-fit test is not performed for row and columns which exhibit perfect separation. Instead, these are assigned an `NA` value. But, where applicable for individual elements, the quadratic-logistic model provides a better means of assessing goodness-of-fit than using a straightforward measure of row or column concentration, since instances where rows and columns concentrate more densely around a central tendency will be penalized less than those that are uniformly distributed between their maximum and minimum ranking.

For rows and columns which are not perfectly separated, the goodness-of-fit test uses the familiar likelihood ratio test of the deviance of the model against that of a null hypothesis [10, 35], where deviance is

$$\operatorname{dev}(M) = 2[\log L(\hat{\mu}_S; y) - \log L(\hat{\mu}_M; y)] \tag{3}$$

$\log L(\hat{\mu}_M; y)$ is the log-likelihood under the model, and $\log L(\hat{\mu}_S; y)$ is the saturated model, and where $y = i$ or j, depending on whether rows or columns are being tested. The deviance test is here not being used to test different models on the rows or columns, but rather in an exploratory fashion to assess which rows and columns have better or worse fit using the quadratic-logistic model. Let H_0 be the null hypothesis with corresponding saturated model M_0, and let H_1 be the alternative hypothesis with M_1 being the fitted model. Given that the degrees of freedom will be one less than the total number of column or row elements (depending on whether row or columns are being tested) under the null hypothesis, and two less under the fitted model, the difference

$$\operatorname{dev}(M_0) - \operatorname{dev}(M_1) \tag{4}$$

will be chi-squared distributed with d.f. = 1. The resulting p value indicates the probability of obtaining a fit which is as or greater than the one obtained for that particular row or column. A table in the Lakhesis Calculator shows the row or elements with the highest p values, i.e., which fit the quadratic-logistic model more poorly and hence might not be conducive to seriation, and which the investigator may want to suppress or remove from the plot.

5. Application

In order to assess the quality of the method advocated here, nine different methods of seriating the same data were used on the same data set, `quattrofontanili` (Table 1). In addition to the original published seriation [9], the seriations were obtained by Procrustes-fit CA (Sec. 2) and by “Lakhesizing” three investigator-selected strands, contained in the `qfStrands` data object in the package (Sec. 3). Six other seriations were obtained using preexisting methods. DCA was performed using the `decorana()` function from `vegan`, PCA, PCA (Angular Distance), and the Bond-Energy Algorithm with Traveling Salesperson Problem Solver (BEA-TSP) were implemented using `seriate()` from `seriation`, nMDS was implemented using `monoMDS` from vegan, and a latent variable model using Bayesian ordination was performed using `boral`. A replication script has been included for this application.

To assess the optimality of each of these methods, the concentration coefficient coefficient κ (Eq. 1) is incorporated alongside other measures of optimality from the `criterion()` function in `seriation`: the measure of effectiveness (ME), weighted correlation coefficient (Cor_R), and two stress measures, Moore and Neumann [20, 5-6]. The first two of these are merit measures (indicating larger values are more

Seriation	Method
Original	Original seriation by Close-Brooks and Ridgway [9].
CA Procrustes	CA with Procrustes fitting to reference curve using `lakhesis`.
Lakhesis	Consensus seriation of three investigator-selected partial strands using Procrustes-fit CA using `lakhesis`.
DCA	Detrended correspondence analysis using `decorana()` in `vegan`.
PCA	Principal component analysis using `seriation`.
PCA Angle	Principal component analysis using angular distance using `seriation`.
BEA-TSP	Bond-energy algorithm with a traveling salesperson problem solver using `seriation`.
nMDS	Non-metric multidimensional scaling using `monoMDS()` in `vegan`.
LVA	Latent variable model using Bayesian ordination, taking the projection along the first axis, using `boral`.

Table 1. Nine methods of obtaining a seriated matrix (both row and column) for the `quattrofontanili` data.

Method	Effectiveness	Cor_R	Moore	Neumann	Concentration (κ)
Original	0.6311	270	4520	2196	4.771
CA Procrustes	0.9211	251	4810	2342	3.115
Lakhesis	0.9345	255	4826	2326	2.968
DCA	0.9394	233	4966	2414	2.996
PCA	-0.7387	200	4950	2454	6.090
PCA Angle	-0.4317	228	4968	2408	4.420
BEA-TSP	-0.5059	411	3848	1700	5.302
nMDS	0.1719	162	5380	2596	6.611
LVA	-0.8668	217	4948	2386	4.314

Table 2. Criteria of the quality of the seriated matrix obtained for nine different methods including that of the published original. The first two columns are merit measures (higher values are more optimal) and the last three columns are loss measures (lower values are more optimal).

optimal) and the last two are loss measures (lower values are more optimal). The concentration coefficient κ is a loss measure.

The results of these criteria are shown in Table 2. The results of Lakhesis analysis on three investigator-defined strands contained in `qfStrands` elicited the most optimal concentration coefficient $\kappa = 2.968$, with DCA close behind at $\kappa = 2.996$. DCA presented a slightly better measure of effectiveness (ME), with Lakhesis second. Lakhesis resulted in the third-highest weighted correlation (Cor_R), and came in third or fourth in terms of Moore and Neumann stress. Using Procrustes-fit CA resulted in the third-lowest concentration coefficient and ME, and was third or fourth in terms of Cor_R and Moore and Neuman stress. BEA-TSP resulted in the most optimal seriations in terms of Cor_R and Moore and Neumann stress, while that method performed poorly in terms of the other criteria, with DCA eliciting middling scores for those measures. Generally, PCA and nMDS produced the least optimal seriations. In sum, however, different measures of optimality appeared to favor different methods. BEA-TSP especially was outperforming all methods in terms of weighted correlation and the stress measures. It should however be noted that stress measures were developed for gradient, frequency seriations, and so may prove less reliable for binary data.

In order to investigate whether methods each producing their own particular se-

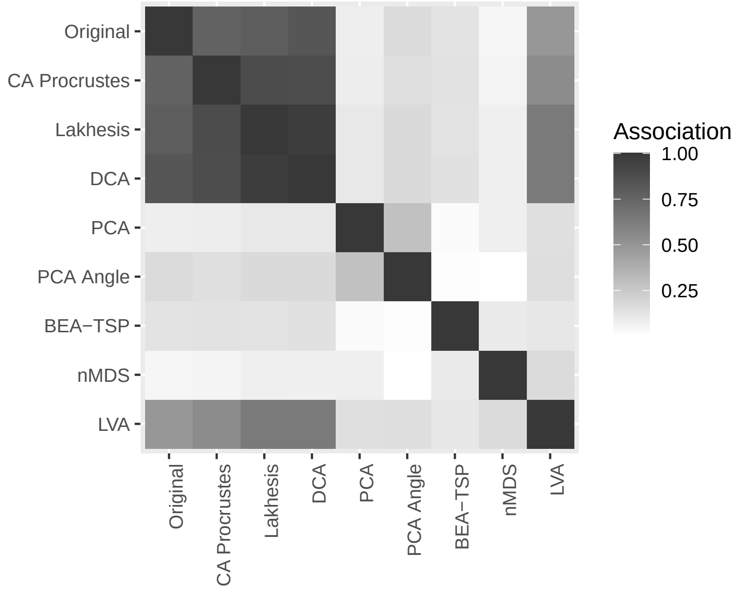

Figure 3. Coefficients of agreement, $\rho_r^2\rho_c^2$, among the methods tested in Table 1.

riation, or whether they were all more or less consonant with one another, a matrix plot of the agreement coefficients, $\rho_r^2\rho_c^2$, between each method was produced (Fig. 3). It is clear that the original seriation, Procrustes-fit CA, Lakhesis analysis, and DCA all have comparatively high coefficients of association, and are generally producing the seriations which agree with one another; these methods have the highest concentration and effectiveness measures as well. Lakhesis had coefficient of association of $\rho_r^2\rho_c^2 = 0.97$ with DCA, 0.89 with Procrustes-fit CA, and 0.79 with the original seriation. Beyond these highly associated methods, LVA via `boral` had an association of 0.64 with Lakhesis and DCA both. The other methods scored much lower measures of association. In particular, BEA-TSP had its largest measure of association of only 0.14 with DCA, indicating its seriation was highly idiosyncratic. Seriations attained through PCA, PCA Angle, and nMDS were all likewise highly particular, in addition to having generally low optimality measures.

There is accordingly a clear neighborhood of optimal solutions produced by correspondence analysis-affiliated methods (DCA and the two methods introduced here, Procrustes-fit CA and consensus seriation via Lakhesis). By providing an exploratory framework in which to identify investigator-selected seriations, Lakhesis analysis affords the ability to harmonize different well-seriated selections of elements into an optimal ordering that compares well with other established methods. Given that incidence matrices might have multiple potential seriations with high optimality, such coefficients of agreement are useful in determining whether there is a single, highly optimal neighborhood of solutions, generated by affiliated methods, as here.

6. Conclusion and Discussion

Seriation or ordination remains very much an art in terms of the choice of methods used to find an optimal ordering for matrix rows and columns. Procrustes-fitting correspondence analysis scores to a well-seriated reference curve represents another choice of methodology, which performs well in comparison with other methods, and even outperforms in terms of the concentration principle. However, successful seriations depend ultimately on the data themselves. Accordingly, this paper offers investigators a more rapid means of graphically selecting score points and re-running correspondence analysis, selecting their own seriated sequences, which can then be harmonized into a single consensus. This strategy is useful especially in cases where an investigator has a large matrix, whose plot may initially be difficult to read but which may contain a well-seriated subset of rows and columns. This is especially the case given flexible sampling frameworks, where the rows and columns of an incidence matrix may admit the incorporation of more or fewer elements.

Data Availability Statement

The `lakhesis` package containing the manual of functions and data to perform the results outlined in this paper is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/`. These results were generated using the development version (v. 0.0.2) hosted at `https://github.com/scollinselliott/lakhesis`, which will be submitted to CRAN along with this paper as the supporting citation, if accepted for publication.

Disclosure statement

The author reports no competing interests to declare.

Supplementary Material Online

The "Guide to Lakhesis" submitted with this paper is also included within the `lakehsis` package as a vignette.

The replication script includes the R code for producing the results in Section 5.

References

[1] P. Arabie and L. Hubert, *The Bond Energy Algorithm Revisited*, IEEE Transactions on Systems, Man, and Cybernetics: Systems 20 (1990), pp. 268–274.

[2] L. Bellanger, R. Tomassone, and P. Husi, *A Statistical Approach for Dating Archaeological Contexts*, Journal of Data Science 6 (2008), pp. 135–154.

[3] J.P. Benzécri, *L'Analyse des Données. Volume II. L'Analyse des Correspondances*, Dunod, Paris, 1973.

[4] E. Bølviken, E. Helskog, K. Helskog, I. Holm-Olsen, L. Solheim, and R. Bertelsen, *Correspondence Analysis: An Alternative to Principal Components*, World Archaeology 14 (1982), pp. 41–60.

[5] D. Carlson, *Quantitative Methods in Archaeology Using R*, Cambridge University Press, Cambridge, 2017.

[6] W. Chang and B. Borges Ribeiro, *shinydashboard: Create Dashboards with 'shiny'* (2021). Available at https://CRAN.R-project.org/package=shinydashboard.

[7] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges, *shiny: Web Application Framework for R* (2024). Available at https://shiny.posit.co, R package version 1.8.1.9001; https://github.com/rstudio/shiny.

[8] C. Chen, *Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices*, Statistica Sinica 12 (2002), pp. 7–29.

[9] J. Close-Brooks and D. Ridgway, *Veii in the Iron Age*, in *Italy Before the Romans*, D. Ridgway and F. Ridgway, eds., Academic Press, London, 1979, pp. 95–127.

[10] J. Cohen, P. Cohen, S. West, and L. Aiken, *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*, 3rd ed., Lawrence Erlbaum Associates, Mahwah, NJ, 2003.

[11] C. Czado, *Analyzing Dependent Data with Vine Copulas: A Practical Guide With R*, Springer, New York, 2019.

[12] J. Doran, *Computer Analysis of Data from the la Tène Cemetry at Münsingen-Rain*, in *Mathematics in the Archaeological and Historical Sciences*, F. Hodson, D. Kendall, and P. Táutu, eds., Edinburgh University Press, Edinburgh, 1971, pp. 422–431.

[13] A. Feuerverger, Y. He, and S. Khatri, *Statistical Significance of the Netflix Challenge*, Statistical Science 27 (2012), pp. 202–231.

[14] H. Gauch Jr. and G. Chase, *Fitting the Gaussian Curve to Ecological Data*, Ecology 55 (1974), pp. 1377–1381.

[15] H. Gauch Jr., G. Chase, and R. Whittaker, *Ordination of Vegetation Samples by Gaussian Species Distributions*, Ecology 55 (1974), pp. 1382–1390.

[16] D. Goodall and R. Johnson, *Maximum-Likelihood Ordination: Some Improvements and Further Tests*, Vegetatio 73 (1987), pp. 3–12.

[17] J. Gower and G. Dijksterhuis, *Procrustes Problems*, Oxford University Press, Oxford, 2004.

[18] M. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, London, 1984.

[19] M. Greenacre, *Correspondence Analysis in Practice*, Chapman and Hall/CRC, Boca Raton, FL, 2007.

[20] M. Hahsler, K. Hornik, and C. Buchcta, *Getting Things in Order: An Introduction to the R Package seriation*, Journal of Statistical Software 25 (2008), pp. 1–34.

[21] M. Hill, *Correspondence Analysis: A neglected multivariate technique*, Journal of the Royal Statistical Society Series C: Applied Statistics 23 (1974), pp. 340–354.

[22] M. Hill and H. Gauch, *Detrended Correspondence Analysis: an Improved Ordination Technique*, Vegetatio 42 (1980), pp. 47–58.

[23] H. Hirschfeld, *A Connection between Correlation and Contingency*, Mathematical Proceedings of the Cambridge Philosophical Society 31 (1935), pp. 520–524.

[24] F. Hui, *boral: Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R*, Methods in Ecology and Evolution 7 (2016), pp. 744–750.

[25] F. Hui, S. Taskinen, S. Pledger, S. Foster, and D. Warton, *Model-Based Approaches to Unconstrained Ordination*, Methods in Ecology and Evolution 6 (2015), pp. 399–411.

[26] P. Ihm, *A Contribution to the History of Seriation in Archaeology*, in *Classification – The Ubiquitous Challenge*, C. Weihs and W. Gaul, eds., Springer, Berlin, 2005, pp. 307–316.

[27] P. Ihm and H. Van Groenewoud, *Correspondence Analysis and Gaussian Ordination*, COMPSTAT Lectures 3 (1984), pp. 5–60.

[28] D. Kendall, *A Statistical Approach to Flinders Petrie's Sequence Dating*, Bulletin of the International Statistical Institute 40 (1963), pp. 657–680.

[29] D. Kendall, *A Mathematical Approach to Seriation*, Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences 269 (1970), pp. 125–134.

[30] D. Kendall, *Seriation from Abundance Matrices*, in *Mathematics in the Archaeological and Historical Sciences*, F. Hodson, D. Kendall, and P. Táutu, eds., Edinburgh University Press, Edinburgh, 1971, pp. 215–252.

[31] J. Kruskal, *Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis*, Psychometrika 29 (1964), pp. 1–27.

[32] C. Lipo, M. Madsen, and R. Dunnell, *A Theoretically-Sufficient and Computationally-Practical Technique for Deterministic Frequency Seriation*, PLOS ONE 10 (2015), p.

e0124942.

[33] K. Lockyear, *Simulation, Seriation and the Dating of Roman Republican Coins*, Journal of Computer Applications in Archaeology 5 (2022), pp. 1–18.

[34] W. McCormick, P. Schweitzer, and T. White, *Problem Decomposition and Data Reorganization by a Clustering Technique*, Operations Research 20 (1972), pp. 993–1009.

[35] P. McCullagh, *Generalized Linear Models*, 2nd ed., Routledge, Boca Raton, 1989.

[36] P. Murrell, *R Graphics*, Chapman and Hall/CRC, Boca Raton, 2005.

[37] O. Nenadic and M. Greenacre, *Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package*, Journal of Statistical Software 20 (2007), pp. 1–13.

[38] J. Oksanen, G. Simpson, F. Blanchet, R. Kindt, P. Legendre, P. Minchin, R.B. O'Hara, P. Solymos, M. Stevens, E. Szoecs, H. Wagner, M. Barbour, M. Bedward, B. Bolker, D. Borcard, G. Carvalho, M. Chirico, M. De Caceres, S. Durand, H. Evangelista, R. FitzJohn, M. Friendly, B. Furneaux, G. Hannigan, M. Hill, L. Lahti, D. McGlinn, M.H. Ouellette, E. Cunha, T. Smith, A. Stier, C. ter Braak, and J. Weedon, *vegan: Community Ecology Package* (2024). Available at https://CRAN.R-project.org/package=vegan.

[39] M. Peeples and G. Schachner, *Refining Correspondence Analysis-Based Ceramic Seriation of Regional Data Sets*, Journal of Archaeological Science 39 (2012), pp. 2818–2827.

[40] J. Poblome and J. Groenen, *Constrained Correspondence Analysis for Seriation of Sagalassos Tablewares*, in *The Digital Heritage of Archaeology: Computer Applications and Quantitative Methods in Archaeology*, M. Doerr and A. Sarris, eds., Hellenic Ministry of Culture, Athens, 2003, pp. 301–306.

[41] G. Popovic, F. Hui, and D. Warton, *Fast Model-Based Ordination with Copulas*, Methods in Ecology and Evolution 13 (2022), pp. 194–202.

[42] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna (2024). Available at http://www.r-project.org/.

[43] D. Roberts, *Comparison of Distance-Based and Model-Based Ordinations*, Ecology 101 (2020), p. e02908.

[44] W. Robinson, *A Method for Chronologically Ordering Archaeological Deposits*, American Antiquity 16 (1951), pp. 293–301.

[45] M. Schulze, *A New Monotonic, Clone-Independent, Reversal Symmetric, and Condorcet-Consistent Single-Winner Election Method*, Social Choice and Welfare 36 (2011), pp. 267–303.

[46] C. Sievert, J. Cheng, and G. Aden-Buie, *bslib: Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'* (2024). Available at https://rstudio.github.io/bslib/, R package version 0.7.0, https://github.com/rstudio/bslib.

[47] C. Spearman, *The Proof and Measurement of Association between Two Things*, American Journal of Psychology 15 (1904), pp. 72–101.

[48] C. ter Braak, *Canonical Correspondence Analysis: A New Eigenvector Method for Multivariate Direct Gradient Analysis*, Ecology 67 (1986), pp. 1167–1179.

[49] C. ter Braak and C. Looman, *Weighted Averaging, Logistic Regression and the Gaussian Response Model*, Vegetatio 65 (1986), pp. 3–11.

[50] C. ter Braak and I. Prentice, *A Theory of Gradient Analysis*, Advances in Ecological Research 18 (1988), pp. 271–317.

[51] M. van de Velden, P. Groenen, and J. Poblome, *Seriation by Constrained Correspondence Analysis: A Simulation Study*, Computational Statistics & Data Analysis 53 (2009), pp. 3129–3138.

[52] W. Venables and B. Ripley, *Modern Applied Statistics with S*, 4th ed., Springer, New York, 2002.

[53] H. Von Wehrden, J. Hanspach, H. Bruelheide, and K. Wesche, *Pluralism and Diversity: Trends in the Use and Application of Ordination Methods 1990-2007*, Journal of Vegetation Science 20 (2009), pp. 695–705.

[54] D. Wartenberg, S. Ferson, and F. Rohlf, *Putting Things in Order: A Critique of Detrended Correspondence Analysis*, The American Naturalist 129 (1987), pp. 434–448.

[55] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, New York (2016).

[56] H. Wickham, J. Hester, and J. Bryan, *readr: Read Rectangular Text Data* (2024). Available at https://readr.tidyverse.org, R package version 2.1.5, https://github.com/tidyverse/readr.

[57] T. Yee, *A New Technique for Maximum-Likelihood Canonical Gaussian Ordination*, Ecological Monographs 74 (2004), pp. 685–701.

[58] T. Yee, *VGAM: Vector Generalized Linear and Additive Models* (2024). Available at https://CRAN.R-project.org/package=VGAM.