Seriation by constrained correspondence analysis: a simulation study

Michel van de Velden,* Patrick J.F. Groenen,† and Jeroen Poblome‡

October 12, 2007

Econometric Institute Report EI 2007-40

Abstract

One of the many areas in which Correspondence Analysis (CA) is an effective method, concerns ordination problems. For example, CA is a well-known technique for the seriation of archaeological assemblages. A problem with the CA seriation solution, however, is that only a relative ordering of the assemblages is obtained. To improve the usual CA solution, a constrained CA approach that incorporates additional information in the form of equality and inequality constraints concerning the time points of the assemblages may be considered. Using such constraints, explicit dates can be assigned to the seriation solution. In this paper, we extend the set of constraints that can be used in CA by introducing interval constraints. That is, constraints that put the CA solution within a specific time-frame. Moreover, we study the quality of the constrained CA solution in a simulation study. In particular, by means of the simulation study we are able to assess how well ordinary and constrained CA can recover the true time order. Furthermore, for the constrained approach, we see how well the true dates are retrieved. The simulation study is set up in such a way that it mimics the data of a series of ceramic assemblages consisting of the locally produced tableware from Sagalassos (SW Turkey). We find that the dating of the assemblages on the basis of constraints appears to work quite well.

1 Introduction

An important classification problem concerns the ordination of objects in time. For example, in archaeology one is often interested in ordering a set of assemblages on

*Corresponding author: Michel van de Velden, Econometric Institute, Erasmus Universiteit Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, the Netherlands; vandevelden@few.eur.nl

†Patrick Groenen, Econometric Institute, Erasmus Universiteit Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, the Netherlands; groenen@few.eur.nl

‡Jeroen Poblome, Sagalassos Archaeological Research Project, Katholieke Universiteit Leuven, Blijde Inkomststraat 21-bus 3314, B-3000 Leuven, Belgie; jeroen.poblome@arts.kuleuven.be

the basis of collected artefacts. The basic assumption underlying such an ordination is the single-peakedness of the distribution of the artefacts over time. That is, artefacts, for example, pots, shards, coins, etc., have a life cycle that goes from non-existence, to popularity to disappearance. A popular statistical method that can be used to solve the ordination problem is correspondence analysis (see, e.g., Greenacre, 1984).

Correspondence analysis (CA) renders, simultaneously, an ordination of the artefacts and assemblages on the basis of the archaeological data. For this purpose, data are gathered in such a way that, for example, rows of the data matrix correspond to assemblages and columns to artefacts. The cell elements of the data matrix then either indicate the presence of a type of artefact in an assemblage, or the frequency of artefacts encountered at a certain assemblage. The first type of data, where presences are typically denoted by ones and absences by zeros, is usually referred to as incidence data. The second type is called abundance data. In this study, we focus on seriation of abundance data.

A CA seriation solution only provides a relative ordering of the assemblages. Without additional information it is impossible to determine the actual order in time. That is, we cannot say which assemblages are older or newer. Typically, additional archaeological information is present or can be inferred from the data, so that the direction of the order can be determined. In an achaeological setting, it is not uncommon that explicit information concerning dates of certain assemblages is available. For example, adjacent assemblages at the same depth should most likely be attributed to the same period and an assemblage that is physically below another is older than one above it. Also, for some assemblages additional information may be present (such as the find of dated objects) that make it possible to assign an exact date. This additional information is ignored in the standard CA approach. However, by using a constrained CA approach, the CA solution is forced to be in accordance with the additional information. For example, if it is known that two assemblages are from the same date, we can constrain the CA solution in such a way that the CA scores for these assemblages are equal.

The importance and usefulness of introducing constraints in CA has been recognized and discussed by several authors. For example, Böckenholt and Böckenholt (1990), Takane et al. (1991), Böckenholt and Takane (1994), consider several approaches for incorporating linear constraints in CA. Ritov and Gilula (1993) and Groenen and Poblome (2003) also consider inequality constraints, that is, constraints that impose a certain order on the CA scores. Moreover, Groenen and Poblome (2003) show that by using linear constraints it becomes possible to assign explicit dates to all assemblages in an archeaological study, if the exact dates of at least two assemblages are known . In an empirical study concerning tableware from Sagalassos (SW Turkey), Groenen and Poblome (2003) and Poblome and Groenen (2003) showed that results of such a constrained CA approach appeared to be plausible. However, little can be said about the accuracy of the method. As the exact dates of most assemblages in an archeaological study are unknown, there is no way of knowing how well constrained CA is able to reconstruct the underlying time axis. To overcome this problem we study the performance of the constrained CA approach

under various conditions by using a simulation study. By mimicking archaeological data, we study the quality of the explicit dating obtained using constrained CA.

The remainder of this paper is organized as follows. In the next section, we set off with a brief introduction of CA and constrained CA. Then, in Section 3, we describe the design of the simulation study, which is tailored after the Sagalassos ceramic tableware data. Results of the simulation study are described in Section 4 and we conclude with a discussion.

2 Correspondence Analysis

In CA, scores are obtained for row and column variables of a contingency table in such a way that the deviations from the independence model are best approximated. There exist several excellent expositions of CA such as Greenacre (1984). For a description of the method in the context of archaeology we refer to Shennan (1988) and Cool and Baxter (1999).

Mathematically the correspondence analysis objective can be expressed as

$$\min_{\mathbf{a},\mathbf{b}} L(\mathbf{a},\mathbf{b}) = \left\| \mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' - \mathbf{D}_r \mathbf{a}\mathbf{b}'\mathbf{D}_c \right) \mathbf{D}_c^{-1/2} \right\|^2, \tag{1}$$

where

$$\|\mathbf{X}\|^2 = \text{trace}(\mathbf{X}'\mathbf{X})$$

denotes the sum of squared elements of $\mathbf{X}$, $\mathbf{P}$ is the so-called $n \times p$ correspondence matrix with as its ijth element the number of artefacts j encountered at assemblage i divided through the total number of observations, $\mathbf{r}$ and $\mathbf{c}$ are vectors of, respectively, row and column totals of $\mathbf{P}$, and $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices with diagonal elements the vectors $\mathbf{r}$ and $\mathbf{c}$. Note that the sum of all elements of $\mathbf{P}$ equals one, i.e., $\sum_{i,j} p_{ij} = 1$. To identify $\mathbf{a}$ and $\mathbf{b}$, we standardize $\mathbf{b}$ so that

$$\mathbf{b}^{'}\mathbf{D}_c\mathbf{b} = 1.$$

A solution for $\mathbf{a}$ and $\mathbf{b}$ can be obtained by using the singular value decomposition of $\mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{D}_c^{-1/2}$. Moreover, the solutions for the rows and columns, that is the vectors $\mathbf{a}$ and $\mathbf{b}$ are related in the following way:

$$\mathbf{a} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{b}, \tag{2}$$

and

$$\mathbf{b} = \frac{1}{\lambda}\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{a}, \tag{3}$$

where λ is the largest singular value of $\mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{D}_c^{-1/2}$. The score vectors for the rows satisfy $\mathbf{a}^{'}\mathbf{D}_r\mathbf{a} = \lambda^2$. For details on these properties, see, for example, Greenacre (1984).

In an archaeological setting, the vectors $\mathbf{a}$ and $\mathbf{b}$ may represent the score vectors for the assemblages and artefacts respectively. In such a case, the ordering in time

of the assemblages can be inferred from the elements of $\mathbf{a}$. However, if $(\mathbf{a}, \mathbf{b})$ is a solution then $(-\mathbf{a}, -\mathbf{b})$ is also a solution. Hence, CA does not define the direction of the scale and only a relative order is obtained.

3 Constrained Correspondence Analysis

Groenen and Poblome (2003) proposed a method that allows explicit dating of the assemblages by incorporating time constraints. That is, by using information provided by the archaeologist concerning dates of some assemblages, the score vector obtained from CA is related to explicit dates. We distinguish four different types of constraints:

1. For some assemblages the exact dates are known.
2. For some assemblages it is known that they are from the same date.
3. For some assemblages the order in time is known.
4. For some assemblages it is known that they are from before or after a specific date.

The first three types of constraints were used in the study of Sagalassos Tableware described in Groenen and Poblome (2003). The fourth type of constraint is new and has not been applied before.

For the first type of constraints (Type 1), Groenen and Poblome restricted the specific dates to the assemblages in a linear fashion. These linear constraints can be formulated as follows. Let y_i denote the date (year) corresponding to assemblage i. We then restrict the CA score corresponding to the ith assemblage to be linearly related to the date. That is, we impose: $a_i = d_0 + d_1 y_i$, where d_0 is an unknown constant and d_1 gives the slope of the line. Both the constant d_0 and slope d_1 must be estimated.

Using the linear constraint, it becomes possible to assign dates to all assemblages. For example, suppose that the dates of three assemblages, say A1, A2, and A6, are known. Then, constrained CA ensures that the scores corresponding to these points (i.e., a_1, a_2, and a_6) are linearly related to the known dates. Thus, we can draw a line going through these three points. Then, as is illustrated in Figure 1, the date for assemblage A3 can be obtained by considering the point on the line corresponding to the obtained constrained CA score a_3.

For the second type of constraint (Type 2) we merely require certain scores to be equal to each other. Hence, if assemblage i must have the same dat as assemblage j (for $i \neq j$) we impose $a_i = a_j$. Hence, these equality constraints are also linear constraints on $\mathbf{a}$.

Both the equality constraints and the linear date constraints can be expressed algebraically as $\mathbf{Hd} = \mathbf{a}$, where $\mathbf{H}$ is a design matrix (consisting of ones and known

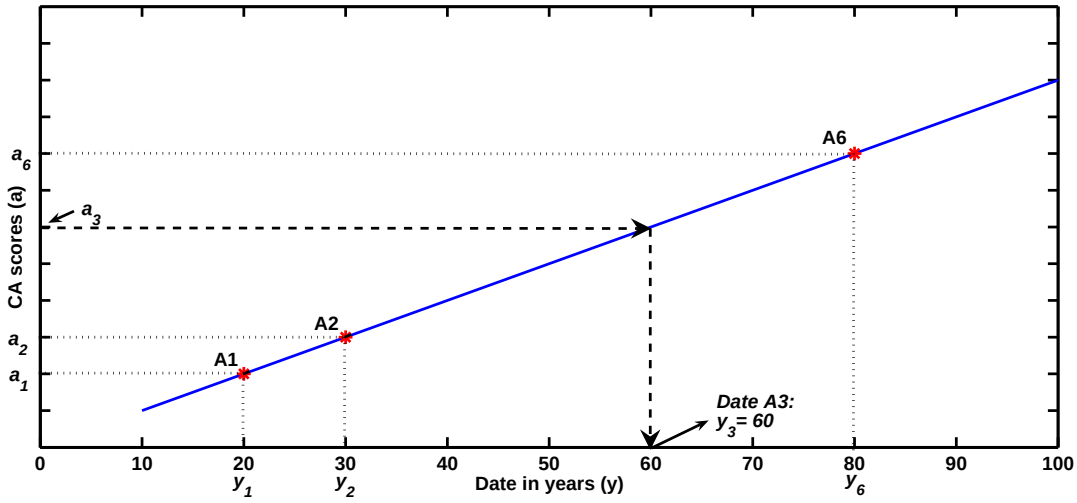

Figure 1: Assigning dates to scores and vice versa. A1, A2, and A3 have known dates and CA coordinates are linearly related to these dates. The date of A3 is not known but can be reconstructed from its CA analysis score a3 and the dotted line linking the CA score to an actual date.

years and columns of dummy variables, one for each group of assemblages), $\mathbf{d}$ is a vector of coefficients (that is, a vector consisting of the constant term d_0, the slope d_1 and coefficients used to impose the equality constraints) and $\mathbf{a}$ is the vector of the constrained correspondence analysis scores. To illustrate how the equality and linear constraints work, consider the following example. The dates of the assemblages A1, A2 and A6 are known to be, respectively, 20, 30, and 80. Furthermore, it is known that assemblages A4 and A5 stem from the same date. Then, if there are only 6 assemblages in total, we let

$$\mathbf{H} = \begin{pmatrix} 1 & 20 & 0 & 0 \\ 1 & 30 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 80 & 0 & 0 \end{pmatrix} \text{ and } \mathbf{d} = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \end{pmatrix},$$

so that

$$\mathbf{Hd} = \begin{pmatrix} d_0 + 20d_1 \\ d_0 + 30d_1 \\ d_2 \\ d_2 \\ d_3 \\ d_0 + 80d_1 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix} = \mathbf{a}.$$

Böckenholt and Takane (1994), showed that solving the correspondence analysis objective (1) subject to linear constraints $\mathbf{Hd} = \mathbf{a}$, can be reformulated by using

the so-called null-space approach. In this approach, we first calculate the null-space of $\mathbf{H}'$, that is we obtain $\mathbf{H}_0$ satisfying $\mathbf{H}'\mathbf{H}_0 = \mathbf{0}$. Then, the restrictions $\mathbf{Hd} = \mathbf{a}$ can be simplified by premultiplying both sides by $\mathbf{H}_0'$ so that $\mathbf{H}_0'\mathbf{a} = \mathbf{0}$. Rephrasing the constraints in this fashion has as advantage that we do not need to explicitly calculate the parameter vector $\mathbf{d}$.

The third type of constraints (Type 3) concerns order restrictions. If it is known that assemblage i is from a later date than assemblage j, the CA scores can be restricted to satisfy this relationship as well, that is, $a_i \geq a_j$. This type of inequality constraint can be expressed as $\mathbf{Ga} \geq \mathbf{0}$, where $\mathbf{G}$ is a design matrix whose rows correspond to an inequality. The column elements of $\mathbf{G}$ correspond to assemblages involved in the inequality. They are minus one (for the older assemblage) and plus one (for the newer assemblage). For example, if it is know that assemblage A3 is from a date earlier that assemblage A5, $\mathbf{G} = \begin{pmatrix} 0 & 0 & -1 & 0 & 1 & 0 \end{pmatrix}$ so that $\mathbf{Ga} = a_5 - a_3 \geq 0$.

Finally, the fourth type of constraint (Type 4) requires that certain assemblages are from before or after a specified date. This constraint can be implemented only if we have Type 1 constraints. In that case, the linear relationship between scores and dates makes it possible to link actual dates to CA scores and vice versa. Hence, using this relationship we are able to re-express inequality restrictions for the dates as inequality restrictions for the scores. These inequality constraints can then be implemented in a similar fashion as Type 3 constraints. For example, if it is known that assemblage A4, stems from before the year 70 it follows that the CA score must be lower than the score corresponding to the year 70. To find the appropriate score we need two pairs of dates and years. Consider again the situation sketched in Figure 1, where A1 and A2 are known to be from the years 20 and 30 respectively. Then, it is not difficult to see that

$$\begin{aligned} a_4 &\leq a_1 + (70-20)\frac{(a_2 - a_1)}{(30-20)}, \\ 0 &\leq a_1 + 5(a_2 - a_1) - a_4, \\ 0 &\leq -4a_1 + 5a_2 - a_4. \end{aligned}$$

Hence, for the general case where y_i and y_j, with $y_j > y_i$, denote the known years corresponding to the ith and jth assemblages, the restriction that the kth assemblage is from before y_k can be expressed as

$$0 \leq (1 - \gamma_k)a_i + \gamma_k a_j - a_k,$$

where $\gamma_k = (y_k - y_i)/(y_j - y_i)$. Thus, Type 4 constraints can be expressed as $\mathbf{Za} \geq \mathbf{0}$, where $\mathbf{Z}$ is a design matrix with elements $1, -1, \gamma_k$ and $-\gamma_k$ in the appropriate places. We refere to these constraints as *interval* constraints. Defining

$$\mathbf{G}^* = \begin{pmatrix} \mathbf{G} \\ \mathbf{Z} \end{pmatrix},$$

allows us to express all inequality constraints as $\mathbf{G}^*\mathbf{a} \geq \mathbf{0}$. For ease of notation, we drop the superscribed * from here on so that the matrix $\mathbf{G}$ denotes the design matrix for all inequality constraints.

The objective of constrained CA that incorporates all types of constraints is to minimize (1) subject to the restrictions $\mathbf{H}_0'\mathbf{a} = \mathbf{0}$ and $\mathbf{G}\mathbf{a} \geq \mathbf{0}$. We can solve this problem by using a so-called alternating least-squares algorithm that alternates between solving the objective for $\mathbf{a}$ for fixed column-score vector $\mathbf{b}$, and for $\mathbf{b}$ for fixed row-score vector $\mathbf{a}$. The solution for $\mathbf{b}$ for fixed $\mathbf{a}$ can be obtained as follows. First, note that, for the unconstrained CA solution the solution for the row and column scores are related through formula's (2) and (3). As we do not impose any restriction on $\mathbf{b}$ other than its standardization, the optimal column-scores can simply be obtained from (3).

To calculate the row-score vector $\mathbf{a}$ given a certain column score vector $\mathbf{b}$ we must solve the following optimization problem:

$$\min_{\mathbf{a}} \left\| \mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' - \mathbf{D}_r \mathbf{a}\mathbf{b}'\mathbf{D}_c \right) \mathbf{D}_c^{-1/2} \right\|^2, \tag{4}$$

$$s.t.\ \mathbf{H}_0'\mathbf{a} = \mathbf{0} \text{ and } \mathbf{G}\mathbf{a} \geq \mathbf{0}.$$

However, if $\mathbf{b}$ is (assumed) known and $\mathbf{b}'\mathbf{D}_c\mathbf{b} = 1$, we can rewrite the problem as follows:

$$\begin{aligned}
&\left\| \mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' - \mathbf{D}_r \mathbf{a}\mathbf{b}'\mathbf{D}_c \right) \mathbf{D}_c^{-\frac{1}{2}} \right\|^2 = \\
&\operatorname{trace}\left[\left(\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{D}_c^{-\frac{1}{2}} - \mathbf{D}_r^{\frac{1}{2}} \mathbf{a}\mathbf{b}'\mathbf{D}_c^{\frac{1}{2}} \right)' \left(\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{D}_c^{-\frac{1}{2}} - \mathbf{D}_r^{\frac{1}{2}} \mathbf{a}\mathbf{b}'\mathbf{D}_c^{\frac{1}{2}} \right) \right] = \\
&\operatorname{trace}\left(\mathbf{a}'\mathbf{D}_r\mathbf{a} \right) - 2 \operatorname{trace}\left(\mathbf{b}' \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right)' \mathbf{a} \right) + \operatorname{trace}\left(\mathbf{D}_c^{-1} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right)' \mathbf{D}_r^{-1} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \right) = \\
&\operatorname{trace}\left[\left(\mathbf{D}_r^{\frac{1}{2}} \mathbf{a} - \mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{b} \right)' \left(\mathbf{D}_r^{\frac{1}{2}} \mathbf{a} - \mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{b} \right) \right] + e = \\
&\left\| \mathbf{D}_r^{\frac{1}{2}} \mathbf{a} - \mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{b} \right\|^2 + e
\end{aligned}$$

where

$$e = \operatorname{trace}\left(\mathbf{D}_c^{-1} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right)' \mathbf{D}_r^{-1} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \right) - \operatorname{trace}\left(\mathbf{b}' \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right)' \mathbf{D}_r^{-1} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{b} \right).$$

As e is constant for fixed $\mathbf{b}$, the minimization problem (4) is equivalent to

$$\min_{\mathbf{a}} \left\| \mathbf{D}_r^{1/2} \mathbf{a} - \mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{b} \right\|^2,$$

$$s.t.\ \mathbf{H}_0'\mathbf{a} = \mathbf{0} \text{ and } \mathbf{G}\mathbf{a} \geq \mathbf{0}.$$

This is a least-squares problem with linear equality and inequality constraints. A solution to this type of problem can be obtained by transforming it to a nonnegative least-squares problem (see, Lawson and Hanson, 1974).

Combining the results we propose the following algorithm:

1. For fixed $\mathbf{b}$, satisfying the standardization constraint $\mathbf{b}'\mathbf{D}_c\mathbf{b} = 1$ use the algorithm described in Lawson and Hanson (1974; pp168-169) to solve

$$\min_{\mathbf{a}} \left\| \mathbf{D}_r^{1/2} \mathbf{a} - \mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}' \right) \mathbf{b} \right\|^2,$$

$$s.t.\ \mathbf{H}_0'\mathbf{a} = \mathbf{0} \text{ and } \mathbf{G}\mathbf{a} \geq \mathbf{0}.$$

2. Let $\mathbf{b}^* = \mathbf{D}_c^{-1}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^{'}\right)^{'}\mathbf{a}$, where $\mathbf{a}$ is the score vector obtained in step 1. Rescale $\mathbf{b}^*$ as $\mathbf{b} = \rho\mathbf{D}_c^{-1/2}\mathbf{b}^*$, where $\rho = (\mathbf{b}^{*\prime}\mathbf{D}_c\mathbf{b}^*)^{-1/2}$ so that $\mathbf{b}'\mathbf{D}_c\mathbf{b} = 1$. Insert the thus obtained score vector $\mathbf{b}$ in the first step of this algorithm and repeat until subsequent solutions remain constant up to a (small) pre-specified constant.

By alternating between these two steps $L(\mathbf{a}, \mathbf{b})$ decreases monotonically. Unfortunately, it can not be guaranteed that the thus obtained minimum is a global minimum. Therefore, it is better to use several random starts and select the thus obtained best solution.

4 The Simulation Study

Constrained CA applied to Sagalassos tableware yielded results that were compatible with the archaeological knowledge (Poblome and Groenen, 2003 compared to Poblome, 1999). However, as the true dates of most assemblages are unknown, it is difficult to ascertain how well the new method works. It is also unclear how accurate the explicit dating is for the assemblages that had no exact dating. To consider the performance of the constrained CA approach we use a simulation study. The idea of using a simulation study in the context of archaeology is not new (Graham et al., 1976; Herzog and Scollar, 1988; Lockyear, 1991). Some of these studies involve the simulation of very specific data, for example, simulation of coin hoard formation (Lockyear, 1991) or cemetery data (Graham et al.,1976). Herzog and Scollar (1988) provide a general simulation framework that also allows the generation of abundance data. Their method is incorporated in the Bonn Archaeological Software Package (BASP, http://www.uni-koeln.de/∼al001/basp.html). As BASP does not allow simulation of the constraints, we developed our own simulation study.

4.1 Data Generating Process

We consider an archaeological setting that covers data for a certain pre-determined timespan. We divide the complete timespan into equal-length time-intervals. For each time-interval, we randomly determine the number of assemblages that correspond to that period. Each of these assemblages is represented as a row in the data table. Then, for each time-interval, we simulate the number of different types of artefacts that are introduced during the interval. Each artefact type corresponds to a column in our data table. Next, for each of the artefact types, we randomly determine the total number of artefacts that was produced. These totals are the column marginals of the complete table. Now we have a table where the rows represent assemblages associated to certain periods. The columns represent artefacts corresponding to certain time-periods and the column totals are known. The next step is to simulate, for each observed artefact, the time between introduction and disappearance. Then, by adding this time to the introduction time, we can assign the artefacts to appropriate assemblages. That is, an assemblage that corresponds

to the time-interval containing the time it disappeared. Simulating data in this manner yields a table in which we have several assemblages and several artefacts for each time-period. The cell elements of the table are the number of times that artefact j was found at assemblage i. We call this table the complete data table as it contains all assemblages and artefacts.

Note that, the complete data table contains all assemblages and all types of artefacts. However, in practice, the archaeologist will not have access to all assemblages, and not all artefacts are retrieved. Instead, typically, only a fraction of the true assemblages and artefacts will be present on the archaeological site. Therefore, we randomly select observed assemblages and artefacts, that is, we draw rows and columns from the complete data table, to determine the observed data matrix.

4.1.1 Model Parameters

We model our simulation study after the Sagalassos tableware study by setting the simulation parameters in such a way that the simulated data matrix resembles the Sagalassos data. We consider a total time span of 1000 years that we divide into 50 periods of 20 years. For each period, we randomly determine the number of assemblages between 1 and 5. This means that on average there will be 3 assemblages corresponding to the same interval. The total number of assemblages (rows) thus lies between 50 (one assemblage in each time-period) and 250 (five assemblages in each time-period). Similarly, corresponding to each period we randomly draw a number between 1 and 10 that gives the number of types of artefacts that were introduced. Hence, on average 5.5 different types of artefacts are introduced during the same time-interval. The total number of artefact types (columns) lies between 50 (exactly one artefact in each time period) and 500 (10 types of artefacts in each time period).

Next, we simulate for each artefact, the total number that was produced (the column marginals of the complete table). Typically in archeological settings, there are many artefacts for which we have a low total number of observations and few artefacts that are frequently observed. To mimic a distribution that has a high peak for low values while covering a large range of values, we use a Gamma distribution with parameters 1/3 and 1500. The simulated numbers are rounded to obtain an integer representing the number of introduced artefacts of a certain type. Figure 2 shows a histogram of the sample distribution based on 100 draws.

To determine the lifetime for an individual artefact we again use a Gamma distribution. For the choice of parameters we consider two scenarios. In the first scenario, the average lifetime of an artefact is approximately one generation: 30 years. We achieve this by choosing a Gamma distribution with parameters 1.5 and 20. For the second scenario we choose a Gamma distribution with parameters 1.5 and 10 so that the average lifetime equals 15 years. The density functions for these two scenarios are plotted in Figure 3. On the basis of the introduction time and the randomly determined lifetime, an artefact can be assigned to an appropriate assemblage. If more than one assemblage exists for the same time period, the artefact is randomly assigned to one of the assemblages.

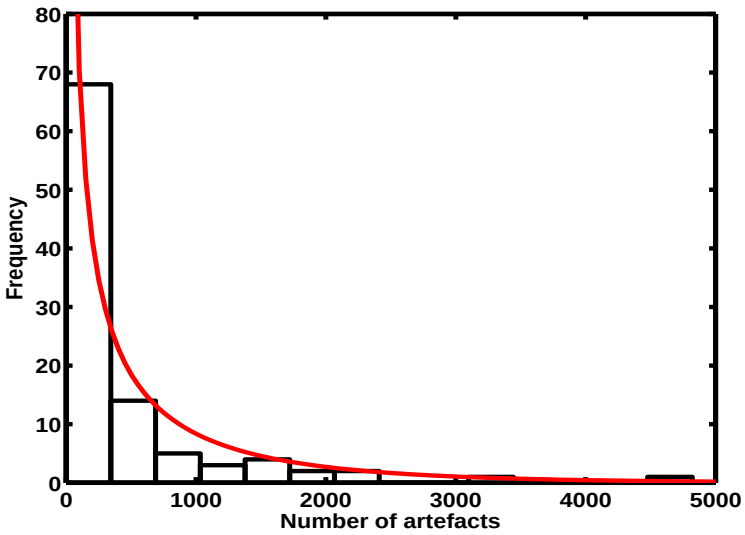

Figure 2: Histogram of 100 draws from a Gamma(1/3,1500) distribution and the theoretical distribution.

After assigning all artefacts to appropriate assemblages, we obtain a matrix in which we have observations for all assemblages and all artefacts. To determine the observed artefacts and assemblages we randomly draw from this complete set. We consider two cases:

1. From the total set of assemblages (which ranges between 50 and 250),we draw 20 assemblages.

2. From the total set of assemblages we draw 30 assemblages.

In both cases, the number of artefact types (the number of columns) is determined by randomly drawing 100 types from the complete set. After the selection of assemblages (rows) and artefact types (columns) the situation may occur that certain rows or columns have zero observations. Such rows and columns are deleted from the observation matrix.

Note that the correct time interval for each pottery type and assemblage is known. This knowledge can be used to generate constraints and to assess the quality of our approximation. For Type 1 constraints (linear time constraints) we consider two cases: three dates are known or four dates are known. For Type 2 constraints (equality constraints), we take the number of existing equalities if there are three or less equalities. If there are more than three equalities we randomly determine the number of constraints between three and the total number of equalities minus one. For Type 3 constraints (inequality constraints) we consider two cases: 5% or 10% of all inequalities are given. Finally, we consider the situation where it is known that all dates lie in the period -100 and 1100 years, thereby incorporating the interval constraints introduced in Section 3. Table 1 summarizes the factors and their levels

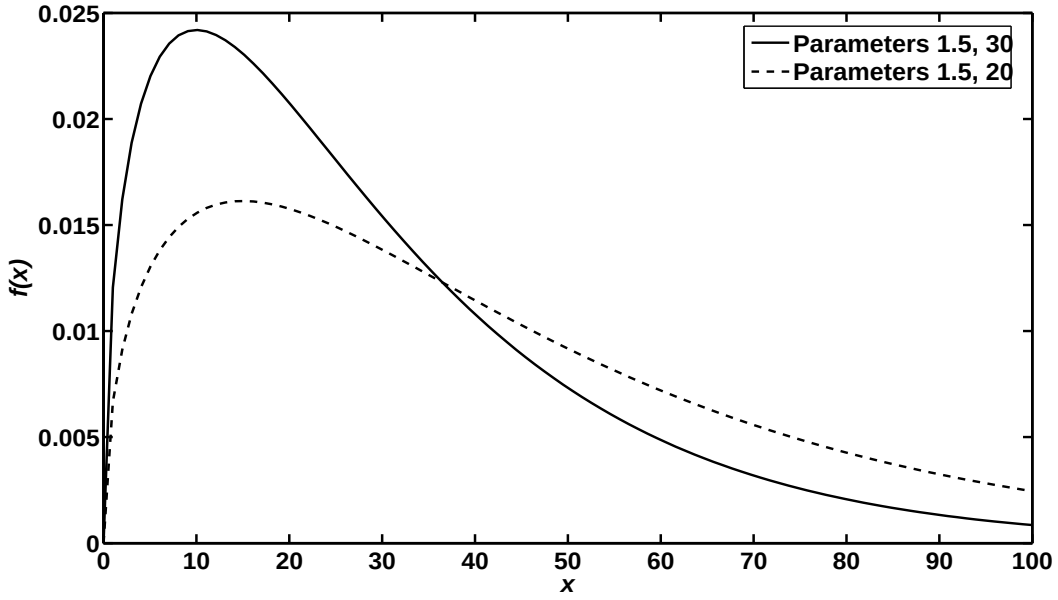

Figure 3: Gamma density functions used to simulate the lifetimes

Table 1: Factors varied in the simulation study

Factors	Levels
Number of Assemblages	{20,30}
Average Lifetime of Artefacts	{15,30}
Number of Known Years	{3,4}
Percentage of Known Inequalities	{5,10}
Interval Constraints	{No,Yes}

varied in our simulation study. For each combination of factor levels, 500 simulations were performed.

5 Results

To assess the performance of the constrained CA approach we consider various measures. First of all, a comparison is made regarding the Spearman rank correlation coefficient between the true orders and the retrieved orders for both methods. Note that the unconstrained approach does not yield actual dates and a comparison other than the ordering in time of the two methods is all that we can achieve. Furthermore, the unconstrained approach does not give a unique direction. We therefore used the absolute value of the correlations implicitly assuming that the direction of the ordination is correct in the unconstrained approach. The average Spearman rank correlations for the unconstrained and constrained approach can be found in

Table 2: Spearman rank correlations. Mean lifetime of artefacts is 15 years

	Known			20 Assemblages		30 Assemblages	
Method	years	Ineq.	Interval	Mean	Std	Mean	Std
CA	-	-	-	0.46	0.27	0.55	0.30
CCA	3	5%	No	0.89	0.19	0.93	0.16
CCA	3	5%	Yes	0.93	0.09	0.95	0.10
CCA	3	10%	No	0.92	0.15	0.96	0.12
CCA	3	10%	Yes	0.95	0.07	0.97	0.08
CCA	4	5%	No	0.93	0.14	0.96	0.17
CCA	4	5%	Yes	0.95	0.08	0.97	0.08
CCA	4	10%	No	0.96	0.09	0.98	0.05
CCA	4	10%	Yes	0.97	0.05	0.98	0.06

Table 3: Spearman rank correlations. Mean lifetime of artefacts is 30 years

	Known			20 Assemblages		30 Assemblages	
Method	years	Ineq.	Interval	Mean	Std	Mean	Std
CA	-	-	-	0.81	0.30	0.94	0.18
CCA	3	5%	No	0.90	0.26	0.93	0.21
CCA	3	5%	Yes	0.96	0.19	0.96	0.14
CCA	3	10%	No	0.91	0.22	0.95	0.16
CCA	3	10%	Yes	0.96	0.11	0.98	0.09
CCA	4	5%	No	0.94	0.17	0.95	0.17
CCA	4	5%	Yes	0.97	0.06	0.98	0.10
CCA	4	10%	No	0.96	0.14	0.97	0.12
CCA	4	10%	Yes	0.98	0.05	0.98	0.06

Tables 2 and 3. We see that the results for the unconstrained approach (Method: CA) are severely affected by the lifetime distribution. If artefacts have shorter lifetimes, correspondence analysis has more difficulty in reconstructing the correct order. When the lifetimes increase, the method performs quite well. Note also that the number of assemblages plays an important role. More assemblages lead to higher correlations. Also, for almost all settings, the average rank correlation is higher for the constrained approach (Method: CCA) than for the unconstrained approach, and it is typically quite high. Apparently, the constrained approach is successful in achieving the correct ordering. Adding interval constraints has a positive effect on the correlation coefficients. Note that, in contrast to the results for the unconstrained approach, the lifetime of artefacts does not appear to affect the correlation coefficient.

The prime objective of this paper is to assess how well the constrained CA approach is able to retrieve actual dates. Therefore, we calculate, for each simulated data set, the mean absolute difference between true dates and predicted dates. That is, for all assemblages, the absolute difference between true and predicted dates is

Table 4: Median of the mean absolute difference between reconstructed and true periods of assemblages. The mean lifetime of artefacts is 15 years

	Known			20 Assemblages		30 Assemblages	
Method	years	Ineq.	Interval	Median	IQR	Median	IQR
CA	-	-	-	-	-	-	-
CCA	3	5%	No	60.65	51.07	49.28	41.16
CCA	3	5%	Yes	51.81	34.79	45.83	28.65
CCA	3	10%	No	53.71	43.62	42.22	27.81
CCA	3	10%	Yes	45.09	32.78	39.72	23.00
CCA	4	5%	No	44.37	35.83	38.19	25.08
CCA	4	5%	Yes	40.33	29.53	35.12	20.39
CCA	4	10%	No	38.15	28.81	34.12	19.88
CCA	4	10%	Yes	34.38	23.74	31.21	16.21

calculated and averaged over the number of assemblages. As the true dates are approximated up to the time-interval that they are assigned to (recall that assemblages represent a period of 20 years), we calculate the difference between the predicted date and the interval boundaries. If the predicted date falls in the interval, the difference is zero. In Tables 4 and 5, the median and interquartile range of the mean absolute difference for each parameter setting is given. We use the median rather than the mean to account for possible outliers. In constrained CA without interval constraints, extreme outliers occasionally occur. In such cases, dates are assigned to assemblages that are far beyond the range of plausible values (tens of thousands years before or after the actual dates). In practice, such a solution will easily be discarded. Such outliers are avoided by the interval restrictions that we impose. Comparing Tables 4 and 5 we see that the results for the scenario with mean lifetimes of 15 years are consistently better than those with a mean lifetime of 30 years. This concerns both the location (the median) but also the spread as represented by the interquartile range (IQR). Also, we see that by introducing the interval constraints the median and spread become smaller. This is due to the fact that the interval constraint eliminates solutions with very large deviations. The effect of the number of inequality constraints is limited when the mean lifetime is 30 years. However, when the mean lifetime is 15 years, adding inequality constraints appears to decrease the mean absolute differences by approximately 5 years. Also, note that the number of assemblages does not appear to have a large effect on the median in the 30 years scenario, whereas it leads to a considerable improvement when the mean lifetime of the artefacts is 15 years.

As an alternative measure for the fit of the solutions, we calculate, for each data set, the percentage of assemblages with a difference between the true and predicted dates of less than 60 years (i.e., two generations). Tables 6 and 7 give the medians for the percentages of predictions within two generations for the 15 and 30 years scenarios. We see that these percentages lie between 60%, for the 30 year lifetimes with the lowest amount of restrictions imposed, and 83% for the 15 years scenario

Table 5: Median of the mean absolute difference between reconstructed and true periods of assemblages. The mean lifetime of artefacts is 30 years

	Known			20 Assemblages		30 Assemblages	
Method	years	Ineq.	Interval	Median	IQR	Median	IQR
CA	-	-	-	-	-	-	-
CCA	3	5%	No	66.07	77.73	67.44	76.78
CCA	3	5%	Yes	58.93	45.94	56.92	47.29
CCA	3	10%	No	63.99	78.39	66.81	75.67
CCA	3	10%	Yes	57.68	42.19	56.03	46.93
CCA	4	5%	No	49.29	47.20	50.47	49.25
CCA	4	5%	Yes	43.56	28.84	45.93	34.30
CCA	4	10%	No	47.79	47.56	49.91	48.53
CCA	4	10%	Yes	43.85	28.39	45.75	34.48

Table 6: Median of the mean percentage of assemblages correctly classified within two generations. The mean lifetime of artefacts is 15 years

	Known			20 Assemblages		30 Assemblages	
Method	years	Ineq.	Interval	Median	IQR	Median	IQR
CA	-	-	-	-	-	-	-
CCA	3	5%	No	0.65	0.20	0.68	0.21
CCA	3	5%	Yes	0.70	0.23	0.70	0.21
CCA	3	10%	No	0.70	0.25	0.73	0.23
CCA	3	10%	Yes	0.70	0.20	0.73	0.20
CCA	4	5%	No	0.75	0.20	0.77	0.19
CCA	4	5%	Yes	0.75	0.20	0.77	0.17
CCA	4	10%	No	0.75	0.20	0.80	0.17
CCA	4	10%	Yes	0.80	0.15	0.83	0.18

Table 7: Median of the mean percentage of assemblages correctly classified within two generations. The mean lifetime of artefacts is 30 years

	Known			20 Assemblages		30 Assemblages	
Method	years	Ineq.	Interval	Median	IQR	Median	IQR
CA	-	-	-	-	-	-	-
CCA	3	5%	No	0.60	0.20	0.60	0.27
CCA	3	5%	Yes	0.60	0.25	0.63	0.26
CCA	3	10%	No	0.60	0.25	0.63	0.27
CCA	3	10%	Yes	0.63	0.25	0.63	0.25
CCA	4	5%	No	0.70	0.20	0.68	0.23
CCA	4	5%	Yes	0.70	0.20	0.70	0.20
CCA	4	10%	No	0.70	0.20	0.69	0.23
CCA	4	10%	Yes	0.70	0.20	0.70	0.23

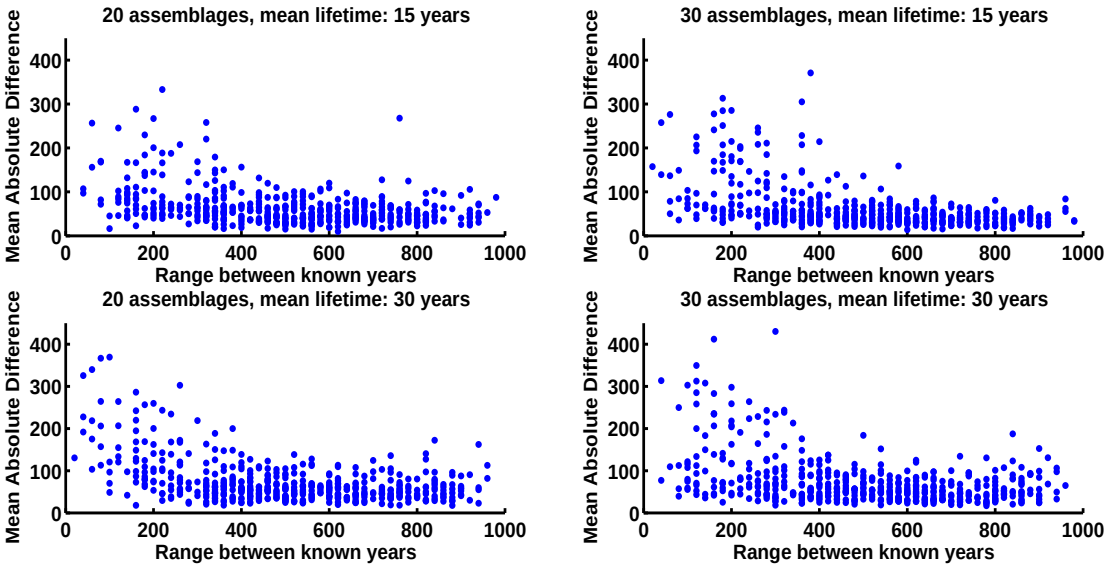

Figure 4: Relationship between the Mean Absolute Difference and the range between known years.

with the maximum amount of information imposed by means of the restriction. Again, the solutions when the average lifetime is 15 years are consistently better than for the 30 years scenario. Furthermore, from Tables 6 and 7 it is clear that the addition constraints generally leads to an improvement of the solution.

The constraints in the simulation study are drawn randomly on the basis of the observed assemblages. Therefore, situations may occur in which the known dates are close to each other. In that case, the prediction of other dates is based on a small interval and possibly less precise than if it was based on a larger interval. To see if this is indeed the case, we plotted the average absolute deviations versus the length of the time interval defined by the oldest and newest constraint dates for the following settings. The patterns for the different settings are remarkably similar. The most important difference concerns the situation without interval constraints. In this case, some extremely bad solutions (mean difference in tens of thousands) occur. These extreme cases occur mostly when the interval between known years is small. To illustrate the similarity in the patters we have plotted the mean absolute difference versus the range between known years for the setting in which 3 years and 5% of the inequalities where known and interval constraints were employed. We see that, as expected, the average absolute deviations decrease as the intervals between the known dates becomes larger. Thus, constrained CA tends to perform better if the range of the known dates is larger.

6 Discussion

In this paper, we studied the performance of constrained CA as proposed by Groenen and Poblome (2003) by means of a simulation study. The results of the simulations study showed that a constrained approach clearly outperforms the unconstrained approach. This is especially true when the artefact lifetime is assumed to be relatively small on average. In practice, this would be a more realistic setting in most archaeological settings as artefacts do not typically have high lifetimes. We also saw that by imposing linear year constraints, the predicted dates were quite good. As the constrained CA approach makes its prediction based on the known years, the range between known years plays an important role in the quality of the solution. We saw that, in general, a large interval between known years leads to better predictions. We also observed that in some cases extreme outliers occurred. This, however, could be remedied by employing interval constraints to place all dates in a realistic time-frame.

The objective of this simulation study was to get insight into the performance of CA and constrained CA as seriation techniques. We were particularly interested to see whether the linear year constraints lead to accurate predictions. To achieve this, we have used a very simple design of the data generation process. The parameter choices were made primarily to resemble the data from the Sagalassos tableware study. Of course, different data generating processes could be studied as well. Other features in the simulation study that could be varied in future studies are the following. In our present study, we separately treated two different (but similar) distributions for the lifetime. In practice, however, a mixture of several distributions for the lifetime of artefacts will be more realistic. Also, instead of considering equally spaced time-periods, the length of the periods for different assemblages could be determined randomly as proposed by Herzog and Scollar, 1988. We believe that the present study shows that constrained CA can be a useful tool for archeologists that improves datign of assemblages.

References

Böckenholt, U. and Böckenholt, I., 1990. Canonical analysis of contingency tables with linear constraints. *Psychometrika* 55, 633-639.

Böckenholt, U. and Takane, Y., 1994. Linear constraints in correspondence analysis. In Greenacre, M. and Blasius, J. (eds.), *Correspondence analysis in the social sciences*. London: Academic Press.

Cool, H. E. M. and Baxter, M.J., 1999. Peeling the onion: an approach to comparing vessel glass assemblages. *Journal of Roman Archaeology* 12, 72-100.

Graham, I., Galloway, P. and Scollar, I., 1976. Model studies in computer seriation. *Journal of Archaeological Science* 3, 1-30.

Greenacre, M.J., 1984. *Theory and applications of correspondence analysis*. New York: Academic Press.

Groenen, P.J.F. and Poblome, J., 2003. Constrained correspondence analysis for seriation in archaeology applied to Sagalassos ceramic tablewares. In: Schwaiger, M. and Opitz, O. (eds.), *Exploratory Data Analysis in Empirical Research.* Berlin, Springer, 90-97.

Herzog, I. and Scollar, I., 1988. A mathematical basis for simulation of seriatable data. In Rahtz, S.P.Q. (ed.), *Computer and Quantitative Methods in Archaeology 1988,* British Archaeological Reports International Series 446, 53-62.

Lawson,C.L. and Hanson, R.J., 1974. *Solving least squares problems*. Englewood Cliffs, NJ: Prentice Hall.

Lockyear, K., 1991. Simulating coin hoard formation. In Lockyear, K. and Rahtz S.P.Q. (eds.), *Computer Applications and Quantitative Methods in Archaeology 1990*, British Archaeological Reports International Series 565, 195-206.

Poblome, J., 1999. *Sagalassos red slip ware. Typology and chronology (studies in Eastern Mediterranean Archaeology 2)*. Turnhout: Brepols Publishers.

Poblome, J. and Groenen, P.J.F., 2003. Constrained correspondence analysis for seriation of Sagalassos tablewares. In: M. Doerr and A. Sarris (eds.), *Computer Applications and Quantitative Methods in Archaeology*. Hellinic Ministry of Culture, 301-306.

Ritov, Y. and Gilula, Z., 1993. Analysis of contingency tables by correspondence models subject to order constraints. *Journal of the American Statistical Association* 88, 1380-1387.

Shennan, S., 1988. *Quantifying archaeology.* Edinburgh, Edinburgh University.

Takane, Y., Yanai, H. and Mayekawa, S., 1991. Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika* 56, 667-684.