Seriation of Archaeological Artifacts by Mathematics and Statistics

Prof. dr. Patrick J.F. Groenen*, dr. Michel van de Velde, and Prof. dr. Jeroen Poblome**

* Econometrics Institute, Erasmus University Rotterdam, The Netherlands

** Sagalassos Archaeological Research Project, Catholic University Leuven, Leuven, Belgium

Summary:

1. The seriation problem
2. Correspondence analysis (CA)
3. Algebra of CA
4. The Sagalassos data
5. Constrained correspondence analysis
6. Reconstructing the dates
7. Results
8. Conclusions and discussion

1. The seriation problem

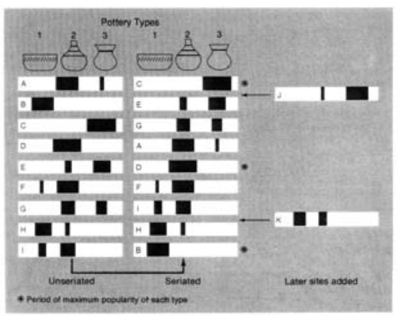
- Data consists of archaeological artifacts collected at several sites (i.e., graves, settlements, etc.)
- Objective of seriation:
  Reconstruct the unknown temporal ordering of the sites.
- Basic assumptions of distribution of artifacts over time:
  - + First an artifact is not used.
  - + At some point it becomes popular.
  - + Then, the artifact is not used anymore.
- Thus the distribution of artifacts over time is single-peaked.

- Consider binary matrix of artifacts by sites, with
  - 1 = presence of artifact *i* in site *j*
  - 0 = absence of artifact *i* in site *j*
- Each column (artifact) shows single-peakedness over the sites (which are ordered here in time).
- Such a structure is called a Petrie matrix (de Petrie, 1899)
- In Psychometrics, this structure is called parallelogram (Coombs, 1964).

	Artifacts					
Sites	a	b	c	d	e	f
1	1	0	0	0	0	0
2	1	1	0	0	0	0
3	1	1	1	0	0	0
4	1	1	1	1	0	0
5	1	1	1	1	1	0
6	0	1	1	1	1	1
7	0	0	1	1	1	1
8	0	0	0	1	1	1
9	0	0	0	0	1	1
10	0	0	0	0	0	1

- Consider frequency matrix of artifacts by sites
- Again, each column (artifact) shows single-peakedness over the sites.
- Battleships (Ford, 1962):

Figure 8.5 At the left, nine excavated sites (A to I) contain different percentages of three distinct pottery types. At the right, the nine sites have been seriated by rearranging the bars of type percentages into battleship curve order. At the far right, later excavations are eventually fitted into the sequence.

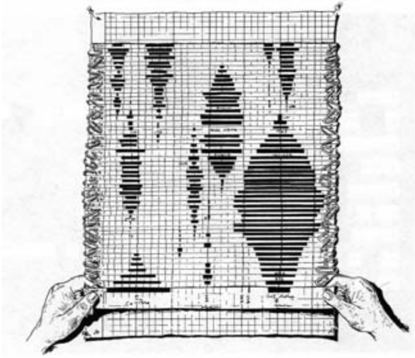


	Artifacts					
Sites	a	b	c	d	e	f
1	5	0	0	0	0	0
2	24	3	0	0	0	0
3	112	14	10	0	0	0
4	84	26	35	80	0	0
5	20	8	12	235	3	0
6	6	1	8	352	8	22
7	0	0	2	176	15	36
8	0	0	0	74	7	68
9	0	0	0	0	1	59
10	0	0	0	0	0	23

- Seriation can be seen a technique that orders sites *in time* such that all distributions of artifacts are single-peaked.

- Procedure by paper and hand for seriation (B.M. Fagan, [1981]. *In the beginning: An introduction into archaeology*):
  - make a strip of paper for each site with every artifact in a column (the width indicates the frequency).
  - position strips manually with paperclips such that battleship forms arise.

Figure 6.6 A seriation graph in the making. Each strip of paper represents a statigraphic unit; the ten columns of black bars are different pottery types. Each strip has the pottery counts for the level plotted on it in bar graph form. The strips, placed in position with paper clips onto graph paper pinned to a backboard, produce the most viable seriated sequence. (The diagram is almost complete.)



- Why not use automated seriation procedures to find the unknown time axis?

2. Correspondence analysis (CA)

- Input data correspondence analysis:
  - two-way data (artifact by assemblage)
  - frequencies of artifact per assemblage

- Geometric idea of correspondence analysis:
  + Compute proportions of artifacts per assemblage.
  + Compute the weighted Euclidean distances between the assemblage.
  + (Weights are the inverse of the square roots of the frequency of the artifacts)
  + Approximate these distances in one dimension by an eigendecomposition.

- Consider the following (hypothetical) distribution of shards over five assemblages.

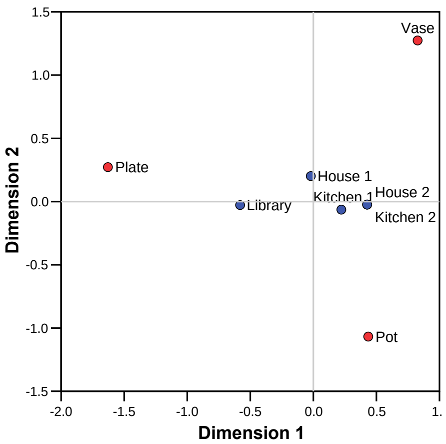
Assemblage * Type Crosstabulation

Count

		Type			
		Pot	Vase	Plate	Total
Assemblage	House 1	8	9	7	24
	House 2	39	30	6	75
	Library	25	11	39	75
	Kitchen 1	3	2	1	6
	Kitchen 2	13	10	2	25
Total		88	62	55	205

- $\chi^2 = 44.9$, df = 8, $p < .001$, thus independence model can be rejected.

- The CA solution looks as follows:



- Note that Kitchen 2 and House 2 are located on top of each other.

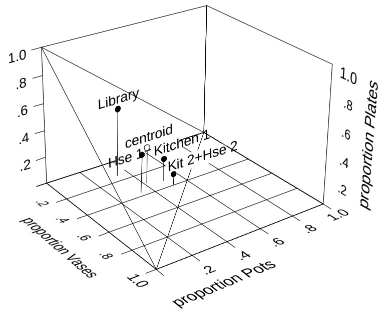
- For the geometric approach, first consider the row proportions:

Assemblage * Type Crosstabulation

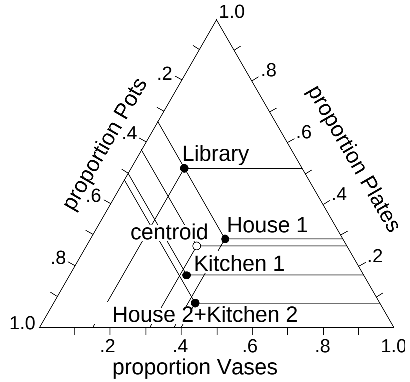
% within Assemblage

		Type			
		Pot	Vase	Plate	Total
Assemblage	House 1	33.3%	37.5%	29.2%	100.0%
	House 2	52.0%	40.0%	8.0%	100.0%
	Library	33.3%	14.7%	52.0%	100.0%
	Kitchen 1	50.0%	33.3%	16.7%	100.0%
	Kitchen 2	52.0%	40.0%	8.0%	100.0%
Total		42.9%	30.2%	26.8%	100.0%

- Then, plot the row proportions as points in a 3D space with the Types as axes

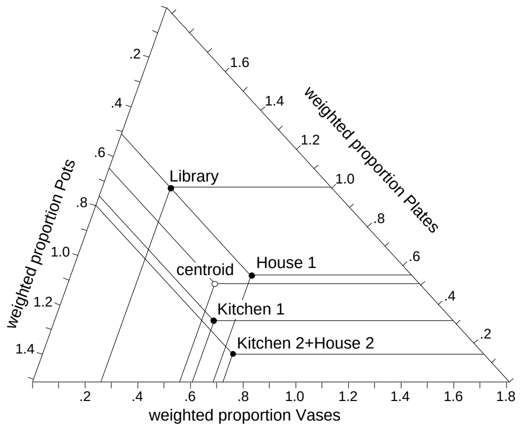


- Because row proportions sum to 1, the span a 2D triangle (with corner points being the parties):



- The next step in CA is stretching the axes by $(f_{+j}/n)^{-1/2}$, with f_{+j}/n the proportion of Type *j*.

Type	f_{+j}/n	$(f_{+j}/n)^{-1/2}$
– Pot	.429	$1/\sqrt{.429} = 1.527$
– Vase	.302	$1/\sqrt{.302} = 1.820$
– Plate	.268	$1/\sqrt{.268} = 1.932$



- Instead of variance accounted for, we use the term Inertia in CA.

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia: Accounted for	Proportion of Inertia: Cumulative	Confidence Singular Value: Standard Deviation	Confidence Singular Value: Correlation 2
1	.462	.214			.975	.975	.060	-.073
2	.074	.005			.025	1.000	.069	
Total		.219	44.917	.000[a]	1.000	1.000		

a. 8 degrees of freedom

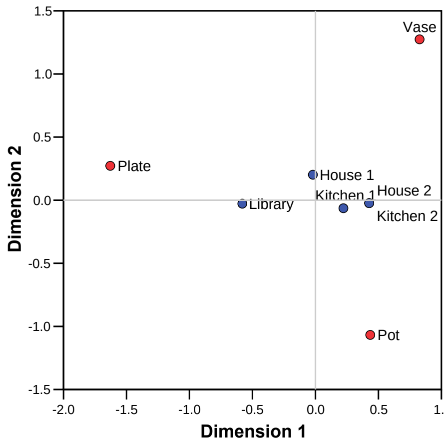
- Inertia φ^2_s for dimension s:
  - It is the equivalent of an eigenvalue of dimension *s*.
  - It is a measure of importance of a dimension.
  - The total inertia is equal to χ^2/n. Here: .219 = 44.917/205.

- The final solution is obtained by rotating the previous plot so that the first dimension explains most of the inertia.

- Permuted table according to first dimension

Permuted Correspondence Table According to Dimension 1

	Type			
Assemblage	Plate	Pot	Vase	Active Margin
Library	39	25	11	75
House 1	7	8	9	24
Kitchen 1	1	3	2	6
House 2	6	39	30	75
Kitchen 2	2	13	10	25
Active Margin	55	88	62	205



3. Algebra of CA

- Let
  - **F** be the matrix with frequencies.
  - **E** be the matrix with expected frequencies under the independence model $e_{ij} = (f_{i+} f_{+j}) / n$.
  - $\mathbf{D}_r$ the diagonal matrix of row sums (thus with diagonal elements f_{i+}).
  - $\mathbf{D}_c$ the diagonal matrix of column sums (thus with diagonal elements f_{+j}).
- Then, CA amounts to the singular value decomposition (SVD) of

$$\mathbf{D}_r^{-1/2}(\mathbf{F}-\mathbf{E})\mathbf{D}_c^{-1/2} = \mathbf{P\Phi Q}'$$

with

  - $\mathbf{P}'\mathbf{P} = \mathbf{I}$,
  - $\mathbf{Q}'\mathbf{Q} = \mathbf{QQ}' = \mathbf{I}$, and
  - $\mathbf{\Phi}$ diagonal with nonnegative singular values φ_s on the diagonal (φ^2_s is the inertia for dimension s).

- Then, the row scores **R** are given by $\mathbf{R} = n^{1/2}\mathbf{D}_r^{-1/2}\mathbf{P\Phi}$.
- The column scores **C** are given by: $\mathbf{C} = n^{1/2}\mathbf{D}_c^{-1/2}\mathbf{Q}$.
- The weighted sum of squares of the row scores equal the inertia: $\mathbf{R}'\mathbf{D}_r\mathbf{R} = \mathbf{\Phi}^2$.
- The weighted sum of squares of the column scores equals: $\mathbf{C}'\mathbf{D}_c\mathbf{C} = n\mathbf{I}$.
- The marginal frequencies are used as:
  - masses for the row scores (weights that indicate the importance of the row category).
  - stretching for the column scores (indicating the importance of the column dimension).

- Correspondence analysis can also be seen as the minimization of the following quadratic loss function:

$$L(\mathbf{r},\mathbf{c}) = \left\| \mathbf{D}_r^{-1/2}(\mathbf{F}-\mathbf{E}-n^{-1}\mathbf{D}_r\mathbf{rc}'\mathbf{D}_c)\mathbf{D}_c^{-1/2} \right\|^2$$

over **r** and **c**, where

r is the vector of scores of the assemblages,
c is the vector of scores of the pottery types,
F is a frequency matrix of n_r assemblages by n_c pottery types,
$\mathbf{D}_r$ diagonal matrix with row sums of **F,**
$\mathbf{D}_c$ diagonal matrix with column sums of **F,**
E is the matrix with expected frequencies: $\mathbf{E} = n^{-1}\mathbf{D}_r\mathbf{11}'\mathbf{D}_c$,
n is total frequency (f_{++}).

$$\|\mathbf{A}\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2$$

- Note that for seriation we only need a one dimensional solution, hence the notation **r** and **c** instead of **R** and **C**.

4. The Sagalassos data

- Excavated at Sagalassos (south west Turkey)

- Data of red slip ware:
  + 27 assemblages or stratigraphical units
  + 26,166 shards
  + every shard is classified into one of 85 types consisting of 5 subgroups:
    A. Cups
    B. Bowls
    C. Dishes
    D. Plates
    E. Containers

- Data that we use here:
  + quantification by counts of shards per type and assemblage

5. Constrained correspondence analysis

- Problem of correspondence analysis:
  – Correspondence analysis does not use any additional information that the archaeologist may know of.
  – No explicit dating of the assemblages is done, only ordering.
- Solution:

  Constrain correspondence analysis to use the additional information.
- Types of additional information
  1. For some assemblages the exact dates are known.
  2. Some assemblages necessarily have the same date.
  3. Some assemblages are necessarily ordered in time.
- Constrained correspondence analysis:

  minimize $L(\mathbf{r}, \mathbf{c})$ subject to appropriate constraints on the coordinates of the assemblages $\mathbf{r}$.

5.1. Imposing restrictions

- Date restrictions on the assemblages
- Consider assemblages A_1 to A_6 and archaeological findings indicate that the year of
  – A_1 is y_1 = 100 AD,
  – A_4 is y_4 = 425 AD,
  – A_6 is y_6 = 600 AD.
- Then, the linear constraints on the correspondence analysis coordinates r_1, r_4, and r_6 are
  – $r_1 = a + by_1$,
  – $r_4 = a + by_4$, and
  – $r_6 = a + by_6$ where a and b need to be estimated.
- Suppose also that year of A_2 must be equal to that of A_3. This equality constraint indicates that $r_2 = r_3$.

- Both types of constraints are imposed by restricting **r** to be a linear sum of the columns of **H**, i.e., **r** = **Hb**, where

$$\mathbf{H} = \begin{bmatrix} 1 & 100 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 425 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 600 & 0 & 0 \end{bmatrix} \text{ so that } \mathbf{r} = \mathbf{Hb} \text{ implies } \begin{array}{l} r_1 = b_1 + 100b_2 \\ r_2 = b_3 \\ r_3 = b_3 \\ r_4 = b_1 + 425b_2 \\ r_5 = b_4 \\ r_6 = b_1 + 600b_2. \end{array}$$

- A different way of stating that **r** = **Hb** is to require that $\mathbf{H}_0'\mathbf{r} = \mathbf{0}$, where $\mathbf{H}_0$ is the null-space of **H** so that $\mathbf{H}'\mathbf{H}_0 = \mathbf{0}$.

- Imposing $\mathbf{H}_0'\mathbf{r} = \mathbf{0}$ is easier because it only involves **r** and not a new set of parameters **b**.

- Assume order restrictions on the assemblages.
  - A_2 must be younger than A_1,
  - A_2 older than A_4, and
  - A_5 must be older than A_6.

- The ordering restrictions on the assemblages mean that
  - $r_1 \le r_2$,
  - $r_2 \le r_4$, and
  - $r_5 \le r_6$.

- In matrix algebra, these inequalities can be written as $\mathbf{Gr} \ge \mathbf{0}$, where

$$\mathbf{G} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \text{ so that } \mathbf{Gr} \ge \mathbf{0} \text{ implies } \begin{array}{l} r_2 \ge r_1 \\ r_4 \ge r_2 \\ r_6 \ge r_5 \end{array}$$

- Considering everything together:
  - Date and equality restrictions form linear constraints.
  - Inequality restrictions form linear inequality constraints.
  - The optimization task is to minimize $L(\mathbf{r}, \mathbf{c})$ over **r** subject to the restrictions $\mathbf{H}_0'\mathbf{r} = \mathbf{0}$ and $\mathbf{Gr} \ge \mathbf{0}$.

- Rewriting $L(\mathbf{r}, \mathbf{c})$ gives

$$L(\mathbf{r}, \mathbf{c}) = \left\| n^{-1}\mathbf{D}_r^{1/2}\mathbf{r} - \mathbf{t} \right\|^2 - \left\| \mathbf{t} \right\|^2 + \left\| \mathbf{D}_r^{1/2}(\mathbf{F} - \mathbf{E})\mathbf{D}_c^{1/2} \right\|^2,$$

where $\mathbf{t} = \mathbf{D}_r^{-1/2}(\mathbf{F} - \mathbf{E})\mathbf{c}$.

- Thus, for fixed **c**, $L(\mathbf{r}, \mathbf{c})$ is quadratic in **r**.

- This problem of minimizing $L(\mathbf{r}, \mathbf{c})$ subject to $\mathbf{H}_0'\mathbf{r} = \mathbf{0}$ and $\mathbf{Gr} \ge \mathbf{0}$ is called the least-squares problem with linear equality and inequality constraints.

- A scheme of the alternating least squares constrained correspondence analysis (ALS-CCA) algorithm is:

1. Choose initial $\mathbf{c}_0$ (with $\mathbf{c}_0'\mathbf{c}_0=1$) and $\mathbf{r}_0$ satisfying the constraint $\mathbf{H}_0'\mathbf{r}_0 = \mathbf{0}$ and $\mathbf{Gr}_0 \ge \mathbf{0}$. Set iteration counter $k = 0$.
2. $k := k + 1$.
3. Update **c**:
   Set $\mathbf{c} = n^{-1}\mathbf{D}_c^{-1/2}(\mathbf{F} - \mathbf{E})'\mathbf{r}_{k-1}$ and
   compute $\mathbf{c}_k = \mathbf{c}/(\mathbf{c}'\mathbf{c})^{1/2}$.
4. Update **r**:
   Solve $\left\| n^{-1}\mathbf{D}_r^{1/2}\mathbf{r} - \mathbf{t} \right\|^2$ over r subject to $\mathbf{H}_0'\mathbf{r} = \mathbf{0}$ and $\mathbf{Gr} \ge \mathbf{0}$ by using Lawson and Hanson (1974, see pages 168-169) and set $\mathbf{r}_k = \mathbf{r}$.
5. If $L(\mathbf{r}_{k-1}, \mathbf{c}_{k-1}) - L(\mathbf{r}_k, \mathbf{c}_k) > 10^{-6}$ then go to step 2, otherwise stop.

6. Reconstructing the dates

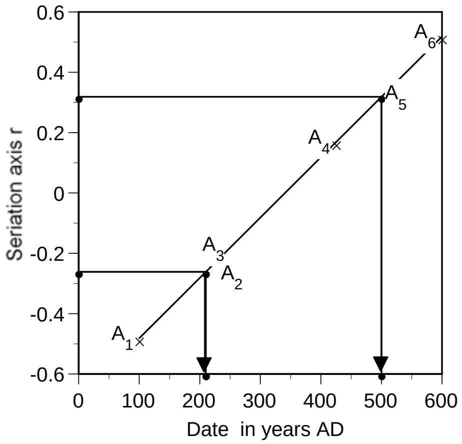
- Two types of assemblages:

  1. Those for which we know the dates and
  2. those for which the dates are unknown

- For the known set (A_1, A_4, A_6) the seriation coordinates in **r** are linearly restrictions to the dates.

- Thus, the date for the unknown set (A_2, A_3, A_5) can be interpolated.

- Constrained correspondence analysis can be used to reconstruct the dates.



7. Results

- 26,166 sherds in 27 assemblages of 85 pottery types.

- We use pottery proportions, because of the large differences in marginal frequencies (1 to 3,384).

- For four assemblages the dates are known:

Assemblage	Date
1	1
4	100
22	410
27	650

- Equality constraints for assemblage pairs 6, 7, and 24, 15.

- Inequality constraints are derived by an *a priori* known ordering for these data into phases.

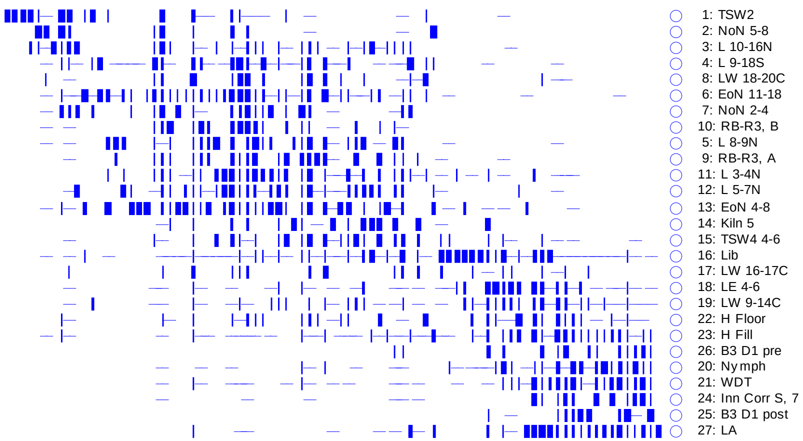
Phase	Assemblages	Suggested dating
1	1, 2, 3	0-50 AD
2	4	50-100 AD
3	5, 6, 7, 8, 9, 10	100-150 AD
4	11, 12, 13	150-200 AD
5	14, 15	200-300 AD
6	16, 17	300-350 AD
7	18, 19	350-450 AD
8	20, 21, 22, 23, 24, 25, 26	450-575 AD
9	27	575-650 AD

- Imposing the restrictions gives only slightly worse fit:
  - 27.17% of χ^2 reconstructed in 1 dim by constrained CA
  - 27.61% of χ^2 reconstructed in 1 dim by ordinary CA

- Results constrained correspondence analysis:

Recon-structed Year	Assemblage nr	Assemblage Label	Phase nr	Phase date	**r**	Contribution to dim	Contribution to total dist
1	1	TSW2	1	0-50	-5.6	.056	.240
29	2	NoN 5-8	1	0-50	-5.1	.045	.136
60	3	L 10-16N	1	0-50	-4.4	.035	.315
100	4	L 9-18S	2	50-100	-3.6	.023	.319
100	5	L 8-9N	3	100-150	-3.6	.023	.301
100	6	EoN 11-18	3	100-150	-3.6	.023	.412
100	7	NoN 2-4	3	100-150	-3.6	.023	.271
100	8	LW 18-20C	3	100-150	-3.6	.023	.245
100	9	RB-R3, A	3	100-150	-3.6	.023	.276
100	10	RB-R3, B	3	100-150	-3.6	.023	.151
102	12	L 5-7N	4	150-200	-3.6	.023	.350
107	13	EoN 4-8	4	150-200	-3.5	.021	.162
126	11	L 3-4N	4	150-200	-3.1	.017	.303
161	14	Kiln 5	5	200-300	-2.4	.010	.027
162	15	TSW4 4-6	5	200-300	-2.4	.010	.149
227	17	LW 16-17C	6	300-350	-1.0	.002	.036
312	16	Lib	6	300-350	0.7	.001	.003
410	18	LE 4-6	7	350-450	2.7	.013	.090
410	19	LW 9-14C	7	350-450	2.7	.013	.237
410	22	H Floor	8	450-575	2.7	.013	.258
556	23	H Fill	8	450-575	5.7	.056	.632
594	26	B3 D1 pre	8	450-575	6.4	.073	.526
613	20	Nymph	8	450-575	6.8	.082	.415
627	21	WDT	8	450-575	7.1	.089	.753
627	24	nn Corr S, 7	8	450-575	7.1	.089	.559
627	25	B3 D1 post	8	450-575	7.1	.089	.293
650	27	LA	9	575-650	7.6	.101	.827

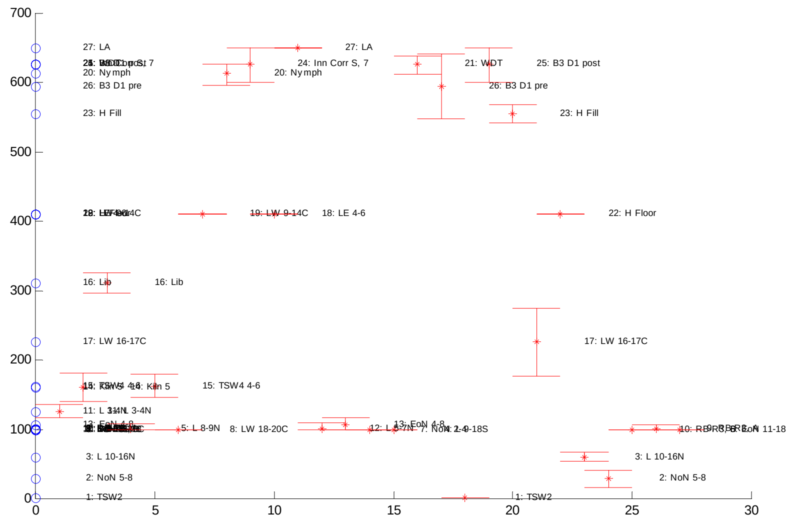
- Results constrained correspondence analysis in 'battleship' figure:



- Stability results by the bootstrap:
  - + Draw B (here B = 5000) bootstrap samples randomly from the original sample.
  - + Compute the solution for each of these bootstrap samples.
  - + Construct a confidence interval for each assemblage covering 95% of the bootstrap points.

- Bootstrap results

Recon-structed Year	95% con-fidence intervals		Assemblage nr	Label	Phase nr	date
1	1	1	1	TSW2	1	0-50
29	16	40	2	NoN 5-8	1	0-50
60	54	67	3	L10-16N	1	0-50
100	100	100	4	L9-18S	2	50-100
100	100	108	5	L8-9N	3	100-150
100	100	100	6	EoN 11-18	3	100-150
100	100	100	7	NoN 2-4	3	100-150
100	100	100	8	LW 18-20C	3	100-150
100	100	107	9	RB-R3, A	3	100-150
100	100	100	10	RB-R3, B	3	100-150
102	100	110	12	L5-7N	4	150-200
107	100	117	13	EoN 4-8	4	150-200
126	117	136	11	L3-4N	4	150-200
161	140	181	14	Kiln 5	5	200-300
162	146	179	15	TSW4 4-6	5	200-300
227	177	275	17	LW 16-17C	6	300-350
312	297	327	16	Lib	6	300-350
410	410	410	18	LE 4-6	7	350-450
410	410	410	19	LW 9-14C	7	350-450
410	410	410	22	HFloor	8	450-575
556	543	568	23	HFill	8	450-575
594	548	641	26	B3 D1 pre	8	450-575
613	597	627	20	Nymph	8	450-575
627	612	639	21	WDT	8	450-575
627	600	650	24	Inn Corr S, 7	8	450-575
627	600	650	25	B3 D1 post	8	450-575
650	650	650	27	LA	9	575-650



8. Conclusions and discussion

- Seriation of frequencies can be performed by reordering the data so that the the distribution of an artifact becomes single peaked.
- Correspondence analysis is one technique to do seriation (available in SPSS).
- Additional archaeological information can be incorporated into constrained correspondence analysis (CCA).
  + Equality of assemblages (by equalities constraints).
  + Partial ordering of assemblages (by inequalities constraints).
  + Dating information (by linear constraints)
- CCA can be used to reconstruct the dating for assemblages with unknown dates.
- Stability of the seriation solution can be assessed by the bootstrap.

- Reconstructed dates have to be interpreted with care. Quality is highly dependent on:
  + the range of the known dates, and
  + the fit of the solution.
- Publications:

– Van de Velden, M., Groenen, P.J.F., & Poblome, J. (2004). Seriation mit bedingter Korrespondenzanalyse: Simulationsexperimente. *Archäologische Informationen*, *26*, 449-455.

– Groenen, P.J.F. & Poblome, J. (2003). Constrained correspondence analysis for seriation in archaeology applied to Sagalassos ceramic tablewares. In: *Exploratory Data Analysis in Empirical Research, Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation e.V.*, University of Munich, March 14-16, 2001, pp. 90-97. Heidelberg: Springer.

– Poblome, J. & 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, Proceedings of the 30th Conference*, Heraklion, Crete, April 2002, pp. 301-306. Hellinic Ministry of Culture.

	1	2	3	4
a	4	28	53	10
b	60	60	10	52
c	32	85	35	13
d	45	4	0	30
e	2	3	20	1