UNIVERSITÄT BERN

Statistical methods for archaeological data analysis I: Basic methods

10 - Correspondence Analysis

Martin Hinz

Institut für Archäologische Wissenschaften, Universität Bern

19.05.2021

Correspondence analysis: idea and basics [1]

Similar things have similar characteristics...[2]

Visual explorative/descriptive method

- Correspondence analysis does not work with significances, therefore it does not 'proof' anything
- Visualization of contingency tables or presence/absence matrices

Idea

- Representation of items (*sites*) and properties (Variables, *species*) in a common space (coordinate system)
- Data that is related to each other is more closely related represented next to each other
- Similarities are calculated using chi-square methods

Prerequisites

A data matrix with at least nominally scaled variables, therefore especially suitable for archaeological questions

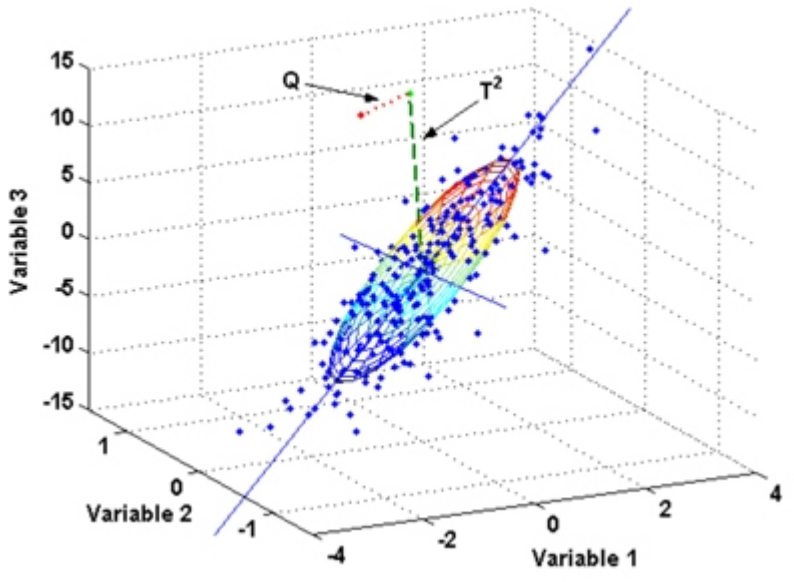
Correspondence analysis: idea and basics [1]

Similar things have similar characteristics...

General procedure

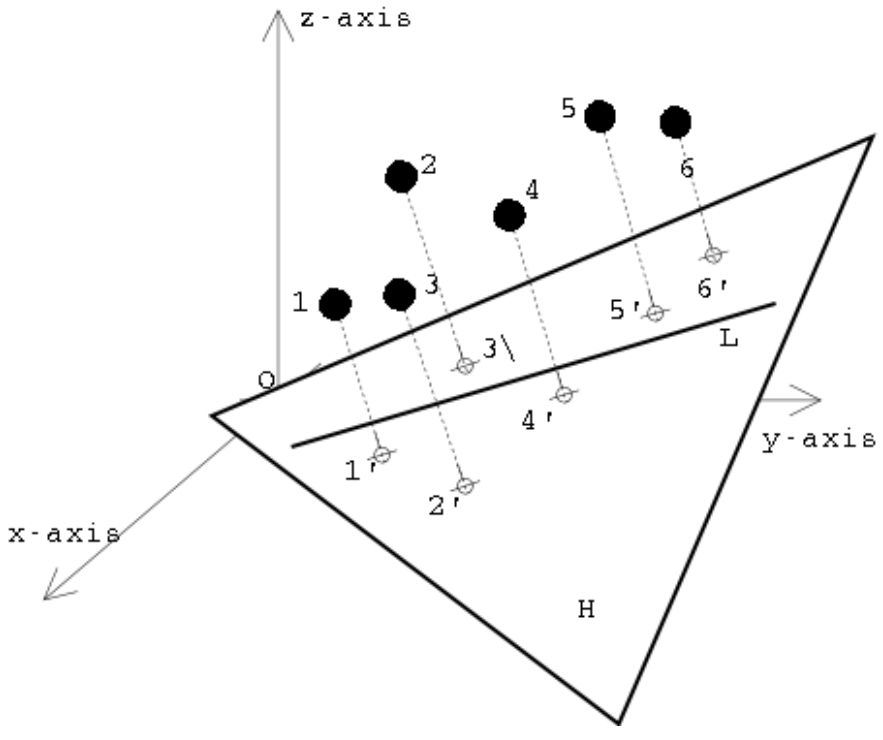
- Standardizing the data to a comparable measure
- "Projection" of the data into a multidimensional variable space
- determining the vectors which stepwise contain most of the information (variability) of the data and are oriented perpendicular to each other
- "Projection" of the data onto these vectors
- Representation of the position of the data on these vectors in a diagram

multidimentional data space



source: http://www.aapspharmscitech.org

projection of points onto a plane



source: http://www.cs.mcgill.ca

Correspondence Analysis: History

General information

- Development in the field of biology and psychology
- Algebrarian Foundations 1940s (Hartley/Guttman)
- First explicit use by Benzéncri in the 1960s linguistic studies
- Further development in various research groups → resulted in different versions and names of the procedure
- 1984 Greenacre basic monograph on the method

In archaeology

- First Seriation: Sir William Flinders-Petrie 1899
- First major trials with seriating methods in Germany Goldman 1979 with reciprocal averaging.
- Wide application of the procedure for chronological sorting of the Rhineland Linear Pottery
- Continuation by institutes Cologne and Kiel (Zimmermann, Müller)

Correspondence Analysis: Procedure

Preparation: contingency table, if necessary

Presence Absence Matrix

Notes the presence or absence of a characteristic for a unit, which is the most widely used base in archaeology

	Pot	Cup	Fibula	Sum
Burial1	1	1	0	2
Burial2	0	1	1	2
Burial3	1	1	1	3
Burial4	1	0	1	2
Sum	3	3	3	9

Prerequisite: total number of filled cells per column at least 2, total per row at least 2

Preparation: contingency table, if necessary

contingency table

Notes the number of a characteristics for a unit or a group of units

	Pot	Cup	Fibula	Sum
Settlements	20	23	40	83
Hoards	23	10	6	39
Burials	10	56	4	70
Sum	53	89	50	192

Also possible: Burt-Matrix, if you want, you can ask me for details after the lecture...

Correspondence analysis: Procedure (using a presence/absence matrix)

Preparation: Standardising to relative frequency

Calculation: Divide each cell by the total sum

	pot	cup	fibula	Sum
burial1	1	1	0	2
burial2	0	1	1	2
burial3	1	1	1	3
burial4	1	0	1	2
Sum	3	3	3	9

	pot	cup	fibula	Sum
burial1	0.11	0.11	0.00	0.22
burial2	0.00	0.11	0.11	0.22
burial3	0.11	0.11	0.11	0.33
burial4	0.11	0.00	0.11	0.22
Sum	0.33	0.33	0.33	1.00

Margins of the table stored for calculation of expectation values and scaling the result later on:

Row profile:

```
## burial1 burial2 burial3 burial4
##    0.22    0.22    0.33    0.22
```

Column profile:

```
##    pot    cup fibula
##   0.33   0.33   0.33
```

Correspondence analysis: Procedure (using a presence/absence matrix)

Preparation: Calculation of expected values

	pot	cup	fibula	Sum
burial1	0.11	0.11	0.00	0.22
burial2	0.00	0.11	0.11	0.22
burial3	0.11	0.11	0.11	0.33
burial4	0.11	0.00	0.11	0.22
Sum	0.33	0.33	0.33	1.00

	pot	cup	fibula	Sum
	0.07	0.07	0.07	0.22
	0.07	0.07	0.07	0.22
	0.11	0.11	0.11	0.33
	0.07	0.07	0.07	0.22
Sum	0.33	0.33	0.33	1.00

Correspondence analysis: Procedure (using a presence/absence matrix)

Preparation: Calculation of standardised values

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$z_{ij} = \frac{(O_i - E_i)}{\sqrt{E_i}}$$

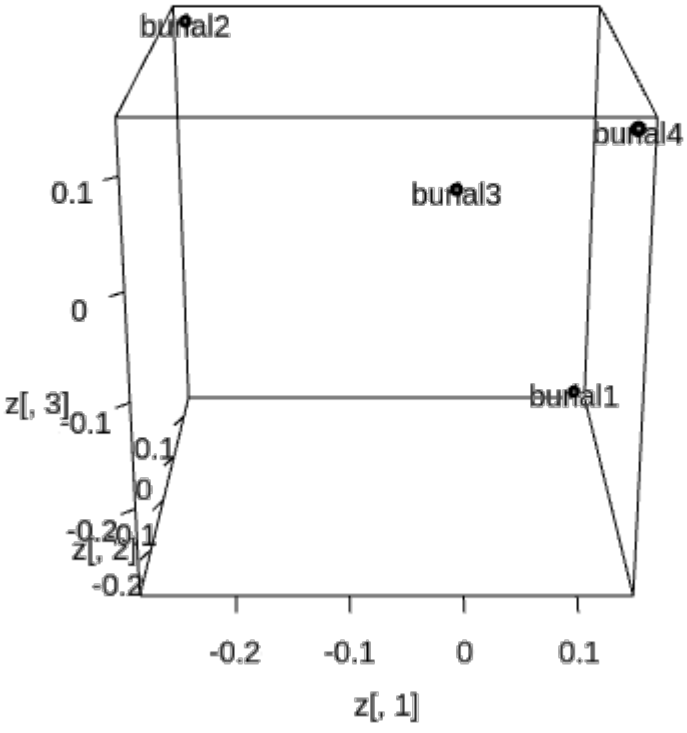
	pot	cup	fibula	Sum
burial1	0.14	0.14	-0.27	0
burial2	-0.27	0.14	0.14	0
burial3	0.00	0.00	0.00	0
burial4	0.14	-0.27	0.14	0
Sum	0.00	0.00	0.00	0

Inertia

Measurement for the spread of the data in relation to the number of cases

$$I = \frac{\chi^2}{n} = \sum_i \sum_j z_{ij}^2$$

Inertia here: 0.3333333


burial2
burial4
burial3
burial1
0.1
0
-0.1
z[, 3]
0.1
0
-0.2
0.1
z[, 2]
-0.2
-0.2
-0.1
0
0.1
z[, 1]

Data normalisation in R

burials.csv

```
burials <- read.csv("burials.csv", row.names = 1)

burials.rel_freq <- burials / sum(burials)
burials.rel_freq
```

```
##                pot       cup    fibula
## burial1 0.1111111 0.1111111 0.0000000
## burial2 0.0000000 0.1111111 0.1111111
## burial3 0.1111111 0.1111111 0.1111111
## burial4 0.1111111 0.0000000 0.1111111
```

Expectation Values in R

Multiply the margins and divide the result by the total number

```
burials.rel_freq.rows <- rowSums(burials.rel_freq)
burials.rel_freq.columns <- colSums(burials.rel_freq)
burials.e <- burials.rel_freq.rows %*% t(burials.rel_freq.columns) /
  sum(burials.rel_freq)^2
burials.e
```

```
##              pot        cup     fibula
## [1,] 0.07407407 0.07407407 0.07407407
## [2,] 0.07407407 0.07407407 0.07407407
## [3,] 0.11111111 0.11111111 0.11111111
## [4,] 0.07407407 0.07407407 0.07407407
```

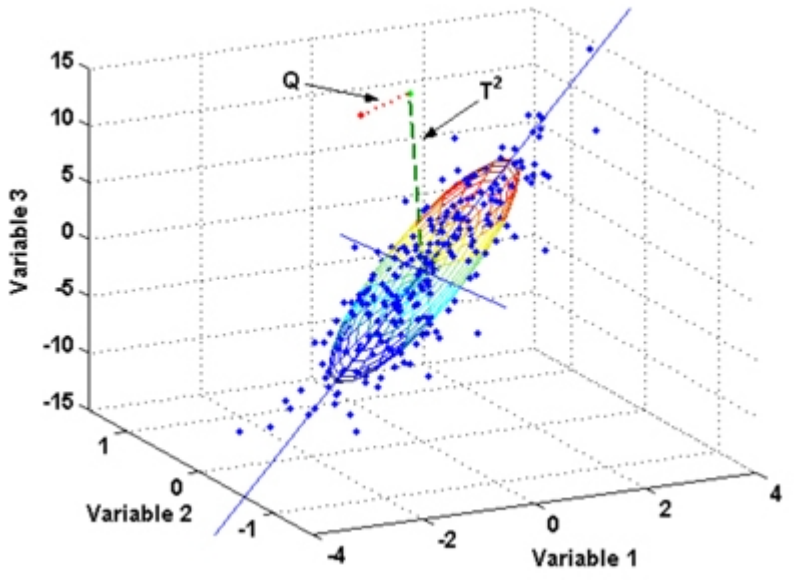
z-values in R

$$z_{ij} = \frac{(O_i - E_i)}{\sqrt{E_i}}$$

```
burials.z <- ( burials.rel_freq - burials.e)/sqrt(burials.e)
burials.z
```

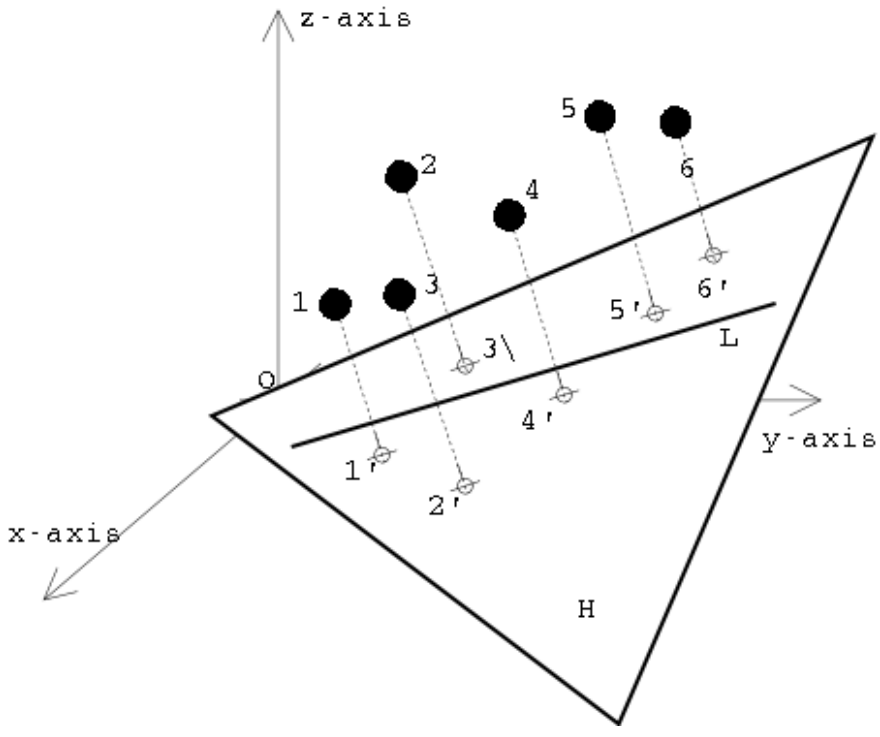
```
##                 pot        cup     fibula
## burial1  0.1360828  0.1360828 -0.2721655
## burial2 -0.2721655  0.1360828  0.1360828
## burial3  0.0000000  0.0000000  0.0000000
## burial4  0.1360828 -0.2721655  0.1360828
```

multidimentional data space



source: http://www.aapspharmscitech.org

projection of points onto a plane



source: http://www.cs.mcgill.ca

Correspondence analysis: Procedure (using a presence/absence matrix)

Extraction of dimensions

SVD

Singular **v**alue **d**ecomposition, method for dimensional reduction with minimal loss of information

$$Z = U * S * V'$$

Z : Matrix with the standardized data

U : Matrix for the row elements

V : Matrix for the column elements

S : Diagonal matrix with the singular values

Gene Golub's license plate, photographed by Professor P. M. Kroonenberg of Leiden University.

Correspondence analysis: Procedure (using a presence/absence matrix)

Extraction of dimensions

SVD in R

```
burials.svd<-svd(burials.z)
burials.svd
```

```
## $d
## [1] 4.082483e-01 4.082483e-01 5.376365e-17
##
## $u
##               [,1]       [,2]       [,3]
## [1,] -7.071068e-01 -0.4082483 -0.5773503
## [2,]  5.551115e-17  0.8164966 -0.5773503
## [3,]  0.000000e+00  0.0000000  0.0000000
## [4,]  7.071068e-01 -0.4082483 -0.5773503
##
## $v
##               [,1]       [,2]      [,3]
## [1,]  1.044525e-16 -0.8164966 0.5773503
## [2,] -7.071068e-01  0.4082483 0.5773503
## [3,]  7.071068e-01  0.4082483 0.5773503
```

SVD and Inertia The singular values (eigenvalues) represent the inertia. The eigenvalues

```
burials.svd$d
```

```
## [1] 4.082483e-01 4.082483e-01 5.376365e-17
```

The squared eigenvalues are the inertia of the individual dimensions

```
burials.svd$d^2
```

```
## [1] 1.666667e-01 1.666667e-01 2.890530e-33
```

The sum of the squared eigenvalues is equal to the total of the intertia.

```
sum(burials.svd$d^2)
```

```
## [1] 0.3333333
```

If the inertia of the individual dimensions is divided by the total inertia, the (eigenvalue) proportion of the dimensions is obtained.

```
burials.svd$d^2/sum(burials.svd$d^2)
```

```
## [1] 5.00000e-01 5.00000e-01 8.67159e-33
```

Correspondence analysis: Procedure (using a presence/absence matrix)

Normalization of coordinates

Scaling of the coordinates in such a way that

The dimensions are weighted according to their proportion of the total inertia.

The rows/columns are weighted according to their proportion of the mass.

Row (*sites*) Points: $r_{ik} = \frac{u_{ik} * \sqrt{s_k}}{\sqrt{p_i}}$

Column (*species*) Points: $c_{jk} = \frac{v_{jk} * \sqrt{s_k}}{\sqrt{p_j}}$

u, v → Matrices of rows/columns from the SVD

s_k → Diagonal matrix

p_i , p_j → Masses of rows/columns from the relative frequency

Normalization of coordinates in R

Scaling of the coordinates in such a way that

The dimensions are weighted according to their proportion of the total inertia.

The rows/columns are weighted according to their proportion of the mass.

Row (*sites*) Points: $r_{ik} = \frac{u_{ik} * \sqrt{s_k}}{\sqrt{p_i}}$

```
rows.scaled <-
  burials.svd$u * sqrt(burials.svd$d) /
  sqrt(burials.rel_freq.rows)
rows.scaled
```

```
##               [,1]          [,2]          [,3]
## [1,] -9.584147e-01 -5.533410e-01 -8.980283e-09
## [2,]  7.523998e-17  1.270004e-08 -7.825423e-01
## [3,]  0.000000e+00  0.000000e+00  0.000000e+00
## [4,]  9.584147e-01 -5.533410e-01 -8.980283e-09
```

Column (*species*) Points: $c_{jk} = \frac{v_{jk} * \sqrt{s_k}}{\sqrt{p_j}}$

```
columns.scaled <-
  burials.svd$v * sqrt(burials.svd$d) /
  sqrt(burials.rel_freq.columns)
columns.scaled
```

```
##               [,1]          [,2]         [,3]
## [1,]  1.155957e-16 -9.036020e-01 6.389431e-01
## [2,] -7.825423e-01  4.518010e-01 6.389431e-01
## [3,]  8.980283e-09  5.184769e-09 7.332370e-09
```

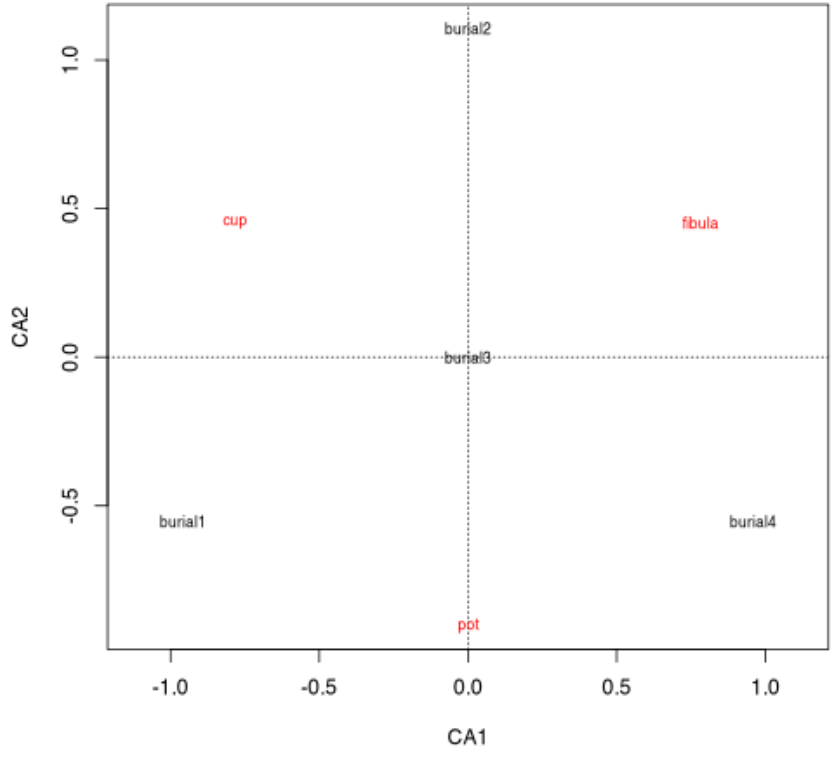
Correspondence analysis: Procedure (using a presence/absence matrix)

Everything in R:

```
library(vegan)

burial <- read.csv("burials.csv",
                   row.names = 1)
burial.cca <- cca(burial)
plot(burial.cca, scaling=3)
```

scaling=3: by default R normalizes only the species (types)

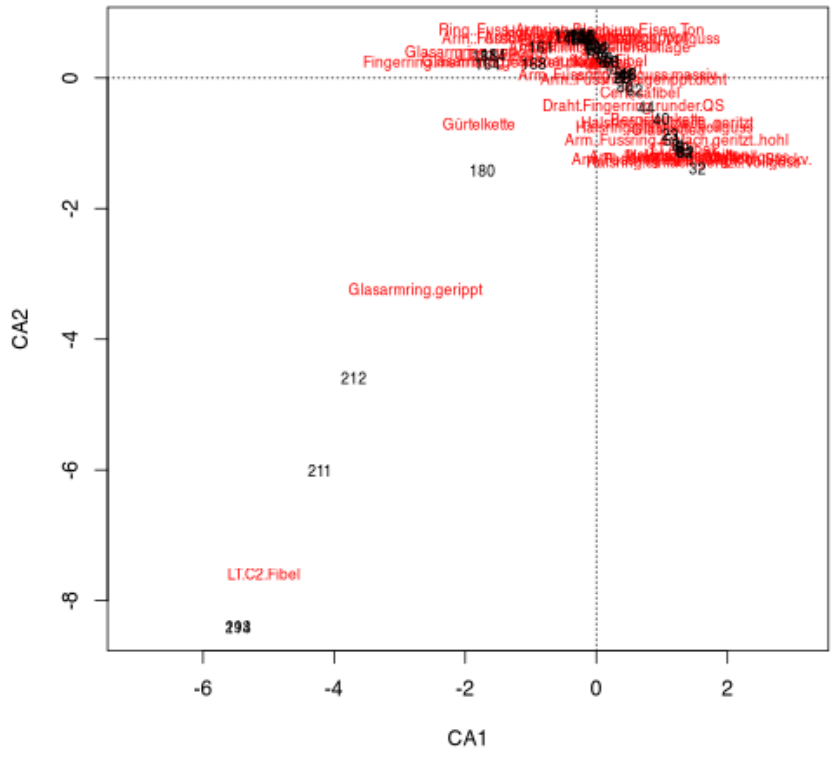
- scaling = 1 : Normalization of sites
- scaling = 2 : Normalization of the Species
- scaling = 3 : Symmetrical normalization of sites and species
- scaling = 0 : No normalization



Correspondence analysis: Real World case

Münsingen Burial Site

```
muensingen <- read.csv("muensingen_ideal.csv",
                       row.names = 1)
muensingen.cca <- cca(muensingen)
plot(muensingen.cca, scaling=3)
```



Correspondence analysis: Real World case

Münsingen Burial Site

```
scores(muensingen.cca, display = "sites")
```

```
##                CA1          CA2
## 32   1.606313e+00 -1.452925953
## 31   1.417566e+00 -1.191711661
## 8b   1.335804e+00 -1.088200431
## 12   1.415720e+00 -1.195513769
## 8a   1.381076e+00 -1.183971647
## 6    1.318179e+00 -1.097469151
## 9    1.305596e+00 -1.130528153
## 23   1.172513e+00 -0.912136067
## 44   7.886929e-01 -0.469460799
## 51   1.207199e+00 -0.998288130
## 40   1.032187e+00 -0.663168035
## 28   4.135180e-01  0.009305833
## 62   6.073775e-01 -0.192755350
## 91   2.594931e-01  0.273907559
## 72   3.852720e-01  0.009685198
## 80   4.578284e-01 -0.135341372
## 46   4.999726e-01  0.062684002
## 48   4.999726e-01  0.062684002
## 49   4.664078e-01  0.040744687
## 68   2.368297e-01  0.259427802
## 79   2.812150e-01  0.075938349
## 61   1.788927e-01  0.267615201
## 102  1.720921e-02  0.473091324
## 81  -5.781215e-02  0.535954589
## 84  -4.796386e-05  0.457809401
## 86   1.289481e-01  0.324469138
## 130 -2.266955e-01  0.659478553
## 136 -1.993537e-01  0.662413700
```

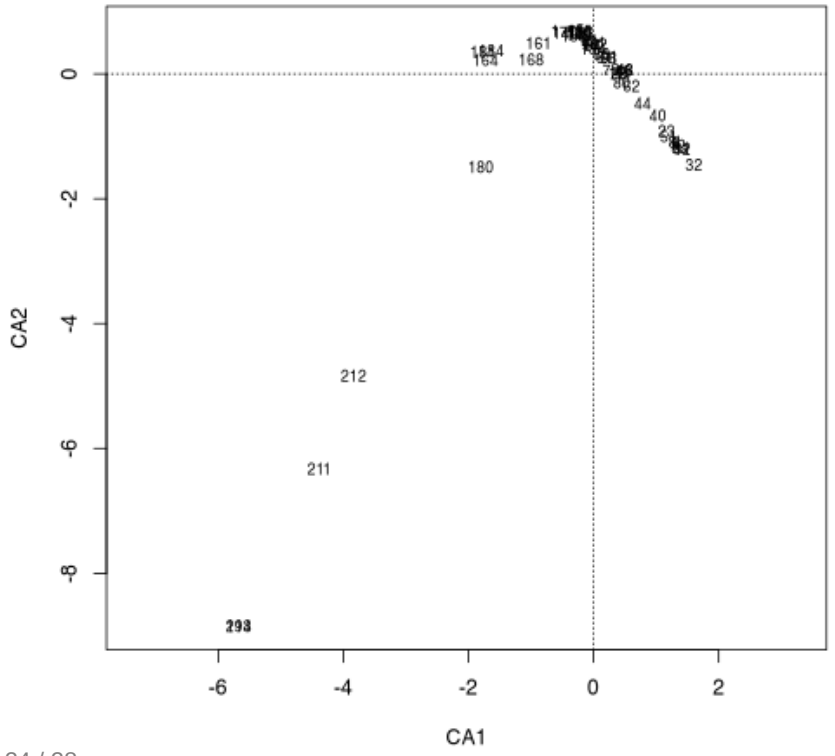
```
scores(muensingen.cca, display = "species")
```

```
##                                                 CA1         CA2
## LT.A.Fibel                                1.26553218 -1.02575507
## Halsring.einfach.geritzt..Vollguss        1.42268370 -1.23712903
## Arm.Fussring.einf...vollg.loch.Steckv.    1.39589208 -1.19799158
## Arm..Fussring.einfach.geritzt..hohl       1.17884979 -0.91744662
## Glasperlen                                1.05767259 -0.76556860
## Bernsteinkette                            0.90576411 -0.59674656
## Arm..Fussring.gerippt.vollguss            1.36065846 -1.16302096
## Hirschgeweih                              1.36694980 -1.14649146
## Halsring.m..Muffen                        1.34962749 -1.14072040
## Armring.mit.Muffen                        1.38107585 -1.18397165
## Draht.Fingerring.runder.QS                0.54905934 -0.42294391
## Arm..Fussring.vollguss.massiv             0.30773335  0.02653237
## Halsring.plastisch..vollguss              0.99794588 -0.73387446
## Halsring.hohlblech..geritzt               1.03218740 -0.66316804
## Arm..Fussringe.gerippt.dicht              0.54345897 -0.03844059
## Certosafibel                              0.63923617 -0.20818750
## Schwert                                   0.26898152  0.14758021
## Kette                                    -0.08973973  0.23569196
## Lanze                                     0.20230485  0.23181940
## LT.B1.Fibel                               0.21181418  0.25693309
## Armreif.mit.Korallenauflage               0.04601620  0.42893018
## Fingerring.flachblech                    -0.09756829  0.46147527
## Schaukelfingerringe                      -0.32647874  0.58102969
## Arm..Fussring.plastisch.gerippt          -0.15984628  0.57814339
## LT.B2.Fibel                              -0.31048691  0.58499166
## Arm..Fussring.genoppt..plastisch..Vollguss -0.23211926  0.55870602
## Ring..Fuss..Armring.Blech.um.Eisen.Ton   -0.36582667  0.68083308
## Hohlbuckelarmringe                       -0.34680816  0.65312662
```

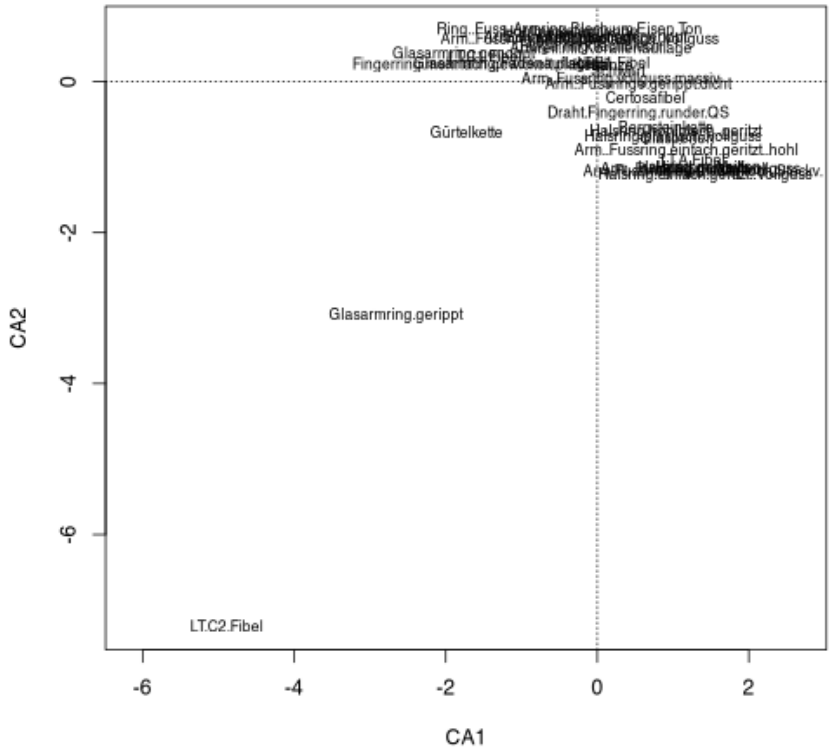
Correspondence analysis: Real World case

Münsingen Burial Site

```
plot(muensingen.cca, display = "sites")
```



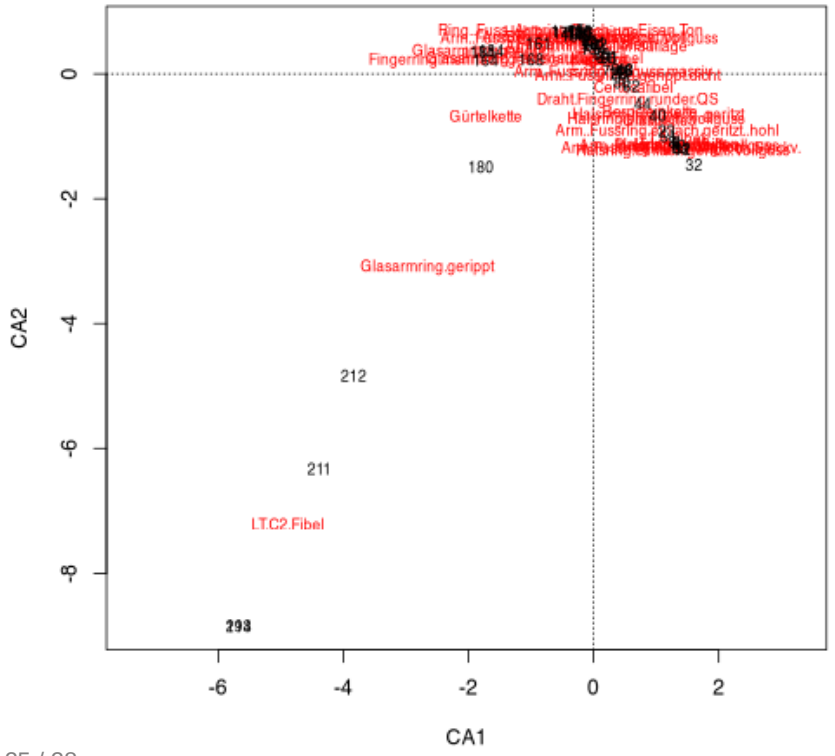

```
plot(muensingen.cca, display = "species")
```



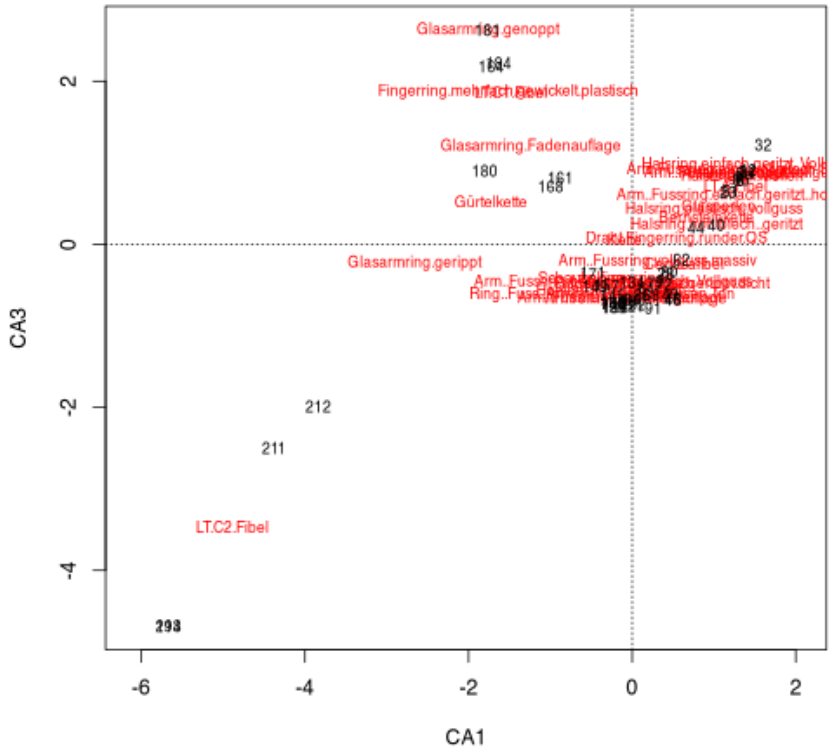
Correspondence analysis: Real World case

Münsingen Burial Site

```
plot(muensingen.cca, choices = c(1,2))
```




```
plot(muensingen.cca, choices = c(1,3))
```



Correspondence analysis: Real World case

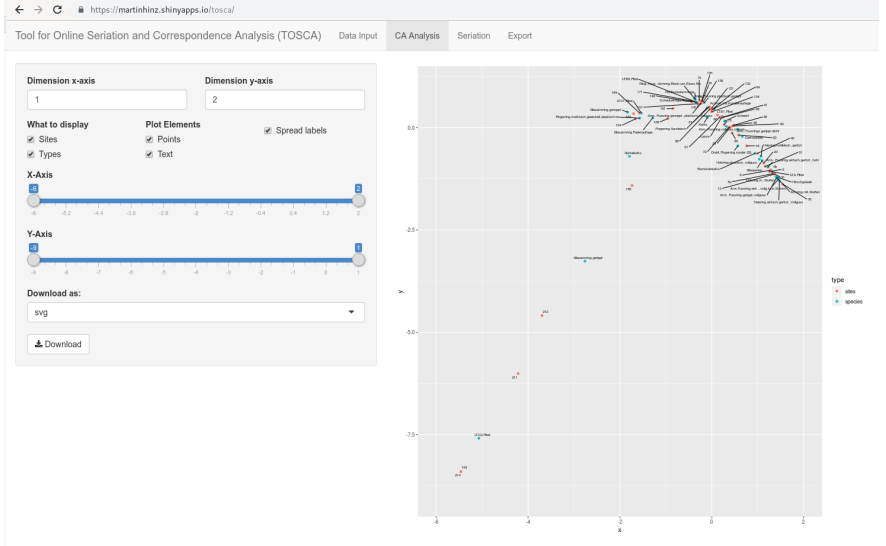
Münsingen Burial Site

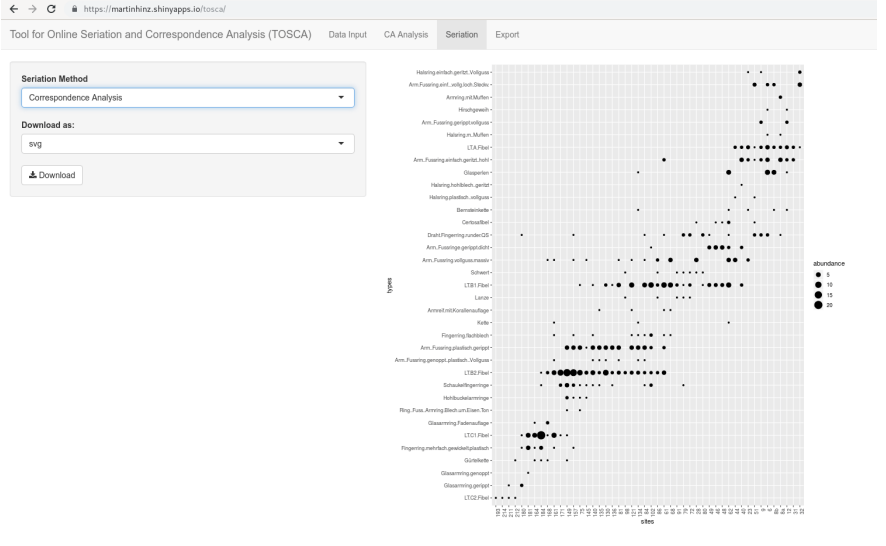
```
library(ggplot2)
library(ggrepel)

muensingen.species <- data.frame(
  scores(muensingen.cca)$species
  )
ggplot(muensingen.species,
       aes(x=CA1,
           y=CA2,
           label=rownames(muensingen.species))) +
  geom_point() + geom_text_repel(size=2)
```

```
## Warning: ggrepel: 13 unlabeled data points (t
## increasing max.overlaps
```

Correspondence analysis: Real World case

Münsingen Burial Site





http://tosca.archaeological.science

Correspondence Analysis: Interpretation

Guttman effect (horseshoe, parabola)

In archaeology, this is often regarded as evidence of a temporal orientation.

The Guttman effect occurs when a process affects the data on multiple levels.

The largest influencing factor, given a longer runtime, is mostly the time, but:

This does not always have to be the case.

Check against other information necessary.