

Correspondence Analysis (CA)

Kristel Uiboaed

Seminar in Methodology and Statistics

24th of May 2011 @RUG

Overview

- Terminology
- An example of dialect variation and word classes
- CA in R
- Syntactic variation of Estonian dialects
- Multiple CA and the study of semantic variation

What is CA?

- Descriptive/exploratory technique
- Analyses two-way and multi-way tables (most often frequency tables)
- Represents tabular data graphically
- Seeks correspondence between rows and columns
- Similar to factor analysis

Definition

„ CA is a method of displaying the rows and columns of a table as points in a spatial map, with a specific geometric interpretation of the positions of the points as a means of interpreting the similarities and differences between rows, the similarities and differences between columns and the association between rows and columns“ (Greenacre 2007)

Profiles / points

| | EAS | NOR | COA | TAR | MUL | VÖR | <i>row totals</i> |
|----------------------|--------------|-------------|--------------|--------------|--------------|--------------|-------------------|
| verbs | 8763 | 3707 | 8618 | 8258 | 6476 | 8124 | <i>43946</i> |
| substantives | 7996 | 3784 | 8016 | 7064 | 5514 | 6843 | <i>39217</i> |
| adjectives | 1239 | 526 | 1005 | 1173 | 876 | 1085 | <i>5904</i> |
| adverbs | 2671 | 1642 | 3563 | 2356 | 2261 | 2638 | <i>15131</i> |
| postp | 716 | 267 | 661 | 638 | 629 | 747 | <i>3658</i> |
| prep | 85 | 39 | 116 | 76 | 63 | 76 | <i>455</i> |
| column totals | <i>21470</i> | <i>9965</i> | <i>21979</i> | <i>19565</i> | <i>15819</i> | <i>19513</i> | |

Row profiles

| | EAS | NOR | COA | TAR | MUL | VÖR | <i>row totals</i> |
|---------------------|------------|------------|------------|------------|------------|------------|-------------------|
| verbs | 19.9 | 8.4 | 19.6 | 18.8 | 14.7 | 18.5 | <i>100</i> |
| substantives | 20.4 | 9.6 | 20.4 | 18.0 | 14.1 | 17.4 | <i>100</i> |
| adjectives | 21.0 | 8.9 | 17.0 | 19.9 | 14.8 | 18.4 | <i>100</i> |
| adverbs | 17.7 | 10.9 | 23.5 | 15.6 | 14.9 | 17.4 | <i>100</i> |
| postp | 19.6 | 7.3 | 18.1 | 17.4 | 17.2 | 20.4 | <i>100</i> |
| prep | 18.7 | 8.6 | 25.5 | 16.7 | 13.8 | 16.7 | <i>100</i> |

Chi-square distance

$$\chi^2 = \sum \frac{(\textit{observed} - \textit{expected})^2}{\textit{expected}}$$

- Measures distance between profiles

Mass / weight

- Weights assigned to the profiles
- Measure of point's importance / influence in the analysis
- Mass of the i -th row = marginal frequency of the i -th row / row total
- Mass of the j -th column = marginal frequency of the j -th column / column total

Inertia / eigenvalue

- Shows variance in the table
- Not dependent on sample size
- Weighted average of the squared χ^2 distance between row profiles and their average profiles
- Inertia is high when the row profiles have large deviation from their average and low when they are close to the average
- High inertia means high row-column association

Principal axes

- Best fitting line – first principal axis
- Direction of spread of points that optimizes the inertia displayed
- Total inertia 0.0038 in the example dataset (very low, which would mean that profiles are almost identical)

(squared) correlation

- Part of the variance of a variable explained by a principal axis
- Independent of the point's mass

Contribution

- Component of inertia accounted for by a certain point on the certain axis
- Calculated with respect to the entire set of variables

CA in R

- Simple CA
 - package *ca*
 - `ca(frequency_table)`
 - package *languageR*
 - `corres.fnc(frequency_table)`
- Multiple CA
 - package *MASS*
 - `mca(categorical_data_table)`
 - package *ca*
 - `mjca(categorical_data_table)`

```
> summary(corres.fnc(words))
```

```
Call:  
corres(words)
```

```
Eigenvalue rates:
```

```
0.7783088 0.1720264 0.03782728 0.01145789 0.0003796231 ...
```

```
Factor 1
```

```
coordinates correlations contributions  
EAS 0.037 0.635 0.093  
NOR -0.089 0.814 0.246  
COA -0.077 0.954 0.405  
TAR 0.059 0.884 0.213  
MUL 0.007 0.028 0.002  
VÖR 0.026 0.476 0.041  
...
```

```
Factor 2
```

```
coordinates correlations contributions  
EAS -0.026 0.314 0.209  
NOR -0.034 0.119 0.163  
COA 0.004 0.003 0.006  
TAR -0.017 0.075 0.082  
MUL 0.039 0.884 0.336  
VÖR 0.027 0.519 0.204  
...
```

```
>
```

```
> summary(ca(words))
```

Principal inertias (eigenvalues):

```
dim  value  % cum%  scree plot
1   0.002951 77.8 77.8 *****
2   0.000652 17.2 95.0 *****
3   0.000143  3.8 98.8 *
4   4.3e-050  1.1 100.0
5   1e-06000  0.0 100.0
```

```
Total: 0.003792 100.0
```

Rows:

```
name  mass  qlt  inr  k=1 cor ctr  k=2 cor ctr
1 | vrbs | 406 926 139 | -33 851 152 | 10 76 61 |
2 | sbst | 362 956  74 |  9 116  11 | -25 839 361 |
3 | adjc |  55 857 114 | -81 815 120 | -18 42 28 |
4 | advr | 140 996 526 | 117 964 651 | 22 32 99 |
5 | pstp |  34 931 128 | -69 328  54 | 93 603 450 |
6 | prep |  4 525  19 | 93 515  12 | 13  9  1 |
```

Columns:

```
name  mass  qlt  inr  k=1 cor ctr  k=2 cor ctr
1 | EAS | 198 949 115 | -37 635  93 | -26 314 209 |
2 | NOR |  92 933 235 |  89 814 246 | -34 119 163 |
3 | COA | 203 957 330 |  77 954 405 |  4  3  6 |
4 | TAR | 181 959 187 | -59 884 213 | -17 75 82 |
5 | MUL | 146 912  65 |  -7  28  2 | 39 884 336 |
6 | VÖR | 180 995  68 | -26 476  41 | 27 519 204 |
```

Warning message:

abbreviate used with non-ASCII chars

```
>
```


Map

- [CA map](#)

Non-finite and finite verb constructions in Estonian dialects

- Variation in dialects
- Frequency lists of non-finite form and finite verb lemma combinations
- Which dialects are more similar in terms of these constructions?
- Are the results compatible with the traditional dialect classifications?

Non-finite forms

- 7 different non-finite forms in Estonian (past and present participles, 2 infinitives, gerund)
- Non-finites have different functions (e.g. English *want to go* and *let go*)
- Can form constructions with finite verbs

Examples

Ta *läks* *jaluta-ma.*

he/she go.3SG.PST walk-2INF

‘He went to walk.’

Ma *oskan* *laul-da.*

I can.1SG.PRS sing-1INF

‘I can sing.’

Fragment of the dataset

| construction | COA | EAS | INS |
|--------------|-----|-----|-----|
| nud_olema | 515 | 512 | 622 |
| tud_saama | 181 | 28 | 192 |
| ma_pidama | 96 | 84 | 176 |

- 10 dialects
- Non-finite form + finite verb lemma
- Frequency counts
- 120 different constructions and their frequencies in all the dialects

Map of constructions

- CA map for verb constructions in Estonian dialects

Multiple CA

- Categorical data
- For more than two categorical variables

Multiple CA applied in semantics

- Semantic variation of verb *run* in Finnish and Estonian (Lehismets & Uibo 2011)
- Around 1000 sentences from both languages
- Annotation of different semantic properties

Fragment of the dataset

| Sentence | Motion | Trajector | Animacy | Landmark | Transitivity | Source | Target |
|--|--------|------------|---------|----------|--------------|-------------|-------------|
| Ei , las Ballangrud jookseb enne ! | concr | concr_traj | anim | time | intr | NA | NA |
| Taas jookseb mõni asjamees puldist stuudiosse ning ütleb pead väänutades : | concr | concr_traj | anim | NA | intr | spat_source | spat_target |
| ja tean , et osa Eestimaa piirist jookseb Venemaaga . | fict | abstr_traj | inanim | NA | intr | NA | NA |
| Tööstuses tehakse imet - see , mida me näpu vahel nikerdame ja pusserdame , jookseb sadade meetritena masinate vahelt ja eksponeeritakse soliidses tekstiilifirma messiboksis paremini kui ühelgi näitusel . | concr | concr_traj | inanim | location | intr | spat_source | NA |
| Eeloleval pühapäeval jookseb New Yorgi Madison Square Gardenis platsile seitse tähtede mängu uustulnukat . | concr | concr_traj | anim | loc_time | intr | NA | spat_target |
| Vene poiss jookseb üle õue koolimajja . | concr | concr_traj | anim | NA | intr | spat_source | spat_target |

CA map of *run*

- CA map for *run*

Some linguistic studies where CA has been applied

- Phonetics and dialectology (Chiocki 2006)
- Syntactic, regional, register, genre and diachronic variation (Ernestus et. al 2007)
- Morphology (Keune et. al 2005)
- Modal verbs (Wilson 2005)
- Semantics and dialectology (Szelid & Geeraerts 2008)
- Semantics (Glynn 2009)

References (1)

- Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.
- Cichocki, W. (2006). Geographic Variation in Acadian French /r/: What Can Correspondence Analysis Contribute toward Explanation?, *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing*, vol. 21, no. 4, pp. 529-541.
- Ernestus, M., Van Mulken, M., & Baayen, R. H. (2007). Ridders en heiligen in tijd en ruimte: Moderne stylometrische technieken toegepast op Oud-Franse teksten. *Taal en Tongval*, 58, 1-83.
- Glynn, D. (2009). Polysemy, Syntax, and Variation. A usage-based method for Cognitive Semantics. V. Evans & S. Pourcel (eds), 77-106. *New Directions in Cognitive Linguistics*. Amsterdam: John Benjamins.

References (2)

- Greenacre, M. (2007). Correspondence analysis in practice, 2nd edn, CRC Press, Boca Raton.
- Keune, K., Ernestus, M., Van Hout, R. and Baayen, R. H. (2005). Social, geographical, and register variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory* 1, 183-223.
- Lebart, L., Salem, A. & Berry, L. (1998). Exploring textual data, Kluwer Academic Publishers, Dordrecht.
- Lehismets, K and Uiboaed, K. (2011). Eesti ja soome keele *jooksma/juosta* semantilisest varieeruvusest. Seminar „Grammatikat lüpsmas“ 1st-2nd of April 2011 in Tartu
- Szelid, Veronika and Dirk Geeraerts (2008). Usage-based dialectology: Emotion concepts in the Southern Csango dialect, [Annual Review of Cognitive Linguistics](#), Ruiz de Mendoza Ibáñez, Francisco José (ed.), 23–49.
- Wilson, A (2005). Modal verbs in written Indian English: a quantitative analysis of the Kolhapur corpus using correspondence analysis. *ICAME Journal* 29: 151-169.