Taking into account semantic similarities in correspondence analysis

Mattia Egloff and François Bavaud

University of Lausanne

UNIL | Université de Lausanne

COMHUM 2018

COMHUM 2018 – Lausanne

I

Recall: text-document matrix, distributional dissimilarities and Correspondence Analysis

Text-document matrix

The term-document matrix $N = (n_{ik})$ counts the occurrences of terms $i = 1, \ldots, n$ in documents $k = 1, \ldots, p$

Exemple: distribution of $n = 643$ verbs (after lemmatisation) in the $p = 11$ chapters of Book I of "An Inquiry into the Nature and Causes of the Wealth of Nations" by Adam Smith (1776)

	abandon	abolish	abound	abridge	abuse	accelerate	...
Chapter 1	0	0	0	3	0	0	...
Chapter 2	0	0	0	0	0	0	...
Chapter 3	0	0	0	0	0	0	...
Chapter 4	0	0	0	0	1	0	...
Chapter 5	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...

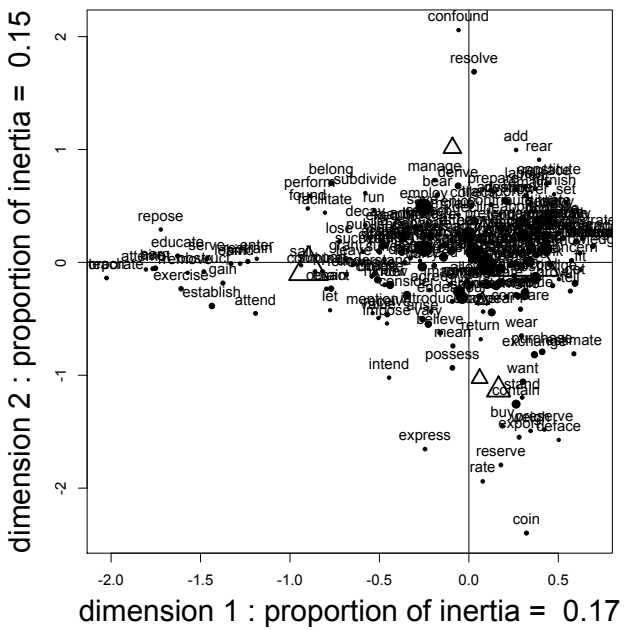
Table: *transpose of the term-document matrix* $N = (n_{ik})$

Omnipresence of the text-document matrix in textual quantitative studies

- **distributional** "chi2" dissimilarities between documents

$$D_{kl}^{\chi}=\sum_{i=1}^{n} f_i(q_{ik}-q_{il})^2 \qquad f_i=\frac{n_{i\bullet}}{n_{\bullet\bullet}} \qquad q_{ik}=\frac{n_{ik}n_{\bullet\bullet}}{n_{i\bullet}n_{\bullet k}}$$

- weighted multidimensional scaling (MDS) on D^{χ} (and dual MDS on dissimilarities between terms): **Correspondence Analysis**



- Latent semantic analysis, topic modelling, etc.

The distributional hypothesis

The success of the methods based on text-document matrix (initiated by Benzecri (1973) and others) may be explained in part by the distributional hypothesis (Harris 1968; but also Bloomfield (1933), de Saussure (1916), and many others).

Distributional hypothesis postulates an association between the **distributional similarity** and **semantic similarity** of terms

II

Semantic similarity between terms

A largely overlooked methodological limitation:

(partial) synonymy

A largely overlooked restriction in term-documents matrices (and categorical variables in general) is to consider that two terms i and j :

- are either completely different
- or completely similar.

Yet, `abandon` and `abolish`, say, are arguably more semantically similar than are `abandon` and `accelerate`, or than are `abolish` and `accelerate`.

Terms being partially similar, the corresponding chi2 dissimilarity between the documents (made of term distributions) should decrease.

Defining semantic similarities between terms

How to define and compute semantic similarities $\mathbb{S} = (\mathbb{s}_{ij})$ or dissimilarities $\mathbb{D} = \mathbb{d}_{ij}$ between pair of terms ij ?

$\rightarrow$ a largely open issue in linguistics, cognitive science, natural language processing, artificial intelligence, etc.

Possible direct strategies:

- use a dictionary of synonyms
- use an ontology, such as *WordNet* (G.Miller, C.Fellbaum and others 1990–)

A sketch of WordNet semantic disssimilarities

- hierarchy between concepts : $c_1 \leq c_2$ means that *concept* c_1 *is an instance of concept* c_2 (ex. $\texttt{cat} \leq \texttt{animal}$)
- $c_1 \vee c_2$ represents the *least general concept* subsuming both c_1 and c_2 (ex. $\texttt{cat} \vee \texttt{dog} = \texttt{animal}$)
- the probability $p(c)$ of a concept c is estimated as the relative frequency (in some reference corpus) of words j whose sense c_j is an instance of concept c:

$$p(c) := \frac{\sum_j n_j \, \mathbf{1}(c_j \leq c)}{\sum_j n_j}$$

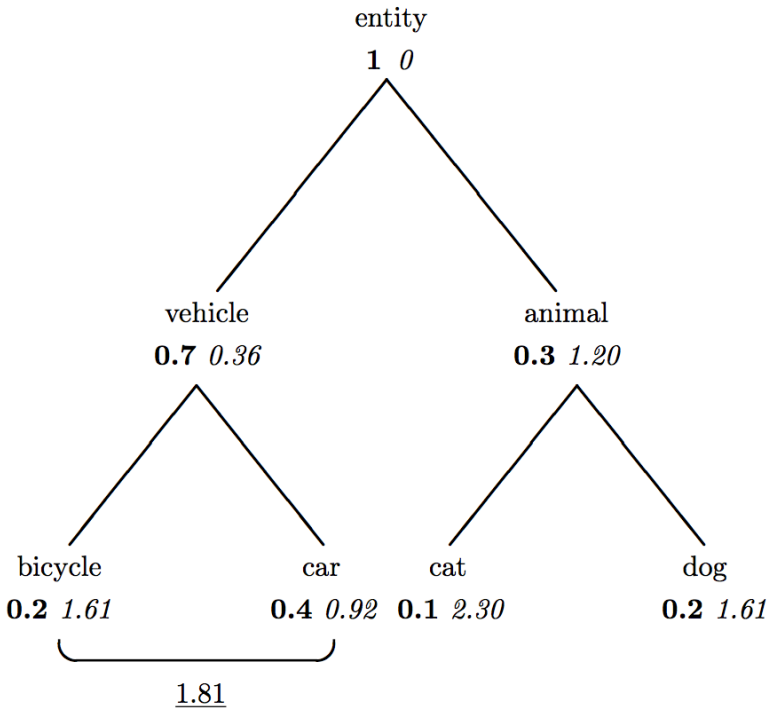
- the *semantic similarity* s_{ij} between pairs of words ij is defined as $s_{ij} = -\ln p(c_i \vee c_j)$, and the corresponding *dissimilarity* can be defined as

$$s_{ij} = -\ln p(c_i \vee c_j) \qquad \mathbb{d}_{ij} = s_{ii} + s_{jj} - 2s_{ij}$$

Implications:

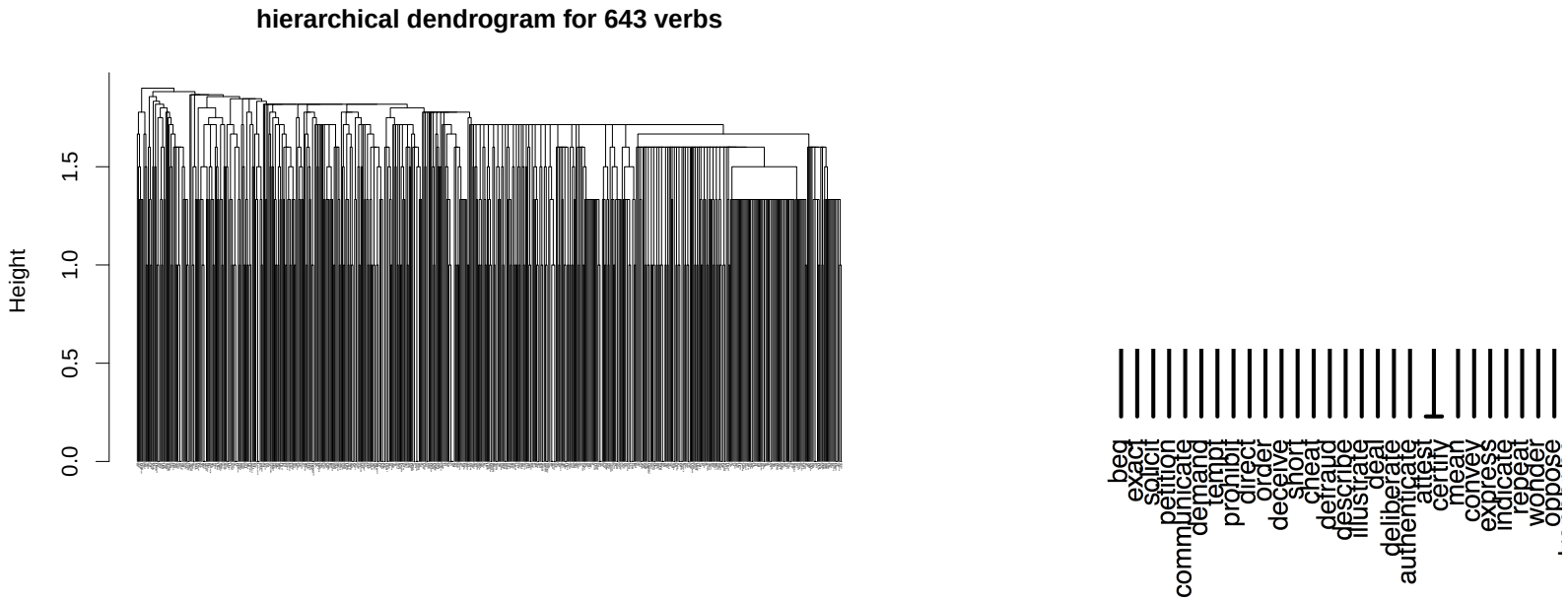
- $s_{ii} = -\ln p(c_i) \geq 1$ measures the rarity of (first) sense of word i
- $\mathbb{d}_{ij}$ *is an* ultrametric *disssimilarity* (and, in particular, squared Euclidean: MDS is feasible).

A toy example (for nouns) in WordNet



Toy WordNet ontology: probabilities $p(c)$ (boldface) and their negative logarithms $-\ln p(c)$ (italic). The resulting dissimilarity between `bicycle` and `car` is $1.61 + 0.92 - 2 \times 0.36 = 1.81$.

WordNet dissimilarities between verbs

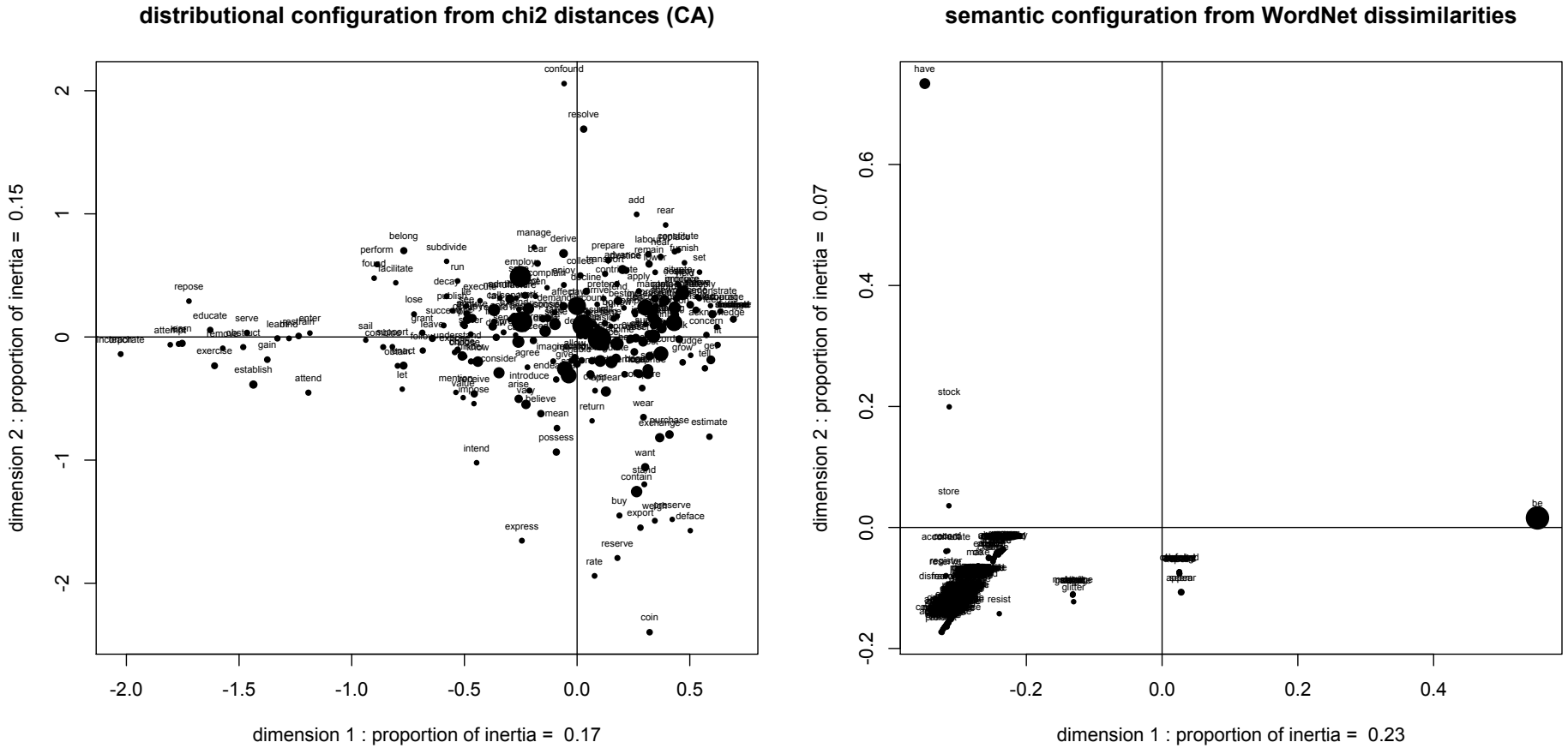


Left: *exact dendrodram* on ultrametric *path-distance* WordNet semantic dissimilarities (first sense only; 20 distinct values for pairs of dissimilarities).

Right: detail of the 28 leftmost verbs

Distributional and semantic dissimilarities

Configurations (in dimensions 1 and 2), for distributional dissimilarities D^{χ}_{ij} (left) and semantic dissimilarities $\mathfrak{d}_{ij}$ (right) between verbs:

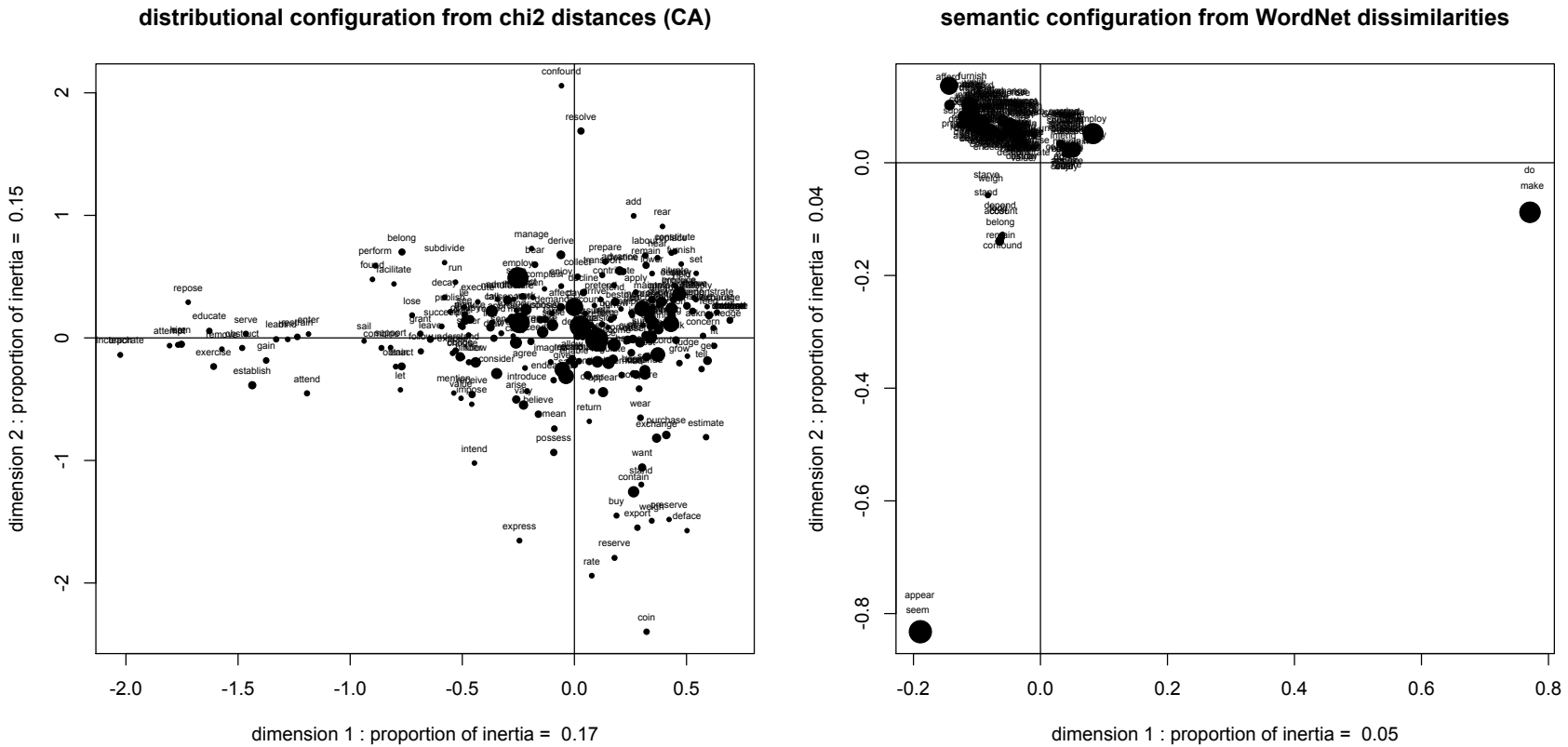


The association between the two configurations (**test of the distributional hypothesis**, derived from the test of spatial autocorrelation) is significative:

$p = 0.00009$

Reduced analysis

Retaining the verbs occurring at least 5 times, and excluding "`have`" and "`be`" yields 234 verbs (instead of 643 verbs):



Statistically significative association between the two configurations : $p = 0.000003$

III

Term similarities → variety reduction

Normalized similarities or similitudes

For various important formal reasons outside the scope of the presentation, it is fruitful to consider *normalized semantic similarities* or *similitudes* $\mathbb{S} = (\mathbb{s}_{ij})$ satisfying

$$0 \;\leq\; \mathbb{s}_{ij} = \mathbb{s}_{ji} \;\leq\; 1 = \mathbb{s}_{ii}$$

Let f_i be the relative frequency of term i. Shannon measure of diversity (entropy)

$$H(f) = -\sum_{i=1}^{n} f_i \ln f_i$$

can be replaced by the *reduced entropy* (inspired from considerations in quantitative ecology)

$$\hat{R}(f) = -\sum_{i=1}^{n} f_i \ln b_i \qquad \text{where} \quad b_i = \sum_{j=1}^{n} \mathbb{s}_{ij} f_j = (\mathbb{S}f)_i \quad \text{is the } \textit{banality} \text{ of term } i \ .$$

By construction, $f_i \leq b_i \leq 1$, and

$\hat{R}(f) \leq H(f)$: variety decrease due to the semantic similarities between terms.

Decrease of the chi-square dissimilarities

The chi2 dissimilarities between documents

$$D^{\chi}_{kl} = \sum_{i=1}^{n} f_i \, (q_{ik} - q_{il})^2$$

can be replaced by the *reduced chi2 dissimilarities* (squared Euclidean as well)

$$\hat{D}_{kl} = \sum_{ij} \mathbb{t}_{ij} (q_{ik} - q_{il})(q_{jk} - q_{jl}) \qquad \text{where} \quad \mathbb{t}_{ij} = \frac{f_i f_j \mathbb{s}_{ij}}{\sqrt{b_i b_j}}$$

In particular, the chi2 measure of dependence can be replaced by the reduced chi2 measure of dependence

$$\Delta = \frac{1}{2} \sum_{kl} \rho_j \rho_k D_{jk} \qquad \rightarrow \qquad \hat{\Delta} = \frac{1}{2} \sum_{kl} \rho_j \rho_k \hat{D}_{jk} \;\; \leq \;\; \Delta$$

MDS on D^{χ} amounts to ordinary Correspondence Analysis.

New proposal: *MDS on the reduced chi2 dissimilarities $\hat{D}$ amounts to Reduced Correspondence Analysis* .

Choice of the semantic normalized similarity

For various important formal reasons outside the scope of the presentation (bis), it is fruitful to define semantic normalized similarities $\mathbb{S} = (\mathbb{s}_{ij})$ from semantic (WordNet) dissimilarities $\mathbb{D} = (\mathbb{d}_{ij})$ as

$$\mathbb{s}_{ij} = \exp(-\beta\, \mathbb{d}_{ij}/\mathbb{\Delta}) \qquad \text{where} \quad \mathbb{\Delta} = \frac{1}{2}\sum_{ij} f_i f_j \mathbb{d}_{ij} \quad \text{is the } \textit{semantic inertia.}$$

$\beta > 0$ is a *bandwidth parameter*, which controls the *paradigmatic sensitivity* of the linguistic subject:

- the higher β, the larger the semantic distances between documents, and the larger the spread of the factorial cloud as measured by reduced inertia $\hat{\Delta}(\beta)$
- a low β can model an illiterate person, sadly unable to discriminate between documents, which look all alike. In particular,

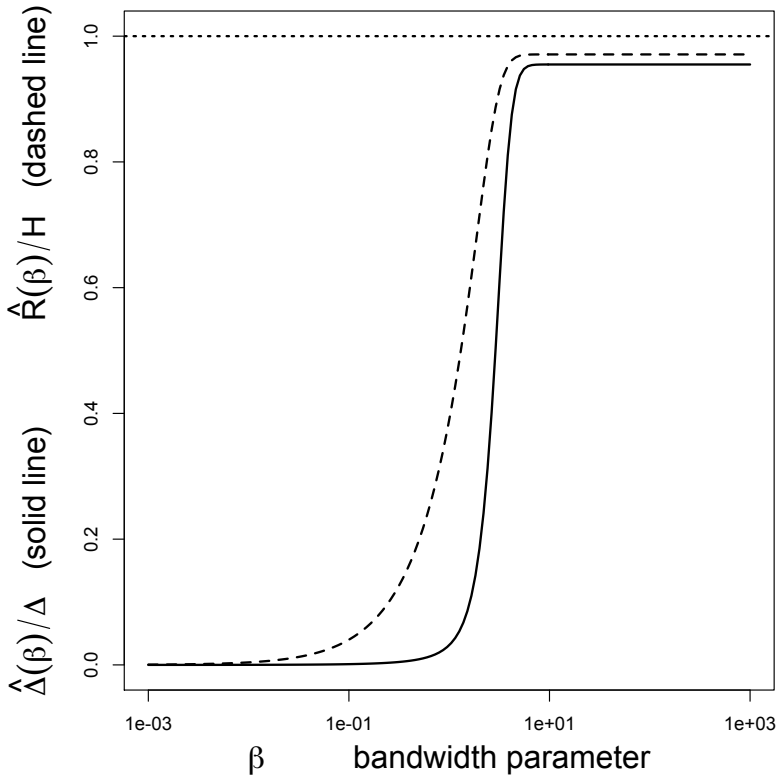
$$\lim_{\beta \to 0} \hat{\Delta}(\beta) = 0 \qquad \lim_{\beta \to 0} \hat{R}(\beta) = 0 \ .$$

Entropy and inertia reduction

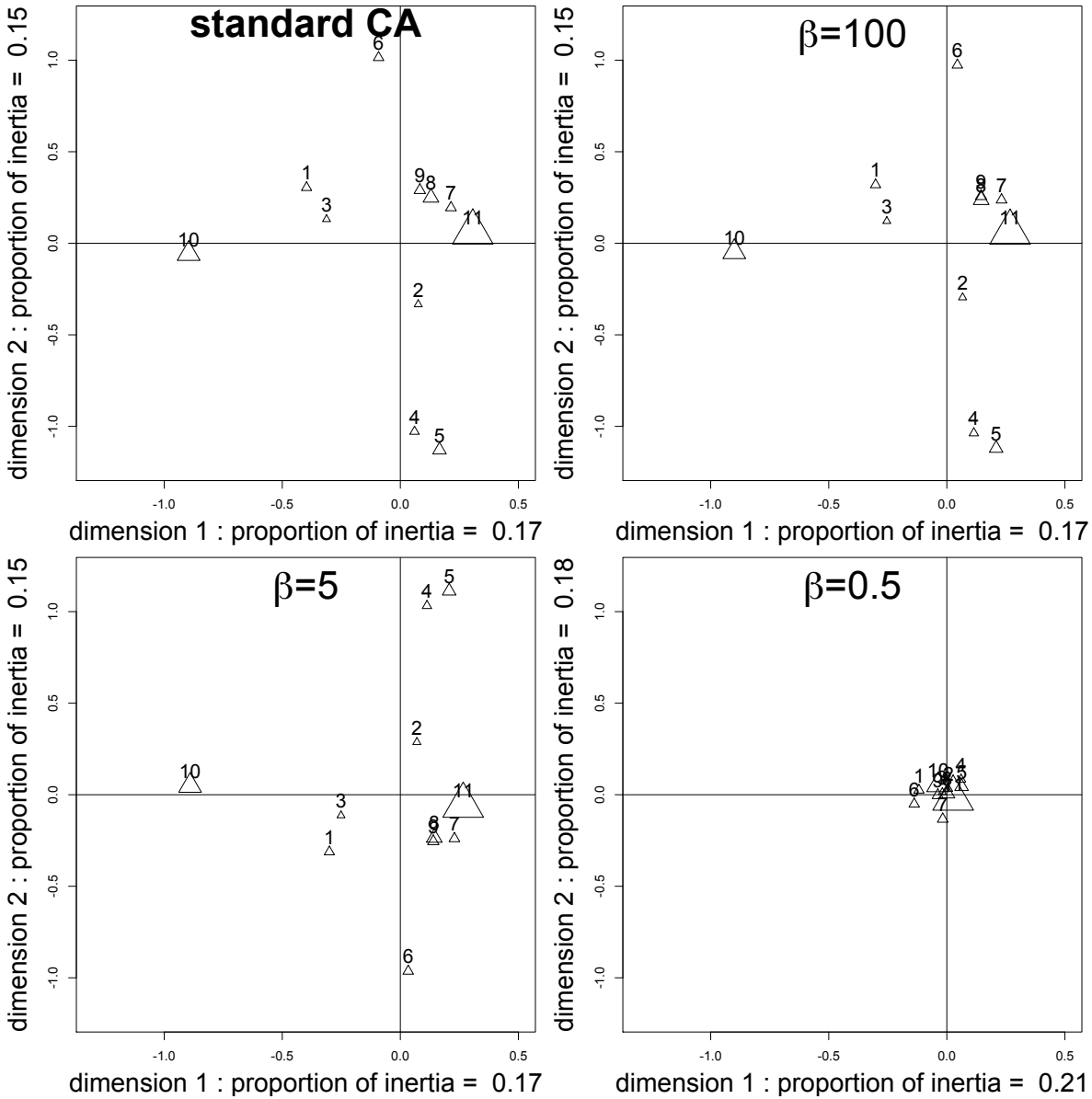
When all the terms are semantically distinct (i.e. $\mathbb{d}_{ij} > 0$ for $i \neq j$), then

$$\lim_{\beta \to \infty} \hat{\Delta}(\beta) = \Delta$$
$$\lim_{\beta \to \infty} \hat{R}(\beta) = H$$

This behavior does not hold in the example: the $n = 234$ verbs display, accordingly to their first sense in WordNet, 15 cliques of size 2 and 3 cliques of size 3: `employ-apply-use`, `set-lay-put` and `supply-furnish-provide`.



Reduced Correspondence Analysis



In guise of conclusion: future work

- attempting to go beyond the first sense in WordNet (the problem of semantic disambiguation)
- other sources of semantic similarities / dissimilarities? (synonym dictionaries such as CRISCO, Université Caen Normandie)
- normalized similarities arise as convex mixtures of binary equivalence relations → strong formal relations with *soft clustering* and *fuzzy partitionning*, defined as convex mixtures of binary equivalence relations.
- systematic recognition of the *underlying similarity between modalities of a nominal variable* in quantitative methods and statistics in general.

  (simple example: some pairs among the `nationality` categories such as {`swiss, italian, spanish, french, british, belgian, dutch, german, austrian, finnish, russian, japanese ...`} may be judged as more similar than others...)
- extend the venerable but very crude *bag of words* representation of documents to more flexible models for textual contexts: *fuzzy neighborhoods*, but also *Markov navigation* on *bibliographic references* or *hyperlinks*, etc.