From Data to the p-Adic or Ultrametric Model

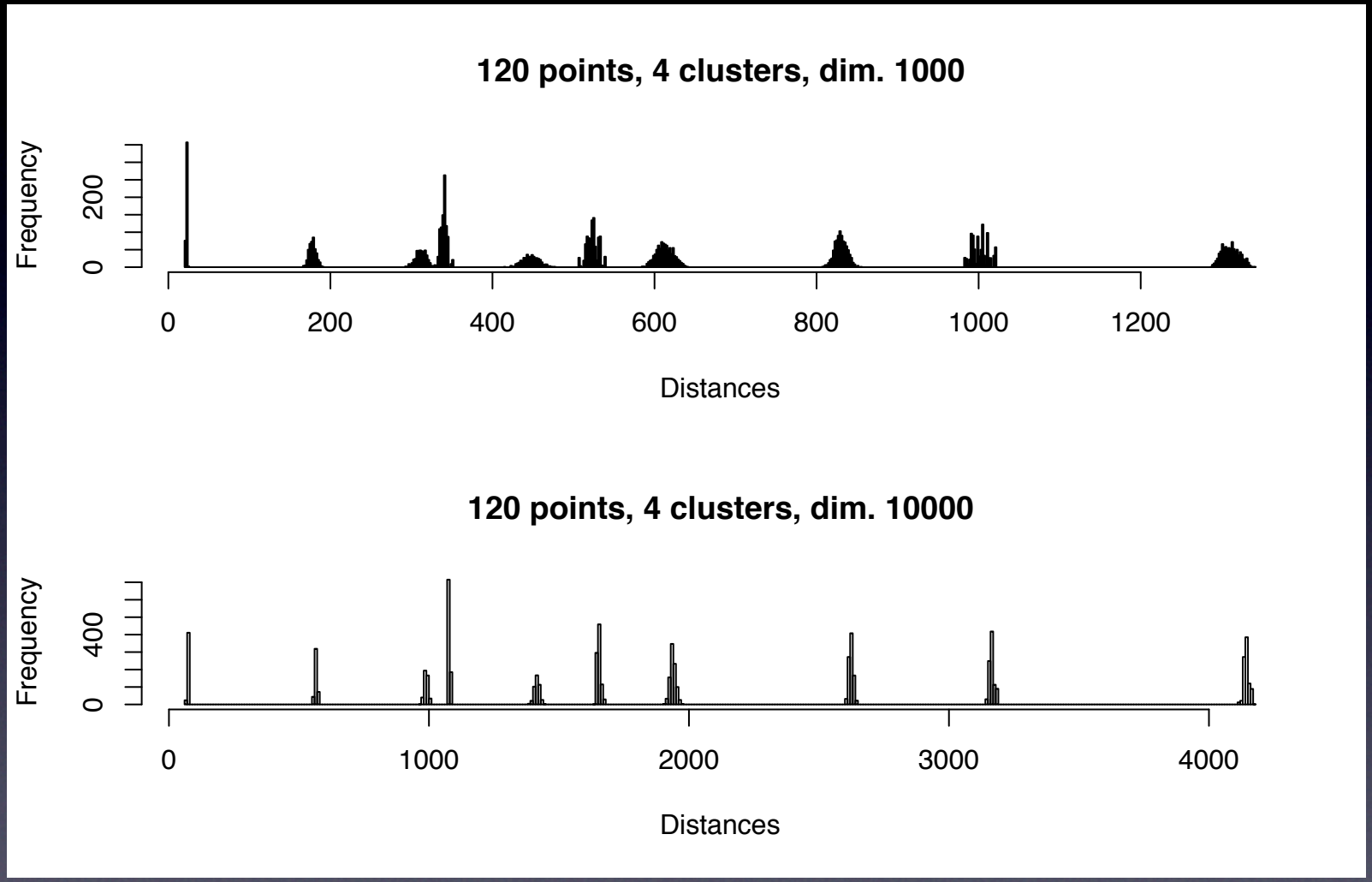
- High dimensional spaces endowed with a scalar product are naturally ultrametric.
- Why scalar product spaces? Because in this work angles, and triangle properties, are a convenient way to characterize ultrametricity.
- In practice, we need to take heterogeneous data, and often frequencies of occurrence, into a "well behaved" scalar product space. For this, Correspondence Analysis provides a versatile tool, and facilitates inducing a hierarchy.
- Hierarchy provides a powerful means of studying anomaly, or innovation, or change.

Remark: Growing interest in handling very high dimensional data in a new way, bypassing Bellman's "curse of dimensionality

- Hall, Marron, Neeman, "Geometric representation of high dimensions, low sample size data", J. Royal Statistical Society B, 67, 427-444, 2005
- Aggarwal et al., "On the surprising behavior of distance metrics in high dimensional spaces", Proc 8th Intl. Conf. on Database Theory, 420-434, 2001
- Breuel, "A note on approximate nearest neighbor methods", arXiv:cs/0703101, 2007
- Donoho, Tanner, "Neighborliness of randomly-projected simplices in high dimension", Proc Natl. Academy of Sciences, 102, 9452-9457, 2005
- Murtagh, "The remarkable simplicity of high dimensional data: application of model-based clustering", sub. to J Classification

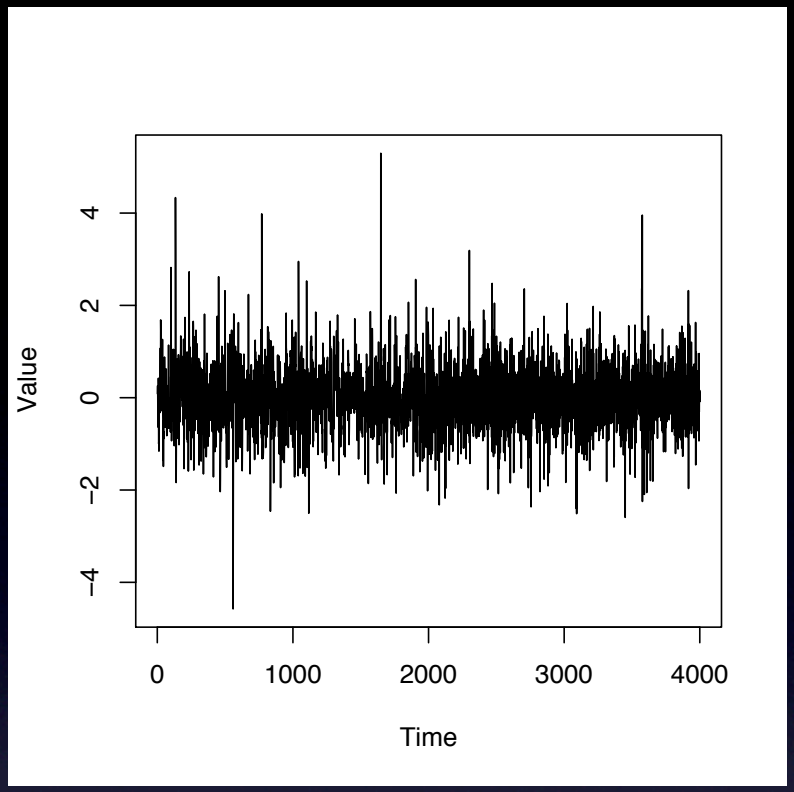
Simulation

- Generate four Gaussian clouds, each of 30 points, with respective means and standard deviations on all coordinates: (10, 0.5); (0, 4); (40, 10); (25, 7). Look at histogram of Euclidean distance as dimensionality increases.
- We know that with such data the inherent ultrametricity increases (based on: relative frequency over all triangles of the two cases of (i) isosceles with small base, and (ii) equilateral). Is latter always the case in very high dimensions? - No!


120 points, 4 clusters, dim. 1000
Frequency
0
200
0
200
400
600
800
1000
1200
Distances
120 points, 4 clusters, dim. 10000
Frequency
0
400
0
1000
2000
3000
4000
Distances

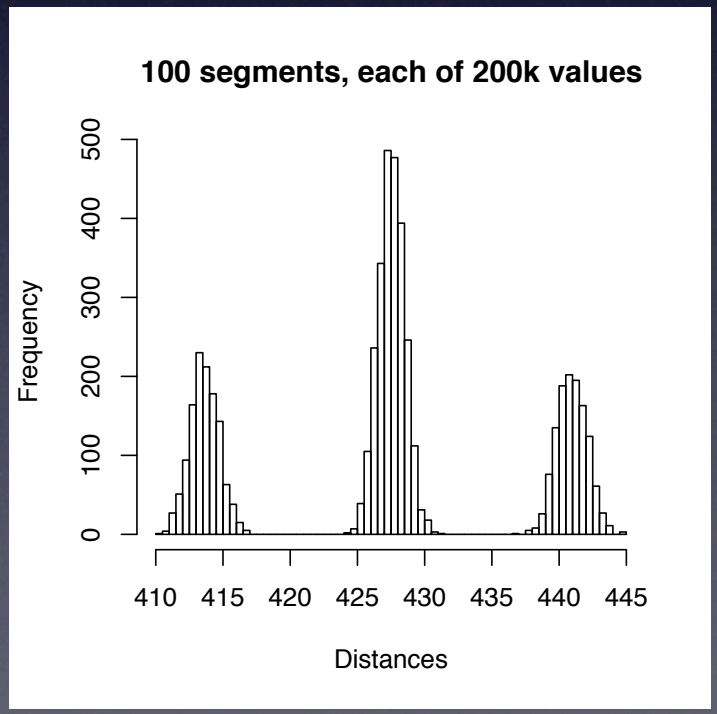
Conclusion and application

- The higher the embedding dimensionality, the easier it is to find clusters, by reading off from such a distance histogram
- Examples of two ARIMA time processes, with Student t innovation, providing “mildly longtailed” distributions



Sample of 2000 values of first signal, followed by 2000 values of second

Taking 100 "windows", of different lengths (embedding dimensionality) and assessing triangles:



Dim.	Isosc.	Equil.	UM
2000	0.17	0.32	0.49
20000	0.15	0.50	0.65
200000	0.03	0.57	0.60

A Challenge Faced by Data Analysis

A Challenge Faced by Data Analysis

- Particular interest of anomalous, atypical, exceptional, innovative cases (rather than the norm or even perhaps what can be encompassed within a statistical model)

A Challenge Faced by Data Analysis

- Particular interest of anomalous, atypical, exceptional, innovative cases (rather than the norm or even perhaps what can be encompassed within a statistical model)
- Data are heterogeneous, often massive in number of items, and of very large representational dimensionality. Proceed as follows, using Correspondence Analysis.

A Challenge Faced by Data Analysis

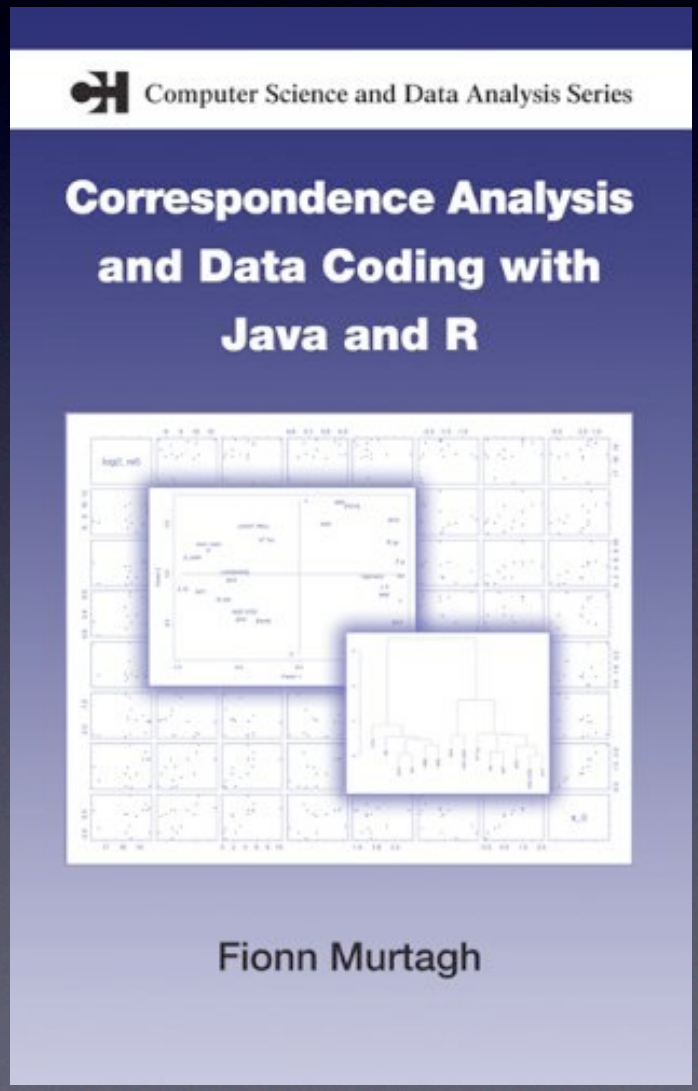
- Particular interest of anomalous, atypical, exceptional, innovative cases (rather than the norm or even perhaps what can be encompassed within a statistical model)
- Data are heterogeneous, often massive in number of items, and of very large representational dimensionality. Proceed as follows, using Correspondence Analysis.
- Firstly, from frequencies of occurrence or other input data: embed the data in a homogeneous space - a full rank Euclidean space

A Challenge Faced by Data Analysis

- Particular interest of anomalous, atypical, exceptional, innovative cases (rather than the norm or even perhaps what can be encompassed within a statistical model)
- Data are heterogeneous, often massive in number of items, and of very large representational dimensionality. Proceed as follows, using Correspondence Analysis.
- Firstly, from frequencies of occurrence or other input data: embed the data in a homogeneous space - a full rank Euclidean space
- Secondly, embed the metric data in an ultrametric space

Correspondence Analysis is A Tale of Three Metrics

1) Chi sqd. metric: profiles of frequencies of occurrence
2) Euclidean, for visualization
3) Ultrametric, hierarchic



Correspondence Analysis takes frequency of occurrence or other input data and embeds the observations and attributes in a Euclidean space

Space $\mathbb{R}^m$:

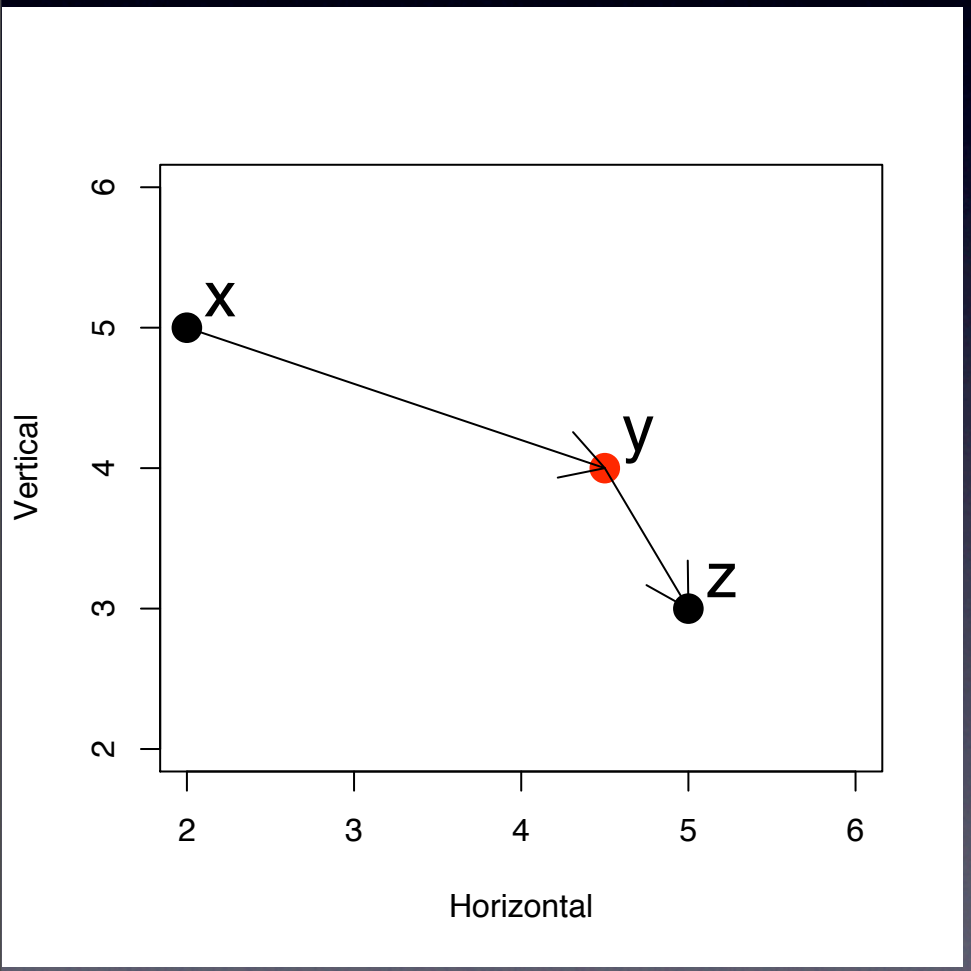
1. n row points, each of m coordinates.
2. The j^{th} coordinate is x_{ij}/x_i.
3. The mass of point i is x_i.
4. The χ^2 distance between row points i and k is:
   $d^2(i,k) = \sum_j \frac{1}{x_j}(\frac{x_{ij}}{x_i} - \frac{x_{kj}}{x_k})^2$.
   Hence this is a Euclidean distance, with respect to the weighting $1/x_j$ (for all j), between *profile* values x_{ij}/x_i etc.
5. The criterion to be optimized: the weighted sum of squares of projections, where the weighting is given by x_i (for all i).

The dual spaces of observations (rows) and attributes (columns) of the input data array are closely related

Space $\mathbb{R}^n$:

1. m column points, each of n coordinates.
2. The i^{th} coordinate is x_{ij}/x_j.
3. The mass of point j is x_j.
4. The χ^2 distance between column points g and j is:
   $d^2(g,j) = \sum_i \frac{1}{x_i}(\frac{x_{ig}}{x_g} - \frac{x_{ij}}{x_j})^2$.
   Hence this is a Euclidean distance, with respect to the weighting $1/x_i$ (for all i), between *profile* values x_{ig}/x_g etc.
5. The criterion to be optimized: the weighted sum of squares of projections, where the weighting is given by x_j (for all j).

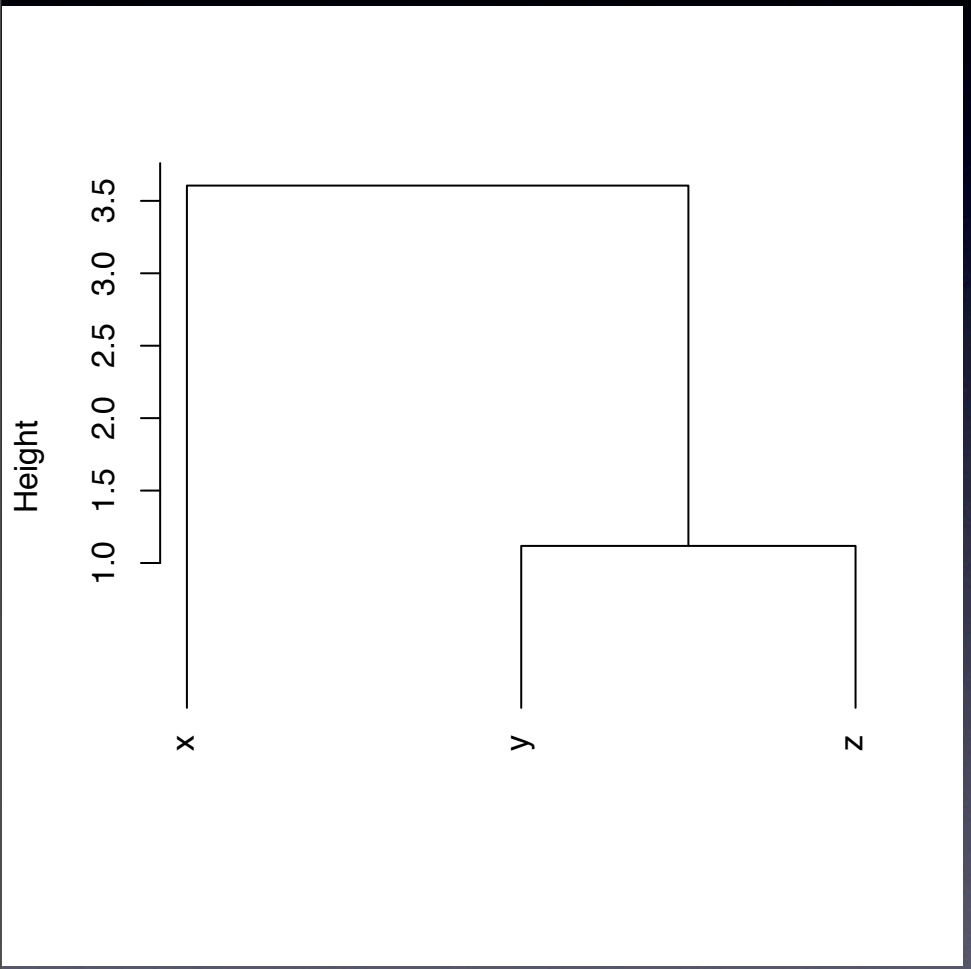
Triangular inequality holds for metrics



Example: Euclidean distance

$$d(x, z) \leq d(x, y) + d(y, z)$$

Strong triangular inequality, or ultrametric inequality, holds for tree distances



$$d(x,z) \leq$$

$$\max\{d(x,y), d(y,z)\}$$

$$d(x,z) = 3.5$$

$$d(x,y) = 3.5$$

$$d(y,z) = 1.0$$

Closest common ancestor distance is an ultrametric

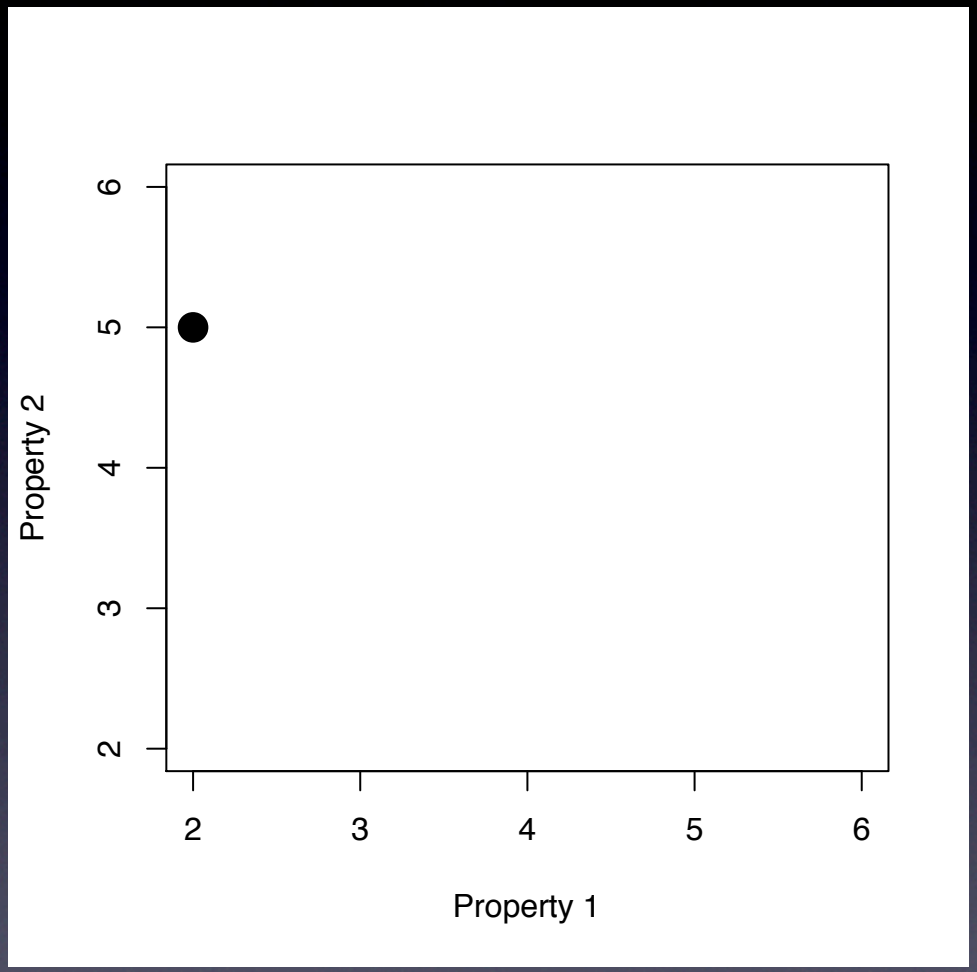
Correspondence Analysis as a versatile data preprocessing tool

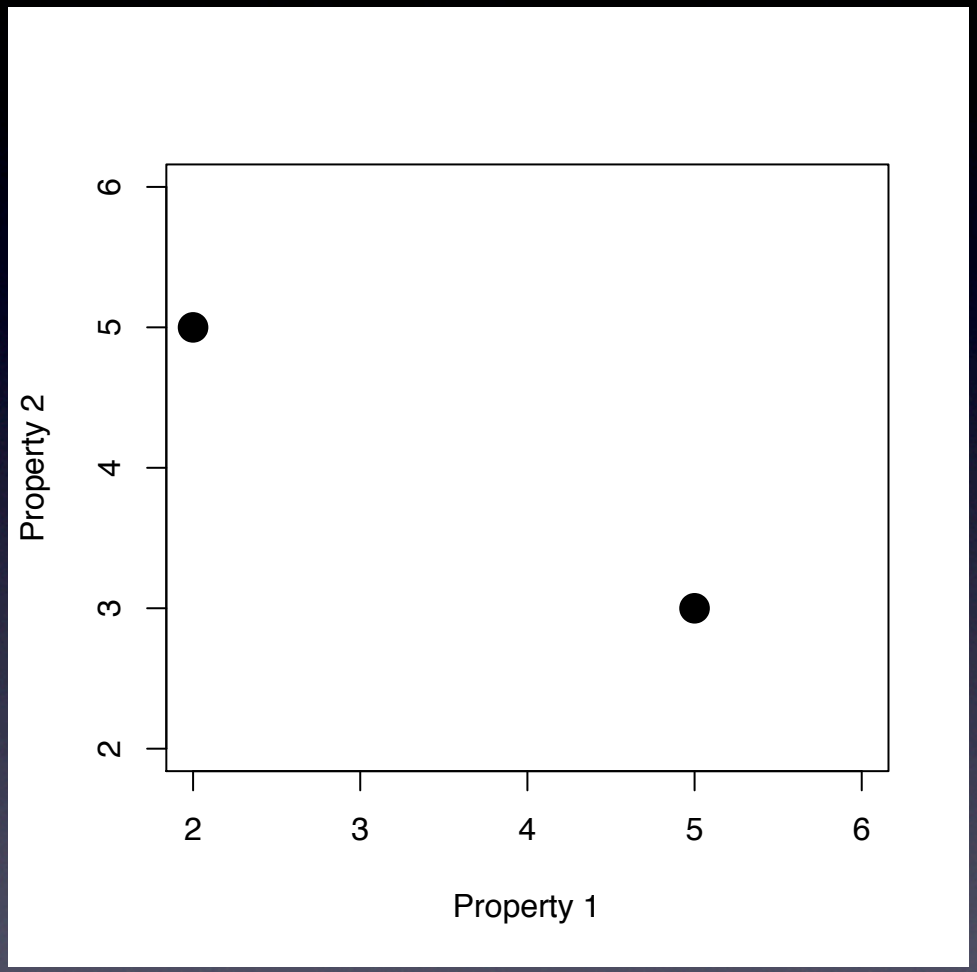
- We know that high dimensional data are increasingly ultrametric as dimensionality increases, so we want to exploit this.
- Problem: we need a scalar product space, in order to look at triangle properties, in order to determine such inherent ultrametricity. In addition, data are heterogeneous to begin with.
- Correspondence Analysis often provides an excellent tool for homogenizing the data (taking care of weighting, normalization, etc.) and furnishing a scalar product space.
- Example of high dimensional data analysis in practice: textual analysis.

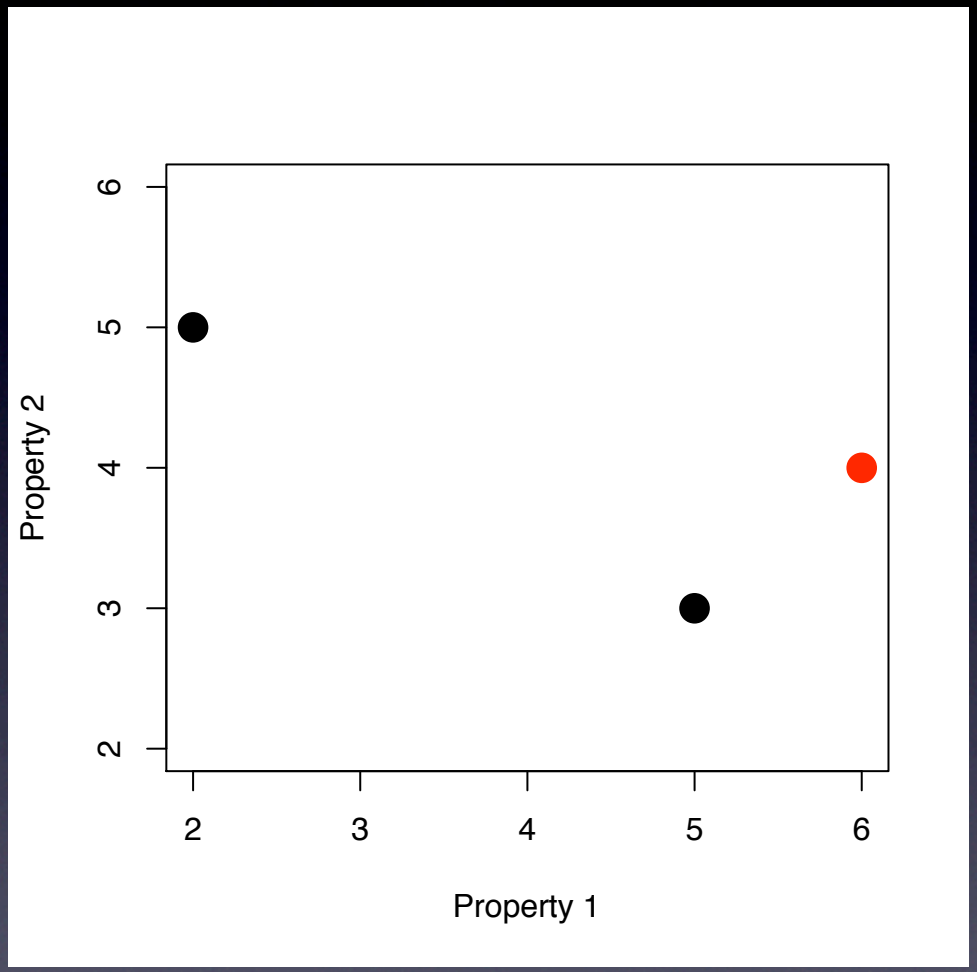
From the standard and typical to the innovative and anomalous

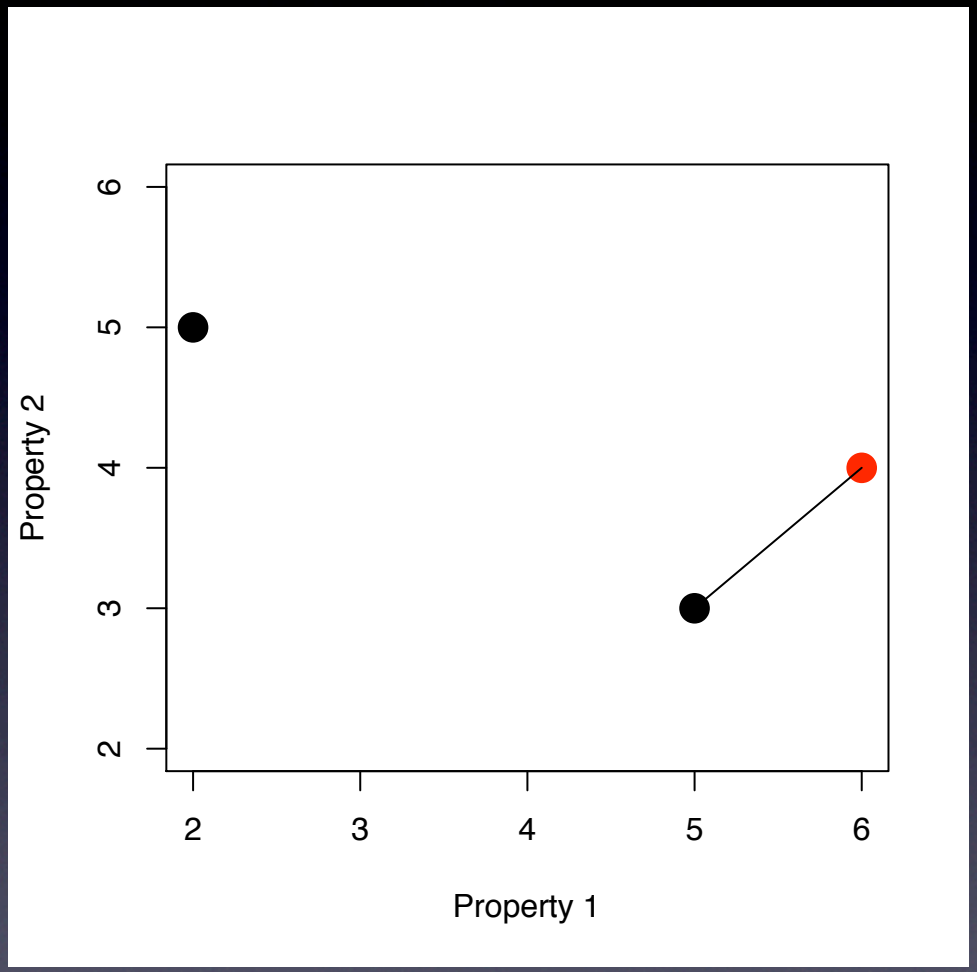
- Euclidean distance makes a lot of sense when the population is homogeneous
- Ultrametric distance makes a lot of sense when the observables are heterogeneous, discontinuous
- Latter is especially useful for determining: anomalous, atypical, innovative cases

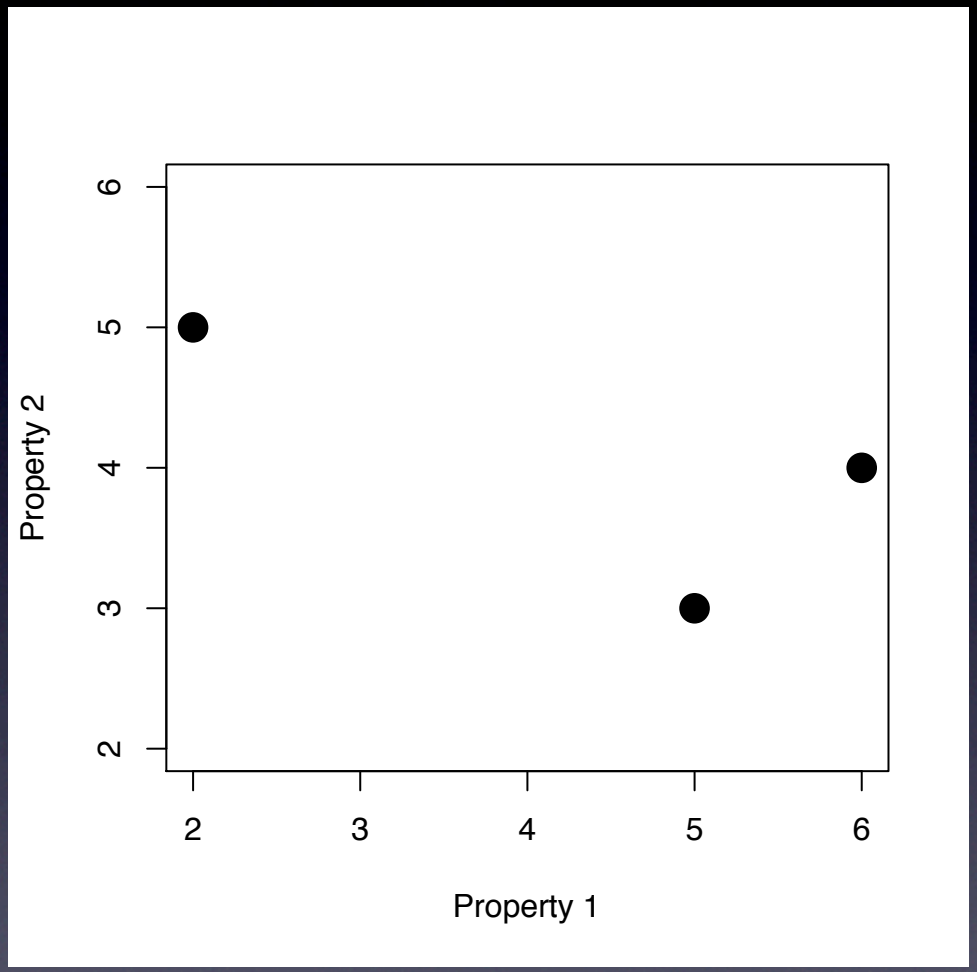
Schematic illustration of how hierarchy encapsulates anomaly or innovation

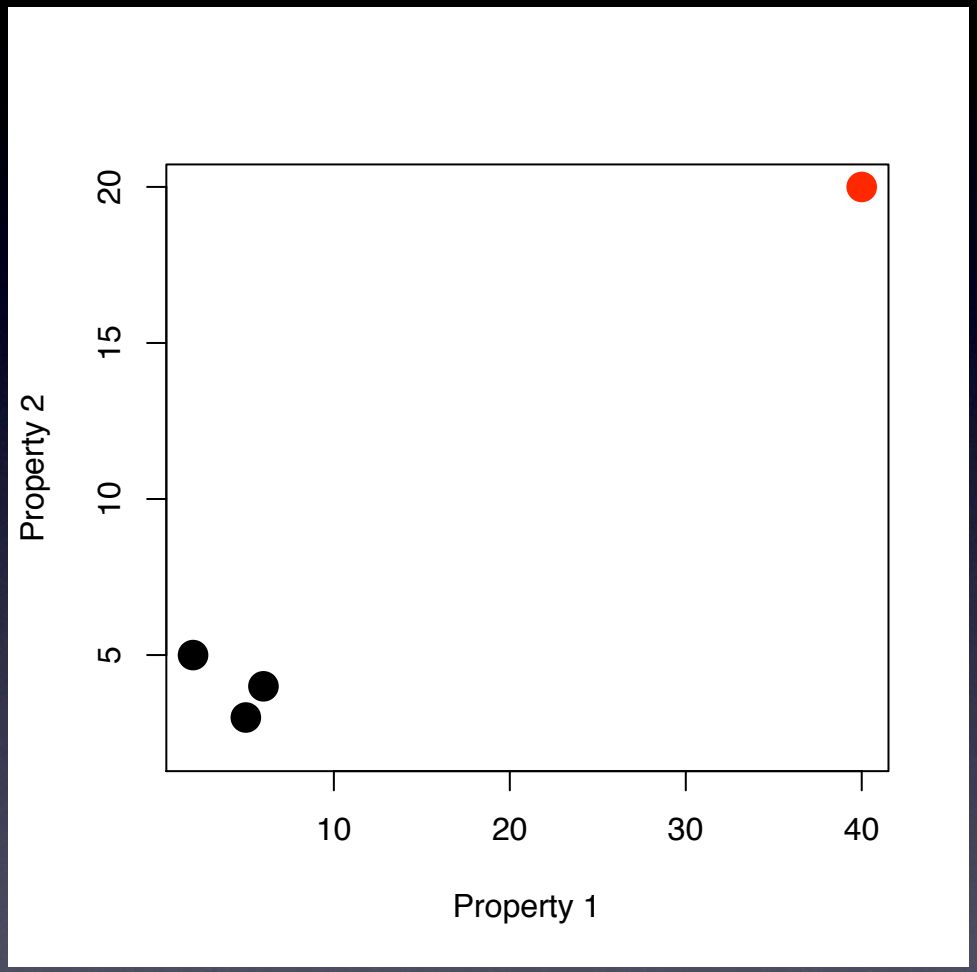


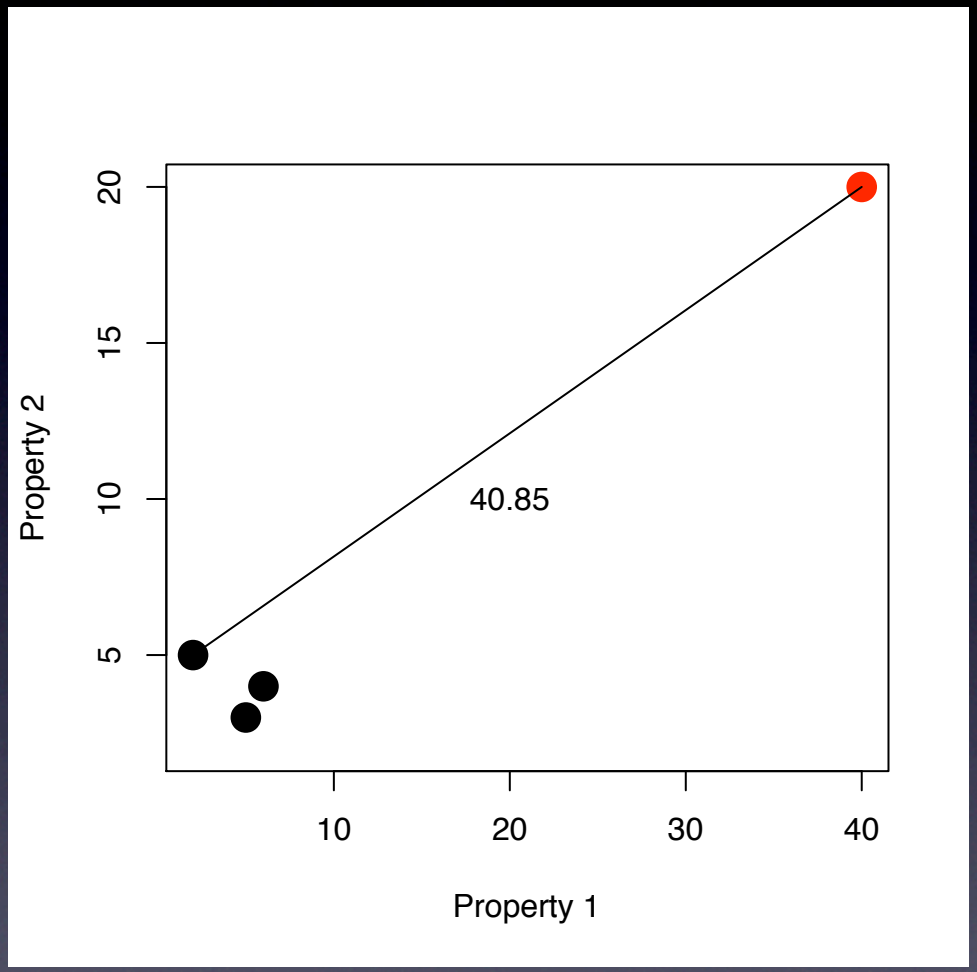

Property 2
6
5
4
3
2
2
3
4
5
6
Property 1

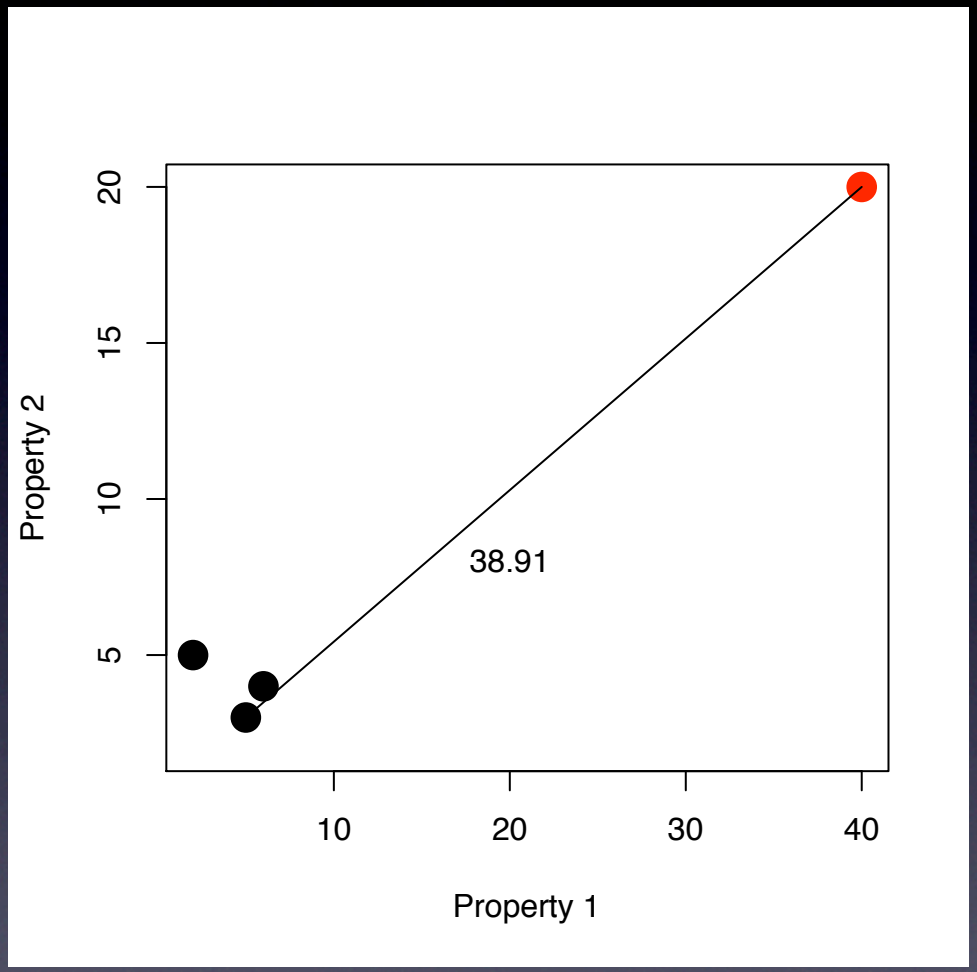

Property 2
6
5
4
3
2
2
3
4
5
6
Property 1

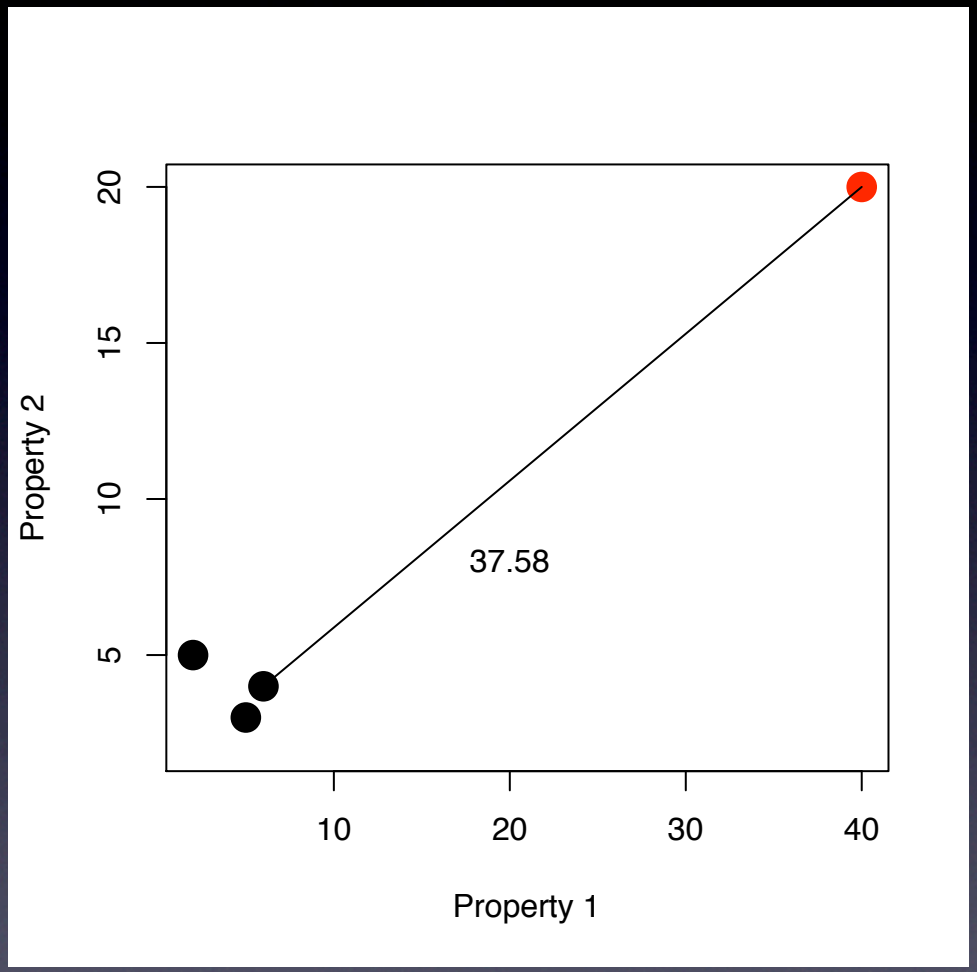

Property 2
6
5
4
3
2
2
3
4
5
6
Property 1


Property 2
6
5
4
3
2
2
3
4
5
6
Property 1


Property 2
5
10
15
20
10
20
30
40
Property 1


Property 2
5
10
15
20
40.85
10
20
30
40
Property 1


Property 2
5
10
15
20
38.91
10
20
30
40
Property 1


Property 2
5
10
15
20
37.58
10
20
30
40
Property 1


Isosceles triangle:
approx equal long sides
Property 2
Property 1
5
10
15
20
10
20
30
40

The Role of Metric and Ultrametric Embeddings in Change Detection and the Finding of Anomaly and Innovation

- Benois-Pineau and Khrennikov, "Significance delta reasoning with p-adic neural networks. Application to change detection in video", Computer Journal, forthcoming, 2007: points to importance of p-adic viewpoint for change detection in data streams.
- Chafe, "The flow of thought and the flow of language", in Syntax and Semantics: Discourse and Syntax, Ed. T. Givón, Academic Press, 1979: uses written or spoken language to map out human thinking in both its linear and hierarchical levels. Latter: hierarchically structured units - memory/story; episode/paragraph; thought/sentence; and focus/phrase.

Next: case study of sequence of texts

- Firstly, texts are scored on specified keywords
- Secondly, alternative: we take free text

A film script offers an excellent analysis framework

- Possible uses: (i) as aid in developing interactive games; (ii) similarly for use in interactive and/or immersive training and learning environments; and (iii) for support of interactive television and virtual digital environments
- Large numbers of film (or television) scripts are publicly available
- Film scripts are semi-structured, subdivided into scenes. Each scene has some metadata, e.g. location, NIGHT/DAY, INTERIOR/EXTERIOR, character names, etc.

Interaction and decision making - Casablanca (1942)

- Script half completed when production began
- Dialog for some scenes written while shooting in progress
- Joint work with Adam Ganz and Stewart McKie, Dept. of Media Arts, RHUL

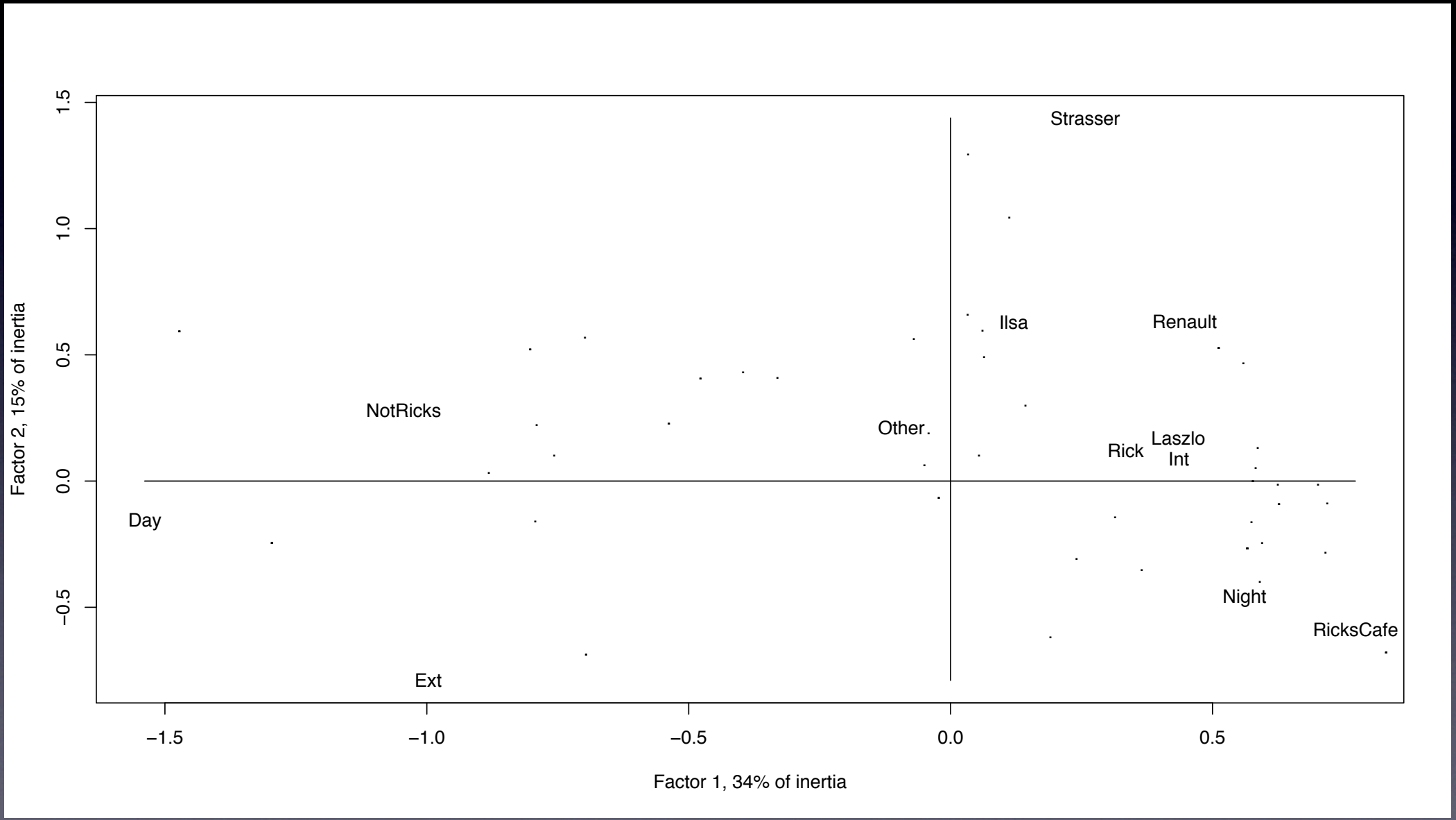
Interaction and decision making - Casablanca (1942)

- Having the story, specified by the film script, how can we make of it an interactive drama?
- Rather than a games approach involving the writing of many alternative trajectories, we seek to leverage popularity of a given script, and to reverse engineer it to support interactivity.
- Consider 77 scenes in Casablanca and seek those with major change. The nodal points, or scene of significant change, are those that best permit branching.
- We use the filmscript.

Interaction and decision making - Casablanca (1942)

- A first data set had 77 successive scenes crossed by attributes - Int, Ext, Day, Night, Rick, Ilsa, Renault, Strasser, Laszlo, Other (i.e. minor character), and 29 locations.
- Many locations were met with just once; Rick's Café was the location of 36 scenes. (We did not distinguish between "Main room", "Office", "Balcony", etc.)

12 attributes displayed; 77 scenes displayed as dots

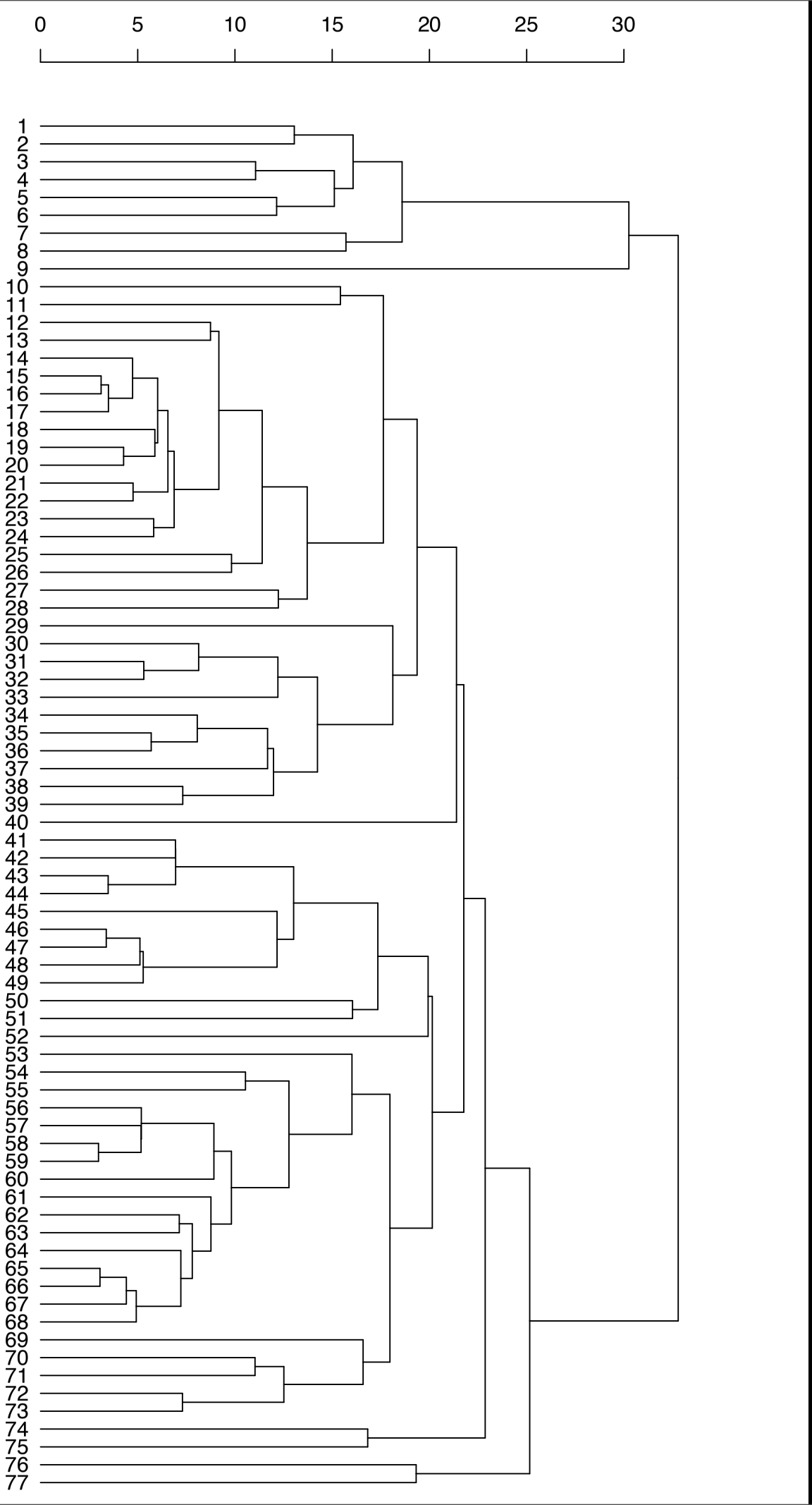


Interaction and decision making - Casablanca (1942)

- A second data set used frequency of occurrence of (in all) 6710 words in the 77 scenes. Metadata and dialog included. Scenes varied between just 5 words, and 1017 words. Lower case used throughout.
- CA is not as interpretable with such data (successive % inertia explained: 3.3, 3.0, 2.9, 2.7, ...), since the many dimensions are not easily visualized in e.g. 2 dimensions.

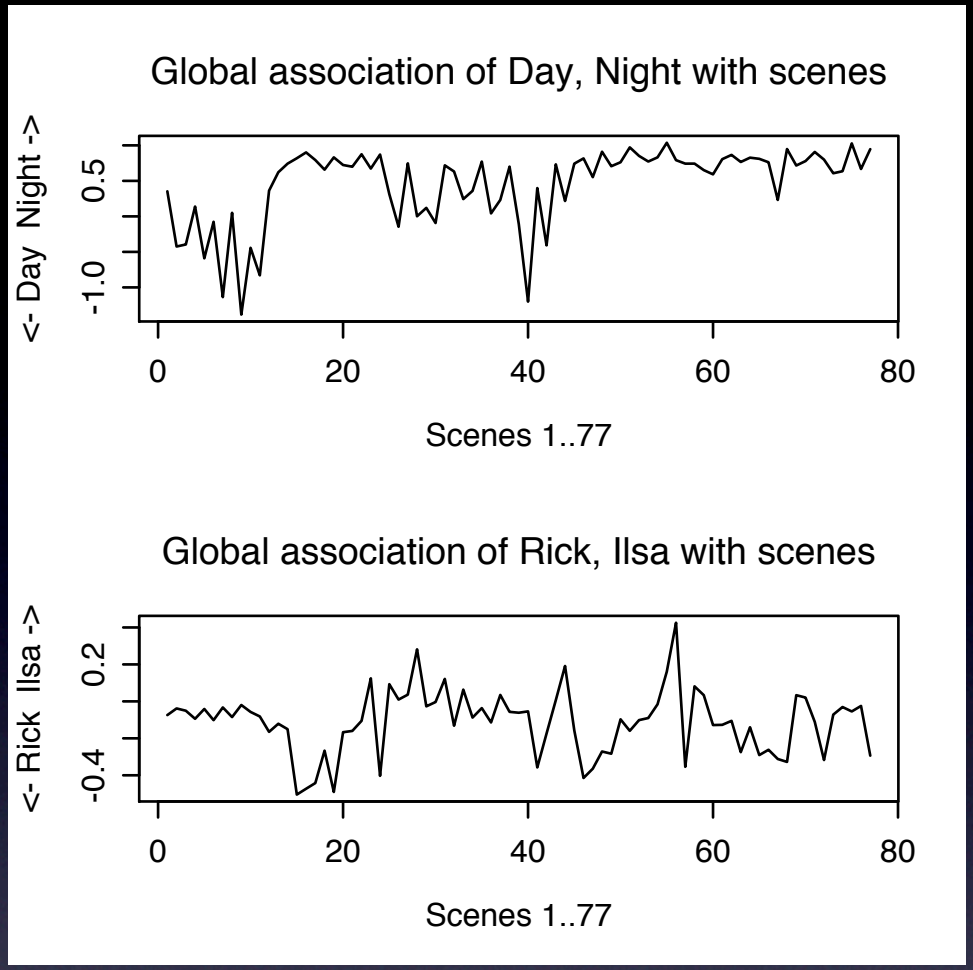
Interaction and decision making - Casablanca (1942)

- A better visualization is provided by the clustering of the scenes which additionally takes the sequence of scenes into account. So we have a contiguity or adjacency relationship between scenes.
- Single link hierarchical clustering on this contiguity graph; or the complete link hierarchical clustering such that agglomerands share at least one contiguous pair, can be shown to be free of inversions. We use the contiguity-constrained complete link hierarchical clustering method.
- For the clustering, we use the factor projections, i.e. the full-dimensionality Euclidean embedding. The clustering therefore uses Euclidean distance. As for any hierarchical clustering, it embeds the metric-endowed data in an ultrametric topology.


0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77

- We have the 77 scenes, and the 6710 words, embedded in the same 76-dimensional factor space, endowed with the Euclidean metric.
- So we can easily choose any word and proceed to determine its distance to any scene, or to clusters of scenes.
- We find very similar outcomes for characters Rick, Ilsa, Laszlo, ... since the distant scenes are bridging or context or extremely short scenes.

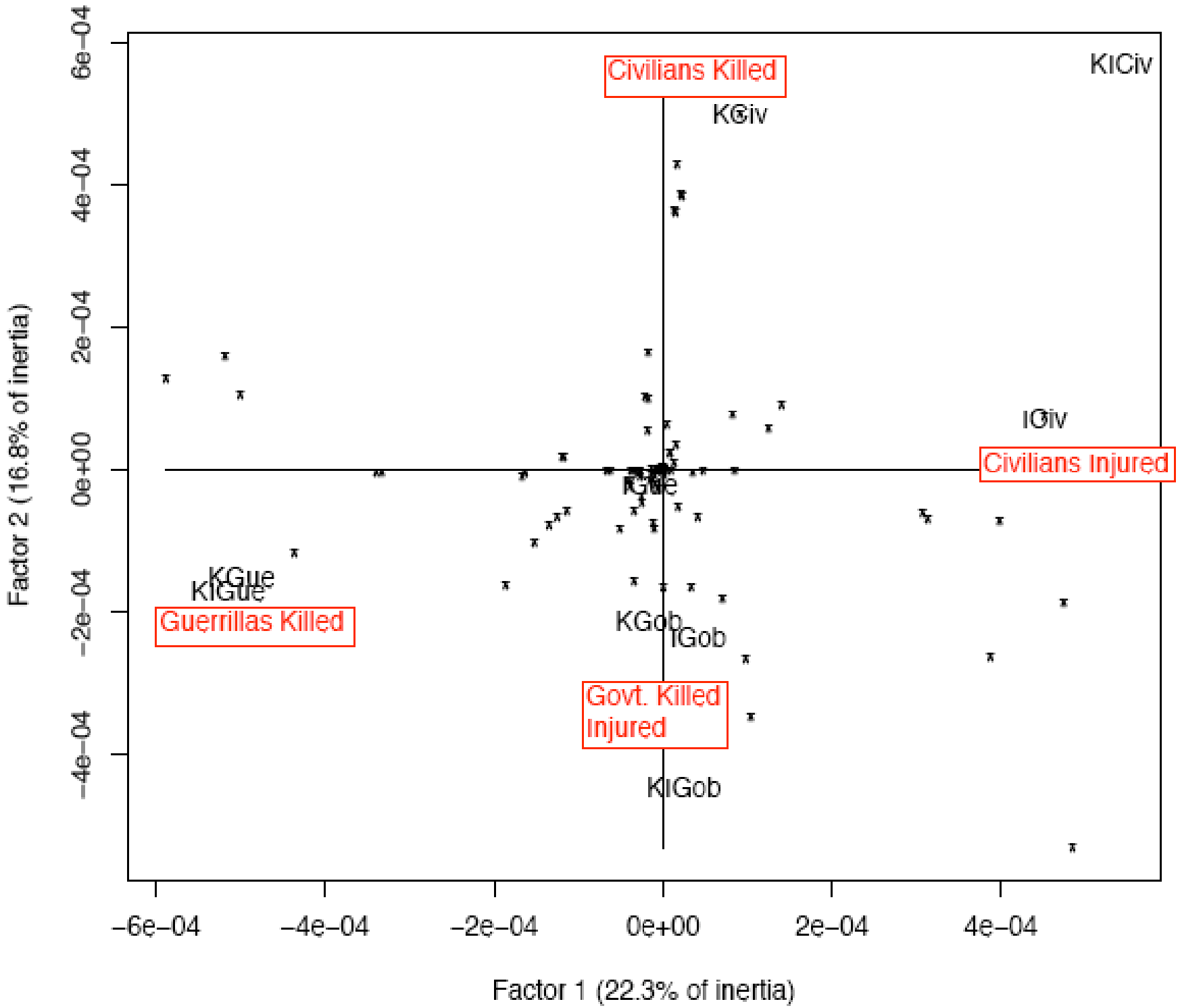
More useful analysis: differences between a word and all 77 scenes



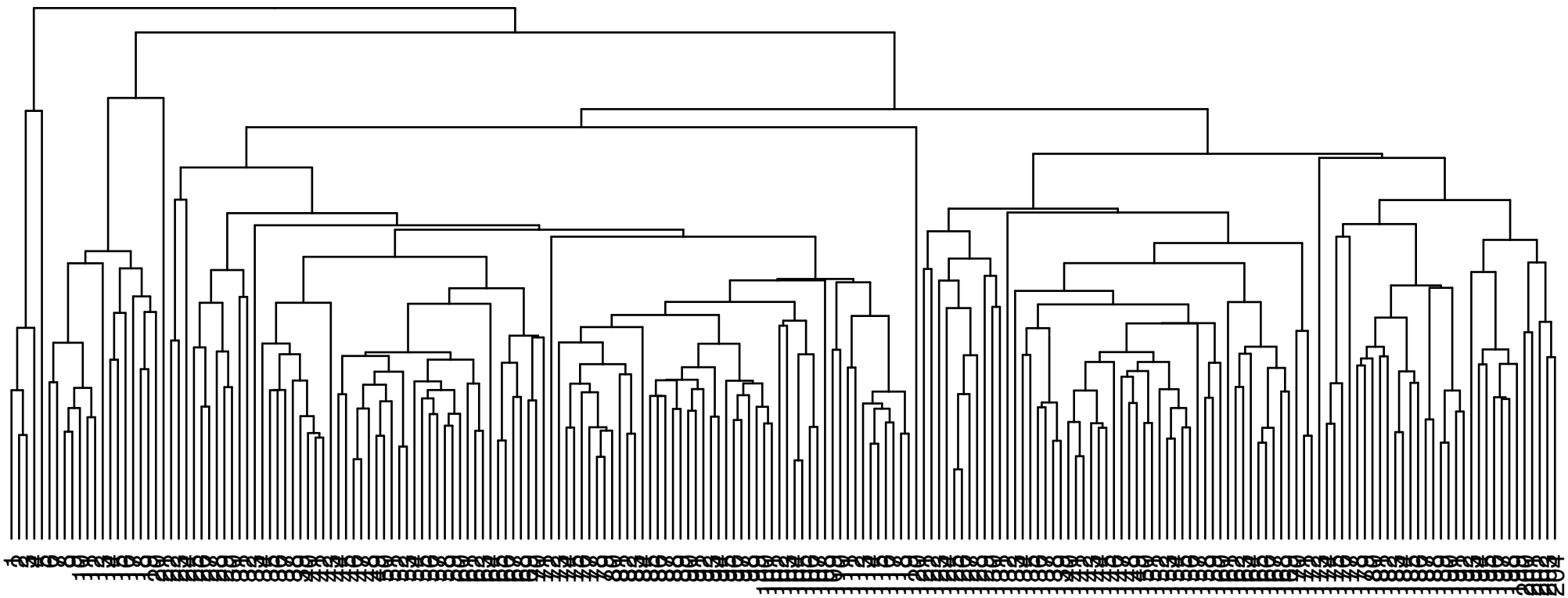
- But any word has a roughly similar set of distances to the scenes - so we look at the relativities between words
- Scenes 9 (overhead shot of plane) and 40 (blue parrot outside café) are strongly linked to Day; majority of scenes are linked to night

Interaction and decision making - Colombian civil conflict

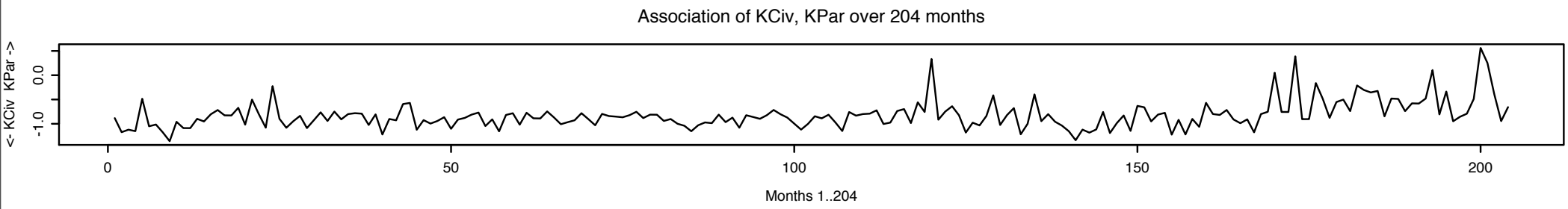
- > 20,000 events, approx 250 attributes
- 1988-2004
- > 3000 casualties per year
- Joint work with Michael Spagat, Dept. of Economics, RHUL

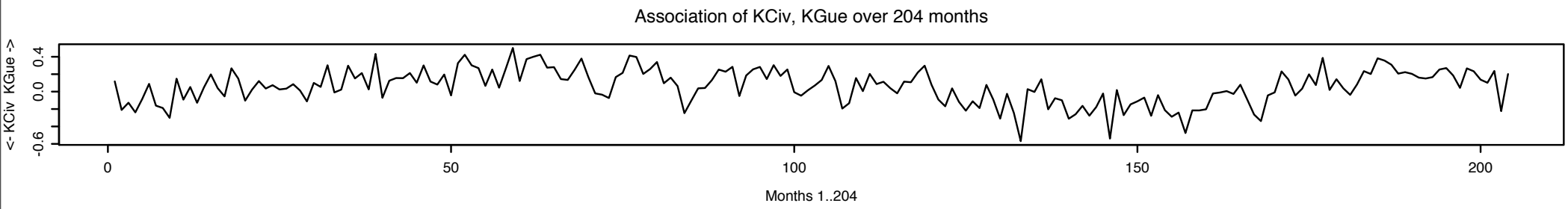

Civilians Killed
KIGiv
KGiv
IGiv
Civilians Injured
KGue
KIGue
KGue
Guerrillas Killed
KGob
IGob
Govt. Killed
Injured
KIGob
6e−04
4e−04
2e−04
0e+00
−2e−04
−4e−04
Factor 2 (16.8% of inertia)
−6e−04
−4e−04
−2e−04
0e+00
2e−04
4e−04
Factor 1 (22.3% of inertia)

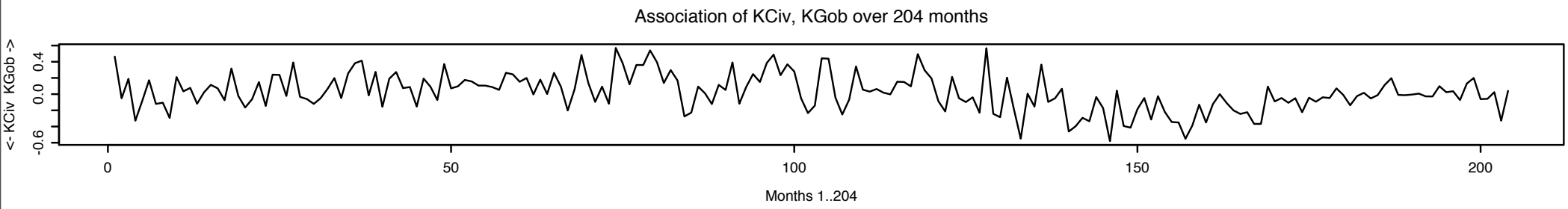
20288 events, aggregated in 204 successive months.
144 numerical features.
Inherent dimensionality found to be 80.
Displayed is a sequence-constrained complete link hierarchy.



Killings of civilians (KCiv) relative to those of: KPar (paramilitaries), KGue (guerillas), KGob (government forces), by successive month over 204 months







Conclusions

- High dimensional spaces endowed with a scalar product are naturally ultrametric
- Hierarchy provides a powerful means of studying anomaly, or innovation, or change