Lecture 2, 3rd Part

# Ultrametric Embedding: Application to Data Fingerprinting

•Consider “hierarchical structure” whenever “ultrametricity” is mentioned.

• Now, clustering is often the search for compact groups. But certainly not always…

• For us, hierarchical structure is targeted. Embedded subsets. Furthermore: local structure.

• Hierarchies are often represented by trees. We use binary rooted trees, termed dendrograms. Such a hierarchy defines an ultrametric topology. There is a close relationship between an ultrametric topology and a p-adic number system (i.e. base p, where p is a prime). (For this, see Lecture 3.)



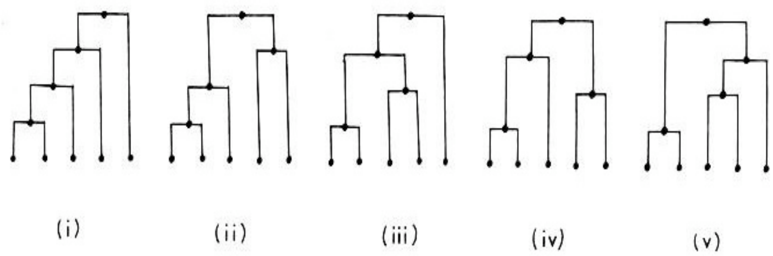


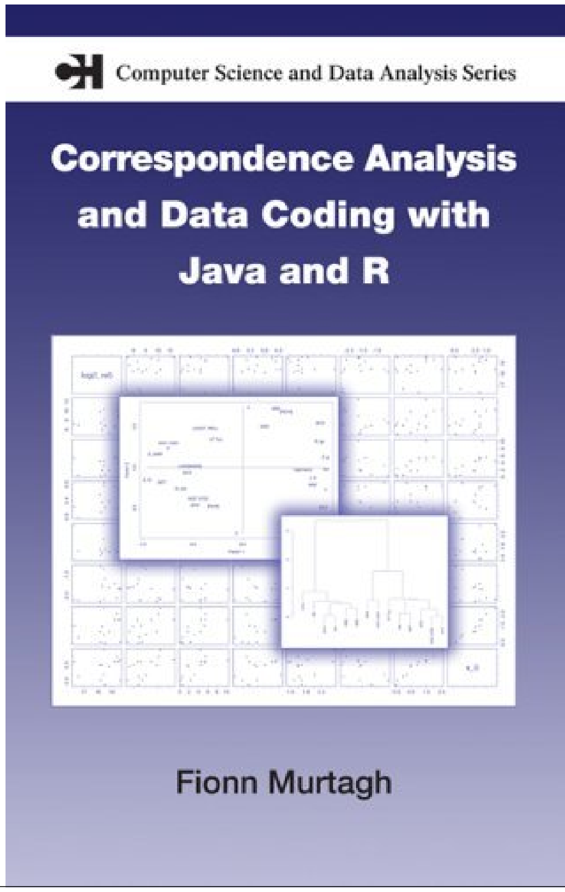

Fig. 2. Five dendrograms on $n = 5$.

- Remark (for data analysts) on the methodology here:
- We do not wish to fit a dendrogram to a data set.
- We want to see if a data set is inherently hierarchical - if so, [most] agglomerative hierarchical clustering criteria will give the same result.
- We do this by looking for local hierarchical structure.

# Some Properties of Ultrametrics

- The distance between two objects -- or two terminals in the tree -- is the lowest rank which dominates them. Lowest or closest common ancestor distance.
- The ultrametric inequality holds for any 3 points (or terminals):
- $d(i, k) \leq \max \{d(i,j), d(j,k)\}$
- Recall: the triangular inequality is: $d(i,k) \leq \{d(i,j) + d(j,k)\}$
- An ultrametric space is weird: (i) all triangles are isosceles with small base, or equilateral; (ii) every point in a ball is its center; (iii) the radius of a ball equals the diameter; (iv) a ball is clopen; (v) an ultrametric space is always topologically 0-dimensional. Etc.





# Data recoding can enhance inherent hierarchical structure

- One early motivation for this work: What is the benefit of data encoding as used in Correspondence Analysis? One answer: it tends to bring about greater ultrametricity in our data.
- Fisher iris data, 150 x 4. We quantify ultrametricity -- inherent hierarchical structure in a way to be described shortly -- and arrive at a value of 0.017 (on a scale of 0 = no ultrametricity, 1 = 100% ultrametricity.
- Now we recode the iris data to 0 and 1 values, furnishing a 150 x 150 array. Actually some columns are all 0-valued, so we remove them, leaving a 150 x 123 array. The ultrametricity now is 0.948.



# Two major implications...

- Data coding is often so far upstream of data analysis that it is just taken for granted.
- Major domain of application: high dimensional data analysis - search for invariants and symmetries.
- Note: high dimensional problems are (very) closely linked to small sample size problems.

## Quantifying ultrametricity – I

- Assume Hilbert space. Consider a triplet of points, that defines a triangle.
- Take smallest internal angle, a, in triangle ≤ 60 deg.
- … and, for the two other internal angles, b and c, if | b – c | < 2 deg. (arbitrary small angle),
- Then this triangle is ultrametric.
- We look for the overall proportion of such triangles in our data.



## Quantifying ultrametricity – II

- So: we take all possible triplets, i, j, k
- We look at their angles, and judge whether or not the ultrametric triangle properties are verified
- If so: #UM-triangles++
- Having examined all possible triangles, our $\alpha$ measure is: #UM-triangles / #triangles
- All triangles respect these ultrametric properties implies $\alpha$ = 1; no triangle does, then = 0
- For n objects, this is computationally prohibitive, so we sample i,j,k in practice (uniformly)

# Other Ways of Quantifying Ultrametricity – III

- Relationship between subdominant ultrametric, and given dissimilarities.
- Rammal, Toulouse and Virasoro, Ultrametricity for physicists, Rev. Mod. Phys., 58, 765-788, 1986.
- Whether interval between median and max rank dissimilarity of every set of triplets is nearly empty. (Taking ranks provides scale invariance.) We will look at Lerman's measure later.
- Lerman, Classification et Analyse Ordinale des Données, Dunod, 1981.



# Pervasive Ultrametricity

- As dimensionality increases, so does ultrametricity.
- In very high dimensional spaces, the ultrametricity approaches being 100%.
- Relative density is important: high dimensional and spatially sparse mean the same in this context.
- We find equilateral polygons which can be analyzed through equivalence classes defined by level sets.

- See: F Murtagh, "On ultrametricity, data coding, and computation", Journal of Classification, 21, 167-184, 2204
- Hall, P., Marron, J.S., and Neeman, A., "Geometric representation of high dimension low sample size data", JRSS B, 67, 427-444, 2005
- F. Delon, Espaces ultramétriques, J. Symbolic Logic, 49, 405-502, 1984

# Fingerprinting Using Ultrametricity

1) Wide range of time series signals
2) Wide range of texts



# Assessing the ultrametricity of time series - I

- Fingerprint the time series signals based on their ultrametricity.
- Approach used: Take "sliding window" of fixed length. Used "window" sizes m = 5, 10, 15, … , 105, 110. Look at distance between each pair of values in the window. Encode as high/low distance. Test ultrametricity of all these indicators of local variability, and accumulate ultrametricity index over all such "windows".
- In "window" code each value as 1 if there is no/small change; and 2 if there is large change (up or down). Small/large defined relative to threshold $\max_{jj'} d_{jj'}^2/2$, $j,j' \in$ "window". Recoded values are metric.

# Ultrametricity of time series - II

- So in a local region (window) we map pairwise dissimilarities onto relative (i.e. local) “change = 2” versus “no change = 1” distance.
- This is our “change/no change” metric.

- Used signals: FTSE, USD/EUR, sunspot, stock, futures, eyegaze, Mississippi, www traffic, EEG/sleep/normal, EEG/petit mal epilepsy, EEG/irreg. epilepsy, quadratic chaotic map, uniform.
- Signals can be clearly distinguished. Extremes are: EEG and uniform.



| | | |
|---|---|---|
| 1 | FTSE | 1326 |
| | FTSE – Financial Tmes Stock Exchange index | |
| 2 | USD/EUR | 1169 |
| | USD/EUR daily foreign exchange rates | |
| 3 | Sunspot | 2739 |
| | Monthly index values of sunspot solar physics activity | |
| 4 | Stock | 1374 |
| | Stock price, unknown origin | |
| 5 | Futures-3080 | 3080 |
| | First 3080 values of futures | |
| 6 | Futures | 6160 |
| | Futures, daily highs | |
| 7 | Eyegaze | 1471 |
| | One coordinate of eyegaze position from eye tracker | |
| 8 | Mississippi-20000 | 20,000 |
| | First 20,000 values of Mississippi data | |
| 9 | Mississippi | 43,829 |
| | Mississippi River daily water levels | |
| 10 | WWW traffic | 34,726 |
| | Bytes transferred per hour by a web server | |
| 11 | EEG-chan4 | 2500 |
| | EEG channel p4, sampled at 250 Hz for 10 seconds | |
| 12 | EEG-chan5 | 2500 |
| | EEG channel o1, sampled at 250 Hz for 10 seconds | |
| 13 | Quadratic map 1 | 2500 |
| | $x_{t+1} = 4x_t(1 - x_t)$, $x_0 = 0.2$ | |
| 14 | Quadratic map 2 | 2500 |
| | $x_{t+1} = 4x_t(1 - x_t)$, $x_0 = 0.37777$ | |
| 15 | Quadratic map 3 | 2500 |
| | $x_{t+1} = 4x_t(1 - x_t)$, $x_0 = 0.451$ | |
| 16 | Sleep EEG chan. 1 | 999 |

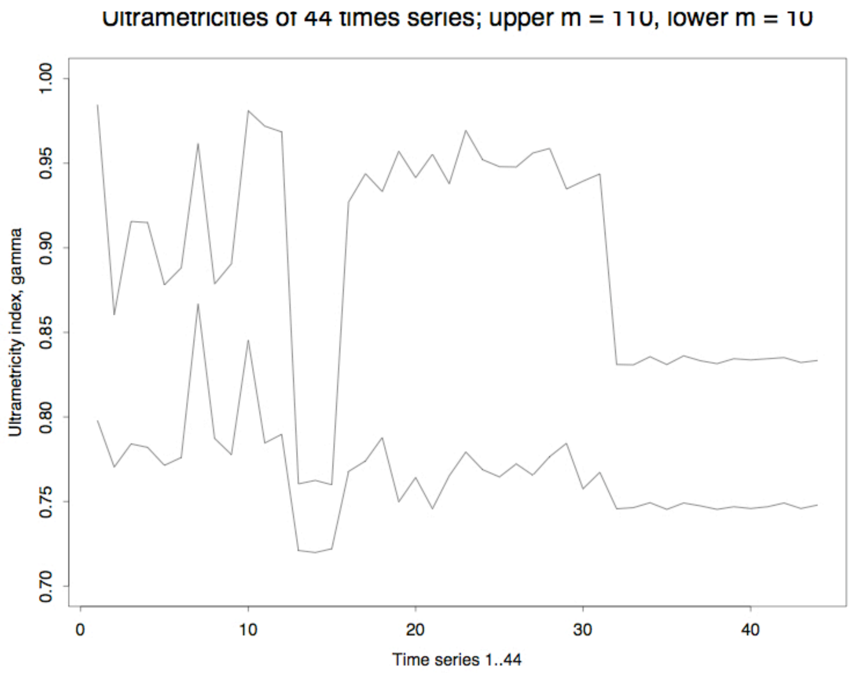



Fig. 3. Investigation of two of the windows (embedding dimensions), $m = 10$ and $m = 110$. Results for 44 time series are shown, with window size $m = 110$ on top and $m = 10$ on bottom. In both cases, an ultrametricity $\gamma$ value is plotted for each time series. Portraying the $\gamma$ values as a continuous curve for all data sets is done for visualization.

# Assessing the ultrametricity of text

- Semantic networks defined from texts, through shared words.
- Used as texts: 209 tales of Brothers Grimm; 266 Jane Austen chapters (full/partial) from 3 novels from 1811, 1813, 1817; 50 air accident reports; 384 dream reports. In all: nearly 1000 texts, over 1 million words.
- Using Benzécri ("bag of words") approach, use words as found (no stemming). Define $\chi^2$ distance between profiles of frequency of occurrence table.
- We "euclideanized" by mapping into correspondence analysis factor space. E.g. for dream reports, 384 texts crossed by 11,441 words.
- Then we determined ultrametricity of text collections in factor space.
- We found dream reports to be highest in ultrametricity (albeit with fairly small coefficient of ultrametricity); and air accident reports similar to Grimm texts.
- Other assessments were carried out on Aristotle's Categories; and James Joyce's Ulysses (304,414 words).

## Ultrametricity (i.e. hierarchical substructure) for various text collections

- 209 Grimm Brothers tales, 209 x 7443, ultrametricity coefficient 0.1147
- 266 Jane Austen chapters or partial chapters, 266 x 9723, ultrametricity coefficient 0.1404
- 50 aviation accident reports, 50 x 4261, ultrametricity coefficient 0.1154
- 385 dream reports, 385 x 11441, ultrametricity coefficient 0.1933
- 171 Barbara Sanders dream reports, 171 x 7044, ultrametricity coefficient 0.2603



## Results quite consistent: Example of Brothers Grimm

| 209 Brothers Grimm fairy tales | | | | |
|---|---|---|---|---|
| Texts | Orig.Dim. | FactorDim. | Alpha, mean | Alpha, sdev. |
| 209 | 1000 | 208 | 0.1236 | 0.0054 |
| 209 | 2000 | 208 | 0.1123 | 0.0065 |
| 209 | 7443 | 208 | 0.1147 | 0.0066 |

# Applications of local ultrametricity

- Application 1 - To characterize the data set
- Application 2 - To help in proximity and related search problems

- Application 1 - This leads to what?
- It serves to determine the data generation process, and the phenomenon or activity represented by the data
- Application 2 - Lecture 3



# Lerman's H-classifiability

- Quantifies how ultrametric a given metric is; useful because it is based on rank orders - so avoids messiness of handling RA, Dec, redshift coordinates.
- Let M(x,y,z) be median pair among {(x,y), (y,z), (x,z)}; and let S(x,y,z) be highest ranked pair in this triplet. J is the set of all possible triplets.
- We consider the open interval ]M(x,y,z), S(x,y,z)[
- If triplet {x,y,z} is such that (x,y) ≤ (y,z) ≤ (x,z) for the preorder defined by the distance used, then the preorder is ultrametric if the interval ]M(x,y,z), S(x,y,z)[ is empty.
- Lerman's approach is based on counting how often this interval is found to be empty. 0 if ultrametric, 1 if very non-ultrametric. (Note: my triangle-based measure was 1 for ultrametric, and 0 for non-ultrametric.)