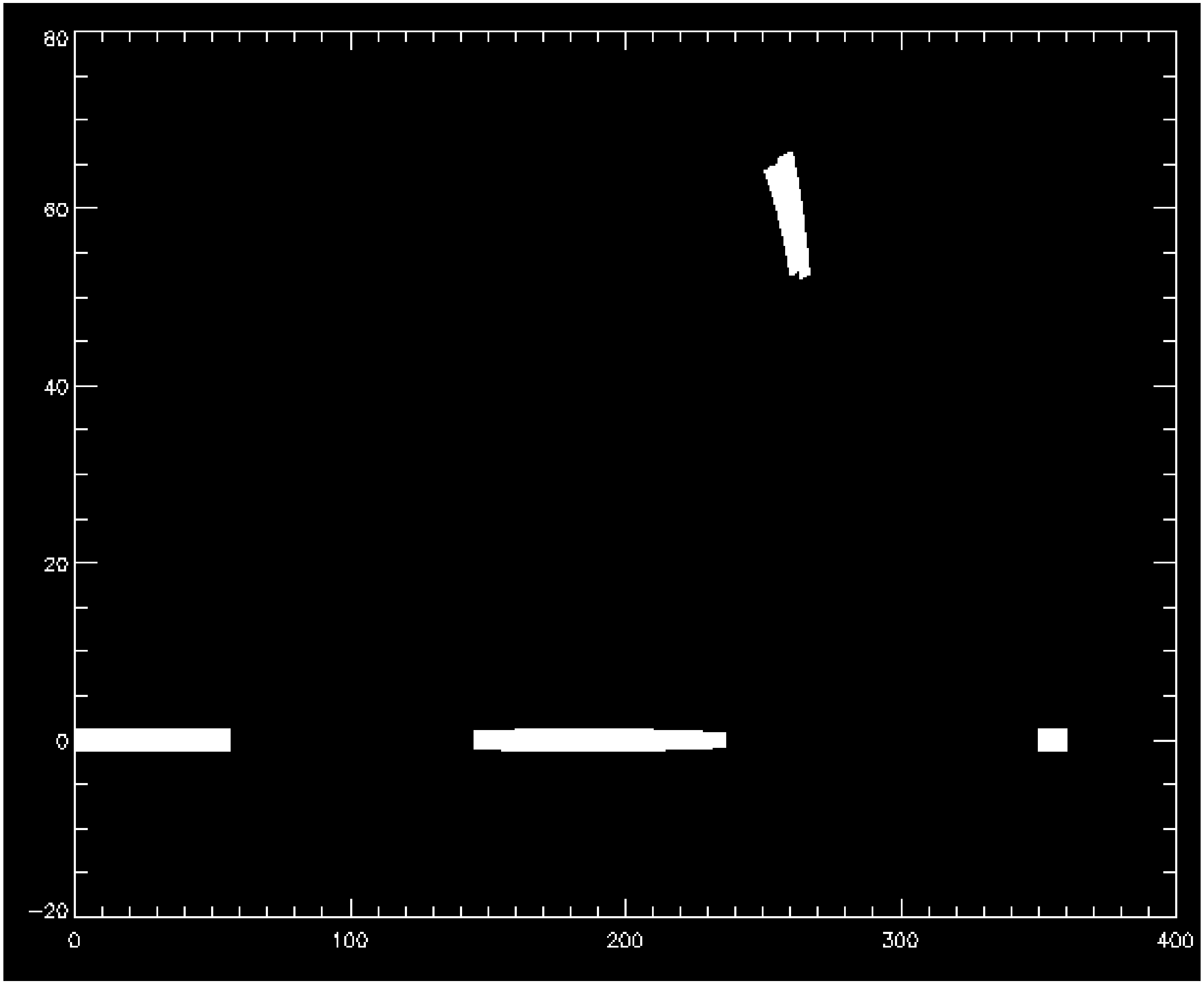
Hierarchic Clustering of 3D Galaxy Distributions

Topics:

- Data
- Hierarchic clustering
- Ultrametric topology
- P-adic algebra
- Practical interest
- Testing for ultrametricity
- Lerman’s H-classifiability
- Conclusion and critique

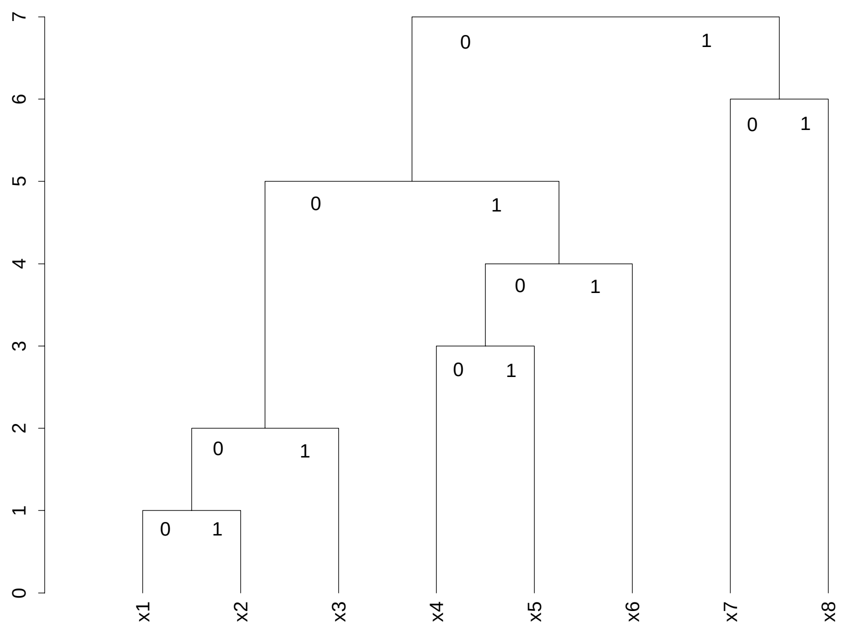
Data

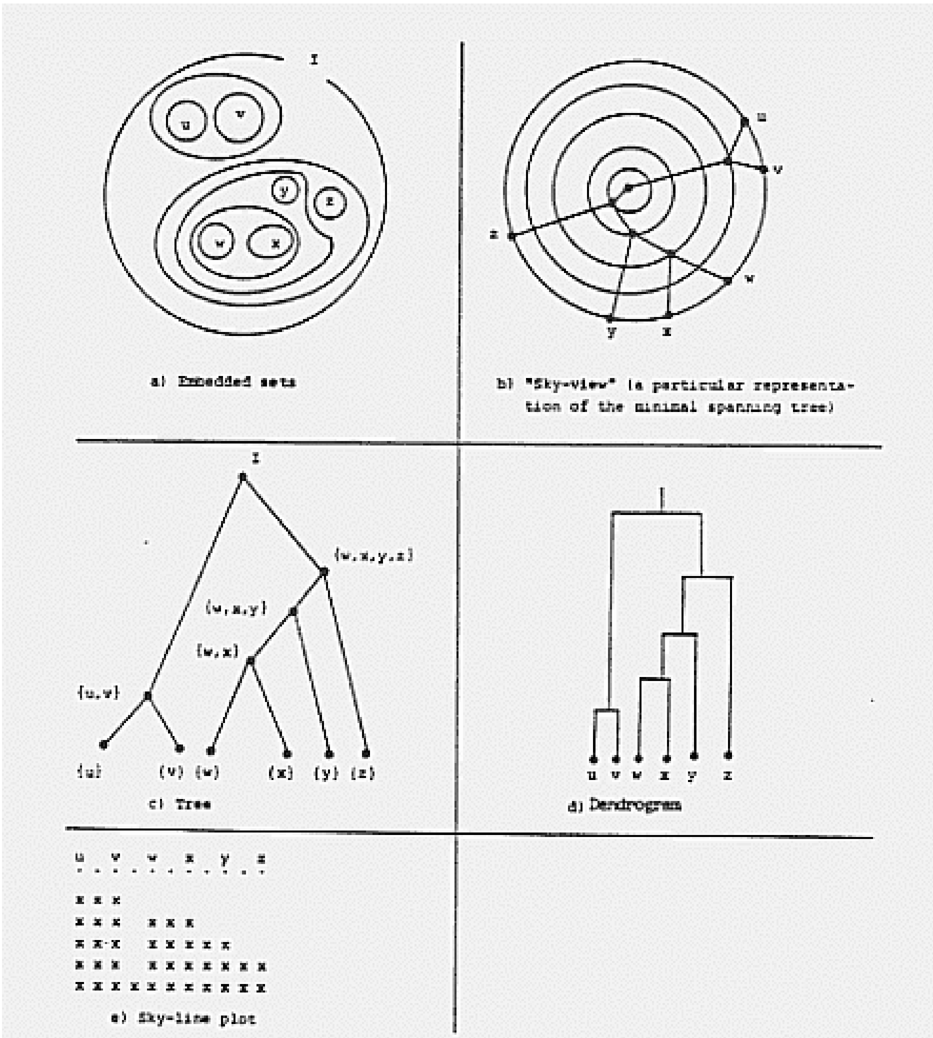
- Sloan Digital Sky Survey data
- RA, Dec, redshift value, reliability indicator
- 345109 galaxies in right ascension and declination, photometric redshift
- In this work we used the low RA, galaxy plane area.


80
60
40
20
0
−20
0
100
200
300
400

Hierarchic Clustering

Labeled, ranked dendrogram on 8 terminal nodes. Branches labeled 0 and 1.




a) Embedded sets
b) "Sky-view" (a particular representation of the minimal spanning tree)
c) Tree
d) Dendrogram
e) Sky-line plot

Hierarchic Clustering: Metric $\Longrightarrow$ Ultrametric

- Hierarchical agglomeration on n observation vectors, $i \in I$
- Series of $1, 2, \ldots, n-1$ pairwise agglomerations of observations or clusters
- Hierarchy $H = \{q | q \in 2^I\}$ such that (i) $I \in H$, (ii) $i \in H \ \forall i$, and (iii) for each $q \in H, q' \in H : q \cap q' \neq \emptyset \Longrightarrow q \subset q'$ or $q' \subset q$.
- Indexed hierarchy is the pair (H, ν) where the positive function defined on H, i.e., $\nu : H \rightarrow \mathbb{R}^+$, satisfies: $\nu(i) = 0$ if $i \in H$ is a singleton; and (ii) $q \subset q' \Longrightarrow \nu(q) < \nu(q')$. Function ν is the agglomeration level.
- Take $q \subset q'$, let $q \subset q''$ and $q' \subset q''$, and let q'' be the lowest level cluster for which this is true. Then if we define $D(q, q') = \nu(q'')$, D is an ultrametric.

Ultrametric Spaces and Properties

- Let (E, d) be a metric space, i.e. a set E and a positive function $E \times E \longrightarrow \mathbb{R}_+$ satisfying
  1. $d(x, y) = d(y, x)$
  2. $d(x, y) = 0$ iff $x = y$
  3. $d(x, z) \leq d(x, y) + d(y, z)$

  A space is ultrametric if in addition we have $d(x, z) \leq \max(d(x, y), d(y, z))$
- A metric space (E, d) is ultrametric iff all its triangles are isosceles, with the length of the base being less than or equal to that of the sides.
- Each point of a circle in E is its center. Each ball in an ultrametric space is both open and closed.
- Two non-disjoint balls are concentric.

P-adic Coding

- For the dendrogram shown in we develop the following p-adic encoding for $p = 2$ of terminal nodes, traversing a path from the root.
- $x_1 = 0 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + 0 \cdot 2^1$;
- $x_2 = 0 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + 1 \cdot 2^1$;
- $x_4 = 0 \cdot 2^7 + 1 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3$;
- $x_6 = 0 \cdot 2^7 + 1 \cdot 2^5 + 1 \cdot 2^4$.
- The decimal equivalents of this p-adic representation of terminal nodes work out as $x_1, x_2, \ldots x_8 = 0, 2, 4, 32, 40, 48, 128, 192$.
- A p-adic encoding for x_i is given by $\sum_1^{n-1} a_k p_k$ where $a_k \in \{0, 1\}$ and $p_k = 2^k$.

P-adic (Algebraic) = Ultrametric (Topology)

- Various terms are used interchangeably for analysis in and over such fields such as p-adic, ultrametric, non-Archimedean, and isosceles.
- The natural geometric ordering of metric valuations is on the real line, whereas in the ultrametric case the natural ordering is a hierarchical tree.
- Ostrowski's theorem: Each non-trivial valuation on the field of the rational numbers is equivalent either to the absolute value function or to some p-adic valuation
- Alternatively: Up to equivalence, the only norms on the rationals are the p-adic norm and the usual norm given by the absolute value.

Practical Interest of Ultrametricity

- Hierarchies arise naturally in language syntax, and (it has been claimed) in financial markets.
- Rammal et al.: Ultrametricity is a natural property of high-dimensional spaces, and ultrametricity emerges as a consequence of randomness and of the law of large numbers.
- Again Rammal et al. and recent work of ours: Sparsely coded data tend to be ultrametric. Examples include: the use of complete disjunctive forms of coding in correspondence analysis; and categorical data coding in genomics and proteomics, speech, and other fields.
- Ultrametricity is considered to hold at low Planck scales, and in superstrings (Brekke and Freund, Phys. Rep., 233, 1–66, 1993).
- Also to be valid for optimization spaces.

Testing for Ultrametricity

- Rammal et al.: determine the subdominant ultrametric (aka single link hierarchic clustering).
- Interesting phase space effects for increase in dimensionality.
- However the subdominant ultrametric gives rise to pathologies.
- E.g. “friends of friends” chaining effect: $d(x, y) \leq r_0, d(y, z) \leq r_0$ then $d(x, z) = 2r_0 - \epsilon$ for arbitrarily small ϵ. Hence $d(x, z)$ can be anomalously large.

Lerman’s H-classifiabilty

- A basic unifying framework for pairs of objects, and the distance valuation on them, is that of a *binary relation*.
- On a set E, a binary relation is a *preorder* if it is reflexive and transitive;
- it is an *equivalence relation* if the binary relation is reflexive, transitive and symmetric;
- and it is an *order* if the binary relation is reflexive, transitive, and anti-symmetric.

Lerman's H-classifiabilty

- Let F denote the set of pairs of distinct units in E. A distance defines a total preorder on F:

$$\forall \{(x, y), (z, t)\} \in F : (x, y) \leq (z, t) \Longleftrightarrow d(x, y) \leq d(z, t)$$

- A preorder is called ultrametric if:

$$\forall x, y, z \in E : \rho(x, y) \leq r \text{ and } \rho(y, z) \leq r \Longrightarrow \rho(x, z) \leq r$$

where r is a given integer and $\rho(x, y)$ denotes the rank of pair (x, y) for $\bar{\omega}$.

- A necessary and sufficient condition for a distance on E to be ultrametric is that the associated preorder (on $E \times E$, or alternatively preordonnance on E) is ultrametric.

Lerman's H-classifiabilty

- We move on now to define Lerman's H-classifiability index, which measures how ultrametric a given metric is.
- Let $M(x, y, z)$ be the median pair among $\{(x, y), (y, z), (x, z)\}$ and let $S(x, y, z)$ be the highest ranked pair among this triplet. J is the set of all such triplets of E.
- Mapping τ of all triplets J into the open interval of all pairs F for the given preorder ω:

$$\tau : J \longrightarrow]M(x, y, z), S(x, y, z)[$$

- Given a triplet $\{x, y, z\}$ for which $(x, y) \leq (y, z) \leq (x, z)$, for preorder ω, the interval $]M(x, y, z), S(x, y, z)[$ is empty if ω is ultrametric. Relative to such a triplet, the preorder ω is "less ultrametric" to the extent that the cardinal of $]M(x, y, z), S(x, y, z)[$, defined on ω, is large.
- $H(\omega) = \sum_J |]M(x, y, z), S(x, y, z)[|/(|F| - 3)|J|$

Lerman's H-classifiabilty

- Data sets that are "more classifiable" in an intuitive way, i.e. they contain "sporadic islands" of more dense regions of points – a prime example is Fisher's iris data contrasted with 150 uniformly distributed values in $\mathbb{R}^4$ – such data sets have a smaller value of $H(\omega)$. For Fisher's data we find $H(\omega) = 0.0899$, whereas for 150 uniformly distributed points in a 4-dimensional hypercube, we find $H(\omega) = 0.1835$.
- Extensive tests carried out have shown that uniform data has values around 0.18 – 0.21. Whereas with more sparsely coded data, etc., one finds values around 0.1 – 0.14.

Lerman’s H-classifiabilty

- We took 3D cylanders defined by RA and Dec within a tight radius of a position, to limit the number of galaxies studied at any given time to around 500.
- We used data in (lower left block in Sloan data) – low RA, near galactic plane.
- Then we used 3D uniformly distributed data to see how different the Lerman index would be.
- For Sloan data: 0.149837, 0.115096, 0.148676.
- For uniform data: 0.187662, 0.179590, 0.171903.
- Numbers in each case: 589, 554, 715.

Conclusions and Critique

- The Sloan data came out as more ultrametric in all cases, compared to uniformly distributed 3D values.
- But a Euclidean distance was used for determining the Lerman index.
- Also the cylandrical volume used in Sloan space may have biased the results (in view of the redshift value).
- Future work: replace the cylander with a cone, and study replacement for the Euclidean distance.