Lecture 4: Hierarchy and Symmetry in Data Analysis – Thinking Ultrametrically

Minicourse, First International Conference on Models of Complex Hierarchic Systems and Non-Archimedean Analysis, Cinvestav – Abacus Center, Mexico. Fionn Murtagh.

Themes

1. Applications
2. Hierarchy and Other Symmetries in Data Analysis
3. Hierarchical Clustering – Terminology
4. A Brief Introduction to p-Adic Numbers
5. Brief Discussion of p-Adic and m-Adic Numbers
6. Ultrametric Space for Representing Hierarchy
7. Ultrametric Topology
8. Some Geometrical Properties of Ultrametric Spaces
9. Clustering Through Matrix Row and Column Permutation

Themes 2/2

Applications 1/2

- In data analysis, both because of the fitting of tree structures and/or visualizations to data sets, to provide a possible way to present a range of partitions to the user, and also to provide for a genealogical model to be fit to data.
- In physics in order to take account of phenomena at very small spatial and time scales, where discreteness of structures is represented well by p-adic number systems; and also for any systems that involve movement between discrete states that are characterized by their energy levels.
- A considerable number of search and discovery algorithms developed in recent years have an interpretation or vantage point in terms of ultrametric topology.

Applications 2/2

- Ultrametric topology and closely associated p-adic number theory as used in a wide range of fields, that all share strong elements of common mathematical and computational underpinnings.
- These include data analysis, including in the "big data" world of massive and high dimensional data sets; physics at very small scales; search and discovery in general information spaces; and in logic and reasoning.

Hierarchy and Other Symmetries in Data Analysis, 1/3

- H. Weyl, *Symmetry*, Princeton University Press, 1983, makes the case for the fundamental importance of symmetry in science, engineering, architecture, art and other areas. As a "guiding principle", "Whenever you have to do with a structure-endowed entity ... try to determine its group of automorphisms, the group of those element-wise transformations which leave all structural relations undisturbed. You can expect to gain a deep insight in the constitution of [the structure-endowed entity] in this way. After that you may start to investigate symmetric configurations of elements, i.e. configurations which are invariant under a certain subgroup of the group of all automorphisms; ..."

Hierarchy and Other Symmetries in Data Analysis, 2/3

- Herbert A. Simon, Nobel Laureate in Economics, originator of "bounded rationality" and of "satisficing", believed in hierarchy as the basis of the human and social sciences, as the following quotation shows: "... my central theme is that complexity frequently takes the form of hierarchy and that hierarchic systems have some common properties independent of their specific content. Hierarchy, I shall argue, is one of the central structural schemes that the architect of complexity uses." (H.A. Simon, *The Sciences of the Artificial*, MIT Press, 1996.)

Hierarchy and Other Symmetries in Data Analysis, 3/3

- Symmetries in data, such that the data represent complex phenomena, and the symmetries provide a model for understanding these complex phenomena. Hierarchy gives rise to a rich expanse of symmetries and we will be concerned mostly with hierarchy.
- Partitioning a set of observations leads to some very simple symmetries. This is one approach to clustering and data mining. But such approaches, often based on optimization, are of less direct interest here. Instead I will pursue the theme pointed to by Simon, namely that the notion of hierarchy is fundamental for interpreting data and the complex reality which the data expresses.

Hierarchical Clustering – Terminology 1/2

- For us here, this is *unsupervised classification*. Having the "data speak for themselves". Also termed automatic classification, clustering.
- Alternative is *supervised classification*, also known as discriminant analysis or (in a general way) machine learning. Here: training set used to learn decision making (class assignment) rules. Then the test set is use to validate the machine learning, followed by generalizion.
- Methodologies include: statistical modelling; graph theory; neural networks; optimization; linear algebra for low dimension decomposition; and other approaches.
- Classifying data, i.e. observations, objects, events, phenomena, etc.

Hierarchical Clustering – Terminology 2/2

- Families of clustering, or unsupervised classification, algorithms, include: (i) array permuting and other visualization approaches; (ii) partitioning to form (discrete or overlapping) clusters through optimization, including graph-based approaches; and – of interest to us in this article – (iii) embedded clusters interrelated in a tree-based way.
- Traditionally, agglomerative clustering, as opposed to divisive. In recent years a range of algorithms have come to the fore, using spatial density, or grids, and these are often divisive.

A Brief Introduction to p-Adic Numbers

- p-Adic numbers were introduced by Kurt Hensel in 1898.
- Rational numbers, $\mathbb{Q}$, expressed as any integer divided by any other (non-zero) integer.
- $\mathbb{Q}$ is "well-behaved" and "tangible". Rationals can be used to make measurements.
- A repeating decimal like $0.33333\ldots$ is easily expressed as a rational. But π and e (base of Naperian logs) are not.
- To go further than rationals using approximation, we need to allow for "continuity" and this presupposes a topology.
- To endow the rationals with a topology, we need a completion of the field $\mathbb{Q}$ of rationals.
- To complete the field $\mathbb{Q}$ of rationals, we need Cauchy sequences and this requires a norm on $\mathbb{Q}$ (because the Cauchy sequence must converge, and a norm is the tool used to show this).

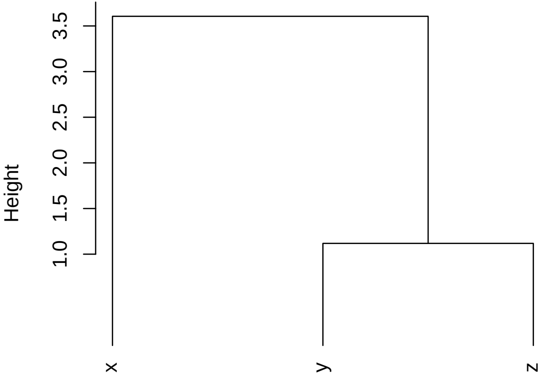
- Archimedean norm such that: $\forall\ x, y \in \mathbb{Q}$, with $|x| < |y|$, then $\exists\ N$ integer, such that $|Nx| > |y|$.
- Write: $|x|_\infty$ for this norm.
- If the completion of the rationals is Archimedean, then we have $\mathbb{R} = \mathbb{Q}_\infty$, the reals. Acceptable if space is taken as commutative and Euclidean.
- Alternatives: all norms are known. Besides the $\mathbb{Q}_\infty$ norm, we have an infinity of norms, $|x|_p$, labelled by primes, p. Ostrowski's theorem these are all the possible norms on $\mathbb{Q}$.
- So we have an unambiguous labelling, via p, of the infinite set of non-Archimedean completions of $\mathbb{Q}$ to a field endowed with a topology.
- In all cases, we obtain locally compact completions, $\mathbb{Q}_p$, of $\mathbb{Q}$. They are the fields of p-adic numbers. All these $\mathbb{Q}_p$ are continua. Being locally compact, they have additive and multiplicative Haar measures. As such we can integrate over them, such as for the reals.

Brief Discussion of p-Adic and m-Adic Numbers

- p: a prime; m: a non-zero positive integer.
- A p-adic number is such that any set of p integers which are in distinct residue classes modulo p may be used as p-adic digits.
- Recall that a ring does not allow division, while a field does. m-Adic numbers form a ring; but p-adic numbers form a field. So a priori, 10-adic numbers form a ring. This provides us with a reason for preferring p-adic over m-adic numbers.
- We can consider various p-adic expansions.
- 1. $\sum_{i=0}^{n} a_i p^i$, which defines positive integers. For a p-adic number, we require $a_i \in 0, 1, ...p - 1$. (In practice: just write the integer in binary form.)
  2. $\sum_{i=-\infty}^{n} a_i p^i$ defines rationals.
  3. $\sum_{i=k}^{\infty} a_i p^i$ where k is an integer, not necessarily positive, defines the field $\mathbb{Q}_p$ of p-adic numbers.
- $\mathbb{Q}_p$, the field of p-adic numbers, is (as seen in these definitions) the field of p-adic expansions.

- The choice of p is a practical issue. Indeed, adelic numbers use all possible values of p.
- A biotechnology example is considered as follows.
- DNA (desoxyribonucleic acid) is encoded using four nucleotides: A, adenine; G, guanine; C, cytosine; and T, thymine.
- In RNA (ribonucleic acid) T is replaced by U, uracil.
- A 5-adic encoding can be used, since 5 is a prime and thereby offers uniqueness.
- Or a 4-adic encoding can be used, or a 2-adic encoding, with the latter based on 2-digit boolean expressions for the four nucleotides (00, 01, 10, 11).
- A default norm has been used, based on a longest common prefix – with p-adic digits from the start or left of the sequence.

Ultrametric Space for Representing Hierarchy



Ultrametric Topology

1. Ultrametric topology was introduced by Marc Krasner in 1944. Ultrametric inequality having been formulated by Hausdorff in 1934.
2. (Schikhof, *Ultrametric Calculus*, 1984.) Real and complex fields gave rise to the idea of studying any field K with a complete valuation $|.|$ comparable to the absolute value function.
3. Such fields satisfy the "strong triangle inequality" $|x + y| \leq \max(|x|, |y|)$.
4. Given a valued field, defining a totally ordered Abelian (i.e. commutative) group, an ultrametric space is induced through $|x - y| = d(x, y)$.
5. Various terms are used interchangeably for analysis in and over such fields such as p-adic, ultrametric, non-Archimedean, and isosceles.
6. Natural geometric ordering of metric valuations: real line. Ultrametric case: a hierarchy or rooted tree.

Some Geometrical Properties of Ultrametric Spaces

See I.C. Lerman, *Classification et Analyse Ordinale des Données*, Dunod, Paris, 1981.

1. In an ultrametric space, all triangles are either isosceles with small base, or equilateral.
2. Every point of a circle in an ultrametric space is a centre of the circle.
3. In an ultrametric topology, every ball is both open and closed (termed *clopen*).
4. An ultrametric space is 0-dimensional.
5. Informally, in an ultrametric space everything "lives" in a hierarchy expressed by a tree.

For an $n \times n$ matrix of positive reals, symmetric with respect to the principal diagonal, to be a matrix of distances associated with an ultrametric distance on X, a sufficient and necessary condition is that a permutation of rows and columns satisfies the following form of the matrix:

1. Above the diagonal term, equal to 0, the elements of the same row are non-decreasing.
2. For every index k, if

$$d(k, k+1) = d(k, k+2) = \cdots = d(k, k+\ell+1)$$

then

$$d(k+1, j) \leq d(k, j) \text{ for } k+1 < j \leq k+\ell+1$$

and

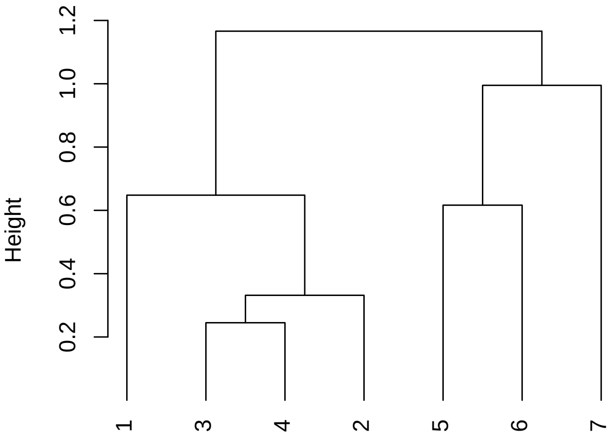
$$d(k+1, j) = d(k, j) \text{ for } j > k+\ell+1$$

Under these circumstances, $\ell \geq 0$ is the length of the section beginning, beyond the principal diagonal, the interval of columns of equal terms in row k.

To illustrate the ultrametric matrix format, consider the small data set shown in the table below. A dendrogram produced from this is in the figure (to follow). The ultrametric matrix that can be read off this dendrogram is shown in the next table. Finally a visualization of this matrix, illustrating the ultrametric matrix properties discussed, is presented.

Table : Input data: 8 iris flowers characterized by sepal and petal widths and lengths. From Fisher's iris data (Fisher, 1936).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
iris1	5.1	3.5	1.4	0.2
iris2	4.9	3.0	1.4	0.2
iris3	4.7	3.2	1.3	0.2
iris4	4.6	3.1	1.5	0.2
iris5	5.0	3.6	1.4	0.2
iris6	5.4	3.9	1.7	0.4
iris7	4.6	3.4	1.4	0.3


Height
0.2
0.4
0.6
0.8
1.0
1.2
1
3
4
2
5
6
7

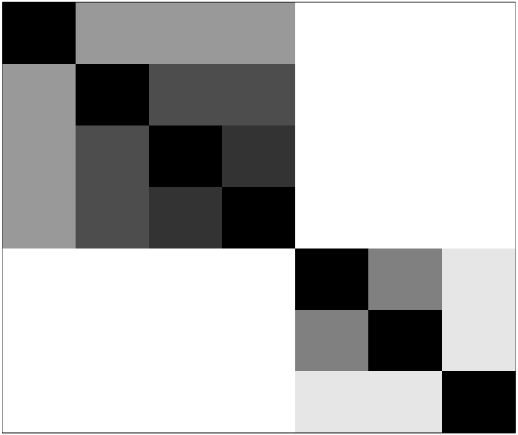
Ultrametric matrix

Table : Ultrametric matrix derived from the dendrogram in the previous figure.

	iris1	iris2	iris3	iris4	iris5	iris6	iris7
iris1	0	0.6480741	0.6480741	0.6480741	1.1661904	1.1661904	1.1661904
iris2	0.6480741	0	0.3316625	0.3316625	1.1661904	1.1661904	1.1661904
iris3	0.6480741	0.3316625	0	0.2449490	1.1661904	1.1661904	1.1661904
iris4	0.6480741	0.3316625	0.2449490	0	1.1661904	1.1661904	1.1661904
iris5	1.1661904	1.1661904	1.1661904	1.1661904	0	0.6164414	0.9949874
iris6	1.1661904	1.1661904	1.1661904	1.1661904	0.6164414	0	0.9949874
iris7	1.1661904	1.1661904	1.1661904	1.1661904	0.9949874	0.9949874	0

Ultrametric matrix visualization

A visualization of the ultrametric matrix of the Table, where bright or white = highest value, and black = lowest value.



Clustering Through Matrix Row and Column Permutation

1. Clustering by visualization.
2. Referred to also as *block modelling*.
3. There are optimized ways to carry this out (e.g. based on the graph traversal algorithm known as the travelling salesman problem).
4. For such approaches, underpinning them are row and column permutations, that can be expressed in terms of the permutation group, S_n, on n elements.

The Generalized Ultrametric

1. Ultrametric defined on the power set or join semilattice.
2. Typically hierarchical clustering is based on a distance (which can be relaxed often to a dissimilarity, not respecting the triangular inequality, and *mutatis mutandis* to a similarity), defined on all pairs of the object set: $d : X \times X \to \mathbb{R}^+$. I.e., a distance is a positive real value. Usually we require that a distance cannot be 0-valued unless the objects are identical.
3. A different form of ultrametrization is achieved from a dissimilarity defined on the power set of attributes characterizing the observations (objects, individuals, etc.) X.
4. Here we have: $d : X \times X \longrightarrow 2^J$, where J indexes the attribute (variables, characteristics, properties, etc.) set.
5. This gives rise to a different notion of distance, that maps pairs of objects onto elements of a join semilattice.
6. The latter can represent all subsets of the attribute set, J. That is to say, it can represent the power set, commonly denoted 2^J, of J.

Link with Formal Concept Analysis

1. As an example, consider, say, $n = 5$ objects characterized by 3 boolean (presence/absence) attributes, shown in the figure (next slide; top).
2. Define dissimilarity between a pair of objects in this table as a *set* of 3 components, corresponding to the 3 attributes, such that if both components are 0, we have 1; if either component is 1 and the other 0, we have 1; and if both components are 1 we get 0.
3. We get then $d(a, b) = 1, 1, 0$ which we will call d1,d2. Then, $d(a, c) = 0, 1, 0$ which we will call d2. Etc. With the latter we create lattice nodes as shown in the middle part of the figure (next slide).
4. Rule for each component: 0,0 $\longrightarrow$ 1; 1,0 $\longrightarrow$ 1; 0,1 $\longrightarrow$ 1; 1,1 $\longrightarrow$ 0.

Example: 5 objects, boolean attributes, associated lattice

	v_1	v_2	v_3
a	1	0	1
b	0	1	1
c	1	0	1
e	1	0	0
f	0	0	1

```
Potential lattice vertices      Lattice vertices found      Level

        d1,d2,d3                       d1,d2,d3               3
                                        /    \
                                       /      \
  d1,d2    d2,d3    d1,d3          d1,d2      d2,d3           2
                                       \      /
                                        \    /
   d1       d2       d3                   d2                  1
```

The set d1,d2,d3 corresponds to: $d(b, e)$ and $d(e, f)$
The subset d1,d2 corresponds to: $d(a, b)$, $d(a, f)$, $d(b, c)$, $d(b, f)$, and $d(c, f)$
The subset d2,d3 corresponds to: $d(a, e)$ and $d(c, e)$
The subset d2 corresponds to: $d(a, c)$

Clusters defined by all pairwise linkage at level ≤ 2:
a, b, c, f
a, c, e

Clusters defined by all pairwise linkage at level ≤ 3:
a, b, c, e, f

1. In Formal Concept Analysis it is the lattice itself which is of primary interest.
2. Traditional hierarchical cluster analysis is based on $d : I \times I \to \mathbb{R}^+$. The following was considered by Janowitz: hierarchical cluster analysis "based on abstract posets" (a poset is a partially ordered set), based on $d : I \times I \to 2^J$.
3. Also Mel Janowitz: cluster, then summarize (implies traditional hierarchical clustering).
4. Alternatively: summarize, then cluster (implies Formal Concept Analysis).

Applications of Generalized Ultrametrics

1. Usual ultrametric for a set I, $d : I \times I \longrightarrow \mathbb{R}^+$.
2. Generalized ultrametric where the range is a subset of the power set: $d : I \times I \longrightarrow \Gamma$, where Γ is a partially ordered set.
3. Applications include:
4. Non-monotonic reasoning, application of a succession of conditionals (sometimes called consequence relations). Negation or multiple valued logic (i.e. encompassing intermediate truth and falsehood) require support for non-monotonic reasoning.
5. The convergence to fixed points that are based on a generalized ultrametric system is precisely the study of spherically complete systems and expansive automorphisms.
6. As expansive automorphisms we see here again an example of symmetry at work.

Hierarchy, Ultrametric Topology and the p-Adic Number System

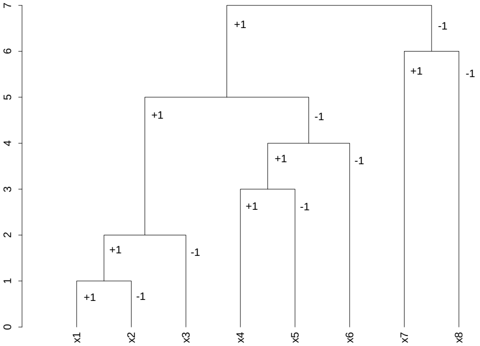
1. Importance of p-adic representation for physics on very small scales: see work of Igor Volovich from the 1980s.
2. Volovich (2010) poses the general principle that the fundamental physical laws should be invariant under the change of the number field. This leads to the following ambitious statement: "If these ideas are true then number theory and the corresponding branches of algebraic geometry are ... the ultimate and unified physical theory".
3. Hierarchy, as a branching process, is a good means of expressing suboptimal and/or discrete energy states or levels.
4. Genealogy, evolutionary processes, etc.
5. Sequence of partitions. Partial order on the clusters/nodes, total order on the partitions.

p-Adic Encoding of a Dendrogram

1. We introduce now the one-to-one mapping of clusters (including singletons) in a dendrogram H into a set of p-adically expressed integers (*a fortiori*, rationals, or $\mathbb{Q}_p$).
2. The field of p-adic numbers is the most important example of ultrametric spaces.
3. Addition and multiplication of p-adic integers, $\mathbb{Z}_p$, are well-defined. Inverses exist and no zero-divisors exist.

Terminal-to-root traversal in a dendrogram or binary rooted tree

- Use the path $x \subset q \subset q' \subset q'' \subset \ldots q_{n-1}$, where x is a given object specifying a given terminal, and $q, q', q'', \ldots$ are the embedded classes along this path, specifying nodes in the dendrogram.
- The root node is specified by the class q_{n-1} comprising all objects.



Labelled, ranked dendrogram on 8 terminal nodes, $x_1, x_2, \ldots, x_8$. Branches are labelled $+1$ and -1. Clusters are: $q_1 = \{x_1, x_2\}, q_2 = \{x_1, x_2, x_3\}, q_3 = \{x_4, x_5\}, q_4 = \{x_4, x_5, x_6\}, q_5 = \{x_1, x_2, x_3, x_4, x_5, x_6\}, q_6 = \{x_7, x_8\}, q_7 = \{x_1, x_2, \ldots, x_7, x_8\}$.

Using terminal-to-root traversals, define p-adic encoding of terminal nodes, and hence objects

$$
\begin{aligned}
x_1 &: \quad +1 \cdot p^1 + 1 \cdot p^2 + 1 \cdot p^5 + 1 \cdot p^7 \\
x_2 &: \quad -1 \cdot p^1 + 1 \cdot p^2 + 1 \cdot p^5 + 1 \cdot p^7 \\
x_3 &: \quad -1 \cdot p^2 + 1 \cdot p^5 + 1 \cdot p^7 \\
x_4 &: \quad +1 \cdot p^3 + 1 \cdot p^4 - 1 \cdot p^5 + 1 \cdot p^7 \\
x_5 &: \quad -1 \cdot p^3 + 1 \cdot p^4 - 1 \cdot p^5 + 1 \cdot p^7 \\
x_6 &: \quad -1 \cdot p^4 - 1 \cdot p^5 + 1 \cdot p^7 \\
x_7 &: \quad +1 \cdot p^6 - 1 \cdot p^7 \\
x_8 &: \quad -1 \cdot p^6 - 1 \cdot p^7
\end{aligned}
$$

Remarks on this Encoding 1/2

1. If $p = 2$ the resulting decimal equivalents could be the same: cf. contributions based on $+1 \cdot p^1$ and $-1 \cdot p^1 + 1 \cdot p^2$.
2. Coding based on $p = 3$ is required to avoid ambiguity among decimal equivalents.
3. A precise definition of the tree considered here is: labelled ranked binary tree.
4. We require the labels $+1$ and -1 for the two branches at any node. Of course we could interchange these labels, and have these $+1$ and -1 labels reversed at any node. By doing so we will have different p-adic codes for the objects, x_i.
5. The following properties hold: (i) *Unique encoding:* the decimal codes for each x_i (lexicographically ordered) are unique for $p \geq 3$; and (ii) *Reversibility:* the dendrogram can be uniquely reconstructed from any such set of unique codes.

Remarks on this Encoding 2/2

1. The p-adic encoding defined for any object set can be expressed as follows for any object x associated with a terminal node:

$$x = \sum_{j=1}^{n-1} c_j p^j \text{ where } c_j \in \{-1, 0, +1\} \tag{1}$$

2. In greater detail we have:

$$x_i = \sum_{j=1}^{n-1} c_{ij} p^j \text{ where } c_{ij} \in \{-1, 0, +1\} \tag{2}$$

   Here j is the level or rank (root: $n-1$; terminal: 1), and i is an object index.

3. We have used: $c_j = +1$ for a left branch, $= -1$ for a right branch, and $= 0$ when the node is not on the path from that particular terminal to the root.

Matrix form of p-Adic Encoding 1/3

1. A matrix form of this encoding is as follows, where $\{\cdot\}^t$ denotes the transpose of the vector.
2. Let $\mathbf{x}$ be the column vector $\{x_1 \ x_2 \ \ldots x_n\}^t$.
   Let $\mathbf{p}$ be the column vector $\{p^1 \ p^2 \ \ldots p^{n-1}\}^t$.
   Define a characteristic matrix C of the branching codes, $+1$ and -1, and an absent or non-existent branching given by 0, as a set of values c_{ij} where $i \in I$, the indices of the object set; and $j \in \{1, 2, \ldots, n-1\}$, the indices of the dendrogram levels or nodes ordered increasingly. For the dendrogram used, we therefore have as follows.

Matrix form of p-Adic Enoding 2/3

$$C = \{c_{ij}\} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{pmatrix} \tag{3}$$

For given level j, $\forall i$, the absolute values $|c_{ij}|$ give the membership function either by node, j, which is therefore read off columnwise; or by object index, i, which is therefore read off rowwise.

Matrix form of p-Adic Enoding 3/3

The matrix form of the p-adic encoding used in equations seen is:

$$\mathbf{x} = C\mathbf{p} \tag{4}$$

Here, $\mathbf{x}$ is the decimal encoding, C is the matrix with dendrogram branching codes (cf. example shown in expression (3)), and $\mathbf{p}$ is the vector of powers of a fixed prime p.

Remarks on Matrix form of p-Adic Encoding 1/2

1. Labels $+1$ and -1 were required (as opposed to the choice of 0 and 1, which might have been our first thought) to fully cater for the ranked nodes (i.e. the total order, as opposed to a partial order, on the nodes).
2. We can consider the objects that we are dealing with to have equivalent integer values. I.e., decimal equivalents of the p-adic expressions used above for $x_1, x_2, \dots$.
3. We have equivalence between: a p-adic number; a p-adic expansion; and an element of $\mathbb{Z}_p$ (the p-adic integers).
4. The coefficients used to specify a p-adic number, F.Q. Gouvêa (2003) notes, "must be taken in a set of representatives of the class modulo p. The numbers between 0 and $p-1$ are only the most obvious choice for these representatives. There are situations, however, where other choices are expedient."

Remarks on Matrix form of p-Adic Encoding 2/2

1. (F. Critchley and W. Heiser, *Journal of Classification*, 1988.) Matrix C used there. Title: "Hierarchical trees can be perfectly scaled in one dimension". My (somewhat trivial) view: p-adic numbering is feasible, and hence a one dimensional representation of terminal nodes is easily arranged through expressing each p-adic number with a real number equivalent.
2. (B. Mirkin, "Linear embedding of binary hierarchies and applications", B. Mirkin, F. McMorris, F. Roberts, and A. Rzhetsky (Eds.) Mathematical Hierarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, V. 37, AMS: Providence, 331-356, 1997.) A *nest* (i.e. cluster nesting) indicator function is defined, based on the set $\{a_w, -b_w, 0\}, a_w, b_w \in \mathbb{R}^+$ in the same way that the set $\{1, -1, 0\}$ is used above for the matrix C. Orthonormality properties of the nest indicator functions are studied.

p-Adic Distance on a Dendrogram 1/4

1. We will now induce a metric topology on the p-adically encoded dendrogram, H. It leads to various symmetries relative to identical norms, for instance, or identical tree distances.
2. We use the following longest common subsequence, starting at the root: we look for the term p^r in the p-adic codes of the two objects, where r is the lowest level such that the values of the coefficients of p^r are equal.
3. Use our case study set of p-adic codes for $x_1, x_2, \ldots$ and relations, as an example,

p-Adic Distance on a Dendrogram 2/4

Refer to slides 31, 32.

For x_1 and x_2, we find the term we are looking for to be p^1, and so $r = 1$.

For x_1 and x_5, we find the term we are looking for to be p^5, and so $r = 5$.

For x_5 and x_8, we find the term we are looking for to be p^7, and so $r = 7$.

Having found the value r, the distance is defined as p^{-r}

p-Adic Distance on a Dendrogram 3/4

1. This longest common prefix metric is also known as the Baire distance.
2. In topology the Baire metric is defined on infinite strings.
3. It is more than just a distance: it is an ultrametric bounded from above by 1, and its *infimum* is 0 which is relevant for very long sequences, or in the limit for infinite-length sequences. T
4. The longest common prefix metric leads directly to Patrick Erik Bradley's *p-adic hierarchical classification*.
5. This is a special case of the "fast" hierarchical clustering that we have developed (see other lectures).
6. Relative to the longest common prefix metric, there are other related forms of metric, and simultaneously ultrametric.
7. E.g. Metric defined via the integer part of a real number.

p-Adic Distance on a Dendrogram 3/4

1. Or, for integers x, y consider: $d(x, y) = 2^{-\text{order}_p(x-y)}$ where p is prime, and $\text{order}_p(i)$ is the exponent (non-negative integer) of p in the prime decomposition of an integer.
2. Furthermore let $S(x)$ be a series: $S(x) = \sum_{i \in \mathbb{N}} a_i x^i$. ($\mathbb{N}$ are the natural numbers.) The order of $S(i)$ is the rank of its first non-zero term: $\text{order}(S) = \inf\{i : i \in \mathbb{N}; a_i \neq 0\}$. (The series that is all zero is of order infinity.) Then the ultrametric similarity between series is: $d(S, S') = 2^{-\text{order}(S-S')}$.

Scale-Related Symmetry: The Dilation Operator 1/3

1. Motivation for name: dilation operator is used in the wavelet transform.
2. This operator is p-adic multiplication by $1/p$.
3. Consider the set of objects $\{x_i | i \in I\}$ with its p-adic coding.
4. Take $p = 2$. (Non-uniqueness of corresponding decimal codes is not of concern to us now, and taking this value for p is without any loss of generality.)
5. Multiplication of $x_1 = +1 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^5 + 1 \cdot 2^7$ by $1/p = 1/2$ gives: $+1 \cdot 2^1 + 1 \cdot 2^4 + 1 \cdot 2^6$.
6. Each level has decreased by one, and the lowest level has been lost.
7. Subject to the lowest level of the tree being lost, the form of the tree remains the same.
8. By carrying out the multiplication-by-$1/p$ operation on all objects, it is seen that the effect is to rise in the hierarchy by one level.

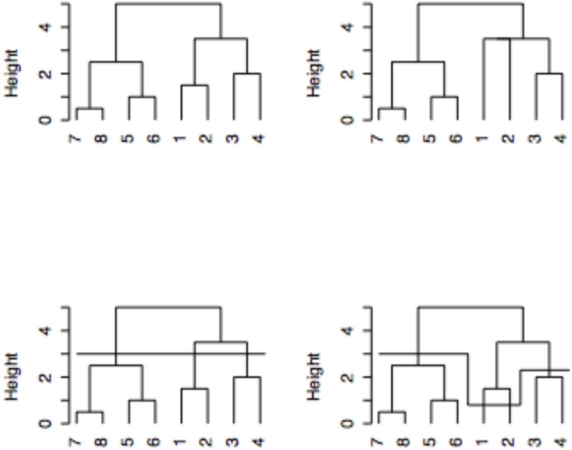
Scale-Related Symmetry: The Dilation Operator 2/3

1. Let us call product with $1/p$ the operator A.
2. The effect of losing the bottom level of the dendrogram means that either (i) each cluster (possibly singleton) remains the same; or (ii) two clusters are merged.
3. Therefore the application of A to all q implies a subset relationship between the set of clusters $\{q\}$ and the result of applying A, $\{Aq\}$.
4. Repeated application of the operator A gives Aq, A^2q, A^3q, $\dots$.
5. Starting with any singleton, $i \in I$, this gives a path from the terminal to the root node in the tree.
6. Each such path ends with the null element, which we define to be the p-adic encoding corresponding to the root node of the tree.
7. Therefore the intersection of the paths equals the null element.

Scale-Related Symmetry: The Dilation Operator 3/3

1. Benedetto and Benedetto (2004) discuss A as an expansive automorphism of I, i.e. form-preserving, and locally expansive.
2. Implications: For any q, let us take $q, Aq, A^2q, \ldots$ as a sequence of open subgroups of I, with $q \subset Aq \subset A^2q \subset \ldots$, and $I = \bigcup\{q, Aq, A^2q, \ldots\}$. This is termed an inductive sequence of I, and I itself is the inductive limit. (H. Reiter and J.D. Stegemen, *Classical Harmonic Analysis and Locally Compact Groups*, OUP, 2000.)
3. Each path defined by application of the expansive automorphism defines a spherically complete system. (W.H. Schikhof, *Ultrametric Calculus*, CUP, 1984; A.C.M. Van Rooij, *Non-Archimedean Functional Analysis*, Marcel Dekker, 1978.)
4. This is a formalization of well-defined subset embeddedness. Such a methodological framework finds application in multi-valued and non-monotonic reasoning (noted earlier).

Haar Wavelet Transform on a Hierarchy, with Hard Thresholding



Haar Wavelet Transform on a Hierarchy 2/2

- Hierarchy on upper left is wavelet transformed; small wavelet coefficients are set to zero (hard thresholding); hierarchy is re-constituted (inverse wavelet transform); result is on upper right.
- Net result can also be seen in lower right
- This can be contrasted with usual partitioning on lower left.
- We have a piecewise or clusterwise smooth of our data.
- (See Murtagh, Journal of Classification, The Haar wavelet transform of a dendrogram, 24, 3-32, 2007.)

Remarkable Symmetries in Very High Dimensional Spaces 1/2

1. As ambient dimensionality increases, distances became more and more ultrametric. (Rammal et al., Murtagh, Hall, Donoho, etc.)
2. That is to say, a hierarchical embedding becomes more and more immediate and direct as dimensionality increases.
3. Hence there is inherent hierarchical structure in high dimensional data spaces.
4. Points in high dimensional spaces become increasingly equidistant with increase in dimensionality.
5. Donoho: "not only are the points [of a Gaussian cloud in very high dimensional space] on the convex hull, but all reasonable-sized subsets span faces of the convex hull. This is wildly different than the behavior that would be expected by traditional low-dimensional thinking".

Remarkable Symmetries in Very High Dimensional Spaces 2/2

1. Very simple structures in very high dimensions are not necessarily trivial.
2. Even very simple structures (hence with many symmetries) can be used to support fast and perhaps even constant time worst case proximity search.
3. In the machine learning framework, there are important implications ensuing from the simple high dimensional structures.
4. Very high dimensional clustered data contain symmetries that in fact can be exploited to "read off" the clusters in a computationally efficient way.
5. What we might want to look for in contexts of considerable symmetry are the "impurities" or small irregularities that detract from the overall dominant picture.

Example of the change of topological properties as ambient dimensionality increases 1/2

1. Table will show typical results, based on 300 sampled triangles from triplets of points.
2. For uniform, the data are generated on $[0, 1]^m$; hypercube vertices are in $\{0, 1\}^m$, and for Gaussian on each dimension, the data are of mean 0, and variance 1.
3. Dimen. is the ambient dimensionality.
4. Isosc. is the number of isosceles triangles with small base, as a proportion of all triangles sampled.
5. Equil. is the number of equilateral triangles as a proportion of triangles sampled.
6. UM is the proportion of ultrametricity-respecting triangles (= 1 for all ultrametric).

Example of the change of topological properties as ambient dimensionality increases 2/2

No. points	Dimen.	Isosc.	Equil.	UM
Uniform				
100	20	0.10	0.03	0.13
100	200	0.16	0.20	0.36
100	2000	0.01	0.83	0.84
100	20000	0	0.94	0.94
Hypercube				
100	20	0.14	0.02	0.16
100	200	0.16	0.21	0.36
100	2000	0.01	0.86	0.87
100	20000	0	0.96	0.96
Gaussian				
100	20	0.12	0.01	0.13
100	200	0.23	0.14	0.36
100	2000	0.04	0.77	0.80
100	20000	0	0.98	0.98

Application to Segmentation of Financial Time Series

1. Financial futures, circa March 2007, denominated in euros from the DAX exchange. Data stream is at the millisecond rate. Comprises about 382,860 records. Each record includes: 5 bid and 5 asking prices, together with bid and asking sizes in all cases, action.
2. Extracted one symbol (commodity) with 95,011 single bid values.
3. "Sliding windows", embeddings that were: 100, 1000, 10000 values in length. Defined points in 100-, 1000-, or 10000-dimensional space.
4. Histograms of distances were studied using model-based clustering.
5. This included model identification using the Bayesian Information Criterion, BIC.
6. From the clusters, the segments were inferred.

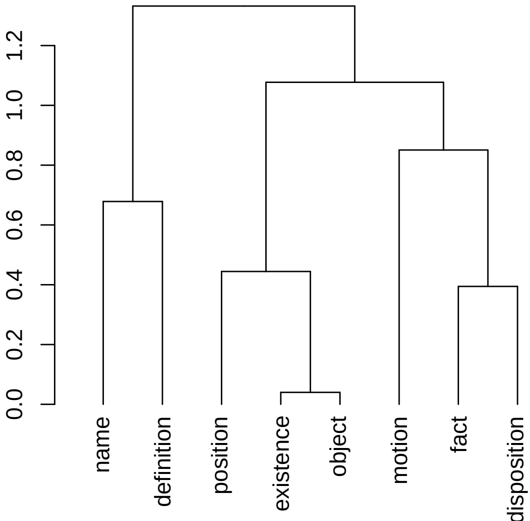
Partial Ultrametric Embedding: Notes 1/3

1. Permutation representations of a data stream. Hierarchies can also be represented as permutations. Hence we can associate data streams with hierarchies. (Early computational work on hierarchical clustering used permutation representation to great effect ("packed representation" used by Robin Sibson in the 1970s.)
2. Murtagh: "Identifying the ultrametricity of time series", *European Physical Journal B*, 43, 573–579, 2005: ultrametric embedding of time-varying signals, including biomedical, meteorological, financial and other. At issue: inherent hierarchical properties in the data.
3. Most non-ultrametric time series: chaotic.
4. Eyegaze trace data: remarkably high in ultrametricity, due to extreme saccade movements.
5. Some questions raised in regard to the EEG data used, for sleeping, petit mal and irregular epilepsy cases.

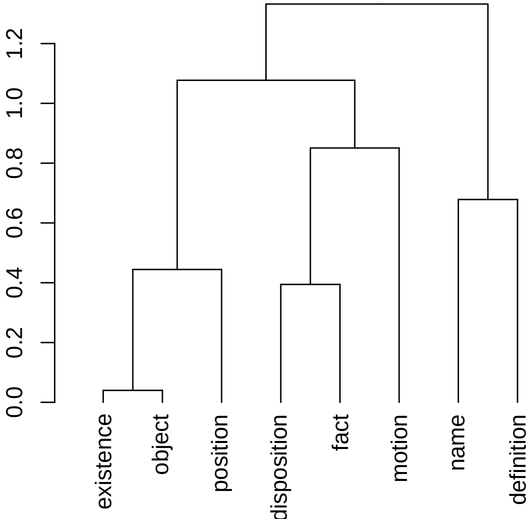
Sibson's "Packed Representation" of a Dendrogram 1/4

1. Put a lower ranked subtree always to the left; and read off the oriented binary tree on non-terminal nodes (see example).
2. Then for any terminal node indexed by i with the exception of the rightmost which will always be n, define $p(i)$ as the rank at which the terminal node is first united with some terminal node to its right.
3. Inorder traversal of the oriented binary tree.
4. For the dendrogram to follows we find the unique permutation: (13625748).

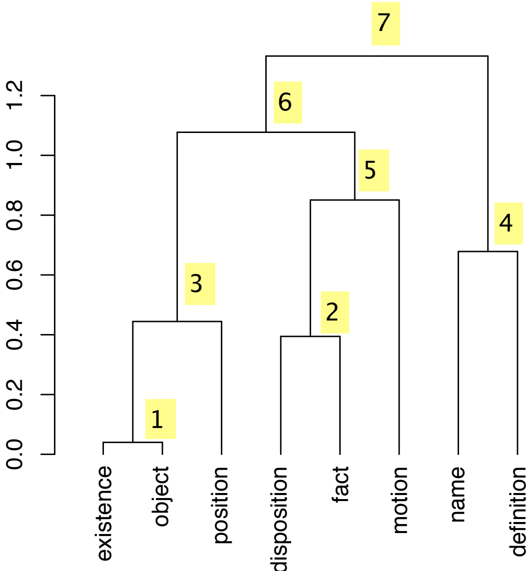
Sibson's "Packed Representation" of a Dendrogram 1/4



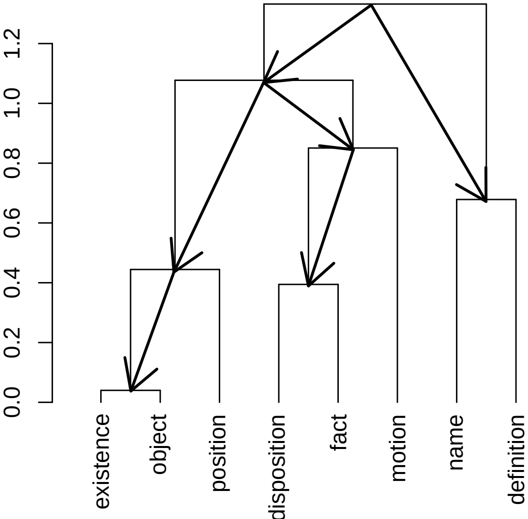
Sibson's "Packed Representation" of a Dendrogram 2/4



Sibson's "Packed Representation" of a Dendrogram 3/4



Sibson's "Packed Representation" of a Dendrogram 4/4



Partial Ultrametric Embedding: Notes 2/3

1. Khrennikov et al.: modelling multi-agent systems.
2. Bose-Einstein and Fermi-Dirac statistical distributions (derived from quantum statistics of energy states of bosons and fermions, i.e. elementary particles with integer, and half odd integer, spin).
3. Multi-agent behaviours modelled using such energy distributions.
4. Framework: urn model, balls can move, loss of energy over time, possibility to receive input energy, but potentially shared with other balls.
5. Monte Carlo system. Sequences of actions (and moves), viz. their histories, coded such that triangle properties can be investigated. Leads to characterization of how ultrametrically-embeddable the data is.

Partial Ultrametric Embedding: Notes 3/3

1. Andrei Khrennikov: case presented for such analysis of behavioural histories being important for study of social and economic complexity.
2. Van Rijsbergen's *The Geometry of Information Retrieval*, CUP, 2004: quantum physics formalism for information retrieval.
3. Text: tales from the Brothers Grimm, Jane Austen novels, dream reports, air accident reports, and James Joyce's Ulysses.

Ultrametric Baire Space and Distance 1/3

1. Baire space: countably infinite sequences; metric defined in terms of the longest common prefix: the longer the common prefix, the closer a pair of sequences.
2. Univariate values: let base $m = 10$; x and $y \in [0, 1]$. Maximum precision, $|K|$; ordered sets x_k, y_k for $k \in K$, or, for $k = 1, 2, \ldots, |K|$.
3. Consider $x = 0.478$; and $y = 0.472$. 1st, 2nd digits the same, 3rd different.
4. $$d_B(x, y) \equiv d_B(x_K, y_K) = \begin{cases} 1 & \text{if } \; x_1 \neq y_1 \\ \inf \; 10^{-k} & \quad x_k = y_k, \quad 1 \leq k \leq |K| \end{cases} \tag{5}$$
5. This Baire distance is an ultrametric.
6. For a metric, require $d(x, y) = 0$ iff $x = y$ whereas for the Baire distance this reflexivity property is relaxed by having the 0 value replaced by the definably minimal value.

Ultrametric Baire Space and Distance 2/3

1. Application: This distance splits a unidimensional string of decimal values into a 10-way hierarchy (base $m = 10$, in which each leaf is associated with a grid cell.
2. Pairwise distances of points assigned to the same cell are the same.
3. Relative to agglomerative hierarchical clustering, in Baire-based hierarchy each node of this tree is associated with a grid (more strictly, in what we have described, interval) cell.
4. Grid cardinality defines local density.
5. Top-down hierarchy construction.
6. By having target data structure, regular 10-way tree, we can cluster data in a single scan. Hence linear time computational complexity.
7. Higher dimensionality: random projections.

Ultrametric Baire Space and Distance 3/3

1. Random matrix R and project the $d \times N$ data matrix X into the k dimensions is of order $O(dkN)$. If X is sparse with c nonzero entries per column, the complexity is of order $O(ckN)$.
2. Random projection: class of hashing function. If two points (p, q) are close, they will have a very small $\|p - q\|$ (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability.
3. See results for: chemoinformatics, astronomy, text retrieval.

Approximating an Ultrametric for Similarity Metric Space Searching

1. Fast nearest neighbour searching in metric spaces often uses heuristics. Ultrametric spaces give rise instead to a unifying view.
2. Fast nearest neighbour finding often makes use of pivots to establish bounds on points to be searched, and points to be bypassed as infeasible.
3. E.g. feasibility bounds: determine a neigbour and the distance to it from our given point; have pre-stored distances to a fixed point; many other points can be excluded through use of the triangular inequality.
4. Can be shown to be: "stretching the triangular inequality" (transform locally) to be the ultrametric inequality.
5. Another heuristic: embedding the given metric space points in lower dimensional spaces.

Symmetry: Examples Seen

1. p-Adic encoding and also tree and embedded clusters: expansive automorphism of object set I, i.e. form-preserving, and locally expansive.
2. Dendrogram: invariant relative to rotation (alternatively: permutation) of left and right child nodes. Denote the permutation at level ν by P_ν. Then the automorphism group is given by:

$$G = P_{n-1} \text{ wr } P_{n-2} \text{ wr } \ldots \text{ wr } P_2 \text{ wr } P_1$$

   where wr denotes the wreath product.
3. We considered labelled, ranked, binary trees. If we considere non-labelled, ranked, binary trees, then these are isomorphic to either down-up permutations, or up-down permutations (both on $n-1$ elements). Thus we are dealing with the symmetry group of these permutations.

Final Word

1. "My thesis has been that one path to the construction of a nontrivial theory of complex systems is by way of a theory of hierarchy." (Herbert Simon)
2. "... my central theme is that complexity frequently takes the form of hierarchy and that hierarchic systems have some common properties independent of their specic content. Hierarchy, I shall argue, is one of the central structural schemes that the architect of complexity uses." (Herbert Simon)
3. "Human thinking (as well as many other information processes) is fundamentally a hierarchical process. ... In our information modeling the main distinguishing feature of p-adic numbers is the treelike hierarchical structure." (Andrei Khrennikov)

Final Word 2/2

4. "Whenever you have to do with a structure-endowed entity ... try to determine its group of automorphisms, the group of those element-wise transformations which leave all structural relations undisturbed. You can expect to gain a deep insight in the constitution of [the structure-endowed entity] in this way. After that you may start to investigate symmetric configurations of elements, i.e. configurations which are invariant under a certain subgroup of the group of all automorphisms; ..." (Hermann Weyl)