Symmetry in Data Mining and Analysis: A Unifying View based on Hierarchy

Fionn Murtagh
Department of Computer Science
Royal Holloway, University of London
Egham TW20 0EX, UK
fmurtagh@acm.org

May 20, 2008

Abstract

Data analysis and data mining are concerned with unsupervised pattern finding and structure determination in data sets. The data sets themselves are explicitly linked as a form of representation to an observational or otherwise empirical domain of interest. "Structure" has long been understood as symmetry which can take many forms with respect to any transformation, including point, translational, rotational, and many others. Beginning with the role of number theory in expressing data, we show how we can naturally proceed to hierarchical structures. We show how this both encapsulates traditional paradigms in data analysis, and also opens up new perspectives towards issues that are on the order of the day, including data mining of massive, high dimensional, heterogeneous data sets. Linkages with other fields are also discussed including computational logic and symbolic dynamics.

Keywords: Multivariate data analysis, pattern recognition, information storage and retrieval, clustering, hierarchy, ultrametric topology

> "... 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." ([76], p. 184.)

1 Introduction

Herbert A. Simon, Nobel Laureate in Economics, originator of "bounded rationality" and of "satisficing", believed in hierarchy at the basis of the human and social sciences, as our opening quotation shows.

Partitioning a set of observations [78, 79, 52] leads to some very simple symmetries. This is one approach to clustering and data mining. But such approaches, often based on optimization, are really not of direct interest to us here. Instead we 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. Our work is very different too from the marvellous view of the development of mathematical group theory – but viewed in its own right as a complex, evolving system – presented by Foote [20].

1.1 Structure in Observed Data

Weyl [83] 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; ..." ([83], p. 144).

"Symmetry is a vast subject, significant in art and nature.", Weyl states (p. 145), and no better example of the "mathematical intellect" at work. "Although the mathematics of group theory and the physics of symmetries were not fully developed simultaneously – as in the case of calculus and mechanics by Newton – the intimate relationship between the two was fully realized and clearly formulated by Wigner and Weyl, among others, before 1930." ([80], p. 1.) Powerful impetus was given to this (mathematical) group view of study and exploration of symmetry in art and nature by Felix Klein's 1872 Erlangen Program [41] which proposed that geometry was at heart group theory: geometry is the study of groups of transformations, and their invariants. Klein's Erlangen Program is at the cross-roads of mathematics and physics. The purpose of this article is to locate symmetry and group theory at the cross-roads of data analysis and data mining too.

1.2 About this Article

In section 2 we will look at number systems, as a first step in encoding or representing that which we seek to study. So our numeric representation is the point of departure, and is prior to the search for, or characterization of, symmetries or related structures.

In section 3 we look at the topological viewpoint, instead of the number theoretic (or algebraic). Here we are quickly brought face to face with practical application, – for topology is defined by distance and other such relationships. We review not just the application to analysis of masses of high dimensional data, but also touch on other fields, e.g. computational logic.

P-adic numbers, which are hierarchically-based number schemes, can directly encompass symmetry in data, as will be shown in section 4.

In section 5 we look at a particular group action on trees (or hierarchies).

In section 6 we look at permutation group action on a particular representation of trees.

In section 7, we draw an interesting practical conclusion of considerable significance from this work. We note how the handling of very high dimensional data may, counter-intuitively, be easier than the handling of low dimensional data.

The central theme of this article is that symmetry underpins a great deal of practical data analysis. Ancillary to this are other themes:

1. If symmetry is immediately apparent then it is quite likely to be uninteresting. So it is more beneficial to strike out at new and insightful ways of analyzing data. In this article we show how many different vantage points on data and its analysis all are underpinned with forms of symmetry.

2. Measurement and observation go hand in hand in science. We show that one small enhancement of this statement, namely that measurement can be carried out in a p-adic number system, has far-reaching implications. It implies that hierarchy, defined by a partial order, is every bit as reasonable as the real number system, defined by a linear or total order, for the task of observation and measurement.

3. We could describe hierarchy built from pairwise dissimilarities as a "precision tool" for data mining; and hierarchies built from the generalized ultrametric (see section 3.5) as leading to a "power tool" for data mining. The former is (without special algorithmic speedups) typically quadratic or $O(n^2)$ in its computational requirement. The latter can be linear or $O(n)$ in its computation. Here n relates to number of observations.

4. In logic, chains of implications or conditionals have to be analyzed. When we consider a partial order of conditionals, then the framework of spherical (ultrametric) completeness or inductive limit (sections 3.5.2 and especially 4.5) become very useful indeed.

5. In sections 4.5, and 5 especially, we touch on quite similar problems and issues relating to hierarchical structuring of data in image and signal processing applications, for which we provide citations.

6. The permutation representations of hierarchies – hence the permutation representation of data – explored in section 6 provide linkages, via data encoding schemes, with analysis of signal and time series dynamics.

1.3 A Brief Introduction to Hierarchical Clustering

For the reader new to analysis of data a very short introduction is now provided on hierarchical clustering. Along with other families of algorithm, the

objective is automatic classification, for the purposes of data mining, or knowledge discovery. Classification, after all, is fundamental in human thinking, and machine-based decision making. But we draw attention to the fact that our objective is *unsupervised*, as opposed to *supervised* classification, also known as discriminant analysis or (in a general way) machine learning. So here we are *not* concerned with generalizing the decision making capability of training data, nor are we concerned with fitting statistical models to data so that these models can play a role in generalizing and predicting. Instead we are concerned with having "data speak for themselves". That this unsupervised objective of classifying data (observations, objects, events, phenomena, etc.) is a huge task in our society is unquestionably true. One may think of situations when precedents are very limited, for instance.

Among families of clustering, or unsupervised classification, algorithms, we can distinguish the following: (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.

For the last-mentioned family of algorithm, agglomerative building of the hierarchy from consideration of object pairwise distances has been the most common approach adopted. As comprehensive background texts, see [51, 29, 84, 30].

2 p-Adic and Real Number Systems: Both Completions of the Rationals

In this section we will discuss:

1. How p-adic numbers are good alternatives compared to the real numbers for precise measurement.
2. We will show that they are different from the reals.
3. We will provide a reason for using a prime, p, rather than a more general integer, m.
4. We will not be able to say why one p should be preferred over another. In practice, for parsimony, we use p = 2 or p = 3.
5. Finally, we will show how p-adic numbers serve us well in regard to symmetry in data.

This then leads to the practical side: using an ultrametric topology, associated with the p-adic number system.

2.1 Reals and p-Adic Numbers: Equally Useful Alternatives

Whether we deal with Euclidean or with non-Euclidean geometry, we are (nearly) always dealing with reals. But the reals start with the natural numbers, and from associating observational facts and details with such numbers we begin the process of measurement. From the natural numbers, we proceed to the rationals, allowing fractions to be taken into consideration.

The following view of how we do science or carry out other quantitative study has been put forward by Freund [23]. We can always use rationals to make measurements. But they will be approximate, in general. It is better therefore to allow for observables being "continuous, i.e. endow them with a topology". Therefore 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). There is the Archimedean norm, defined as follows: for any $x, y \in \mathbb{Q}$, with $|x| < |y|$, then there exists an integer N such that $|Nx| > |y|$. For convenience, we write: $|x|_\infty$ for this norm. So if this completion is Archimedean, then we have $\mathbb{R} = \mathbb{Q}_\infty$, the reals. That is fine if space is taken as commutative and Euclidean.

What of alternatives? Remarkably all norms are known. Besides the $\mathbb{Q}_\infty$ norm, we have an infinity of norms, $|x|_p$, labeled by primes, p. By Ostrowski's theorem, to be further discussed in section 2.4, these are all the possible norms on $\mathbb{Q}$. So we have an unambiguous labeling, 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.

2.2 Short Discussion of the Real Numbers

There are varying, standard ways to define real numbers and their arithmetic operations from rationals or the integers (following [49]): (i) Dedekind cuts of rational numbers are a standard viewpoint in analysis; (ii) equivalence classes of Cauchy sequences of rational numbers are a standard viewpoint in topology; and (iii) base-10 digit sequences is a development of real numbers as infinite decimal expansions, and is an ordinary, everyday approach. (That the latter can be quite dangerous is shown in [18]. Using IEEE standard floating point arithmetic, based on sign, mantissa and exponent notation, the following expressions should give the same result but they do not: $10^{20} + 20 - 10 - 10^{20} = 0; 10^{20} + 20 - 10^{20} - 10 = -10; 10^{20} - 10 - 10^{20} + 20 = 20$. Further examples are provided in [18].) Of course, to the standard ways to define real numbers can be added binary (which we use, in fact, in section 2.3) or hexadecimal number systems, or continued fractions.

The real numbers form the unique linear order that is dense, without endpoints, Dedekind-complete, and separable.

The real numbers may also be viewed as a space of branches of an infinite tree. Trees are partial orders and we can define chains as paths in the sense of the partial order. Different infinite trees give rise to different linear orders. This serves as an introduction to the next subsection, where we will pursue such a tree representation. Further reading on "measurements and numbers" can be found in [38], chap. 1.

In section 3.5.2, we will look at how, [17], a "*computable real number* is ... the lub [least upper bound] of a shrinking sequence of rational intervals which is generated by a master program", and therefore how a real number is computable "in the interval approach to computability on the real line".

2.3 p-Adic and m-Adic Numbers

We will use p to denote a prime, and m to denote 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. (Cf. remark below, at the end of section 4.1, quoting from [26]. It makes the point that this opens up a range of alternative notation options in practice.) Recall that a ring does not allow division, while a field does. An m-adic number is a ring; but a p-adic number is a field. So a priori, a 10-adic number is 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 elements of $\mathbb{Q}_p$ can be viewed as Cauchy sequences, and the application of this will be looked at next.

2.4 Completion

The real numbers can be defined as equivalence classes of Cauchy sequences of rational numbers; this allows us to, for example, write 1 as 1.000... = 0.9999... However, the definition of a Cauchy sequence relies on the metric chosen and, by choosing a different one, numbers other than the real numbers can be constructed. The usual metric which yields the real numbers is called the Euclidean metric. In fact we start with a valuation ("a notion of *size*", [26]) on $\mathbb{Q}$, which is used to define an absolute value, and this in turn gives rise to a metric [64]. Once we have a metric, we then consider the convergence of sequences. Cauchy

sequences are sequences whose terms become arbitrarily close once one goes far enough, i.e. a sequence converges if it is Cauchy.

The reals and the p-adic numbers are the completions of the rationals. Field $\mathbb{Q}_p$ of p-adic numbers is defined as the completion of the metric space $(\mathbb{Q}, d_p)$ where d_p is a metric on $\mathbb{Q}$. The elements of $\mathbb{Q}_p$ are equivalence classes of Cauchy sequences, where two sequences are called equivalent if their difference converges to zero. In this way, we obtain a complete metric space which is also a field and contains $\mathbb{Q}$.

Two absolute values are equivalent if they induce the same topology ([63]; [72], p. 22; [71], p. 3; [26], p. 46).

Ostrowski's theorem: Let $|.|$ be an absolute value on $\mathbb{Q}$, then $|.|$ is equivalent to one of the following:

1. the trivial absolute value, which sends $|0|$ to 0 and $|x| = 1, \quad \forall x \neq 0$,
2. the usual Euclidean absolute value, denoted $|.|_\infty$, or
3. the p-adic absolute value $|.|_p$ for some prime p.

Thus each norm on $\mathbb{Q}$ is equivalent either to the Euclidean norm, the discrete norm, or to one of the p-adic norms for some prime p.

Product formula ([26], p. 48): If $a \neq 0, a \in \mathbb{Q}$, then $\prod_{p \leq \infty} |a|_p = 1$. (This is a natural starting point for adeles, numbers which take all possible number bases, p, into account simultaneously.)

It follows from Ostrowski's result that the real numbers and the p-adic numbers are the only completions of the rational numbers with respect to a metric defined by an absolute value. Going beyond $\mathbb{R}$, which is complete, we need to adjoin a complex element to it to get $\mathbb{C}$, its algebraic closure. Similarly, with $\mathbb{Q}_p$.

To further elucidate the distinction between $\mathbb{Q}_p$ and $\mathbb{R}$, we have the following examples [67]. Since $\sqrt{p} \notin \mathbb{Q}_p$, we know that $\mathbb{Q}_p$ does not contain all real numbers. However $\mathbb{Q}_p$ contains square roots of -1 whenever -1 is a square modulo p. This is the case in $\mathbb{Q}_5$. But other examples exist when -1 is not a square mod p. The overall conclusion from this is that clearly $\mathbb{Q}_p$ is not the same as $\mathbb{R}$, but can furnish a replacement for $\mathbb{R}$.

2.5 The Most Appropriate p

The choice of p is a practical issue. Adelic numbers use all possible values of p (see [9] for extensive use and discussion of the adelic number framework). Consider the choice of most useful p in practice. It should not be an overly direct and immediate representing of observational data in p-adic form (analogous to the case of using an overly immediate representation as reals). Consider [16, 40]. 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. In [16] a 5-adic encoding is used, since 5 is a prime and thereby offers uniqueness. In [40] a 4-adic encoding is used, and a 2-adic encoding, with

the latter based on 2-digit boolean expressions for the four nucleotides (00, 01, 10, 11). A default norm is used, based on a longest common prefix – with p-adic digits from the start or left of the sequence (see section 3.5 below where this longest common prefix norm or distance is used). This is something of a straightjacket on the representation, since after all we have no particular reason to think that the start of such genetic strings (sequences, where start typically means left hand side) is particularly relevant for establishing a correspondence or otherwise characterizing the strings. (But see [18] where there is a justification based on computation on floating point numbers.) Instead we think it crucial for applications to first structure the data in accordance with our perspective – visual, based on our expertise, or whatever – and then, later in the analysis pipeline seek a p-adic encoding using this perspective. This is our point of departure in section 4.1 and elsewhere below.

2.6 Symmetry

In the following sections we will show how p-adic number representation provides a powerful means of picking out symmetries in data. We suggest that they represent the simultaneously most general and most important symmetries.

Some examples of such symmetries include:

- $\mathbb{Q}_p$ acts on some rooted tree, e.g. a tree such that each node has a constant number of (say, p−1) child nodes. Various options are feasible, just as the choice for p is not fixed a priori, as we have seen in section 2.5. We will look concretely at convenient and practical perspectives in section 4.1 below. Each element of $\mathbb{Q}_p$ corresponds to a path in such a tree. In the general case, this path may be of infinite length. In this way we identify each of the terminal nodes of the tree with a p-adic number. There is symmetry therefore vis-à-vis the set of terminals in the tree.

- An intriguing point symmetry is found relative to the root node of the tree in section 4.2. In a p-adic encoding, we show how the root node can be viewed as an origin, or 0 point. This allows us to consider symmetry relative to an origin defined in this way.

- By multiplying two p-adic numbers together one gets another p-adic number. Given a fixed $x \in \mathbb{Q}_p$, multiplication by x represents a symmetry of the paths in this tree and hence of the tree itself. See section 4.2 below.

- Section 5 deals with encodings in terms of wreath product groups, expressing the symmetries available in a tree. Section 6 deals with permutation symmetries.

- In sections 3.4 and 3.5 we will look at two distinct ways to induce an ultrametric topology and so we will, as such, have two distinct ways to define a norm and an (ultra)metric. Equidistance implies another level of symmetries.

3 Ultrametric Topology

3.1 From p-Adics to Ultrametrics

Having a metric on $\mathbb{Q}_p$, which follows from the valuation, allows us to consider a topology. The topology associated with the non-Archimedean valuation used on $\mathbb{Q}_p$ is an ultrametric one.

p-Adic numbers were introduced by Kurt Hensel in 1898. The ultrametric topology was introduced by Marc Krasner [42], the ultrametric inequality having been formulated by Hausdorff in 1934. Essential motivation for the study of this area is provided by [72] as follows. Real and complex fields gave rise to the idea of studying any field K with a complete valuation $|.|$ comparable to the absolute value function. Such fields satisfy the "strong triangle inequality" $|x + y| \leq \max(|x|, |y|)$. Given a valued field, defining a totally ordered Abelian (i.e. commutative) group, an ultrametric space is induced through $|x - y| = d(x, y)$. 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.

3.2 Some Properties of Ultrametric Spaces

We see from the following, based on [45] (chapter 0, part IV), that an ultrametric space is quite different from a metric one. In an ultrametric space everything "lives" on a tree.

- In an ultrametric space, all triangles are either isosceles with small base, or equilateral.
- Every point of a circle in an ultrametric space is a center of the circle.
- Two circles of the same radius, that are not disjoint, are overlapping.
- A divisor of the ultrametric space, E, is an equivalence relation D satisfying $\forall a, b, x, y, \in E : \quad aDb \;\text{ and }\; (d(x, y) \leq d(a, b)) \Longleftrightarrow xDy$.
  - Circles of the same radius form a partition of the ultrametric set. The corresponding equivalence is a divisor of the space.
  - A valuation of a divisor D of the space E is the number $\nu(D) = \sup_{xDy} d(x, y)$.
  - If D and D' are two divisors in E, a finite metric space, verifying $D \leq D'$, then $\nu(D) \leq \nu(D')$ and reciprocally.
  - If C and C' are disjoint circles in E, the distance $d(x, y)$ of an $x \in C$ and of any $y \in C'$ depends on C and C' only, and not on x and y.
  - The quotient E/D of an ultrametric space by a divisor is an ultrametric space. The distance between two of its points is strictly greater than $\nu(D)$ in the finite case.

- An ultrametric proximity is a positive (possibly infinite) function $s : E \times E \longrightarrow \mathbb{R}^+ \cup \{+\infty\}$, verifying (i) $s(y, x) = s(x, y)$, (ii) $s(x, y) = +\infty$ iff $x = y$; and (iii) $s(x, z) \geq \min(s(x, y), s(y, z))$.

  - If d is an ultrametric distance, then $-\log d$ is an ultrametric proximity. If s is an ultrametric proximity, then $\exp(-s)$ is an ultrametric distance.

- 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 E, 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

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

     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.

  See below, section 3.3, for further discussion of this expression in matrix terms of ultrametric distances.

- In an ultrametric topology, every ball is both open and closed (termed clopen).

  - The empty set and the universal set are both clopen. The complement of a clopen set is clopen. Finite unions and intersections of clopen sets are clopen.

- An ultrametric space is 0-dimensional. Informally, a set of points is of necessity 0-dimensional. Formally, from [10, 71]: a topology is 0-dimensional if it has a basis consisting of clopen sets. I.e., if for every $a \in X$ and for every closed $A \subset X$ that does not contain a, there exists a clopen set U such that $a \in U$, and $A \subset X \backslash U$. Apart from having a basis of clopen sets, a 0-dimensional topology is discrete and totally disconnected. That an ultrametric space is 0-dimensional is "in a sense the non-Archimedean analog of *completely regular*" ([71], p. 37).

3.3 Clustering Through Row/Column Permutation

In section 3.2 we have noted how an ultrametric distance allows a certain structure to be visible, subject to row and column permuting, in a matrix defined from the set of all distances. For set X, then, this matrix expresses the distance mapping of the Cartesian product, $d : X \times X \longrightarrow \mathbb{R}+$. A priori the rows and columns of the function of the Cartesian product set X with itself could be in any order. The result in section 3.2 establishes what is possible when the distance is an ultrametric one. Because the matrix (a 2-way data object) involves one *mode* (due to set X being crossed with itself; more typical is the 2-mode case where an observation set is crossed by an attribute set) it is clear that both rows and columns can be permuted to yield the *same* order on X. A property of the form of the matrix is that large values are at or near the principal diagonal.

One can adopt a matrix row/column permutation perspective on matrix diagonalization, which solves the equation $xA = \lambda x$ for vectors x, λ and matrix A. Hence $xA = \text{diag}(\lambda)x$ where diag is a diagonal matrix. The vector of eigenvalues λ can be ordered, and hence the matrix diag(λ) can be ordered by row and column, with the same order for both. One can approximate, "non-destructively" but not necessarily well, this diagonal matrix by taking A and reordering iteratively so that large elements are concentrated on the central diagonal. See [14, 57].

Two generalizations are clear for this sort of clustering by visualization scheme. Firstly, we can directly apply row and column permuting to 2-mode data, i.e. to the rows and columns of a matrix crossing indices I by attributes J, $a : I \times J \longrightarrow R$. Function a gives us a matrix, say a of terms $a(i, j)$ where here, as an example, each such term is real-valued. Secondly we can generalize the principle of permuting such that large values are on or near the principal diagonal to instead allow large values to be near one another, and thereby to facilitate visualization.

An optimized way to do this was pursued in [11, 48].

Comprehensive surveys of clustering algorithms in this area, including objective functions, visualization schemes, optimization approaches, presence of constraints, and applications, can be found in [50, 47].

For all these approaches, underpinning them are row and column permutations, that can be expressed in terms of the permutation group, S_n, on n elements.

3.4 Ultrametric Topology induced from Pairwise Dissimilarities

3.4.1 Pairwise Dissimilarities

Given an observation set, X, we define dissimilarities as the mapping $d : X \times X \longrightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ are the positive reals. A dissimilarity is a positive, definite, symmetric measure (i.e., $d(x, y) \geq 0; d(x, y) = 0$ if $x = y; d(x, y) = d(y, x)$). If in addition the triangular inequality is satisfied (i.e., $d(x, y) \leq$

$d(x,z) + d(z,y), \forall x, y, z \in X$) then the dissimilarity is a distance.

3.4.2 From Dissimilarities to an Ultrametric

If X is endowed with a metric, then we now describe how this metric is mapped onto an ultrametric. In practice, there is no need for X to be endowed with a metric. Instead a dissimilarity is satisfactory.

A hierarchy, H, is defined as a binary, rooted, node-ranked tree, also termed a dendrogram [7, 33, 45, 57]. A hierarchy defines a set of embedded subsets of a given set of objects X, indexed by the set I. These subsets are totally ordered by an index function ν, which is a stronger condition than the partial order required by the subset relation. A bijection exists between a hierarchy and an ultrametric space.

Let us show these equivalences between embedded subsets, hierarchy, and binary tree, through the constructive approach of inducing H on a set I.

Hierarchical agglomeration on n observation vectors with indices $i \in I$ involves a series of $1, 2, \ldots, n-1$ pairwise agglomerations of observations or clusters, with the following properties. A 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 $q' \subset q$. Here we have denoted the power set of set I by 2^I. An 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) $q \subset q' \Longrightarrow \nu(q) < \nu(q')$. Here we have denoted the positive reals, including 0, by $\mathbb{R}^+$. 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. In practice, we start with a Euclidean or alternative dissimilarity, use some criterion such as minimizing the change in variance resulting from the agglomerations, and then define $\nu(q)$ as the dissimilarity associated with the agglomeration carried out.

3.4.3 Remarks on Hierarchical Agglomerative Clustering

Since pairwise dissimilarities are used in constructing the hierarchy, the computation complexity of hierarchical clustering is $O(n^2)$. As the closest clusters (including singletons) are agglomerated at each of $n-1$ agglomerations (card X = card $I = n$), the newly created cluster must be related to others. This is part and parcel of the agglomeration criterion, and can be viewed either as the cluster update rule, or the agglomerative criterion (e.g., based on compactness, or connectivity).

The most efficient algorithms are based on nearest neighbor chains, which by definition end in a pair of agglomerable reciprocal nearest neighbors. The uniqueness and acceptability of on-the-fly agglomeration based on reciprocal nearest neighbors can be proven (resp. disproven) for the given agglomerative criterion. The reciprocal nearest neighbor algorithm was first proposed in two articles in the journal *Les Cahiers de l'Analyse des Données* in 1980 and 1982,

Table 1: Example dataset: 5 objects, boolean 3 attributes.

	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

and are now used in software packages such as Clustan and R. Further information can be found in [54, 55, 57, 58].

3.5 Generalized Ultrametric

In this subsection, we consider an ultrametric defined on the powerset or join semilattice. Comprehensive background on ordered sets and lattices can be found in [13]. A comprehensive review of generalized distances and ultrametrics can be found in [74].

3.5.1 Link with Formal Concept Analysis

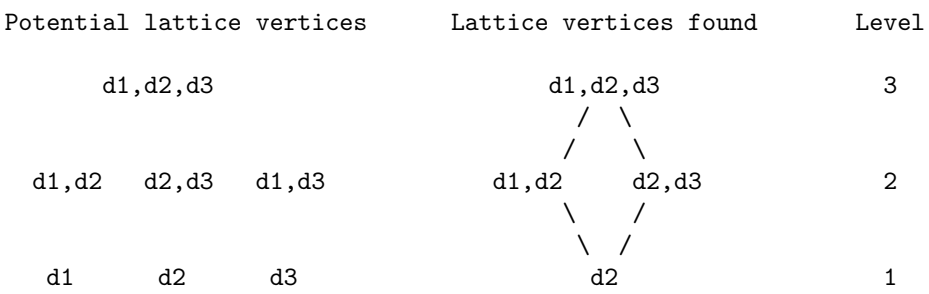
As noted in section 3.4, typically hierarchical clustering is based on a distance (which can be relaxed often to a dissimilarity, not respecting the triangular inequality, and *mutatis mutandis* to a similarity), defined on all pairs of the object set: $d : I \times I \rightarrow \mathbb{R}^+$. I.e., a distance is a positive real value. Usually we require that a distance cannot be 0-valued unless the objects are identical. That is the traditional approach.

A different form of ultrametrization is achieved from a dissimilarity defined on the power set of attributes characterizing the observations (objects, individuals, etc.) X. Here we have: $d : X \times X \longrightarrow 2^J$, where J indexes the attribute (variables, characteristics, properties, etc.) set.

We consider a different notion of distance, that maps pairs of objects onto elements of a join semilattice. 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.

As an example, consider, say, $n = 5$ objects characterized by 3 boolean (presence/absence) attributes, shown in Table 1.

Define dissimilarity between a pair of objects in Table 1 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. This is the simple matching coefficient [32]. We could use, e.g., Euclidean distance for each of the values sought; but we prefer to treat 0 values in both components as signaling a 0 contribution. We get then:



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, e
c, e

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

Figure 1: Lattice and its interpretation, corresponding to the data shown in Table 1 with the simple matching coefficient used. (See text for details.)

$d(a, b) = 1, 1, 0$
$d(a, c) = 0, 1, 0$
$d(a, e) = 0, 1, 1$
$d(a, f) = 1, 1, 0$
$d(b, c) = 1, 1, 0$
$d(b, e) = 1, 1, 1$
$d(b, f) = 1, 1, 0$
$d(c, e) = 0, 1, 1$
$d(c, f) = 1, 1, 0$
$d(e, f) = 1, 1, 1$

If we take the three components in this distance as $d1, d2, d3$, and considering a lattice representation with linkages between all ordered subsets where the subsets are to be found in our results above (e.g., $d(c, f) = 1, 1, 0$ implies that we have a lattice node associated with the subset $d1, d2$), and finally such that the order is defined on subset cardinality, then we see that the representation shown in Figure 1 suffices.

In Formal Concept Analysis [13, 25], it is the lattice itself which is of primary interest. In [32] there is discussion of, and a range of examples on, the close

relationship between the traditional hierarchical cluster analysis based on $d : I \times I \rightarrow \mathbb{R}^+$, and hierarchical cluster analysis "based on abstract posets" (a poset is a partially ordered set), based on $d : I \times I \rightarrow 2^J$. The latter, leading to clustering based on dissimilarities, was developed initially in [31].

3.5.2 Applications of Generalized Ultrametrics

As noted in the previous subsection, the usual ultrametric is an ultrametric distance, i.e. for a set I, $d : I \times I \longrightarrow \mathbb{R}$ (so the ultrametric distance is a real value). The generalized ultrametric is: $d : I \times I \longrightarrow \Gamma$, where Γ is a partially ordered set. In other words, the *generalized* ultrametric distance is a set. With this set one can have a value, so the usual and the generalized ultrametrics can amount to more or less the same in practice (by ignoring the set and concentrating on its associated value). After all, in a dendrogram one does have a set associated with each ultrametric distance value (and this is most conveniently the terminals dominated by a given node; but we could have other designs, like some representative subset or other, of these terminals). Remember that the set, Γ, is defined from the original attributes (which we denote by the set J); whereas the sets of observations read off a dendrogram are subsets of the observation set (which we label with the index set I). So $\Gamma = 2^J$ (and not 2^I).

In the theory of reasoning, a monotonic operator is rigorous application of a succession of conditionals (sometimes called consequence relations). However: "In order to deal with programs of a more general kind (the so-called disjunctive programs) it became necessary to consider multi-valued mappings", supporting non-monotonic reasoning in the way now to be described ([65], pp. 10, 13). The novelty in the work of [65, 66] is that these authors use the generalized ultrametric as a multivalued mapping. (A more critical view of the usefulness of the generalized ultrametric perspective is presented by [43].)

The generalized ultrametric approach has been motivated [73] as follows. "Situations arise ... in computational logic in the presence of negations which force non-monotonicity of the operators involved". To address non-monotonicity of operators, one approach has been to employ metrics in studying some problematic logic programs. These ideas were taken further in examining quasi-metrics, and generalized ultrametrics i.e. ultrametrics which take values in an arbitrary partially ordered set (not just in the non-negative reals). Seda and Hitzler [73] "consider a natural way of endowing Scott domains [see [13]] with generalized ultrametrics. This step provides a technical tool [for finding fixpoints – hence for analysis] of non-monotonic operators arising out of logic programs and deductive databases and hence to finding models for these."

A further, similar, viewpoint is [28]: "Once one introduces negation, which is certainly implied by the term *enhanced syntax* ... then certain of the important operators are not monotonic (and therefore not continuous), and in consequence the Knaster-Tarski theorem [i.e. for fixed points; again see [13]] is no longer applicable to them. Various ways have been proposed to overcome this problem. One such [approach is to use] syntactic conditions on programs ... Another is to consider different operators ... The third main solution is to introduce

techniques from topology and analysis to augment arguments based on order ... [latter include:] methods based on metrics ... on quasi-metrics ... and finally ... on ultrametric spaces."

The convergence to fixed points that are based on a generalized ultrametric system is precisely the study of spherically complete systems and expansive automorphisms discussed in section 4.5 below. As expansive automorphisms we see here again an example of symmetry at work.

3.5.3 Application to Data Mining

The potentially huge advantage of the generalized ultrametric is that it allows a hierarchy to be read directly off the $I \times J$ input data, and bypasses the $O(n^2)$ consideration of all pairwise distances in agglomerative hierarchical clustering. In [62] we study application to chemoinformatics. Proximity and best match finding is an essential operation in this field. Typically we have one million chemicals upwards, characterized by an approximate 1000-valued attribute encoding.

We are pursuing other work using the Sloan Digital Sky Survey (SDSS) archive, with both (high quality, more costly to collect) spectroscopic and (lower quality, more readily available) photometric redshifts, that will be reported on in due course. Typically in this case we are dealing with millions of objects in a low dimensional attribute space.

Implementation-wise the generalized ultrametric is potentially related to a k-d tree, which forms clusters that are comprised of approximately equal cardinalities (arranged through divisive cluster definition using repeated median splits on coordinate axes), and that have boundaries that run parallel to the given coordinate axes. The k-d, or multidimensional binary search, tree can be used for constant (or $O(1)$) expected time nearest neighbor search [6]. Here we are using clusters of fixed increments on the coordinate axes. This was used for expediting nearest neighbor finding, leading to a computationally very efficient implementation of hierarchical clustering algorithms in [53].

We arrive at an interesting perspective: we have a tree structure associated with a generalized ultrametric, and then we use it to expedite building a traditional ultrametric-based agglomerative hierarchical clustering. Since our motivation lies in finding interpretable and exploitable structure in data, it follows that either, or hybrids of both, forms of ultrametrization that are examined in this section, section 3.5, could be of benefit to us.

4 Dendrogram or Binary, Rooted, Possibly Labeled, Tree

A dendrogram is widely used in hierarchical, agglomerative clustering, and is induced from observed data. In this article, one of our important goals is to show how it lays bare many diverse symmetries in the observed phenomenon represented by the data. In section 3 we have drawn links with the practical

application of dendrograms in data analysis and data mining, and with the abundant literature in this field.

4.1 p-Adic Encoding of a Dendrogram

We will 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 $\mathbb{Q}_p$). The field of p-adic numbers is the most important example of ultrametric spaces. Addition and multiplication of p-adic integers, $\mathbb{Z}_p$ (cf. expression in subsection 2.3), are well-defined. Inverses exist and no zero-divisors exist.

A terminal-to-root traversal in a dendrogram or binary rooted tree is defined as follows. We 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.

A terminal-to-root traversal is the shortest path between the given terminal node and the root node, assuming we preclude repeated traversal (backtrack) of the same path between any two nodes.

By means of terminal-to-root traversals, we define the following p-adic encoding of terminal nodes, and hence objects, in Figure 2.

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

If we choose $p = 2$ the resulting decimal equivalents could be the same: cf. contributions based on $+1 \cdot p^1$ and $-1 \cdot p^1 + 1 \cdot p^2$. Given that the coefficients of the p^j terms ($1 \leq j \leq 7$) are in the set $\{-1, 0, +1\}$ (implying for x_1 the additional terms: $+0 \cdot p^3 + 0 \cdot p^4 + 0 \cdot p^6$), the coding based on $p = 3$ is required to avoid ambiguity among decimal equivalents.

A few general remarks on this encoding follow. For the labeled ranked binary trees that we are considering, 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.

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:*

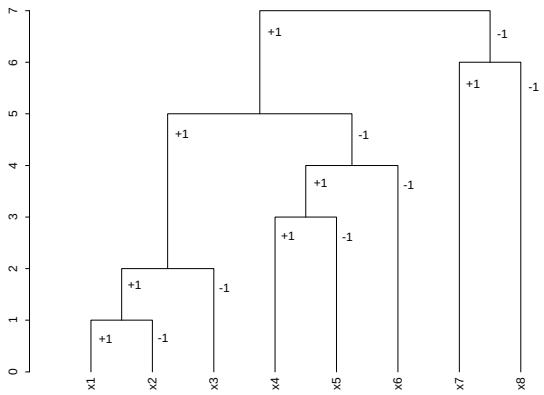


Figure 2: Labeled, ranked dendrogram on 8 terminal nodes, $x_1, x_2, \ldots, x_8$. Branches are labeled $+1$ and -1. Clusters are: $q_1 = \{x_1, x_2\}, q_2 = \{x_1, x_2, x_3\}, q_3 = \{x_4, x_5\}, q_4 = \{x_4, x_5, x_6\}, q_5 = \{x_1, x_2, x_3, x_4, x_5, x_6\}, q_6 = \{x_7, x_8\}, q_7 = \{x_1, x_2, \ldots, x_7, x_8\}$.

the dendrogram can be uniquely reconstructed from any such set of unique codes.

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

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

In greater detail we have:

$$x_i = \sum_{j=1}^{n-1} c_{ij} p^j \text{ where } c_{ij} \in \{-1, 0, +1\} \tag{3}$$

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

In our example we have used: $a_j = +1$ for a left branch (in the sense of Figure 2), $= -1$ for a right branch, and $= 0$ when the node is not on the path from that particular terminal to the root.

A matrix form of this encoding is as follows, where $\{\cdot\}^t$ denotes the transpose of the vector.

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 Figure 2 we therefore have:

$$C = \{c_{ij}\} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{pmatrix} \tag{4}$$

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.

The matrix form of the p-adic encoding used in equations (2) or (3) is:

$$\mathbf{x} = C\mathbf{p} \tag{5}$$

Here, $\mathbf{x}$ is the decimal encoding, C is the matrix with dendrogram branching codes (cf. example shown in expression (4)), and $\mathbf{p}$ is the vector of powers of a fixed integer (usually, more restrictively, fixed prime) p.

The tree encoding exemplified in Figure 2, and defined with coefficients in equations (2) or (3), (4) or (5), with labels $+1$ and -1 was 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).

We can consider the objects that we are dealing with to have equivalent integer values. To show that, all we must do is work out decimal equivalents of the p-adic expressions used above for $x_1, x_2, \ldots$. As noted in [26], we have equivalence between: a p-adic number; a p-adic expansion; and an element of $\mathbb{Z}_p$ (the p-adic integers). The coefficients used to specify a p-adic number, [26] notes (p. 69), "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."

4.2 P-adic Dendrogram Addition and Multiplication, Distance and Norm

The addition operation on the group of dendrograms, H, will now be explored. In order to define a group structure on the p-adic encoded objects, we require an addition operation. We do not "carry and add" in the traditional way because this does not make sense in this context. Instead we define the following "average and threshold" operation for any coefficients (of values of $\mathbf{p}$, as used

in expressions (3) or (5) above). We define the following compositions for such coefficients.

$$\begin{array}{cccccc} + & 1 & + & 1 & \longrightarrow & +1 \\ - & 1 & - & 1 & \longrightarrow & -1 \\ + & 1 & - & 1 & \longrightarrow & 0 \\ - & 1 & + & 1 & \longrightarrow & 0 \\ + & 1 & \pm & 0 & \longrightarrow & 0 \\ - & 1 & \pm & 0 & \longrightarrow & 0 \end{array} \tag{6}$$

Examples from the encoding defined above for $x_1, x_2, \ldots$ (again with reference to Figure 2, and equations (2) or (3), (4) or (5)) follow. The compositions of (6) lead to the $\oplus$ addition operation on the objects, $\{x_i, i \in I\}$.

$x_1 \oplus x_2 = +1 \cdot p^2 + 1 \cdot p^5 + 1 \cdot p^7$
$x_1 \oplus x_3 = +1 \cdot p^5 + 1 \cdot p^7$
$x_1 \oplus x_7 = 0$
$x_3 \oplus x_6 = +1 \cdot p^7$
$x_5 \oplus x_8 = 0$

Informally: in the tree, this addition operation only retains non-zero terms for nodes in the tree strictly *above* the first (i.e. lowest level) cluster within which the two objects find themselves. This means that if the two objects only find themselves together for the first time in the same cluster that contains all objects then the result of the addition operation is 0.

In this p-adic perspective we have the following symmetry result. All terms are symmetric vis-à-vis the root node.

Let us use our "average and threshold" operation, which we are using as a customized addition, to define clusters. We will do so by example, taking Figure 2 as our case study. We will call the clusters, ranked by increasing node level, $q_1, q_2, \ldots$ as used in the caption of Figure 2.

$q_1 = x_1 \oplus x_2 = +1 \cdot p^2 + 1 \cdot p^5 + 1 \cdot p^7$
$q_2 = q_1 \oplus x_3 = +1 \cdot p^5 + 1 \cdot p^7$
$q_3 = x_4 \oplus x_5 = +1 \cdot p^4 - 1 \cdot p^5 + 1 \cdot p^7$
$q_4 = q_3 \oplus x_6 = -1 \cdot p^5 + 1 \cdot p^7$
$q_5 = q_2 \oplus q_4 = +1 \cdot p^7$
$q_6 = x_7 \oplus x_8 = -1 \cdot p^7$
$q_7 = 0$

The trivial cluster containing all n objects, q_{n-1}, is of value 0 in this representation.

Definition of Null Element: On the dendrogram H, the set $q_{n-1} = I$ is the null element when using our p-adic encoding (given in definitions (3) and (5)) and addition operation (6).

Defining p-adic notation for clusters in this way allows us to define norms of clusters; or to define p-adic distances between clusters; or indeed to define

p-adic distances between clusters and objects (singletons, terminals). We will look at these in subsection 4.3 below.

For completeness we will provide a definition of p-adic dendrogram multiplication. Take $x = \sum_j c_j p^j$ and let $y = \sum_j c'_j p^j$. The product operation is defined on the formal (Laurent) power series as:

$$xy = \left(\sum_j c_j p^j \right) \left(\sum_{j'} c'_{j'} p^{j'} \right) = \sum_{jj'} c_j c'_{j'} p^{j+j'} \tag{7}$$

with restriction to the term in p^{n-1}. P-adic dendrogram multiplication will be used below in the definition of the expansive operator: this is multiplication by $1/p$.

An upshot of the discussion in this subsection is that addition and multiplication on trees is possible, so we therefore have symmetries relative to application of these operations.

4.3 P-adic Distance and Norm on a Dendrogram

Thus far, we have been concerned with an analytic framework. Now we will 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.

To find the p-adic distance, 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 absolute values of the coefficients of p^r are equal.

Let us look at the set of p-adic codes for $x_1, x_2, \ldots$ above (Figure 2), to give some examples of this.

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} [7, 26].
Examples based on Figure 2 now follow.

$|x_1 - x_2|_p = |x_2 - x_1|_p = p^{-1}$ since $r = 1$.
$|x_1 - x_4|_p = |x_4 - x_1|_p = p^{-5}$ since $r = 5$.
$|x_3 - x_6|_p = |x_6 - x_3|_p = p^{-5}$ since $r = 5$.

Examples for clusters from Figure 2:

$|q_1 - q_3|_p = |q_3 - q_1|_p = p^{-5}$.
$|q_2 - q_6|_p = |q_6 - q_2|_p = p^{-7}$.

We take for a singleton object $r = 0$, and so the norm of an object is always 1. We therefore define the p-adic norm, $|.|_p$, of an object corresponding to a terminal node in the following way: for any object, x, $|x|_p = 1$.

The norm of a non-singleton cluster is defined analogously. It is seen to be strictly smaller. We have: $|q_2|_p = p^{-2}; |q_4|_p = p^{-4}$. So $|q|_p \leq 1$ with equality only if q is a singleton.

The p-adic norm, or p-adic valuation, satisfies the following properties [72]: (1) $|x|_p \geq 0; |x|_p = 0$ iff $x = 0$; (2) $|x + y|_p \leq \max(|x|_p, |y|_p)$; and (3) $|xy|_p = |x|_p|y|_p$.

4.4 Application to p-Adic Clustering

In the previous subsection we have seen that encoding the data p-adically amounts to having a hierarchical, or rooted tree, representation of the data.

An alternative way of looking at this, from the p-adic expansions listed in relations (2), is as follows. Consider the longest common sequence of coefficients using terms of the expansion, and starting at the root. Determine the p^r term at which the coefficients first differ. Then the distance is defined as p^{-r}.

This longest common prefix metric is also known as the Baire distance. In topology the Baire metric is defined on infinite strings [46]. 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. The use of this Baire metric is pursued in [62] based on random projections [81], and providing computational benefits over the classical $O(n^2)$ hierarchical clustering based on all pairwise distances.

The longest common prefix metric leads directly to a *p-adic hierarchical classification* (cf. [8]). This is a special case of the "fast" hierarchical clustering discussed in section 3.5. The following summary is provided by [8]: "... p-adic classification is algorithmically much simpler than its classical counterpart [i.e., based on all pairwise distances]. The consequence for data mining lies in the shift from classification to data encoding."

Compared to the longest common prefix metric, there are other closely related forms of metric, and simultaneously ultrametric.

In [24], the metric is defined via the integer part of a real number: $d(x, y) = \inf\{2^{-r} : r \in \mathbb{Z}, 2^i(x - e) = 2^i(y - e)\}$ where e is any irrational number and d is an ultrametric.

In [7], for integers x, y we have: $d(x, y) = 2^{-\text{order}_p(x-y)}$ where p is prime, and $\text{order}_p(i)$ is the exponent (non-negative integer) of p in the prime decomposition of an integer.

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 distance between series is: $d(S, S') = 2^{-\text{order}(S-S')}$.

These [24, 7] (ultra)metrics are all viable alternatives but, from a practical standpoint (cf. [62]), the longest common prefix is just fine for us, as an ultrametric yielding a p-adic clustering, and in view of its direct applicability.

4.5 Dilation Operation: p-Adic Multiplication by $1/p$

Scale-related symmetry is very important in practice. In this subsection we introduce an operator that provides this symmetry. We also term it a dilation operator, because of its role in the wavelet transform on trees (see [60] for discussion and examples).

Consider the set of objects $\{x_i | i \in I\}$ with its p-adic coding considered above. 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.) 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: $+1 \cdot 2^1 + 1 \cdot 2^4 + 1 \cdot 2^6$. Each level has decreased by one, and the lowest level has been lost. Subject to the lowest level of the tree being lost, the form of the tree remains the same. 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.

Let us call product with $1/p$ the operator A. 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. Therefore the application of A to all q implies a subset relationship between the set of clusters $\{q\}$ and the result of applying A, $\{Aq\}$.

Repeated application of the operator A gives Aq, A^2q, A^3q, ... Starting with any singleton, $i \in I$, this gives a path from the terminal to the root node in the tree. Each such path ends with the null element, as a result of the Null Element definition (subsection 4.2). Therefore the intersection of the paths equals the null element.

Benedetto and Benedetto [4, 5] discuss A as an expansive automorphism of I, i.e. form-preserving, and locally expansive. Some implications [4] of the expansive automorphism follow. 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 ([70], p. 131).

Each path defined by application of the expansive automorphism defines a spherically complete system [72, 24, 71], which is a formalization of well-defined subset embeddedness.

5 Tree Symmetries through the Wreath Product Group

A dendrogram like that shown in Figure 2 is invariant relative to rotation (alternatively, here: permutation) of left and right child nodes. These rotation (or permutation) symmetries are defined by the wreath product group (see [21, 22, 19] for an introduction and applications in signal and image processing), and can be used with any m-ary tree, although we will treat the binary case here.

For the group actions, with respect to which we will seek invariance, we consider independent cyclic shifts of the subnodes of a given node (hence, at each level). Equivalently these actions are adjacency preserving permutations

of subnodes of a given node (i.e., for given q, with $q = q' \cup q''$, the permutations of $\{q', q''\}$). We have therefore cyclic group actions at each node, where the cyclic group is of order 2.

The symmetries of H are given by structured permutations of the terminals. The terminals will be denoted here by Term H. The full group of symmetries is summarized by the following generative algorithm:

1. For level $l = n - 1$ down to 1 do:
2. Selected node, $\nu \longleftarrow$ node at level l.
3. And permute subnodes of ν.

Subnode ν is the root of subtree H_ν. We denote H_{n-1} simply by H. For a subnode ν' undergoing a relocation action in step 3, the internal structure of subtree $H_{\nu'}$ is not altered.

The algorithm described defines the automorphism group which is a wreath product of the symmetric group. 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.

Call Term H_ν the terminals that descend from the node at level ν. So these are the terminals of the subtree H_ν with its root node at level ν. We can alternatively call Term H_ν the cluster associated with level ν.

We will now look at shift invariance under the group action. This amounts to the requirement for a constant function defined on Term $H_\nu, \forall \nu$. A convenient way to do this is to define such a function on the set Term H_ν via the root node alone, ν. By definition then we have a constant function on the set Term H_ν.

Let us call V_ν a space of functions that are constant on Term H_ν. Possible bases of V_ν that were considered in [60] are:

1. Basis vector with $|\text{Term}H_{n-1}|$ components, with 0 values except for value 1 for component i.
2. Set (of cardinality $n = |\text{Term}H_{n-1}|$) of m-dimensional observation vectors.

The constant function for each node or level ν is:

$$L : \text{Term}H_\nu \longrightarrow V_\nu$$

Consider the resolution scheme arising from moving from
Term $H_{\nu'}$, Term $H_{\nu''}\}$ to Term H_ν. From the hierarchical clustering point of view it is clear what this represents, simply, an agglomeration of two clusters called Term $H_{\nu'}$ and Term $H_{\nu''}$, replacing them with a new cluster, Term H_ν.

Let the spaces of constant functions corresponding to the two cluster agglomerands be denoted $V_{\nu'}$ and $V_{\nu''}$. These two clusters are disjoint initially,

which motivates us taking the two spaces as a couple: $(V_{\nu'}, V_{\nu''})$. In the same way, let the space of constant functions corresponding to node ν be denoted V_ν.

Let us exemplify a case that satisfies all that has been defined in the context of the wreath product invariance that we are targeting. It is the algorithm discussed in depth in [60]. Take the constant function on $V_{\nu'}$ to be $f_{\nu'}$. Take the constant function on $V_{\nu''}$ to be $f_{\nu''}$. Then define the constant function, the *scaling function*, on V_ν to be $(f_{\nu'} + f_{\nu''})/2$. Next define the zero mean function, $(w_{\nu'} + w_{\nu''})/2 = 0$, the *wavelet function*, as follows:

$$w_{\nu'} = (f_{\nu'} + f_{\nu''})/2 - f_{\nu'}$$

in the support interval of $V_{\nu'}$, i.e. Term $H_{\nu'}$, and

$$w_{\nu''} = (f_{\nu'} + f_{\nu''})/2 - f_{\nu''}$$

in the support interval of $V_{\nu''}$, i.e. Term $H_{\nu''}$.

Since $w_{\nu'} = -w_{\nu''}$ we have the zero mean requirement.

6 Tree Symmetries through Permutation Groups

6.1 Ordinal Encodings in Symbolic Dynamics

In symbolic dynamics, we seek to extract symmetries in the data based on topology alone, before considering metric properties. For example, instead of listing a sequence of iterates, $\{x_i\}$, we may symbolically encode the sequence in terms of up or down, or north, south, east and west moves. This provides a sequence of symbols, and their patterns in a phase space, where the interest of the data analyst lies in a partition of the phase space. Patterns or templates are sought in this topology. Sequence analysis is tantamount to a sort of topological time series analysis.

Thus, in symbolic dynamics, the data values in a stream or sequence are replaced by symbols to facilitate pattern-finding, in the first instance, through topology of the symbol sequence. This can be very helpful for analysis of a range of dyamical systems, including chaotic, stochastic, and deterministic-regular time series. Through measure-theoretic or Kolmogorov-Sinai entropy of the dynamical system, it can be shown that the maximum entropy conditional on past values is consistent with the requirement that the symbol sequence retains as much of the original data information as possible. Alternative approaches to quantifying complexity of the data, expressing the dynamical system, is through Lyapanov exponents and fractal dimensions, and there are close relationships between all of these approaches [44].

From the viewpoint of practical and real-world data analysis, however, many problems and open issues remain. Firstly, noise in the data stream means that reproducibility of results can break down [2]. Secondly, the symbol sequence, and derived partitions that are the basis for the study of the symbolic dynamic topology, are not easy to determine. Hence [2] enunciate a pragmatic principle,

whereby the symbol sequence should come as naturally as possible from the data, with as little as possible by way of further model assumptions. Their approach is to define the symbol sequence through (i) comparison of neighboring data values, and (ii) up-down or down-up movements in the data stream.

Taking into account all up-down and down-up movements in a signal allows a permutation representation.

Examples of such symbol sequences from [2] follow. They consider the data stream $(x_1, x_2, \ldots, x_7) = (4, 7, 9, 10, 6, 11, 3)$. Take the order as 3, i.e. consider the up-down and down-up properties of successive triplets. $(4, 7, 9) \longrightarrow 012; (7, 9, 10) \longrightarrow 012; (9, 10, 6) \longrightarrow 201; (6, 11, 3) \longrightarrow 201; (10, 6, 11) \longrightarrow 102$. (In the last, for instance, we have $x_{t+1} < x_t < x_{t+2}$, yielding the symbolic sequence 102.) In addition to the order, here 3, we may also consider the delay, here 1. In general, for delay τ, the neighborhood consists of data values indexed by $t, t-\tau, t-2\tau, t-3\tau, \ldots, t-d\tau$ where d is the order. Thus, in the example used here, we have the symbolic representation 012012201201102. The symbol sequence (or "itinerary") defines a partition – a separation of phase space into disjoint regions (here, with three equivalence classes, 012, 201, and 102), which facilitates finding an "organizing template" or set of topological relationships [82]. The problem is described in [34] as one of studying the qualitative behavior of the dynamical system, through use of a "very coarse-grained" description, that divides the state space (or phase space) into a small number of regions, and codes each by a different symbol.

Different encodings are feasible and [37, 36] use the following. Again consider the data stream $(x_1, x_2, \ldots, x_7) = (4, 7, 9, 10, 6, 11, 3)$. Now given a delay, $\tau = 1$, we can represent the above by $(x_{6\tau}, x_{5\tau}, x_{4\tau}, x_{3\tau}, x_{2\tau}, x_\tau, x_0)$. Now look at rank order and note that: $x_\tau > x_{3\tau} > x_{4\tau} > x_{5\tau} > x_{2\tau} > x_{6\tau} > x_0$. We read off the final permutation representation as (1345260). There are many ways of defining such a permutation, none of them best, as [37] acknowledge. We see too that our m-valued input stream is a point in $\mathbb{R}^m$, and our output is a permutation $\pi \in S_m$, i.e. a member of the permutation group.

Keller and Sinn [37] explore invariance properties of the permutations expressing the ordinal, symbolic coding. Resolution scale is introduced through the delay, τ. (An alternative approach to incorporating resolution scale is used in [12], where consecutive, sliding-window based, binned or averaged versions of the time series are used. This is not entirely satisfactory: it is not robust and is very dependent on data properties such as dynamic range.) Application is to EEG (univariate) signals (with some discussion of magnetic resonance imaging data) [35]. Statistical properties of the ordinal transformed data are studied in [3], in particular through the S_3 symmetry group. We have noted the symbolic dynamics motivation for this work; in [1] and other work, motivation is provided in terms of rank order time series analysis, in turn motivated by the need for robustness in time series data analysis.

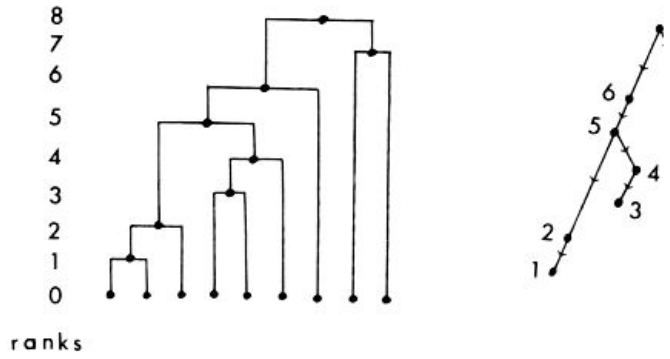


Figure 3: Left: dendrogram with lower ranked subtree always to the left. Right: oriented binary tree associated with the non-terminal nodes.

6.2 Permutation Representation of a Hierarchy

There is an isomorphism between the class of hierarchic structures, termed unlabeled, ranked, binary, rooted trees, and the class of permutations used in symbolic dynamics. Each non-terminal node in the tree shown in Figure 3 has two child nodes. This is a dendrogram, representing a set of $n-1$ agglomerations based on n initial data vectors. A packed representation [75] or permutation representation of a dendrogram is derived as follows. Put a lower ranked subtree always to the left; and read off the oriented binary tree on non-terminal nodes (see left and then right parts of Figure 3). 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. For the dendrogram shown, the packed representation is: (125346879). This is also an inorder traversal of the oriented binary tree. The packed representation is a uniquely defined permutation of $1 \ldots n$. Dendrograms (on n terminals) of the sort shown in Figure 3, referred to as non-labeled, ranked (NL-R) in [56], are isomorphic to either down-up permutations, or up-down permutations (both on $n-1$ elements). For its combinatorial properties see A000111 at [77].

We see therefore how we are dealing with the group of up-down or down-up permutations.

7 Remarkable Symmetries in Very High Dimensional Spaces

In the work of [68, 69] it was shown how as ambient dimensionality increased distances became more and more ultrametric. That is to say, a hierarchical embedding becomes more and more immediate and direct as dimensionality in-

creases. A better way of quantifying this phenomenon was developed in [59]. What this means is that there is inherent hierarchical structure in high dimensional data spaces.

It was shown in [68, 69, 59] how this is somewhat trivial: points in high dimensional spaces become increasingly equidistant with increase in dimensionality. Both [27] and [15] study Gaussian clouds in very high dimensions. The latter finds that "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".

That very simple structures come about in very high dimensions is not as trivial as it might appear at first sight. Firstly, even very simple structures (hence with many symmetries) can be used to support fast and perhaps even constant time worst case proximity search [59]. Secondly, as shown in the machine learning framework by [27], there are important implications ensuing from the simple high dimensional structures. Thirdly, [61] shows that very high dimensional clustered data contain symmetries that in fact can be exploited to "read off" the clusters in a computatationally efficient way.

8 Conclusions

"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." ([76], p. 216.) Or again: "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. ... [the work] is devoted to classical and quantum models of flows of hierarchically ordered information." ([39], pp. xiii, xv.)

We have noted symmetry in many guises in the representations used, in the transformations applied, and in the transformed outputs. These symmetries are non-trivial too, in a way that would not be the case were we simply to look at classes of a partition and claim that cluster members were mutually similar in some way. We have seen how the p-adic or ultrametric framework provides significant focus and commonality of viewpoint.

In seeking (in a general way) and in determining (in a focused way) structure and regularity in data, we see that, in line with the insights and achievements of Klein, Weyl and Wigner, in data mining and data analysis we seek and determine symmetries in the data that express observed and measured reality.

References

[1] C. Bandt. Ordinal time series analysis. *Ecological Modelling*, 182:229–238, 2005.

[2] C. Bandt and B. Pompe. Permutation entropy: a natural complexity measure for time series. *Physical Review Letters*, 88:174102(4), 2002.

[3] C. Bandt and F. Shiha. Order patterns in time series. Technical report, 2005. Preprint 3/2005, Institute of Mathematics, Greifswald, www.math-inf.uni-greifswald.de/∼bandt/pub.html.

[4] J.J. Benedetto and R.L. Benedetto. A wavelet theory for local fields and related groups. *The Journal of Geometric Analysis*, 14:423–456, 2004.

[5] R.L. Benedetto. Examples of wavelets for local fields. In D. Larson C. Heil, P. Jorgensen, editor, *Wavelets, Frames, and Operator Theory, Contemporary Mathematics Vol. 345*, pages 27–47. 2004.

[6] J.L. Bentley, B.W. Weide, and A.C. Yao. Optimal expected time algorithms for closest point problems. *ACM Transactions on Mathematical Software*, 6:563–580, 1980.

[7] J.-P. Benzécri. *La Taxinomie*. Dunod, Paris, 2nd edition, 1979.

[8] P.E. Bradley. Mumford dendrograms. *Computer Journal*, 2007. submitted.

[9] L. Brekke and P.G.O. Freund. p-Adic numbers in physics. *Physics Reports*, 233:1–66, 1993.

[10] P. Chakraborty. Looking through newly to the amazing irrationals. Technical report, 2005. arXiv: math.HO/0502049v1.

[11] W.T. Cormick, P.J. Schweitzer, and T.J. White. Problem decomposition and data reorganization by a clustering technique. *Operations Research*, 20:993–1009, 1982.

[12] M. Costa, A.L. Goldberger, and C.-K. Peng. Multiscale entropy analysis of biological signals. *Physical Review E*, 71:021906(18), 2005.

[13] B.A. Davey and H.A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2nd edition, 2002.

[14] S.B. Deutsch and J.J. Martin. An ordering algorithm for analysis of data arrays. *Operations Research*, 19:1350–1362, 1971.

[15] D.L. Donoho and J. Tanner. Neighborliness of randomly-projected simplices in high dimensions. *Proceedings of the National Academy of Sciences*, 102:9452–9457, 2005.

[16] B. Dragovich and A. Dragovich. A p-adic model of DNA sequence and genetic code. Technical report, 2006. arXiv:q-bio/0607018v1.

[17] A. Edalat. Domains for computation in mathematics, physics and exact real arithmetic. *Bulletin of Symbolic Logic*, 3:401–452, 1997.

[18] A. Edalat. Inaugural lecture. Technical report, 2000. http://www.doc.ic.ac.uk/∼ae/inaugural.ppt.

[19] R. Foote. An algebraic approach to multiresolution analysis. *Transactions of the American Mathematical Society*, 357:5031–5050, 2005.

[20] R. Foote. Mathematics and complex systems. *Science*, 318:410–412, 2007.

[21] R. Foote, G. Mirchandani, D. Rockmore, D. Healy, and T. Olson. A wreath product group approach to signal and image processing: Part I – multiresolution analysis. *IEEE Transactions on Signal Processing*, 48:102–132, 2000.

[22] R. Foote, G. Mirchandani, D. Rockmore, D. Healy, and T. Olson. A wreath product group approach to signal and image processing: Part II – convolution, correlations and applications. *IEEE Transactions on Signal Processing*, 48:749–767, 2000.

[23] P.G.O. Freund. p-Adic strings and their applications. In Z. Rakic B. Dragovich, A. Khrennikov and I. Volovich, editors, *Proc. 2nd International Conference on p-Adic Mathematical Physics*, pages 65–73. American Institute of Physics, 2006.

[24] L. Gajić. On ultrametric space. *Novi Sad Journal of Mathematics*, 31:69–71, 2001.

[25] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999. *Formale Begriffsanalyse. Mathematische Grundlagen*, Springer, 1996.

[26] F.Q. Gouvêa. *p-Adic Numbers: An Introduction*. Springer, 2003.

[27] P. Hall, J.S. Marron, and A. Neeman. Geometric representation of high dimensional, low sample size data. *Journal of the Royal Statistical Society B*, 67:427–444, 2005.

[28] P. Hitzler and A.K. Seda. The fixed-point theorems of Priess-Crampe and Ribenboim in logic programming. *Fields Institute Communications*, 32:219–235, 2002.

[29] A.K. Jain and R.C. Dubes. *Algorithms For Clustering Data*. Prentice-Hall, 1988.

[30] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31:264–323, 1999.

[31] M.F. Janowitz. An order theoretic model for cluster analysis. *SIAM Journal on Applied Mathematics*, 34:55–72, 1978.

[32] M.F. Janowitz. Cluster analysis based on abstract posets. Technical report, 2005–2006. http://dimax.rutgers.edu/∼melj.

[33] S.C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.

[34] K. Keller and H. Lauffer. Symbolic analysis of high-dimensional time series. *International Journal of Bifurcation and Chaos*, 13:2657–2668, 2003.

[35] K. Keller, H. Lauffer, and M. Sinn. Ordinal analysis of EEG time series. *Chaos and Complexity Letters*, 2, 2005.

[36] K. Keller and M. Sinn. Ordinal analysis of time series. *Physica A*, 356:114–120, 2005.

[37] K. Keller and M. Sinn. Ordinal symbolic dynamics. 2005. Technical Report A-05-14, www.math.mu-luebeck.de/publikationen/pub2005.shtml.

[38] A. Khrennikov. *Non-Archimedean Analysis: Quantum Paradoxes, Dynamical Systems and Biological Models*. Kluwer, 1997.

[39] A.Yu. Khrennikov. *Information Dynamics in Cognitive, Psychological, Social and Anomalous Phenomena*. Kluwer, 2004.

[40] A.Yu. Khrennikov. Gene expression from polynomial dynamics in the 2-adic information space. Technical report, 2006. arXiv:q-bio/06110682v2.

[41] F. Klein. A comparative review of recent researches in geometry. *Bull. New York Math. Soc.*, 2:215–249, 1892–1893. Vergleichende Betrachtungen über neuere geometrische Forschungen, 1872, translated by M.W. Haskell.

[42] M. Krasner. Nombres semi-réels et espaces ultramétriques. *Comptes-Rendus de l'Académie des Sciences, Tome II*, 219:433, 1944.

[43] M. Krötzsch. Generalized ultrametric spaces in quantitative domain theory. *Theoretical Computer Science*, 368:30–49, 2006.

[44] V. Latora and M. Baranger. Kolmogorov-Sinai entropy rate versus physical entropy. *Physical Review Letters*, 82:520, 1999.

[45] I.C. Lerman. *Classification et Analyse Ordinale des Données*. Dunod, Paris, 1981.

[46] A. Levy. *Basic Set Theory*. Dover, Mineola, NY, 2002. (Springer, 1979).

[47] S.C. Madeira and A.L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:24–45, 2004.

[48] S.T. March. Techniques for structuring database records. *Computing Surveys*, 15:45–79, 1983.

[49] R. Mayans. Ten ways of looking at real numbers, text and presentation. Technical report, 2005. http://alpha.fdu.edu/∼mayans/mhp_papers.html.

[50] I. Van Mechelen, H.-H. Bock, and P. De Boeck. Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research*, 13:363–394, 2004.

[51] B. Mirkin. *Mathematical Classification and Clustering*. Kluwer, 1996.

[52] B. Mirkin. *Clustering for Data Mining*. Chapman and Hall/CRC, Boca Raton, FL, 2005.

[53] F. Murtagh. Expected time complexity results for hierarchic clustering algorithms which use cluster centers. *Information Processing Letters*, 16:237–241, 1983.

[54] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. *Computer Journal*, 26:354–359, 1983.

[55] F. Murtagh. Complexities of hierarchic clustering algorithms: state of the art. *Computational Statistics Quarterly*, 1:101–113, 1984.

[56] F. Murtagh. Counting dendrograms: a survey. *Discrete Applied Mathematics*, 7:191–199, 1984.

[57] F. Murtagh. *Multidimensional Clustering Algorithms*. Physica-Verlag, Heidelberg and Vienna, 1985.

[58] F. Murtagh. Comments on: Parallel algorithms for hierarchical clustering and cluster validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:1056–1057, 1992.

[59] F. Murtagh. On ultrametricity, data coding, and computation. *Journal of Classification*, 21:167–184, 2004.

[60] F. Murtagh. The Haar wavelet transform of a dendrogram. *Journal of Classification*, 24:3–32, 2007.

[61] F. Murtagh. The remarkable simplicity of very high dimensional data: application to model-based clustering. *Journal of Classification*, 2007. Submitted.

[62] F. Murtagh, G. Downs, and P. Contreras. Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding. *SIAM Journal on Scientific Computing*, 2007. In press.

[63] A. Ostrowski. Über einige Lösungen der Funktionalgleichung $\phi(x) \cdot \phi(y) - \phi(xy)$. *Acta Math.*, 41:271–284, 1918.

[64] J. Preszler. Introduction to p-adic numbers. Technical report, 2005. http://www.math.utah.edu/∼preszler/research/Qp.pdf.

[65] S. Priess-Crampe and P. Ribenboim. Logic programming and ultrametric spaces. *Rendiconti de Matematica, Serie VII*, 19:155–176, 1999.

[66] S. Priess-Crampe and P. Ribenboim. Ultrametric spaces and logic programming. *Journal of Logic Programming*, 42:59–70, 2000.

[67] J. Ramagge. Unreal numbers. the story of p-adic numbers. Technical report, 2005. http://www.ice-em.org.au/pdfs/UnrealNumbers.pdf.

[68] R. Rammal, J.C. Angles d'Auriac, and B. Doucot. On the degree of ultrametricity. *Le Journal de Physique – Lettres*, 46:L–945–L–952, 1985.

[69] R. Rammal, G. Toulouse, and M.A. Virasoro. Ultrametricity for physicists. *Reviews of Modern Physics*, 58:765–788, 1986.

[70] H. Reiter and J.D. Stegeman. *Classical Harmonic Analysis and Locally Compact Groups*. Oxford University Press, Oxford, 2nd edition, 2000.

[71] A.C.M. Van Rooij. *Non-Archimedean Functional Analysis*. Dekker, 1978.

[72] W.H. Schikhof. *Ultrametric Calculus*. Cambridge University Press, Cambridge, 1984. (Chapters 18, 19, 20, 21).

[73] A.K. Seda and P. Hitzler. Generalized ultrametrics, domains and an application to computational logic. *Irish Mathematical Society Bulletin*, 41:31–43, 1998.

[74] A.K. Seda and P. Hitzler. Generalized distance functions in the theory of computation. *Computer Journal*, 2008. In press.

[75] R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *Computer Journal*, 16:30–34, 1980.

[76] H.A. Simon. *The Sciences of the Artificial*. MIT Press, Cambridge, MA, 1996.

[77] N.J.A. Sloane. OEIS – On-Line Encyclopedia of Integer Sequences. Technical report, 2006. http://www.research.att.com/∼njas/sequences/Seis.html, Sequence A000111: http://www.research.att.com/∼njas/sequences/A000111.

[78] D. Steinley. K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59:1–3, 2006.

[79] D. Steinley and M.J. Brusco. Initializing K-means batch clustering: a critical evaluation of several techniques. *Journal of Classification*, 24:99–121, 2007.

[80] Wu-Ki Tung. *Group Theory in Physics*. World Scientific, 1985.

[81] S.S. Vempala. *The Random Projection Method*. American Mathematical Society, 2004. Vol. 65, DIMACS Series in Discrete Mathematics and Theoretical Computer Science.

[82] W. Weckesser. Symbolic dynamics in mathematics, physics, and engineering, based on a talk by N. Tuffilaro. Technical report, 1997. http://www.ima.umn.edu/∼weck/nbt/nbt.ps.

[83] H. Weyl. *Symmetry*. Princeton University Press, 1983.

[84] Rui Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16:645–678, 2005.