On Ultrametricity, Data Coding, and Computation

Fionn Murtagh *

July 31, 2004

Abstract

The triangular inequality is a defining property of a metric space, while the stronger ultrametric inequality is a defining property of an ultrametric space. Ultrametric distance is defined from p-adic valuation. It is known that ultrametricity is a natural property of spaces in the sparse limit. The implications of this are discussed in this article. Experimental results are presented which quantify how ultrametric a given metric space is. We explore the practical meaningfulness of this property of a space being ultrametric. In particular, we examine the computational implications of widely prevalent and perhaps ubiquitous ultrametricity.

Keywords: metric, ultrametric, p-adic, isosceles, non-Archimedean, dendrogram, hierarchy, clustering, classification, unsupervised, binary tree

1 Introduction

The triangular inequality holds for a metric space: $d(x, z) \leq d(x, y) + d(y, z)$ for any triplet of points x, y, z. In addition the properties of symmetry and positive definiteness are respected. The "strong triangular inequality" or ultrametric inequality is: $d(x, z) \leq \max \{d(x, y), d(y, z)\}$ for any triplet x, y, z.

*F. Murtagh was with the School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK. He is now with the Department of Computer Science, Royal Holloway University of London, Egham, Surrey TW20 0EX, England. Email fmurtagh@acm.org

An ultrametric space implies respect for a range of stringent properties. For example, the triangle formed by any triplet is necessarily isosceles, with the two large sides equal; or is equilateral.

Ultrametricity is a natural property of high-dimensional spaces (Rammal, Toulouse and Virasoro, 1986, p. 786); and ultrametricity emerges as a consequence of randomness and of the law of large numbers (Rammal et al., 1986; Ogielski and Stein, 1985).

An ultrametric topology is associated with the p-adic numbers (Mahler, 1981; Gouvêa, 2003). Furthermore, the ultrametric inequality implies non-respect of a relation between a triplet of positive valuations termed the Archimedean inequality. Consequently, ultrametric spaces, p-adic numbers, non-Archimedean numbers, and isosceles spaces all express the same thing.

P-adic numbers were introduced by Kurt Hensel in 1898. The ultrametric topology was introduced by Marc Krasner (Krasner, 1944), the ultrametric inequality having been formulated by Hausdorff in 1934. As is well known, in clustering a bijection is defined between a rooted, binary, ranked, indexed tree, called a dendrogram, and a set of ultrametric distances (Benzécri, 1979, representing work going back to the early 1960s; Johnson, 1967).

Watson (2003) attributes to Mézard, Parisi, Sourlas, Toulouse and Virasoro (1984) the basis for take-off in interest in ultrametrics in statistical mechanics and optimization theory. Mézard et al. (1984) developed a mean-field theory of spin glasses (magnetic materials), showing that the distribution of pure states in a configuration space is ultrametric. "Frustrated optimization problems" are ultrametric, and have been shown as such for spin glass and related special cases. Parisi and Ricci-Tersenghi (2000), considering the spin glass model that has become a basic model for complex systems, state that "ultrametricity implies that the distance between the different states is such that they can be put in a taxonomic or genealogical tree such that the distance among two states is consistent with their position on the tree". An optimization process can be modeled using random walks so if local ultrametricity exists then random walks in ultrametric spaces are important (Ogielski and Stein, 1985). Further historical insight into the recent use of ultrametric spaces is provided by Rammal, Angles d'Auriac and Doucot (1985) and for linguistic research by Roberts (2001).

Essential motivation for the study of this area is provided by Schikhof (1984) 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 group, an ultrametric space is induced through $|x - y| = d(x, y)$. The natural geometric ordering of metric valuations is on the real line, whereas in the ultrametric case the natural ordering is a hierarchical tree.

P-adic numbers, which provide an analytic version of ultrametric topologies, have a crucially important property resulting from Ostrowski's theorem: Each non-trivial valuation on the field of the rational numbers is equivalent either to the absolute value function or to some p-adic valuation (Schikhof, 1984; Gouvêa, 2003). Essentially this theorem states that the rationals can be expressed in terms of reals, or p-adic numbers, and no other alternative system.

Our objectives in this work are the following:

1. We will demonstrate the pervasiveness of ultrametricity, by focusing on the fact that sparse high-dimensional data tend to be ultrametric. For this objective we need to quantify ultrametricity. Note that an algorithmically induced ultrametric determined from a non-ultrametric or near-ultrametric set of points is not necessarily unique, which implies our need to have a measure of ultrametricity that is independent of any hierarchical clustering algorithm.

2. Insofar as ultrametricity is pervasive, we look at two implications which have the goal of bypassing Bellman's (1961) "curse of dimensionality" when carrying out data analysis in very high dimensional spaces. Firstly we look at the computational advantages of carrying out operations like nearest neighbor searching in an ultrametric space. Secondly we present a short review of how and where Euclidean data can be (perhaps subset-wise) effectively ultrametrically embedded, and how this facilitates operations like nearest neighbor or best match searching.

2 Lerman's H-classifiability

2.1 Lerman's H Measure

The principle adopted in any constructive assessment of ultrametricity is to construct an ultrametric on data and see what discrepancy there is between

input data and induced ultrametric data structure. Quantifying ultrametricity using a constructive approach is less than perfect as a solution, given the potential complications arising from known problems, e.g. chaining in single link, and non-uniqueness, or even inversions, with other methods. The conclusion here is that the "measurement tool" used for quantifying ultrametricity itself occupies an overly prominent role relative to that which we seek to measure. For such reasons, we need an independent way to quantify ultrametricity. We begin with Lerman's (1981) H-classifiability index.

From the isosceles triangle principle, where $d(x,y) \neq d(y,z)$ we have $d(x,z) = \max\{d(x,y), d(y,z)\}$, it follows that the largest and second largest of the numbers $d(x,y), d(y,z), d(x,z)$ are equal. Lerman's H-classifiability measure essentially looks at how close these two numbers (largest, second largest) are. So as to avoid influence of distribution of the distance values, Lerman's measure is based on ranks (of these distances) only.

A unifying framework for pairs of objects, and the distance valuation on them, is that of a *binary relation*. On a set E, a binary relation is a *preorder* if it is reflexive and transitive; it is an *equivalence relation* if the binary relation is reflexive, transitive and symmetric; and it is an *order* if the binary relation is reflexive, transitive, and anti-symmetric. Furthermore, in each case, the binary relation can be *total*, i.e. the binary relation is defined for all $(i,j) \in E \times E$, or *partial*, if the binary relation is not defined for all pairs (i,j).

An order has defined maximum, minimum, sup and inf. An equivalence relation is used to model the classes of a partition. A preorder is used to model a set of partitions, using a lattice. An example of a preorder is the relation "Is created before or at the same time as" defined on H, a hierarchy of classes of E, i.e. $H \subset 2^E$, where 2^E is the power set of E.

With a preorder, we can associate a *preordonnance* as follows. A preordonnance on E is associated with a preorder on $E \times E$ satisfying $(i,j) = (j,i) \quad \forall (i,j) \in E \times E$. An extensive review of these topics can be found in Cailliez and Pagès (1976).

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

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

This preorder will be denoted ω_d. Two distances are equivalent on a given set E iff the preordonnances associated with each on E are identical. A total

preorder is equivalent to the definition of a partition (defining an equivalence relation on F), and to a total order on the set of classes.

A preorder $\bar{\omega}$ is called ultrametric if:

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

where r is a given integer and $\rho(x, y)$ denotes the rank of pair (x, y) for $\bar{\omega}$, defined by non-decreasing values of the distance used. A necessary and sufficient condition for a distance on E to be ultrametric is that the associated preorder (on $E \times E$, or alternatively preordonnance on E) is ultrametric. Looking again at the link between a preorder and classes defining a partition, $\forall x, y, z \in E$ s.t. $(x, y) \leq (y, z) \leq (x, z)$ we must have: $(x, z) \leq (y, z)$, i.e. (x, z) and (y, z) are in the same class of a preorder $\bar{\omega}$.

We move on now to define Lerman's H-classifiability index (Lerman, 1981), which measures how ultrametric a given metric is. Let $M(x, y, z)$ be the median pair among $\{(x, y), (y, z), (x, z)\}$ and let $S(x, y, z)$ be the highest ranked pair among this triplet. J is the set of all such triplets of E. We consider the mapping τ of all triplets J into the open interval of all pairs F for the given preorder ω defined as:

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

The range is the number of ranks bounded by, and excluding, the median and maximum. A measure of the discrepancy between preorder ω and an ultrametric preorder will be defined from a measure on all pairs F that is dependent on ω.

Given a triplet $\{x, y, z\}$ for which $(x, y) \leq (y, z) \leq (x, z)$, for preorder ω, the interval $]M(x, y, z), S(x, y, z)[$ is empty if ω is ultrametric. Relative to such a triplet, the preorder ω is "less ultrametric" to the extent that the cardinal of $]M(x, y, z), S(x, y, z)[$, defined on ω, is large. In practice we ensure that ties in the ranks, due to identically-valued distances, are taken into account, by counting ranks that are strictly between M and S.

We take J into account in order to define discrepancy between the structure of ω and the structure of an ultrametric preordonnance where $|.|$ denotes cardinality:

$$H(\omega) = \sum_J |]M(x, y, z), S(x, y, z)[|/(|F| - 3)|J|$$

The value 3 subtracted from $|F|(= n(n-1)/2$ if $|E| = n)$ takes account of the presence of the least, median and maximum distances. (Here, we differ marginally from Lerman, 1981. The subtraction of 3 will not cater though for tied values.) If ω is ultrametric then $H(\omega) = 0$.

As shown in simple cases by Lerman (1981, p. 218), data sets that are "more classifiable" in an intuitive way, i.e. they contain "sporadic islands" of more dense regions of points – a prime example is Fisher's (1936) iris data contrasted with 150 uniformly distributed values in $\mathbb{R}^4$ – such data sets have a smaller value of $H(\omega)$. For Fisher's data we find $H(\omega) = 0.0899$, whereas for 150 uniformly distributed points in a 4-dimensional hypercube, we find $H(\omega) = 0.1835$.

Generating all unique triplets is computationally intensive: for n points, $n(n-1)(n-2)/6$ triplets have to be considered. Hence, in practice, we must draw triangles randomly from the given point set. For integer indices i, j, k, we draw $i \sim [1 \ldots n-2], j \sim [i+1 \ldots n-1], k \sim [\max\,(i,j)+1 \ldots n]$ where sampling is uniform.

2.2 Rammal's Measure based on the Subdominant Ultrametric

The quantifying of how ultrametric a data set is by Rammal et al. (1985, 1986) was influential for us in this work. The Rammal ultrametricity index is given by $\sum_{x,y}(d(x,y) - d_c(x,y))/\sum_{x,y} d(x,y)$ where d is the metric distance being assessed, and d_c is the subdominant ultrametric. The Rammal index is bounded by 0 (= ultrametric) and 1. As pointed out in Rammal et al. (1985, 1986), this index suffers from "the chaining effect and from sensitivity to fluctuations". The single link hierarchical clustering method, yielding the subdominant ultrametric, is, as is well known, subject to such difficulties. For this reason we prefer the Lerman index. The latter is unbounded and, given the definition used above, we have found maximum values (i.e. greatest non-ultrametricity) in the region of 0.24. For assessing *relative* degree of ultrametricity it does its job well.

Rammal et al. (1985, 1986) discuss a range of important data analysis cases, all of which are characterized by potential sparseness of point occupancy in the ambient space: a set of n binary words, randomly defined among the 2^k possible words of k bits; and n words of k letters extracted from an alphabet of size K. For binary words, $K = 2$; for nucleic acids,

four nucleotids give $K = 4$; for proteins, twenty amino acids give $K = 20$; and for spoken words, typically around 40 phonemes give $K = 40$. Using the Rammal ultrametricity index, experimental findings demonstrate that random data are increasingly ultrametric as the spatial dimensionality and sparseness increase.

2.3 Ultrametricity as a Function of Sparseness and Dimensionality

In this article our use of the term "sparseness" has the geometrical sense (relating to spread of points, and not the sense of zero values in a data array).

We use uniformly distributed data and also uniformly distributed hypercube vertex positions. The latter is used to simulate the multivalued words considered by Rammal et al., as described in the previous subsection. Random values are converted to hypercube vertex locations by use of complete disjunctive data coding (Benzécri 1992). Say a variable has maximum and minimum values $x_{\max}$ and $x_{\min}$. Say, further, that $K = 3$. We set thresholds at the values $x_{\min}, x_{\min}+(x_{\max}-x_{\min})*0.25, *0.5$, and $*0.75$. A value of x falling in the first category receives a 4-valued set: $1, 0, 0, 0$; a value of x falling in the second category receives the 4-valued set: $0, 1, 0, 0$; and so on. Such complete disjunctive coding is widely used in correspondence analysis. It is easily verified that the row marginals are constant. In this important case, Lerman (1981) develops an analytic probability density function for the H-classifiability index.

The results of Table 1 are summarized in Figure 1. We note the following findings:

- There is no increase in H-classifiability, i.e. departure from ultrametricity, for increasing numbers of points, n, at least for the range used here: $n = 1000, 2000, 3000, 4000, 5000$.
- There is increase in H-classifiability, i.e. departure from ultrametricity, for increasing dimensionality. Again this holds for the dimensionalities examined here: $m = 50, 100, 250, 500$.
- Random hypercube vertex data are "more classifiable", i.e. such data has smaller H-classifiability and is more ultrametric, compared to uniformly distributed data.

In our experimentation we chose data sets with no a priori clustering. These data sets were random, being either

- uniformly distributed, or
- sparsely coded as hypercube vertices.

We have shown that the latter is consistently more ultrametric than the former.

Our results point to the importance of the "type" of data used or, better expressed, how the data are coded. Binary data representing any categorical (qualitative) variables are consistently more ultrametric than uniformly distributed data.

3 A New Ultrametricity Measure

3.1 Motivation

There are two problems with Lerman's index. Firstly, ultrametricity is associated with $H = 0$ but non-ultrametricity is not bounded. In extensive experimentation, we have found maximum values for H in the region of 0.24. The second problem with Lerman's index is that for floating point coordinate values, especially in high dimensions, the strict equality necessitated for an equilateral triangle is nearly impossible to achieve. However our belief is that approximate equilateral triangles are very likely to arise in important cases of high-dimensional spaces with data points at hypercube vertex locations. We would prefer therefore that the quantifying of ultrametricity should "gracefully" take account of triplets which are "close to" equilateral. Note that for some authors, the equilateral case is considered to be "trivial" or a "trivial limit" (Treves, 1997). For us, however, it is an important case, together with the other important case of ultrametricity (i.e., isosceles with small base).

3.2 Distance-Based Measures

Treves (1997) considers triplets of points giving rise to minimal, median and maximal distances. In the plot of $d_{\text{min}}/d_{\text{max}}$ against $d_{\text{med}}/d_{\text{max}}$, the triangular inequality, the ultrametric inequality, and the "trivial limit" of equilateral triangles, occupy definable regions.

Hartmann (1998) considers $d_{\max} - d_{\text{med}}$. Now, Lerman (1981) uses ranks in order to give (translation, scale, etc.) invariance to the sensitivity (i.e., instability, lack of robustness) of distances. Hartmann instead fixes the remaining distance $d_{\min}$.

3.3 A New Measure Based on Angles

We seek to avoid, as far as possible, lack of invariance due to use of distances. We seek to quantify both isosceles with small base configurations, as well as equilateral configurations. Finally, we seek a measure of ultrametricity bounded by 0 and 1. We will therefore use a coefficient of ultrametricity – we will term it α – which is specified algorithmically as follows.

1. All triplets of points are considered, with a distance (by default, Euclidean) defined on these points. Since for a large number of points, n, the number of triplets, $n(n-1)(n-2)/6$ would be computationally prohibitive, we instead randomly (uniformly) sample coordinates $(i \sim \{1..n\}, j \sim \{1..n\}, k \sim \{1..n\})$.

2. We check for possible alignments (implying degenerate triangles) and exclude such cases.

3. Next we select the smallest angle as less than or equal to 60 degrees. (We use the well-known definition of the cosine of the angle facing side of length x as: $(y^2 + z^2 - x^2)/2yz$.) This is our first necessary property for being a strictly isosceles (< 60 degrees) or equilateral ($= 60$ degrees) ultrametric triangle.

4. For the two other angles subtended at the triangle base, we seek an angular difference of strictly less than 2 degrees (0.03490656 radians). This condition is an approximation to the ultrametric configuration, based on an arbitrary choice of small angle. This condition is targeting a configuration that may not be exactly ultrametric but nonetheless is very close to ultrametric.

5. Among all triplets (1) satisfying our exact properties (2, 3) and close approximation property (4), we define our ultrametricity coefficient as the relative proportion of these triplets. Approximately ultrametric

data will yield a value of 1. On the other hand, data that is non-ultrametric in the sense of not respecting conditions 3 and 4 will yield a low value, potentially reaching 0.

In summary, the α index is defined in this way:

Consider a triplet of points, that defines a triangle. If the smallest internal angle, a, in this triangle is ≤ 60 degrees, and, for the two other internal angles, b and c, if $|b - c| < 2$ degrees, then this triangle is an ultrametric one. We look for the overall proportion of such ultrametric triangles in our data.

The Fisher iris data (150×4) gives $\alpha = 0.0162$, indicating some, very limited, ultrametricity by this measure. By recoding the four iris variables into discrete (zero or one) categories, we find the following. Firstly, with two discrete categories (data now: 150×8), we find $\alpha = 0.0949$. For four discrete categories (data now: 150×16), we find $\alpha = 0.477327$. For eight discrete categories (data now: 150×32), we find $\alpha = 0.741361$. This shows how increasing dimensionality, and sparseness, lead to greater ultrametricity by this measure.

3.4 Ultrametricity Scaling with Data Size, Dimensionality, and Sparseness

We use uniformly distributed data and also uniformly distributed hypercube vertex positions, as in subsection 2.3. The latter is used to simulate the multivalued words considered by Rammal et al. (see subsection 2.2). Again, random values are converted to hypercube vertex locations by use of complete disjunctive data coding (Benzécri 1992).

- As for the Lerman H-classifiability index, we find surprising independence of α relative to n, the number of points. Consider the following: we generate uniformly distributed data points in $\mathbb{R}^{10}$. For $n = 1000, 5000, 10000, 15000, 20000, 25000$, we find $\alpha = 0.096386$, 0.078000, 0.077077, 0.075075, 0.079000, 0.071000. There appears to be a small decrease in ultrametricity due to increasing density of points.

- Ultrametricity increases with sparsity of coding. We will show this by comparing uniformly distributed points, and points at hypercube vertex locations. We will again take the number of points, $n = 1000, 5000, 10000, 15000, 20000, 25000$. We will also use a 10-dimensional space

with, on this occassion, the points at the vertices of a hypercube. (We do this by generating uniformly in $\mathbb{R}^5$ and then quantizing each of the 5 variables to two discrete categories. See discussion at the start of this section). We find, respectively: $\alpha = 0.271630, 0.247495, 0.260563, 0.264056, 0.269076, 0.275275$. Therefore, with sparsity we again find very little dependence on n. For varying n, these α results are quite similar. However we see a very big relative difference in value of α between points in $\mathbb{R}^{10}$ (discussed under the previous bullet point) and points at the vertices of a 10-dimensional hypercube (discussed under this bullet point).

- Ultrametricity increases with dimensionality. Using $n = 5000$ real-valued points, uniformly distributed in space of dimensionality $m = 50, 100, 500, 1000, 5000$, we find: $\alpha = 0.183183, 0.271000, 0.544000, 0.707708, 0.979000$. (See Figure 2.)

- Dimensionality and (spatial) sparsity, combined, force the tendency towards ultrametricity, but the compounding of these two data properties is not as pronounced as we might have expected. Again we take the number of points, $n = 5000$. Using uniform data in real spaces of dimensions 25, 50, 250, 500 and 2500, and then quantizing to two discrete response categories, gives us dimensionalities $m = 50, 100, 500, 1000, 5000$. Our n points are now at the vertices of hypercubes in spaces of dimensionality m. We find $\alpha = 0.179179, 0.172172, 0.454910, 0.588000, 0.934$ (Again see Figure 2.)

We have found the α measure of ultrametricity to most convincingly demonstrate that sparse spaces become very ultrametric with increase in space dimensionality. We stress also that these experiments were carried out on data which are as "un-clustered" as possible.

4 Computational Costs of Operations in an Ultrametric Space

Given that sparse forms of coding are considered for how complex stimuli are represented in the cortex (see Young and Yamane, 1992), the ultrametricity of such spaces becomes important because of this sparseness of coding.

Among other implications, this points to the possibility that semantic pattern matching is best accomplished through ultrametric computation.

A convenient data structure for points in an ultrametric space is a dendrogram. We define a dendrogram as a rooted, labeled, ranked, binary tree (Murtagh, 1984a). Therefore for n points there are precisely $n-1$ levels. With each level there is an associated rank $1, 2, \ldots, n-1$; and also the ultrametric distance which is a mapping into the positive reals.

Operations on binary trees are often based on tree traversal between root and terminal. Hence computational cost of such operations is dependent on root-to-terminal(s) path length. The total path length of a root-to-terminal traversal varies for each terminal (or point in the corresponding ultrametric space). It will be simplest to consider path length in terms of level or tree node rank (and if it is necessary to avail of path length in terms of ultrametric distances, then constant computational time, only, is needed for table lookup). A dendrogram's root-to-terminal path length can vary from close to $\log_2 n$ ("close to" because the path length has to be an integer) to $n-1$ (Murtagh, 1984b). Let us call this computational cost of a tree traversal $O(t)$.

Most operations that we will now consider make use of a dendrogram data structure. Hence the cost of building a dendrogram is important. For the problem in general, see Křivánek and Morávek (1984, 1986) and Day (1996). For $O(n^2)$ implementations of most commonly used hierarchical clustering algorithms, see Murtagh (1983, 1985).

To place a new point (from an ultrametric space) into a dendrogram, we need to find its nearest neighbor. We can do this, in order to write the new terminal into the dendrogram, using a root-to-terminal traversal in the current version of a dendrogram. This leads to our first proposition.

Proposition 1: The computational complexity of adding a new terminal to a dendrogram is $O(t)$, where t is one traversal from root to terminals in the dendrogram.

Proposition 2: The computational complexity of finding the ultrametric distance between two terminal nodes is twice the length of a traversal from root to terminals in the dendrogram. Therefore distance is computed in $O(t)$ time.

Informally: we potentially have to traverse from each terminal to the root in order to find the common, "parent" node.

Proposition 3: The traversal length from dendrogram root to dendrogram terminals is best case 1, and worst case $n-1$. When the dendrogram is

optimally balanced or structured, the traversal length from root to terminals is $\lfloor \log_2 n \rfloor$. Hence $1 \geq O(t) \geq n - 1$, and for a balanced tree $O(t) = \log_2 n$.

Depending on the agglomerative criterion used, we can approximate the balanced or structured dendrogram – and hence favorable case – quite well in practice (Murtagh, 1984b).

Proposition 4: Nearest neighbor search in ultrametric space can be carried out in $O(1)$ or constant time.

This results from the following: the nearest neighbor pair must be in the same tightest cluster that contains them both. There is only one candidate to check for in a dendrogram. Hence nearest neighbor finding results in firstly finding the lowest level cluster containing the given terminal; followed by finding the other terminal in this cluster. Two operations are therefore required.

5 Approximating an Ultrametric for Similarity Metric Space Searching

In data analysis we usually begin with the Euclidean distance or some other non-ultrametric distance (e.g. some other Minkowski distance, or the squared Euclidean distance, or the χ^2 distance, or some normalized or standardized Euclidean distance, etc.). The previous section has discussed some of the advantages of nearest neighbor (also known as best match) searching if such an operation is carried out in an ultrametric space. One way to arrange for this, of course, is simply to map our data into an ultrametric space, using some appropriate hierarchical clustering algorithm (Willett, 1988). But any $O(n^2)$ algorithm is wholely impractical for large n and large dimensionality. In this section we will look at another approach to mapping points into an ultrametric space, followed by use of ultrametric distance for (exact, and not approximate, as we will note below) metric proximity search.

In much work over the years, nearest neighbor searching has been made more efficient through the use of more easily determined feasibility bounds. An early example is Fukunaga and Narendra (1975), a chapter review is in Murtagh (1985), and a recent survey is Chávez, Navarro, Baeza-Yates and Marroquín (2001). In this section we will show that rendering given distances as ultrametric is a powerful way to facilitate nearest neighbor searching. Furthermore "stretching the triangular inequality" (Chávez and Navarro,

2003) so that it becomes the strong triangular inequality, or ultrametric inequality, gives a unifying view of some algorithms of this type.

Bellman's "curse of dimensionality" (Bellman, 1961) can be defined in various ways. Chávez and Navarro (2000, 2003) characterize search dimensionality as the ratio of mean to variance of given metric space distances. A large mean and/or small variance of distances imply exponential increase in nearest neighbor searching, as typifies high dimensional spaces. In high dimensional spaces, this statistic of distance mean divided by variance, $\rho = \mu^2/2\sigma^2$, expresses the fact that the difference between random distances is small. Chávez and Navarro (2003) "exploit the high dimension of the metric space, specifically the fact that the difference between random distances is small compared to a random distance". Now, roughly equal distances is tantamount to equilateral triangles being formed between triplets of points. Thus the Chávez and Navarro principle is to assert that high dimensional spaces become naturally and trivially ultrametric, in that triplets of points form equilateral triangles. Two intuitive properties of increasing values of Chávez and Navarro (2003) ρ are that as the variance decreases, less information is conveyed by distances (i.e., triangles tend towards being equilateral) independently of the spatial dimensionality; and as μ increases a larger search radius is necessitated.

As against these characteristics of metric spaces, ultrametric spaces offer their own potential for fast nearest neighbor finding. The Chávez and Navarro (2003) solution to alleviating the computational difficulties resulting from increasing ρ is through preprocessing which is approximate and probabilistic. We will show that it also amounts to defining ultrametric distances from the given metric space distances.

Fast nearest neighbor finding often makes use of pivots to establish bounds on points to be searched, and points to be bypassed as infeasible (Bustos, Navarro and Chávez, 2003; Chávez et al., 2001). Consider points u which we seek to discard, when searching for nearest neighbors of query q. Pivots, p_i, are used. By the triangle inequality,

$$d(u,p_i) \leq d(u,q) + d(q,p_i) \text{ and } d(q,p_i) \leq d(q,u) + d(u,p_i) \qquad (1)$$

These two relations lead to the following rejection rule: we want to discard all u such that $|d(u,p_i) - d(q,p_i)| > r$ for a threshold r and for some pivot p_i. Then nearest neighbor searching takes place through all u which can *not* be rejected in this way.

Let us look at this rejection rule, $|d(u,p_i) - d(q,p_i)| > r$, a little closer. We are enforcing approximate equality by rejecting u whenever we have departure from equality. What difficulty can departure from equality of these distances cause for us? We have either $d(u,p_i) > d(q,p_i)$ or vice versa $d(q,p_i) > d(u,p_i)$.

Hence the rejection rule results in the following not being allowed in relations 1: $d(u,p_i) > d(q,p_i)$ (left relation in 1) and $d(q,p_i) > d(u,p_i)$ (right relation in 1).

Take the left relation in 1. We have $d(u,p_i) \leq d(u,q) + d(q,p_i)$, and $d(u,p_i) < d(q,p_i)$ consistent with and respecting the rejection rule. Look at the right hand side of the first of these: we have either $d(q,p_i) > d(u,q)$ or $d(q,p_i) < d(u,q)$. If $d(q,p_i)$ is the larger here, then relation 1 (left) is satisfied. If $d(q,p_i)$ is the smaller of the pair here, then from our rejection rule we again find that relation 1 (left) is satisfied. Relation 1 (right) can be shown in the same way. We conclude: from the relations in 1, given the rejection rule, we have as a consequence:

$$d(u,p_i) \leq \max\{d(u,q), d(q,p_i)\} \text{ and}$$

$$d(q,p_i) \leq \max\{d(q,u), d(u,p_i)\} \tag{2}$$

For further discussion of this consequence of the rejection rule applied, and for its use in practice, the references cited here can be referred to. The rejection rule is seen to be a requirement for $d(u,p_i)$ and $d(q,p_i)$ to be similar, in that they are being rejected precisely if they are not similar. In other words, the rejection rule is forcing retained triangles to be isosceles.

The heuristic pursued by Chávez and Navarro (2003) has been shown by these authors to be computationally beneficial: as the search dimensionality ρ grows, the heuristic entails a small and more reliable search radius. It results from our discussion that this property of their algorithm results from transforming a metric space to be ultrametric. In practice, this transforming need only be partial, i.e. subset-wise.

Fast nearest neighbor searching in metric spaces often appeals to heuristics. As shown in the foregoing discussion, the link with ultrametric spaces gives rise instead to a unifying view. Hjaltason and Samet (2003) discuss heuristic nearest neighbor searching in terms of embedding the given metric space points in lower dimensional spaces. From our discussion in this section, we see that there is evidently another alternative direction for facilitating fast

nearest neighbor searching: viz., taking the metric space as an ultrametric one, and if it does not quite fit this perspective then "stretch" (transform) it so that it does so.

In conclusion, we note that the pivot-based approach described in this section allows us to limit the parts of our original data space (let's assume, Euclidean) to be searched. The ultrametric distance relationships are used for this purpose, and it is not necessary nor useful to define this ultrametric space in detail.

6 Conclusions

We have shown that high dimensional and sparse codings tend to be ultrametric. This is an interesting result in its own right. However a far more important result is that certain computational operations can be carried out very efficiently indeed in space endowed with an ultrametric.

Chief among these computational operations, we have shown, is that nearest neighbor finding can be carried out in (worst case) constant computational time. Depending on the structure of the ultrametric space (i.e. if we can build a balanced dendrogram data structure), pairwise distance calculation can be carried out in logarithmic computational time.

We have also reviewed approaches to using ultrametric distances in order to expedite best match, or nearest neighbor, or more generally proximity search. The usual constructive approach, viz. build a hierarchic clustering, is simply not computationally feasible in very high dimensional spaces as are typically found in such fields as speech processing, information retrieval, or genomics and proteomics.

We have noted how forms of sparse coding are considered to be used in the human or animal cortex. We raise the interesting question as to whether human or animal thinking can be computationally efficient precisely because such computation is carried out in an ultrametric space.

References

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*, Princeton, New Jersey: Princeton University Press.

Benzécri, J.P. (1979). *La Taxinomie*, 2nd ed., Paris: Dunod.

Benzécri, J.P. (1992). Transl. T.K. Gopalan. *Correspondence Analysis Handbook*, Basel: Marcel Dekker.

Bustos, D., Navarro, G. and Chávez, E. (2003). "Pivot Selection Techniques for Proximity Searching in Metric Spaces", *Pattern Recognition Letters*, 24, 2357–2366.

Cailliez, F. and Pagès, J.-P. (1976). *Introduction à l'Analyse des Données*, Paris: SMASH (Société de Mathématiques Appliquées et de Sciences Humaines).

Chávez, E. and Navarro, G. (2000). "Measuring the Dimensionality of General Metric Spaces", Technical Report TR/DCC-00-1, Department of Computer Science, University of Chile.

Chávez, E., Navarro, G., Baeza-Yates, R. and Marroquín, J.L. (2001). "Proximity Searching in Metric Spaces", *ACM Computing Surveys*, 33, 273–321.

Chávez, E. and Navarro, G. (2003). "Probabilistic Proximity Search: Fighting the Curse of Dimensionality in Metric Spaces", *Information Processing Letters*, 85, 39–46.

Day, W.H.E. (1996). "Complexity Theory: An Introduction for Practitioners of Classification", in P. Arabie, L.J. Hubert and G. De Soete, Eds., *Clustering and Classification*, Singapore: World Scientific, 199–233.

Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problems", *The Annals of Eugenics*, 7, 179–188.

Fukunaga, K. and Narendra, P.M. (1975). "A Branch and Bound Algorithm for Computing k-Nearest Neighbors", *IEEE Transactions on Computers*, C-24, 750-753.

Gouvêa, F.Q. 2003, *P-Adic Numbers*, New York: Springer-Verlag, 2nd edn., 3rd printing.

Hartmann, A.K. (1998). "Are Ground States of 3D $\pm$J Spin Glasses Ultrametric?", *Europhysics Letters*, 44, 249–254.

Hjaltason, G.R. and Samet, H. (2003). "Properties of Embedding Methods for Similarity Searching in Metric Spaces", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 530–549.

Johnson, S.C. (1967). "Hierarchical Clustering Schemes", *Psychometrika*, 32, 241-254.

Krasner, M. (1944). "Nombres semi-réels et espaces ultramétriques", *Comptes-Rendus de l'Académie des Sciences, Tome II*, 219, 433.

Křivánek, M. and Morávek, J. (1986). "NP-Hard Problems in Hierarchical-Tree Clustering", *Acta Informatica*, 23, 311–323.

Křivánek, M. and Morávek, J. (1984). "On NP-Hardness in Hierarchical Clustering", in T. Havránek, Z. Sidák and M. Novák, Eds., *Compstat 1984: Proceedings in Computational Statistics*, 189–194, Vienna: Physica-Verlag.

Lerman, I.C. (1981). *Classification et Analyse Ordinale des Données*, Paris: Dunod.

Mahler, M. (1981). *P-adic Numbers and Their Functions*, 2nd edn., Cambridge: Cambridge University Press.

Mézard, M., Parisi, G., Sourlas, N., Toulouse, G. and Virasoro, M.A. (1984). "Nature of the Spin-Glass Phase", *Physical Review Letters*, 52, 1156–1159.

Murtagh, F. (1983). "A Survey of Recent Advances in Hierarchical Clustering Algorithms", *The Computer Journal*, 26, 354–359.

Murtagh, F. (1984a). "Counting Dendrograms: a Survey", *Discrete Applied Mathematics*, 7, 191–199.

Murtagh, F. (1984b). "Structures of Hierarchic Clusterings: Implications for Information Retrieval and for Multivariate Data Analysis", *Information Processing and Management*, 20, 611–617.

Murtagh, F. (1985). *Multidimensional Clustering Algorithms*, Würzburg: Physica-Verlag.

Ogielski, A.T. and Stein, D.L. (1985). "Dynamics of Ultrametric Spaces", *Physical Review Letters*, 55, 1634–1637.

Parisi, G. and Ricci-Tersenghi, F. (2000). "On the Origin of Ultrametricity", *Journal of Physics A: Mathematical and General*, 33, 113–129.

Rammal, R., Angles d'Auriac, J.C. and Doucot, B. (1985). "On the Degree of Ultrametricity", *Le Journal de Physique – Lettres*, 46, L-945 – L-952.

Rammal, R., Toulouse, G. and Virasoro, M.A. (1986). "Ultrametricity for Physicists", *Reviews of Modern Physics*, 58, 765-788.

Roberts, M.D. (2001). "Ultrametric Distance in Syntax", http://arXiv.org/abs/cs.CL/9810012

Schikhof, W.H. (1984). *Ultrametric Calculus*, Cambridge: Cambridge University Press.

Treves, A. (1997). "On the Perceptual Structure of Face Space", *BioSystems*, 40, 189–196.

Watson, S. (2003). "The Classification of Metrics and Multivariate Statistical Analysis", preprint, York University, 27 pp.

Willett, P. (1988). "Recent Trends in Hierarchic Document Clustering: A Critical Review", *Information Processing and Management*, 24, 577–597.

Young, M.P. and Yamane, S. (1992). "Sparse Population Coding of Faces in the Inferotemporal Cortex", *Science*, 256, 1327–1331.

Figure Captions

Fig. 1: Upper curve: uniformly distributed values. Lower curve: random hypercube vertex points. A low value of H-classifiability is related to near-ultrametricity. Each point shows an average of different experiments corresponding to numbers of points $n = 1000, 2000, 3000, 4000, 5000$. From values given in Table 1, there is very little variation as a function of n.

Fig. 2: Upper curve: uniformly distributed values. Lower curve: random hypercube vertex points. A value of alpha close to 1 is related to near-ultrametricity. For each dimensionality (50, 100, 500, 1000, 5000) we used number of points $n = 5000$. In other experiments we found very little variation as a function of n.

Table Captions

Table 1: Results using Lerman's H-classifiability measure. The 100-dimensional hypercube vertex point sets were produced by generating 20-dimensional uniformly distributed points and then transforming into complete disjunctive coding using 5 quantile thresholds. Similarly, the 250- and 500-dimensional data used 50- and 100-dimensional uniform data to begin with, followed by transforming with 5 quantile thresholds.

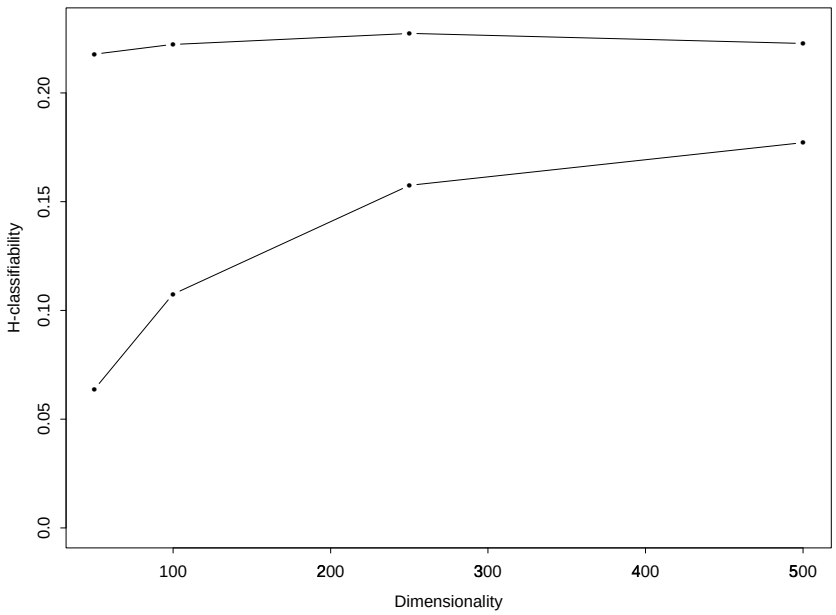

Figure 1: Upper curve: uniformly distributed values. Lower curve: random hypercube vertex points. A low value of H-classifiability is related to near-ultrametricity. Each point shows an average of different experiments corresponding to numbers of points $n = 1000, 2000, 3000, 4000, 5000$. From values given in Table 1, there is very little variation as a function of n.

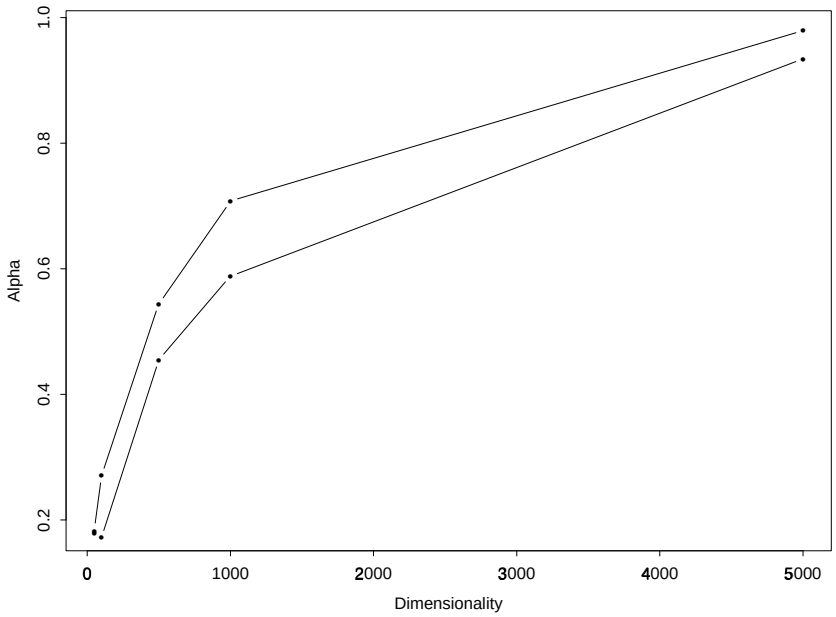

Figure 2: Upper curve: uniformly distributed values. Lower curve: random hypercube vertex points. A value of alpha close to 1 is related to near-ultrametricity. For each dimensionality (50, 100, 500, 1000, 5000) we used number of points $n = 5000$. In other experiments we found very little variation as a function of n.

Table 1: Results using Lerman's H-classifiability measure. The 100-dimensional hypercube vertex point sets were produced by generating 20-dimensional uniformly distributed points and then transforming into complete disjunctive coding using 5 quantile thresholds. Similarly, the 250- and 500-dimensional data used 50- and 100-dimensional uniform data to begin with, followed by transforming with 5 quantile thresholds.

	n	dim.	space	H-classif.
Uniform	1000	50	$\mathbb{R}^{50}$	0.2121
	2000	50	$\mathbb{R}^{50}$	0.2131
	3000	50	$\mathbb{R}^{50}$	0.2271
	4000	50	$\mathbb{R}^{50}$	0.2169
	5000	50	$\mathbb{R}^{50}$	0.2205
Mean		50	$\mathbb{R}^{50}$	0.2179
Hypercube vertex	1000	50	$\{0,1\}^{50}$	0.0622
	2000	50	$\{0,1\}^{50}$	0.0653
	3000	50	$\{0,1\}^{50}$	0.0580
	4000	50	$\{0,1\}^{50}$	0.0592
	5000	50	$\{0,1\}^{50}$	0.0737
Mean		50	$\{0,1\}^{50}$	0.0637
Uniform	1000	100	$\mathbb{R}^{100}$	0.2326
	2000	100	$\mathbb{R}^{100}$	0.2173
	3000	100	$\mathbb{R}^{100}$	0.2241
	4000	100	$\mathbb{R}^{100}$	0.2172
	5000	100	$\mathbb{R}^{100}$	0.2207
Mean		100	$\mathbb{R}^{100}$	0.2224
Hypercube vertex	1000	100	$\{0,1\}^{100}$	0.0954
	2000	100	$\{0,1\}^{100}$	0.1117
	3000	100	$\{0,1\}^{100}$	0.1112
	4000	100	$\{0,1\}^{100}$	0.1122
	5000	100	$\{0,1\}^{100}$	0.1071
Mean		100	$\{0,1\}^{100}$	0.1075
Uniform	1000	250	$\mathbb{R}^{250}$	0.2264
	2000	250	$\mathbb{R}^{250}$	0.2340
	3000	250	$\mathbb{R}^{250}$	0.2231
	4000	250	$\mathbb{R}^{250}$	0.2245
	5000	250	$\mathbb{R}^{250}$	0.2289
Mean		250	$\mathbb{R}^{250}$	0.2274
Hypercube vertex	1000	250	$\{0,1\}^{250}$	0.1547
	2000	250	$\{0,1\}^{250}$	0.1618
	3000	250	$\{0,1\}^{250}$	0.1496
	4000	250	$\{0,1\}^{250}$	0.1636
	5000	250	$\{0,1\}^{250}$	0.1577
Mean		250	$\{0,1\}^{250}$	0.1575
Uniform	1000	500	$\mathbb{R}^{500}$	0.2225
	2000	500	$\mathbb{R}^{500}$	0.2284
	3000	500	$\mathbb{R}^{500}$	0.2320
	4000	500	$\mathbb{R}^{500}$	0.2175
	5000	500	$\mathbb{R}^{500}$	0.2138
Mean		500	$\mathbb{R}^{500}$	0.2228
Hypercube vertex	1000	500	$\{0,1\}^{500}$	0.1771
	2000	500	$\{0,1\}^{500}$	0.1766
	3000	500	$\{0,1\}^{500}$	0.1798
	4000	500	$\{0,1\}^{500}$	0.1711
	5000	500	$\{0,1\}^{500}$	0.1814
Mean		500	$\{0,1\}^{500}$	0.1772