Thinking Ultrametrically

Fionn Murtagh

School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK. f.murtagh@qub.ac.uk

Summary. 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 that are sparse. Here we look at the quantification of ultrametricity. We also look at data compression based on a new ultrametric wavelet transform. We conclude with computational implications of prevalent and perhaps ubiquitous ultrametricity.

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. 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. Ultrametricity is a natural property of high-dimensional spaces (Rammal et al., 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. Important references on ultrametrics in the clustering and classification area are those of Benzécri (1979) representing work going back to 1963, and Johnson (1967).

Watson (2003) attributes to Mézard et al. (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, 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 history of use of ultrametric spaces is provided by Rammal et al. (1985) and for linguistic research by Roberts (2001).

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, p. 22). Essentially this theorem states that the rationals can be expressed in terms of (continuous) reals, or (discrete) p-adic numbers, and no other alternative system.

In this article we will describe a new ultrametric wavelet transform. Our motivation for doing this is to provide analysis capability for situations where our data tends to be ultrametric (e.g., sparse, high-dimensional data). Secondly, in this article, we will present results of Lerman's proposed quantification of ultrametricity in a data set.

2 P-adic Coding from Dendrograms

Dendrograms used in data analysis are usually labeled and ranked: see Figures 1 and 2.

For the ranked dendrogram shown in Figure 1 we develop the following p-adic encoding of terminal nodes, by traversing a path from the root: $x_1 = 0 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + 0 \cdot 2^1$; $x_2 = 0 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + 1 \cdot 2^1$; $x_4 = 0 \cdot 2^7 + 1 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3$; $x_6 = 0 \cdot 2^7 + 1 \cdot 2^5 + 1 \cdot 2^4$. The decimal equivalents of this p-adic representation of terminal nodes work out as $x_1, x_2, \ldots x_8 = 0, 2, 4, 32, 40, 48, 128, 192$.

Distance and norm are defined as follows. $d_p(x, x') = d_p|x - x'| = 2^{-r+1}$ or $2 \cdot 2^{-r}$ where $x = \sum_k a_k 2^k, x' = \sum_k a'_k 2^k, r = \text{argmin}_k\{a_k = a'_k\}$. The norm is defined as $d_p(x, 0) = 2^{-1+1} = 1$.

To find the p-adic distance, we therefore look for the smallest level, r (if ordered from terminal to root as in Figure 1) which is identical in the pair of power series, which yields the result of 2^{-r+1}. We find $|x_1 - x_2|_2 = 2^{-2+1} =$

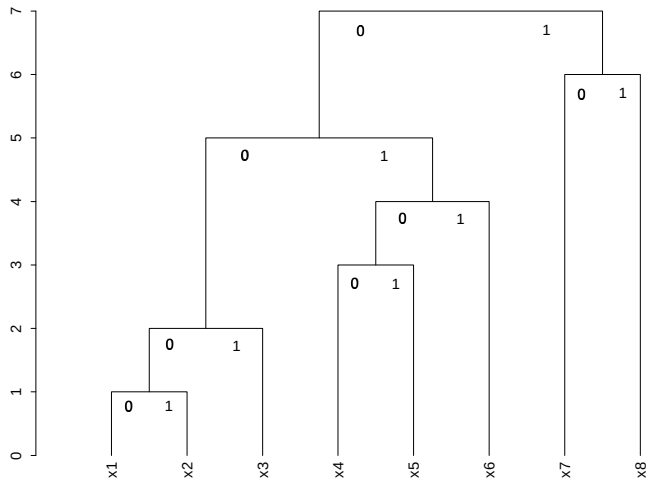

Fig. 1. Labeled, ranked dendrogram on 8 terminal nodes. Branches labeled 0 and 1.

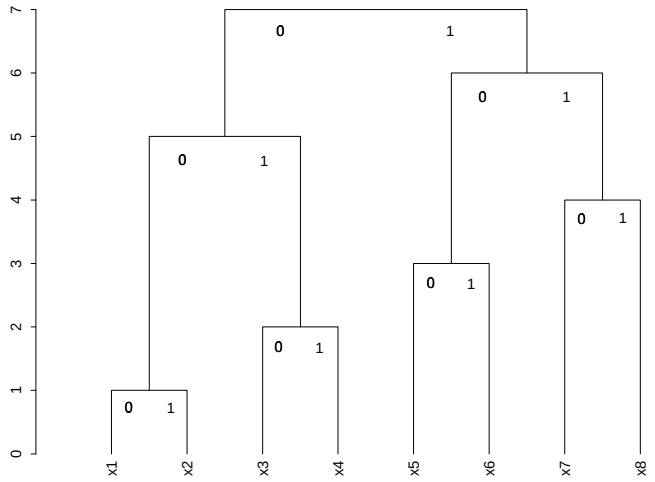

Fig. 2. A structurally balanced, labeled, ranked dendrogram on 8 terminal nodes. Branches labeled 0 and 1.

$\frac{1}{2}$; $|x_1 - x_4|_2 = 2^{-6+1} = \frac{1}{32}$; $|x_1 - x_6|_2 = 2^{-6+1} = \frac{1}{32}$. The smallest p-adic distance between terminals in Figure 1 is seen to be $\frac{1}{128}$.

For Figure 2, we also find the smallest p-adic distance to be $\frac{1}{128}$. In Figure 2, the decimal equivalent of the p-adic number x_8 is 208. If we look for the maximum possible decimal equivalent of the p-adic numbers corresponding to 8 terminal nodes, the answer is $1 \cdot 2^7 + 1 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 = 254$.

The p-adic representation used here is not invariant relative to dendrogram representation. Consider, for example, some alternative representation of Figure 1 such as the representation with terminal nodes in the order: $x_7, x_8, x_1, x_2, x_3, x_4, x_5, x_6$. The dendrogram can be drawn with no crossings, so such a dendrogram representation is perfectly legitimate. With branches labeled 0 = left and 1 = right, as heretofore, we would find the p-adic representation of x_1 to be $1 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + 0 \cdot 2^1$. However if the p-adic representation differs with this new dendrogram representation, a moment's thought shows that both p-adic norm, and p-adic distance, are invariant relative to dendrogram representation. A formal proof can be based on the embedded classes represented by dendrogram nodes.

3 Regression based on Ultrametric Haar Wavelet Transform

The wavelet transform, developed for signal and image processing, has been extended for use on relational data tables and multidimensional data sets (Vitter and Wang, 1999; Joe et al., 2001) for data summarization (micro-aggregation) with the goal of anonymization (or statistical disclosure limitation) and macrodata generation; and data summarization with the goal of computational efficiency, especially in query optimization. There are problems, however, in doing this with direct application of a wavelet transform. Essentially, a relational table is treated in the same way as a 2-dimensional pixelated image, although the former case is invariant under row and column permutation, whereas the latter case is not (Murtagh et al., 2000). Therefore there are immediate problems related to non-uniqueness, and data order dependence. For very small dimensions, for example attributes in a relational data table, a classical application of a wavelet transform is troublesome, and in addition if table dimensionality equal to an integer power of 2 is required, the procedure is burdensome to the point of being counter-productive. Sparse tabular data cannot be treated in the same way as sparse pixelated data (e.g. Sun and Zhou, 2000) if only because row/column permutation invariance causes the outcome to be dominated by sparsity-induced effects. In this article we will develop a different way to wavelet transform tabular data. A range of other input data types are also capable of being treated in this way.

Our motivation for the development of wavelet transforms in ultrametric or hierarchical data structures is to cater for "naturally" or enforced ultrametric data. An example of the former case is questionnaire results with embedded

question sets. An example of the latter case is that of data with already strong ultrametric tendency such as sparsely coded data in speech analysis, genomics and proteomics, and other fields, and complete disjunctive form in correspondence analysis.

The Haar wavelet transform is usually applied to 1D signals, 2D images, and 3D data cubes (see Starck et al. 1998; Starck and Murtagh, 2002). Sweldens (1997) extended the wavelet transform to spherical structures, still in Hilbert space. We extend the wavelet transform to other topological structures, in particular hierarchies which have a natural representation as trees. In regard to the wavelet transform, we focus in particular on the Haar wavelet transform, in its redundant (Soltani et al., 2000; Zheng et al., 1999) and non-redundant versions (e.g., Frazier, 1999).

The Morlet-Grossmann definition of the continuous wavelet transform (Grossmann et al. 1989; Starck and Murtagh 2002) holds for a signal $f(x) \in L^2(\mathbb{R})$, the Hilbert space of all square integrable functions. Hilbert space $L^2(\mathbb{R})$ is isomorphic and isometric relative to its integer analog $l^2(\mathbb{Z})$. The discrete wavelet transform, derived from the continuous wavelet transform, holds for a discrete function $f(x) \in l^2(\mathbb{Z})$.

Our input data is decomposed into a set of band-pass filtered components, the wavelet coefficients, plus a low-pass filtered version of our data, the continuum (or background or residual). We consider a signal, $\{c_{0,l}\}$, defined as the scalar product at samples i of the real function $f(x)$, our input data, with a scaling function $\phi(x)$ which corresponds to a low-pass filter: $c_0(k) = \langle f(x), \phi(x-k) \rangle$.

The wavelet transform is defined as a series expansion of the original signal, c_0, in terms of the wavelet coefficients. The final smoothed signal is added to all the differences: $c_{0,i} = c_{J,i} + \sum_{j=1}^{J} w_{j,i}$. This equation provides a reconstruction formula for the original signal. At each scale j, we obtain a set, which we call a wavelet scale. The wavelet scale has the same number of samples as the signal, i.e. it is redundant, and decimation is not used.

Now consider any hierarchical clustering, H, represented as a binary rooted tree. For each cluster q'' with offspring nodes q and q', we define $s(q'')$ through application of the low-pass filter $\begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$:

$$s(q'') = \frac{1}{2}\left(s(q) + s(q')\right) = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}^t \begin{pmatrix} s(q) \\ s(q') \end{pmatrix} \tag{1}$$

Next for each cluster q'' with offspring nodes q and q', we define detail coefficients $d(q'')$ through application of the band-pass filter $\begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$:

$$d(q'') = \frac{1}{2}(d(q) - d(q')) = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}^t \begin{pmatrix} d(q) \\ d(q') \end{pmatrix} \tag{2}$$

The scheme followed is illustrated in Figure 3, which shows the hierarchy constructed by the median method, using the first 8 observation vectors in Fisher's iris data (Fisher, 1936).

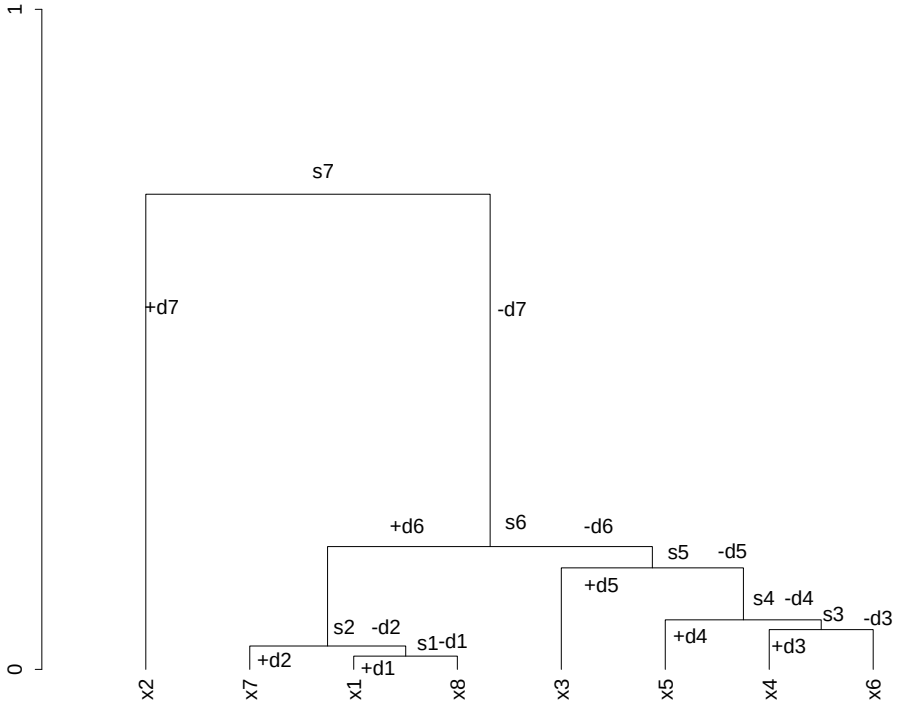

Fig. 3. Dendrogram on 8 terminal nodes constructed from first 8 values of Fisher's iris data. Median method used in this case.

For any $d(q_j)$ we have: $\sum_k d(q_j)_k = 0$, i.e. the detail coefficient vectors are each of zero mean. The inverse transform allows exact reconstruction of the input data. If an observation vector is denoted by x_i, then the ultrametric wavelet transform defines the p-adic encoding for x_i given by $\sum_1^{n-1} a_k p_k$ where $a_k \in \{0, 1\}$ and $p_k = 2^k$. The wavelet transform is defined by $x_i = s_{n-1} + \sum_1^{n-1} a_k d_k$ where s_{n-1} is the final smooth component, and d_k are the detail or wavelet signals (or vectors).

Setting wavelet coefficients to zero and then reconstructing the data is referred to as hard thresholding (in wavelet space) and this is also termed wavelet smoothing or regression. Table 1 shows the excellent results that can

be obtained for Fisher's iris data. Further results on the energy compacting or compression properties of this new ultrametric Haar transform are given in Murtagh (2003b), together with R code for this new transform and its inverse.

Filtering threshold	% coefficients set to zero	mean square error
0	16.95	0
0.1	70.13	0.0098
0.2	91.95	0.0487
0.3	97.15	0.0837
0.4	97.82	0.1040

Table 1. Ultrametric Haar filtering results for Fisher's 150×4 iris data. Filtering is carried out by setting small (less than the threshold) wavelet coefficient values to zero. The data is then reconstructed. The quality of reconstruction between the input data matrix, and the reconstructed data matrix, is measured using mean square error.

4 Lerman's H-classifiability

The work of Rammal et al. (1986) used the discrepancy between the subdominant ultrametric (provided by single link hierarchical clustering) and input metric values as a measure of how ultrametric the given data set was. Their work is further discussed below, in this section. We distrust the single link method in view of its known chaining and other disadvantages. We will now review an alternative measure of ultrametricity in a data set, due to Lerman (1981).

On a set E, a binary relation is a *preorder* if it is reflexive and transitive. Let F denote the set of pairs of distinct units, where a unit is from 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 F 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:

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

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:

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

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 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.

Rammal et al. (1985, 1986) quantify ultrametricity as follows. 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 (1985, 1986), this index suffers from "the chaining effect and from

sensitivity to fluctuations". The single link method, yielding the subdominant ultrametric, is subject to potential pathologies. For this reason the Lerman index is to be preferred. The latter is unbounded and, given the definition used above, we have found maximum values (i.e. greatest non-ultrametricity) in the region of 0.25.

Rammal et al. (1985, 1986) discuss a range of important cases: 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, around 40 phonemes give $K = 40$. Using the Rammal ultrametricity index, experimental findings demonstrate that random data, in the sparse limit (i.e., with increasing dimensionality and with increasing sparseness), are increasingly ultrametric.

Our experimental findings are different, given the very different way we assess ultrametricity, and we contribute some important clarifications in the light of Lerman's H-classifiability to the Rammal et al. discovery that ultrametricity is "a natural property of large spaces".

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. Random values are converted to hypercube vertex locations by use of complete disjunctive data coding (Benzécri, 1992). For example, for $K = 4$ we use four fixed intervals. A value of x falling in the first interval receives a 4-valued set: $1, 0, 0, 0$; a value of x falling in the second interval 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.

In our experiments (see Murtagh, 2003a), we found that 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: $n = 1000, 2000, 3000, 4000, 5000$. In the results shown in Figure 4 we note the following additional findings:

- 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 see that the latter is consistently more ultrametric than the former.

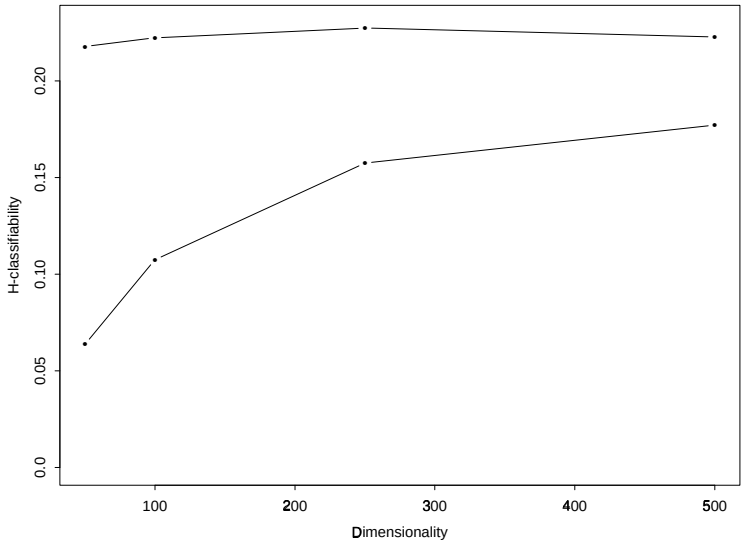

Fig. 4. Upper curve: uniformly distributed values. Lower curve: random hypercube vertex points. A low value of H-classifiability is related to near-ultrametricity. Each point shows an average of different experiments corresponding to numbers of points $n = 1000, 2000, 3000, 4000, 5000$.

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.

Further experiments on quantifying ultrametricity in data can be found in Murtagh (2004).

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. Our justification in indicating this is that once we have a dendrogram data structure nearest neighbor computation is carried out with constant computational complexity, i.e. it is of $O(1)$ computational cost. Other operations can also be carried out with good computational properties once we have a binary rooted tree data structure that defines interrelationships.

5 Conclusion

We have shown that sparse coding tends 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 results is that nearest neighbor finding can be carried out in (worst case) constant computational time. 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.

We have developed a new form of the Haar wavelet transform for topologies associated with hierarchically structured data sets. We have demonstrated the effectiveness of this transform for data filtering and for data compression.

References

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

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

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

Frazier, M.W. (1999). *An Introduction to Wavelets through Linear Algebra*, Springer-Verlag, New York.

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

Grossmann, A., Kronland-Martinet, R., and Morlet, J. (1989). Reading and understanding the continuous wavelet transform, in J. Combes, A. Grossmann and P. Tchamitchian, eds., *Wavelets: Time-Frequency Methods and Phase-Space*, pp. 2–20, Springer-Verlag, New York.

Joe, M.J., Whang, K.-Y., and Kim, S.-W. (2001). Wavelet transformation-based management of integrated summary data for distributed query processing, *Data and Knowledge Engineering*, 39, 293–312.

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.

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., Starck, J.-L., and Berry, M. (2000). Overcoming the curse of dimensionality in clustering by means of the wavelet transform, *The Computer Journal*, 43, 107–120.

Murtagh, F. (2003a). "On ultrametricity, sparse coding and computation", submitted.

Murtagh, F. (2003b). "Hierarchical or ultrametric Haar wavelet transform in multivariate data analysis and data mining", submitted.

Murtagh, F. (2004). "Quantifying ultrametricity", *COMPSTAT 2004*, submitted.

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.

Soltani, S., Boichu, D., Simard, P., and Canu, S. (2000). The long-term memory prediction by multiscale decomposition, *Signal Processing*, 80, 2195–2205.

Starck, J.-L., Murtagh, F., and Bijaoui, A. (1998). *Image and Data Analysis: The Multiscale Approach*, Cambridge University Press, Cambridge.

Starck, J.-L. and Murtagh, F. (2002). *Astronomical Image and Data Analysis*, Springer-Verlag, New York.

Wenchang Sun and Xingwei Zhou (2000). Sampling theorem for wavelet subspaces: error estimate and irregular sampling theorem, *IEEE Transactions on Signal Processing*, 48, 223–226.

Sweldens, W. (1997). The lifting scheme: a construction of second generation wavelets, *SIAM Journal on Mathematical Analysis*, 29, 511–546.

Vitter, J.S., and Wang, M. (1999). Approximate computation of multidimensional aggregates of sparse data using wavelets, in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 193–204.

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

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

Zheng, G., Starck, J.-L., Campbell, J.G., and Murtagh, F. (1999). Multiscale transforms for filtering financial data streams, *Journal of Computational Intelligence in Finance*, 7, 18-35.