

QUANTIFYING ULTRAMETRICITY

Fionn Murtagh

Key words: Classification, clustering, hierarchy, ultrametric, p-adic

COMPSTAT 2004 section: Clustering.

Abstract: The ultrametric properties of hierarchic clustering are well-known. In recent years, there has been interest in ultrametric properties found in statistical mechanics, optimization theory, and physics. It has been shown that sparse, high-dimensional spaces tend to be ultrametric. Given the pervasiveness of ultrametricity, it is important to be able to quantify how close given metric data are to being ultrametric. In this article we assess previously used coefficients of ultrametricity. We present a new coefficient of ultrametricity, and exemplify its properties experimentally. Our immediate objective in this work is to show that sparse, high-dimensional spaces, that are typical of many new data analysis problems in such areas as genomics and proteomics, and speech, tend to be inherently ultrametric.

1 Introduction

Ultrametricity is defined mathematically in section 2.1 below, but can be informally described as follows: there is a natural hierarchical or embedded structure among the data observations under investigation. Hierarchical cluster analysis involves inducing an ultrametric set of relationships on the objects or observations. In the 1980s, ultrametric spaces came under investigation in physics. Some recent work has also used the perspective of ultrametric topology as part of a model of human cognition. An important finding (Rammal et al., 1986) has been that sparse and high-dimensional spaces tend to be ultrametric. This means that such spaces, containing points associated with a set of observations, are characterized by ultrametric (or hierarchical) relationships. The implications of this are far reaching for the analysis of massive, high-dimensional data sets in such fields as speech processing, or proteomics, to name but two.

In this article we will show how sparse, high-dimensional data are found to be ultrametric. Our main focus in this work is the quantifying of ultrametricity.

An initial response to this requirement would be to take a large data set, and construct a hierarchical clustering on it using some suitable clustering criterion. A constructive assessment of ultrametricity is then simply quantifying the discrepancy between input data and induced ultrametric data structure (e.g. through the Euclidean or some other distance between initial pairwise dissimilarities, and induced ultrametric distances).

Examples of such constructive approaches to assessing ultrametricity include: use of any hierarchical clustering algorithm, many of which can be

implemented in a stepwise way based on the Lance-Williams update relationship; specifically for quantifying ultrametricity with a well-defined coefficient, Rammal et al. (1986) use the single link method; and one can of course use a criterion such as the commonly used least squares fit criterion (Chandon and De Soete, 1984; De Soete, 1987; Makarenkov and Leclerc, 1999; De Soete and Carroll, 1996) although it is known that such an approach will only approximate an optimal result (Křivánek and Morávek, 1984, 1986; Day, 1996) for this NP-complete problem.

Quantifying ultrametricity using a constructive approach is less than perfect due to the following:

- Potential complications arising from known problems, e.g. chaining in single link, non-uniqueness of a minimal superior ultrametric (Benzécri, 1976), or inversions (non-compliance with Bruynooghe’s reducibility property: see Murtagh, 1985).
- In the case of the single link method, empirically observed scaling regimes described theoretically and empirically by Rammal et al. (1986).
- Suboptimal solutions and dependency on starting configurations in the case of seeking direct optimization of NP-complete problems.

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.

In section 2, we will give the formal definition of ultrametricity. In section 3, we will look at various direct approaches to quantifying the extent of ultrametricity in a data set. Section 3.1 details an approach due to Lerman, which is based on ranks of dissimilarities. Section 3.2 uses one particular constructive approach, the single link agglomerative method, and the discrepancy between the resulting ultrametric and the initial distances. Section 3.3 describes two other approaches used in the literature. In Section 3.4 then we describe a new coefficient of ultrametricity, and we describe how the properties of this ultrametricity coefficient are advantageous, and outperform previous results. Section 3.5 presents experimental support for this new ultrametricity coefficient.

2 Relevant Ultrametric Axioms and Triangle Properties

2.1 Isosceles Triangles with Base Side Smallest

The ultrametric relation implies that triangles among all triplets are isosceles, with base side of smallest length (Benzécri, 1979; Lerman, 1981).

Consider three points x, y , and z . Without loss of generality let (y, z) be one of the less long sides. (Hence it is either the short base side, or one of the long sides.) Then $d(x, y) \leq \max\{d(y, z), d(x, z)\}$ implies that $d(x, y) \leq d(y, z) \leq d(x, z)$.

Now we permute x and y . But doing this implies that $d(y, x) \leq d(x, z) \leq d(y, z)$. And the only way that we can have simultaneously $d(y, z) \leq d(x, z)$ and $d(x, z) \leq d(y, z)$ is for these to be equal.

Hence we have $d(x, y) \leq d(x, z) = d(y, z)$, QED.

2.2 Ultrametrics, Ultramines and Their Intersection

The ultrametric inequality is: $d(x, y) \leq \max\{d(y, z), d(x, z)\}$ When: $d(x, y) \geq \min\{d(y, z), d(x, z)\}$ then Rizzi (2000) terms this an ultramine. Replacing y with x gives $d(x, x) \geq d(x, z)$, so an ultramine is a similarity measure. Lerman (1981) uses the term ultrametric proximity. Other than the strong triangular inequality, and symmetry, Lerman (1981) gives the remaining property of an ultramine or ultrametric proximity as: $d(x, y) = +\infty$ whenever $x = y$.

As already noted, a metric space is ultrametric iff all triangles are isosceles with base lesser than or equal to the side lengths. A metric space is ultramine iff all triangles are isosceles with base greater than or equal to the side lengths. The intersection of ultrametrics and ultramines is defined by equilateral triangles.

3 Approaches to Quantifying Ultrametricity

3.1 Lerman's H Measure Based on Ranks of Pairwise Dissimilarities

Lerman's measure of ultrametricity is based on ranks of dissimilarities between observations. Use of ranks is for two main reasons: (i) "robustness" is ensured, i.e. limitation of effects of unusually large or unusually small dissimilarities, and (ii) an effective normalization of the dissimilarities results from this, so that comparability between different data sets becomes feasible.

For $x, y, z \in E$: we consider $d(x, y) \leq d(y, z) \leq d(x, z)$. By the ultrametric inequality as seen in section 2.1 we have: $d(x, z) \leq d(y, z)$. Therefore (x, z) and (y, z) must be in the same class of the preorder on $E \times E$.

Hence (Lerman, 1981) for all triples x, y, z , if M is median and S is maximum, consider the open interval $]M(x, y, z), S(x, y, z)[$. If this open interval is empty, then the associated preorder is ultrametric.

Given a triplet $\{x, y, z\} \in J$ 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. We consider the mapping of all triplets J into all pairs F for the given preorder ω . We then define discrepancy between the structure of ω and the structure of an ultrametric preordonnance where $|\cdot|$ denotes cardinality:

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

The value 3 subtracted from $|F|$ takes account of the presence of the least, median and maximum distances. If ω is ultrametric then $H(\omega) = 0$. We are basically saying: the (open) interval between median and maximum of a triplet of distances is examined and the number of distances falling in this interval is counted. The “openness” of the median/maximum interval is important: in practice it means that we do not include the median nor maximum value, nor any values tied with them.

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 (uniformly) from the given point set.

Murtagh (2004) gives empirical results based on Lerman’s H-classifiability. There are two problems with Lerman’s index, however. Firstly, ultrametricity is associated with $H = 0$ but non-ultrametricity is not bounded (nor defined). In 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, and high dimensional, points, 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 high-dimensional spaces, due to increasing sparseness. 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 Discrepancy Between Subdominant Ultrametric and Input Data

In the Introduction we have indicated that creation of a hierarchical clustering, followed by comparison between the ultrametric distances found and the input set of dissimilarities, was an evident way to quantify ultrametricity, but suffered from some disadvantages. The single link hierarchical agglomerative clustering method has some attractive (and some unattractive!) properties. A constructive quantifying of ultrametricity was based on it. This we will now describe.

The quantifying of how ultrametric a data set is by Rammal et al. (1985, 1986) is given as an ultrametricity index: $\sum_{x,y} (d(x,y) - d_c(x,y)) / \sum_{x,y} d(x,y)$ where d is the metric distance being assessed, and d_c is the subdominant ul-

trametric. 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 chaining effect implies that for $d(x, y) \leq r_0, d(y, z) \leq r_0$ then $d(x, z) = |2r_0 - \epsilon$ for arbitrarily small ϵ . Hence $d(x, z)$ can be anomalously large. Another manifestation is the following pathology postulate of Watson (2003). The subdominant ultrametric d_c of a given metric d can be arbitrarily close to zero, even when there is an ultrametric quite close to d in the supremum norm. This is formulated as follows. If d is a metric on a finite set, then there is an ultrametric d_c which minimizes $\sup\{|d(x, y) - d_c(x, y)| : x, y \in X\}$ among those d_c such that $\forall x, y \in X, d_c(x, y) \leq d(x, y)$.

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, are increasingly ultrametric.

3.3 Distance-Based Measures

Treves (1997) considers triplets of points giving rise to minimal, median and maximal distances. In the plot of d_{\min}/d_{\max} against d_{med}/d_{\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} .

3.4 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 defined (by default, Euclidean). 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: $(y^2 + z^2 - xy)/2yz$.) This is our first necessary property for being an 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. This condition is targeting a configuration that is not exactly ultrametric but nonetheless 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.

The Fisher iris data (150×4) gives $\alpha = 0.0162$, indicating some, limited, ultrametricity. 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.

3.5 Ultrametricity Scaling with Data Size, Dimensionality, and Sparseness

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. (see above at end of section 3.2). 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 $K = 4$. We set a series of thresholds at intervals given by $(x_{\max} - x_{\min})/(K - 1)$. 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.

- 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 $\alpha = 0.096386, 0.078000, 0.077077, 0.075075, 0.079000, 0.71000$. There appears to be a small decrease in ultrametricity due to increasing density of points. (We found the same result, i.e. independence relative to n , with Lerman's index: see Murtagh, 2004.)

- Sparsity of coding helps greatly with ultrametricity. 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 occasion, 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 above, earlier in this section). We find, respectively: $\alpha = 0.271630, 0.247495, 0.260563, 0.264056, 0.269076, 0.275275$. 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 change 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).
- Dimensionality helps greatly with ultrametricity. Using $n = 5000$ real-valued points, uniformly distributed in space of dimensionality $m = 50, 100, 500, 1000, 5000$, we find: $\alpha = 0.183183, 0.271000, 0.544000, 0.707708, 0.979000$.
- Dimensionality and 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 $\alpha = 0.179179, 0.172172, 0.454910, 0.588000, 0.934000$.

4 Conclusion

We have clearly shown the dependence of our new ultrametricity coefficient, α , on numbers of points, space dimensionality, and sparsity of this space. Murtagh (2004) describes some of the computational implications of this work, for the processing of massive high-dimensional data sets.

References

- [1] J.P. Benzécri, *La Taxinomie*, 2nd ed., Dunod, Paris, 1979.
- [2] J.P. Benzécri, transl. T.K. Gopalan, *Correspondence Analysis Handbook*, Marcel Dekker, Basel, 1992.

- [3] Chandon, J.L., and De Soete, G., Fitting a least squares ultrametric to dissimilarity data: Approximation versus optimization, in E. Diday, M. Jambu, L. Lebart, J. Pagès, and R. Tomassone (eds.), *Data Analysis and Informatics III*, North-Holland, Amsterdam, 1984, pp. 213–221.
- [4] W.H.E. Day, Complexity theory: an introduction for practitioners of classification, in P. Arabie, L.J. Hubert and G. De Soete, Eds., *Clustering and Classification*, World Scientific, 1996, 199–233.
- [5] De Soete, G.: Least squares algorithms for constructing constrained ultrametric and additive tree representations of symmetric proximity data, *Journal of Classification*, 4, 155–173, 1987.
- [6] G. De Soete and J.D. Carroll, Tree and other network models for representing proximity data, in P. Arabie, L.J. Hubert and G. De Soete, Eds., *Clustering and Classification*, World Scientific, 1996, 157–197.
- [7] A.K. Hartmann, Are ground states of 3D $\pm J$ spin glasses ultrametric?, *Europhysics Letters*, 44, 249–254, 1998.
- [8] M. Krivánek and J. Morávek, NP-hard problems in hierarchical-tree clustering, *Acta Informatica*, 23, 311–323, 1986.
- [9] M. Krivánek and J. Morávek, On NP-hardness in hierarchical clustering, in T. Havránek, Z. Sidák and M. Novák, Eds., *Compstat 1984: Proceedings in Computational Statistics*, 189–194, Physica-Verlag, Vienna.
- [10] I.C. Lerman, *Classification et Analyse Ordinale des Données*, Paris: Dunod, 1981.
- [11] V. Makarenkov and B. Leclerc, An algorithm for fitting a tree metric according to a weighted least-squares criterion, *Journal of Classification*, 16, 3–26, 1999.
- [12] F. Murtagh, *Multidimensional Clustering Algorithms*, Physica-Verlag, Würzburg, 1985.
- [13] F. Murtagh, On ultrametricity, sparse coding, and computation, *Journal of Classification*, submitted, 2004.
- [14] 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.
- [15] R. Rammal, G. Toulouse and M.A. Virasoro, Ultrametricity for physicists, *Reviews of Modern Physics*, 58, 765–788, 1986.
- [16] A. Rizzi, Ultrametrics and p-adic numbers, in W. Gaul, O. Opitz and M. Schader, Eds., *Data Analysis: Scientific Modeling and Practical Application*, Springer-Verlag, 325–324, 2000.
- [17] A. Treves, On the perceptual structure of face space, *BioSystems*, 40, 189–196, 1997.
- [18] S. Watson, The classification of metrics and multivariate statistical analysis, preprint, York University, 27 pp., 2003.

Address: School of Computer Science, Queen’s University Belfast, Belfast BT7 1NN, Northern Ireland, UK

E-mail: f.murtagh@qub.ac.uk