Identifying and Exploiting Ultrametricity

Fionn Murtagh

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

Abstract. We begin with pervasive ultrametricity due to high dimensionality and/or spatial sparsity. How extent or degree of ultrametricity can be quantified leads us to the discussion of varied practical cases when ultrametricity can be partially or locally present in data. We show how the ultrametricity can be assessed in text or document collections, and in time series signals. In our presentation we also discussed applications to chemical information retrieval and to astrophysics, in particular observational cosmology.

1 Introduction

The topology or inherent shape and form of an object is important. In data analysis, the inherent form and structure of data clouds are important. Quite a few models of data form and structure are used in data analysis. One of them is a hierarchically embedded set of clusters, – a hierarchy. It is traditional (since at least the 1960s) to impose such a form on data, and if useful to assess the goodness of fit. Rather than fitting a hierarchical structure to data, our recent work has taken a different orientation: we seek to find (partial or global) inherent hierarchical structure in data. As we will describe in this article, there are interesting findings that result from this, and some very interesting perspectives are opened up for data analysis.

A formal definition of hierarchical structure is provided by ultrametric topology (in turn, related closely to p-adic number theory). We will return to this in section 2 below. First, though, we will summarize some of our findings.

Ultrametricity is a pervasive property of observational data. It arises as a limit case when data dimensionality or sparsity grows. More strictly such a limit case is a regular lattice structure and ultrametricity is one possible representation for it. Notwithstanding alternative representations, ultrametricity offers computational efficiency (related to tree depth/height being logarithmic in number of terminal nodes), linkage with dynamical or related functional properties (phylogenetic interpretation), and processing tools based on well

known p-adic or ultrametric theory (examples: deriving a partition, or applying an ultrametric wavelet transform).

Local ultrametricity is also of importance. Practical data sets (derived from, or observed in, databases and data spaces) present some but not exclusively ultrametric characteristics. This can be used for forensic data exploration (fingerprinting data sets, as we discuss below in section 5). Or, it can be used to expedite search and discovery in information spaces. Indeed we would like to go a lot further, and gain new insights into data (and observed phenomena and events) through ultrametric or p-adic representations. We see this as a program of work for the near future.

2 Quantifying Degree of Ultrametricity

Summarizing a full description in Murtagh (2004) we explored two measures quantifying how ultrametric a data set is, – Lerman's and a new approach based on triangle invariance (respectively, the second and third approaches described in this section).

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 (Lerman (1981)). For example, the triangle formed by any triplet is necessarily isosceles, with the two large sides equal; or is equilateral.

Firstly, Rammal et al. (1986) used discrepancy between each pairwise distance and the corresponding subdominant ultrametric. Now, the subdominant ultrametric is also known as the ultrametric distance resulting from the single linkage agglomerative hierarchical clustering method. Closely related graph structures include the minimal spanning tree, and graph (connected) components. While the subdominant provides a good fit to the given distance (or indeed dissimilarity), it suffers from the "friends of friends" or chaining effect.

Secondly, Lerman (1981) developed a measure of ultrametricity, termed H-classifiability, using ranks of all pairwise given distances (or dissimilarities). The isosceles (with small base) or equilateral requirements of the ultrametric inequality impose constraints on the ranks. The interval between median and maximum rank of every set of triplets must be empty for ultrametricity. We have used extensively Lerman's measure of degree of ultrametricity in a data set. Taking ranks provides scale invariance. But the limitation of Lerman's approach, we find, is that it is not reasonable to study ranks of real-valued distances defined on a large set of points.

Thirdly, our own measure of extent of ultrametricity (Murtagh (2004)) can be described algorithmically. We assume a Euclidean metric. (In view of the use of scalar product in the definition of an angle, we assume a Hilbert space for our data.) We examine triplets of points (exhaustively if possible, or

otherwise through sampling), and determine the three angles formed by the associated triangle. We select the smallest angle formed by the triplet points. Then we check if the other two remaining angles are approximately equal. If they are equal then our triangle is isosceles with small base, or equilateral (when all triangles are equal). The approximation to equality is given by 2 degrees (0.0349 radians). Our motivation for the approximate ("fuzzy") equality is that it makes our approach robust and independent of measurement precision.

Studies are discussed in Murtagh (2004) showing how numbers of points in our clouds of data points are irrelevant; but what counts is the ambient spatial dimensionality. Among cases looked at are statistically uniformly (hence "unclustered", or without structure in a certain sense) distributed points, and statistically uniformly distributed hypercube vertices (so the latter are random 0/1 valued vectors). Using our ultrametricity measure, there is a clear tendency to ultrametricity as the spatial dimensionality (hence spatial sparseness) increases (Murtagh (2004)).

3 Ultrametricity and Dimensionality

3.1 Distance Properties in Very Sparse Spaces

Murtagh (2004), and earlier work by Rammal et al. (1985, 1986), has demonstrated the pervasiveness of ultrametricity, by focusing on the fact that sparse high-dimensional data tend to be ultrametric. One reason for this is as follows.

As dimensionality grows, so too do distances (or indeed dissimilarities, if they do not satisfy the triangular inequality). The least change possible for dissimilarities to become distances has been formulated in terms of the smallest additive constant needed, to be added to all dissimilarities (Torgerson (1958), Cailliez and Pagès (1976), Cailliez (1983), Neuwirth and Reisinger (1982)). Adding a sufficiently large constant to all dissimilarities transforms them into a set of distances. Through addition of a larger constant, it follows that distances become approximately equal, thus verifying a trivial case of the ultrametric or "strong triangular" inequality. Adding to dissimilarities or distances may be a direct consequence of increased dimensionality.

For a close fit or good approximation, the situation is not as simple for taking dissimilarities, or distances, into ultrametric distances. A best fit solution is given by De Soete (1986) (and software is available in R, Hornik (2005)). If we want a close fit to the given dissimilarities then a good choice would avail either of the maximal inferior, or subdominant, ultrametric; or the minimal superior ultrametric. Stepwise algorithms for these are commonly known as, respectively, single linkage hierarchical clustering; and complete link hierarchical clustering. (See Benzécri (1979), Lerman (1981), Murtagh (1985) and other texts on hierarchical clustering.)

3.2 Very High Dimensions are Naturally Ultrametric

Bellman's (1961) "curse of dimensionality" relates to exponential growth of hypervolume, and hence complexity, as a function of dimensionality. Problems become tougher as dimensionality increases. In particular problems related to proximity search in high-dimensional spaces tend to become intractable.

In a way, a "trivial limit" (Treves (1997)) case is reached as dimensionality increases. This makes high dimensional proximity search very different, and given an appropriate data structure – such as a binary hierarchical clustering tree – we can find nearest neighbors in worst case $O(1)$ or constant computational time (Murtagh (2004)). The proof is simple: the tree data structure affords a constant number of edge traversals.

The fact that limit properties are "trivial" makes them no less interesting to study. Let us refer to such "trivial" properties as (structural or geometrical) regularity properties (e.g. all points lie on a regular lattice). First of all, the symmetries of regular structures in our data may be of importance. Secondly, "islands" or clusters in our data, where each "island" is of regular structure, may be exploitable. Thirdly, the mention of exploitability points to the application areas targeted: in this article, we focus on search and matching and show some ways in which ultrametric regularity can be exploited in practice. Fourthly, and finally, regularity by no means implies complete coverage (e.g., existence of all pairwise linkages) so that interesting or revealing structure will be present in real data sets.

Thus we see that in very high dimensions, and/or in very (spatially) sparse data clouds, there is no longer a "curse of dimensionality".

4 Approximating Local Ultrametricity

Now we look at data where some triangles are consistent with ultrametric properties, while others are not.

It has long been known (Chávez et al. (2001), van Rijsbergen (1979)) that forms of data structuring, and more particularly data clustering, can be used to expedite search problems in high dimensions. Some of the work of Chávez and Navarro and their colleagues provides an explanation as to why and how clustering can be exploited for high dimensional proximity search.

In large data sets, i.e. large n or number of observations, a clever way to expedite proximity searching (in particular nearest neighbor finding) in metric spaces is as follows. The metric property implies that the triangular inequality holds. We have a given point and we are looking for its nearest neighbor. We use a third point, called a pivot point. Such a pivot point is carefully selected at the start of the processing, and all necessary distances to it are stored. Through the triangular inequality, we then form a bound on the best potential nearest neighbor distance. Thereby we limit the region within which the search is carried out. See Chávez et al. (2000, 2001, 2003), Bustos et

al. (2003)). As pointed out in Murtagh (2004), the bounding rule, or rejection rule, that ensues, is forcing retained triangles to be isosceles. This is interesting because it can be viewed as finding locally ultrametric relationships.

In Chávez et al. (2000, 2001) the ambient spatial dimension is termed the "representational dimension", or embedding dimension, m. (This is dimensionality, m: we have for example $x \in \mathbb{R}^m$.) Search is subject to the curse of dimensionality when addressed in all generality in $\mathbb{R}^m$. However there is often a smaller "intrinsic dimensionality", or average local dimensionality (e.g. when the data are clustered, or lie on a surface of dimension $< m$). This can be exploited to provide fast proximity searching opportunities. However it is difficult in general to define the intrinsic dimensionality.

These authors (Chávez et al. (2000, 2001)) define intrinsic dimensionality of a metric space as: $\rho = \frac{\mu^2}{2\sigma^2}$ where μ and σ^2 are, respectively, the mean and variance of the distances.

So, firstly, the intrinsic dimensionality grows with the mean distance. We have observed that ultrametricity increases with average distance both by simulations in Murtagh (2004), and also through the argument of a simple additive transformation (in section 3.1 above). Secondly, the intrinsic dimensionality grows with inverse variance. Small variance of distances implies equilateral triangles between point triplets, and therefore implies ultrametricity.

We see therefore that the intrinsic dimensionality of Chávez et al. (2000, 2003) affords another definition of ultrametricity. We have already observed how their fast, pivot-based proximity rule can be interpreted as local enforcement of the ultrametric inequality. We conclude from these observations that local or global ultrametricity (i.e., high values of Chávez and Navarro's ρ, or high local contributions to ρ) permit fast proximity search.

5 Increasing Ultrametricity Through Data Recoding

5.1 Ultrametricity of Time Series

In Murtagh (2005a) we use the following coding to show that chaotic time series are less ultrametric than, say, financial, biomedical or meteorological time series; random generated (uniformly distributed) time series data are remarkably similar in their ultrametric properties; and ultrametricity can be used to distinguish various types of biomedical (EEG) signals. Our methodology is empirical: we took 44 time series, and investigated different user parameters.

A time series can be easily embedded in a space of dimensionality m, by taking successive intervals of length m, or a delay embedding of order m. Thus we define points

$$\mathbf{x}_r = (x_{r-m+1}, x_{r-m+2}, \ldots, x_{r-1}, x_r)^t \in \mathbb{R}^m$$

where t denotes vector transpose.

Given any $\mathbf{x}_r = (x_{r-m+1}, x_{r-m+2}, \ldots, x_{r-1}, x_r)^t \in \mathbb{R}^m$, let us consider the set of s such contiguous intervals determined from the time series of overall size n. For convenience we will take $s = \lfloor n/m \rfloor$ where $\lfloor . \rfloor$ is integer truncation. The contiguous intervals could be overlapping but for exhaustive or near-exhaustive coverage it is acceptable that they be non-overlapping. In our work, the intervals were non-overlapping. The quantification of the ultrametricity of the overall time series is provided by the aggregate over s time intervals of the ultrametricity of each $\mathbf{x}_r$, $1 \leq r \leq s$.

We seek to directly quantify the extent of ultrametricity in time series data. In Rammal et al. (1986) and Murtagh (2004) it was shown how increase in ambient spatial dimensionality leads to greater ultrametricity. However it is not satisfactory from a practical point of view to simply increase the embedding dimensionality m insofar as short memory relationships are of greater practical relevance (especially for prediction). The greatest possible value of m is the total length of the time series, n. Instead we will look for an ultrametricity measurement approach for given and limited sized dimensionality m. Our experimental results for real and for random data sets are for "window" lengths $m = 5, 10, \ldots, 105, 110$.

We seek local ultrametricity, i.e. hierarchical structure, by studying the following: Euclidean distance squared, $d_{jj'} = (x_{rj} - x_{rj'})^2$ for all $1 \leq j, j' \leq m$ in each time window, $\mathbf{x}_r$.

We enforce sparseness (Rammal et al. (1985), Rammal et al. (1986), Murtagh (2004)) on our given distance values, $\{d_{jj'}\}$. We do this by linearly approximating each value $d_{jj'}$, in the range $\max_{jj'} d_{jj'} - \min_{jj'} d_{jj'}$, by an integer in $1, 2, \ldots p$. Note that the range is chosen with reference to the currently considered time series window, $1 \leq j, j' \leq m$. Note too that the value of p must be specified. In our work we set $p = 2$. Thus far, the recoded value, $d'_{jj'}$ is not necessarily a distance. With the extra requirement that $d'_{jj'} \longrightarrow 0$ whenever $j = j'$ it can be shown that $d'_{jj'}$ is a metric (Murtagh, 2005a).

To summarize, in our coding, a small pairwise transition is mapped onto a value of 1; and a large pairwise transition is mapped onto a value of 2. A pairwise transition is defined not just for data values that are successive in time but for any pair of data values in the window considered.

This coding can be considered as (i) taking a local region, defined by the sliding window, and (ii) coding pairwise "change" = 2, versus "no change" = 1, relationships. Then, based on these new distances, we use the ultrametric triangle properties to assess conformity to ultrametricity. The average overall ultrametricity in the time series, quantified in this way, allows us to fingerprint our time series.

5.2 Ultrametricity of Text

In Murtagh (2006a), words appearing in a text (in principle all, but in practice a set of the few hundred most frequent) are used to fingerprint the text. Rare words in a text corpus may be appropriate for querying the corpus for

relevant texts, but such words are of little help for inter-text characterization and comparison. We also use entire words, with no stemming or other preprocessing. A full justification for such an approach to textual data analysis can be found in Murtagh (2005b).

So our methodology for studying a set of texts is to characterize each text with numbers of terms appearing in the text, for a set of terms. The χ^2 distance is an appropriate weighted Euclidean distance for use with such data (Benzécri (1979), Murtagh (2005b)). Consider texts i and i' crossed by words j. Let k_{ij} be the number of occurrences of word j in text i. Then, omitting a constant, the χ^2 distance between texts i and i' is given by $\sum_j 1/k_j(k_{ij}/k_i - k_{i'j}/k_{i'})^2$. The weighting term is $1/k_j$. The weighted Euclidean distance is between the *profile* of text i, viz. k_{ij}/k_i for all j, and the analogous *profile* of text i'. (Our discussion is to within a constant because we actually work on *frequencies* defined from the numbers of occurrences.)

Correspondence analysis allows us to project the space of documents (we could equally well explore the terms in the *same* projected space) into a Euclidean space. It maps the all-pairs χ^2 distance into the corresponding Euclidean distance. In the resulting factor space, we use our triangle-based approach for quantifying how ultrametric the data are.

We did this for a large number of texts (novels – Jane Austen, James Joyce, technical reports – airline accident reports, fairy tales – Brothers Grimm, dream reports, Aristotle's *Categories*, etc.), finding consistent degree of ultrametricity results over texts of the same sort.

Some very intriguing ultrametricity characterizations were found in our work. For example, we found that the technical vocabulary of air accidents did not differ greatly in terms of inherent ultrametricity compared to the Brothers Grimm fairy tales. Secondly we found that novelist Austen's works were distinguishable from the Grimm fairy tales. Thirdly we found dream reports to be have higher ultrametricity level than the other text collections.

5.3 Data Recoding in the Correspondence Analysis Tradition

If the χ^2 distance (see above, section 5.2) is used on data tables with constant marginal sums then it becomes a weighted Euclidean distance. This is important for us, because it means that we can directly influence the analysis by equi-weighting, say, the table rows in the following way: we double the row vector values by including an absence (0 value) whenever there is a presence (1 value) and vice versa. Or for a table of percentages, we take both the original value x and $100 - x$. In the correspondence analysis tradition (Benzécri (1979), Murtagh (2005b)) this is known as *doubling* (*dédoublement*).

More generally, booleanizing, or making qualitative data in this way, for a varying (value-dependent) number of target value categories (or modalities) leads to the form of coding known as *complete disjunctive form*.

Such coding increases the embedding dimension, and data sparseness, and thus may encourage degree of ultrametricity. That it can do more we will now show.

The iris data has been very widely used as a toy data set since Fisher used it in 1936 (taking from a 1935 article by Anderson) to exemplify discriminant analysis. It consists of 150 iris flowers, each characterized by 4 petal and sepal, width and breadth, measurements. On the one hand, therefore, we have the 150 irises in $\mathbb{R}^4$. Next, each variable value was recoded to be a rank (all ranks of a given variable considered) and the rank was boolean-coded (viz., for the top rank variable value, $1000\ldots$, for the second rank variable value, $0100\ldots$, etc.). Following removal of zero total columns, the second data set defined the 150 irises in $\mathbb{R}^{123}$. Actually, this definition of the 150 irises is in fact in $\{0,1\}^{123}$.

Our triangle-based measure of the degree of ultrametricity in a data set (here the set of irises), with 0 = no ultrametricity, and 1 = every triangle an ultrametric-respecting one, gave the following: for irises in $\mathbb{R}^4$, 0.017; and for irises in $\{0,1\}^{123}$: 0.948.

This provides a nice illustration of how recoding can dramatically change the picture provided by one's data. Furthermore it provides justification for data recoding if the ultrametricity can be instrumentalized by us in some way (e.g. to facilitate fast proximity search).

6 Conclusions

It has been our aim in this work to link observed data with an ultrametric topology for such data. The traditional approach in data analysis, of course, is to impose structure on the data. This is done, for example, by using some agglomerative hierarchical clustering algorithm. We can always do this (modulo distance or other ties in the data). Then we can assess the degree of fit of such a (tree or other) structure to our data.

For our purposes, here, this is unsatisfactory.

Firstly, our aim was to show that ultrametricity can be naturally present in our data, globally or locally. We did not want any "measuring tool" such as an agglomerative hierarchical clustering algorithm to overly influence this finding. (Unfortunately Rammal et al. (1986) suffers from precisely this unhelpful influence of the "measuring tool" of the subdominant ultrametric.)

Secondly, let us assume that we did use hierarchical clustering, and then based our discussion around the goodness of fit. This again is a traditional approach used in data analysis, and in statistical data modeling. But such a discussion would have been unnecessary and futile. For, after all, if we have ultrametric properties in our data then many of the widely used hierarchical clustering algorithms will give precisely the same outcome, and furthermore the fit is by definition exact.

In linking data with an ultrametric view of it we have, in this article, proceeded a little in the direction of exploiting this achievement. While some applications, like discrimination between time series signals, or texts, have been covered here, other areas have just been opened up, e.g. search and discovery in massive biochemical databases; hierarchical structures in cosmology (Murtagh (2006b); and automated ontology creation for semantic web applications. In the distance there looms the challenge of analysis of networks of enormous size (internet, or biological). There is a great deal of work to be accomplished.

References

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

BENZÉCRI, J.P. (1979): *L'Analyse des Données, Tome I Taxinomie, Tome II Correspondances*. 2nd ed., Dunod, Paris.

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 de Données*. SMASH (Société de Mathématiques Appliquées et de Sciences Humaines), Paris.

CAILLIEZ, F. (1983): The Analytical Solution of the Additive Constant Problem. *Psychometrika, 48, 305–308.*

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, 39–56.*

DE SOETE, G. (1986): A Least Squares Algorithm for Fitting an Ultrametric Tree to a Dissimilarity Matrix. *Pattern Recognition Letters, 2, 133–137.*

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

HORNIK, K. (2005): A CLUE for CLUster Ensembles. *Journal of Statistical Software, 14, 12.*

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

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

MURTAGH, F. (2004): On Ultrametricity, Data Coding, and Computation. *Journal of Classification, 21, 167–184.*

MURTAGH, F. (2005a): Identifying the Ultrametricity of Time Series. *European Physical Journal B, 43, 573–579.*

MURTAGH, F. (2005b): *Correspondence Analysis and Data Coding with R and Java*. Chapman & Hall/CRC, Florida.

MURTAGH, F. (2006a): A Note on Local Ultrametricity in Text, *Literary and Linguistic Computing*, submitted.

MURTAGH, F. (2006b): From Data to the Physics using Ultrametrics: New Results in High Dimensional Data Analysis. In A.Yu. Khrennikov, Z. Rakić, and I.V. Volovich (Eds.): *p-Adic Mathematical Physics*, American Institute of Physics Conf. Proc. Vol. 826, 151–161.

NEUWIRTH, E. and REISINGER, L. (1982): Dissimilarity and Distance Coefficients in Automation-Supported Thesauri. *Information Systems, 7, 47–52*.

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*.

TORGERSON, W.S. (1958): *Theory and Methods of Scaling*, Wiley, New York.

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

VAN RIJSBERGEN, C.J. (1979): *Information Retrieval*, 2nd ed. Butterworths.