

On the Local Ultrametricity of Finite Metric Data

Patrick Erik Bradley
Karlsruhe Institute of Technology
Institute of Photogrammetry and Remote Sensing
Englerstr. 6
76131 Karlsruhe
Germany

August 15, 2024

Abstract

New local ultrametricity measures for finite metric data are proposed through the viewpoint that their Vietoris-Rips corners are samples from p -adic Mumford curves endowed with a Radon measure coming from a regular differential 1-form. This is experimentally applied to the iris dataset.

Keywords: local ultrametricity, p -adic numbers, finite data, Mumford curves, Vietoris-Rips complex, data analysis

1 Introduction

Ultrametricity is appealing for many reasons, and in particular the simplicity of tree structures encoded in ultrametric spaces seems attractive to data analysts. Because of this, they would like to see how close to an ultrametric space a given data set is in order to extract something meaningful out of a hierarchical classification of the data. With this in mind, ultrametricity indices have been proposed, e.g. by [11] or [9]. F. Murtagh observed experimentally that samples which are sparse and random in hypercubes become more and more ultrametric as dimension increases, using his ultrametricity index [9]. Explanations for this are given in [3] and [15]. Also, ultrametricity can be related to topological data analysis [4], and a corresponding ultrametricity index has a logistic behaviour [5]. This index relies on the Vietoris-Rips complex developed in [12], and which is important in studying the persistent homology of data. Cf. e.g. [14] for a fast construction of the Vietoris-Rips complex.

The p -adic numbers, having an inherent regular hierarchical structure, provide a framework for analysing hierarchical data, and thus p -adic encoding methods were devised, [10] or [2], either in order to bring them closer to ultrametricity or to apply p -adic methods to their already existing hierarchical structure. This leads to the applicability of p -adic analysis outlined e.g. in [13] to the investigation of data.

The scope of this article is to introduce new measures for *local* ultrametricity, arguing that the clusters appearing as connected components of the Vietoris-Rips graphs are likely to be more ultrametric than the whole dataset, which can be seen in an example case taken from the well-known iris dataset. Whether or not this argument is generally valid or not, the viewpoint induced by this approach leads to the idea that data can be seen

as being sampled from Mumford curves. These are p -adic compact algebraic manifolds of dimension 1. Locally, they are holed discs in the p -adic number field, on which there is a natural Haar measure. However, the irregular tree structure of the local data leads to a more natural Radon measure coming from an algebraic regular differential 1-form on the Mumford curve, as constructed in [6]. There, the subdominant ultrametric associated with a finite metric space is used, which can be calculated with the method of [11]. In fact, any ultrametric can be used to approximate the finite metric dataset, i.e. any hierarchical classification method can be used in order to obtain an ultrametric in this approach.

Mumford curves are objects studied in p -adic algebraic and rigid geometry, and are extensively covered in [8] and [7]. What is needed from this relatively deep theory is, however, only the fact that they are algebraic and have an underlying 1-dimensional compact p -adic manifold structure which allows for regular differential 1-forms ω , which are in fact algebraic. Locally, they are of the form

$$\omega(x) = f(x) dx$$

with an analytic p -adic-valued function f defined on the local piece U , and that these give rise to Radon measures on the Mumford curve outside the zeros of ω .

2 Finite locally ultrametric spaces

After defining local ultrametrics in the following subsection, and local p -adic encodings of data via tree embeddings, new invariants of a finite metric space are defined via the Vietoris-Rips graphs and associating Mumford curves and Radon measures to local pieces in the last subsection of this section.

2.1 Local ultrametrics

Let X be a finite set with a metric d on it. Fix $\epsilon > 0$, and let Γ_ϵ be the associated Vietoris-Rips graph with vertex set X . Let

$$d_\epsilon: X \times X \rightarrow \mathbb{R}_{\geq 0}$$

be the partial function which on each connected component C of Γ_ϵ is an ultrametric dominated by d . One can use for d_ϵ e.g. the subdominant corresponding to the distance on X restricted to $C \times C$. But any other ultrametric dominated by d can also be used. Certain hierarchical clustering methods

provide such an ultrametric, among which single-linkage clustering yields the subdominant ultrametric.

Let $\mathcal{C}(\Gamma_\epsilon)$ be the set of connected components of Γ_ϵ . Define a distance d'_ϵ on $\mathcal{C}(\Gamma_\epsilon)$ as

$$d'_\epsilon(C, C') = \min \{ \epsilon' \mid \epsilon' \geq \epsilon : \exists \text{ an edge in } \Gamma_{\epsilon'} \text{ connecting } C \text{ and } C' \}$$

whenever $C \neq C'$. Then define the function

$$\delta_\epsilon: X \times X \rightarrow \mathbb{R}_{\geq 0}, (x, y) \mapsto \begin{cases} d_\epsilon(x, y), & \exists C \in \mathcal{C}(\Gamma_\epsilon): x, y \in C \\ d'_\epsilon(C(x), C(y)), & C(x) \neq C(y) \end{cases}$$

where $C(z) \in \mathcal{C}(\Gamma_\epsilon)$ is the connected component containing $z \in X$. Clearly, δ_ϵ is a distance on X .

Definition 2.1. *The distance δ_ϵ is called a local ultrametric on X . The pair (X, δ_ϵ) is called a locally ultrametric space.*

A criterion for ultrametricity in terms of the Vietoris-Rips graphs is given in [3, Lem. 2.2]: the connected components of the Vietoris-Rips graphs are always cliques iff dataset is ultrametric. The subdominant ultrametric can also be described in terms of the Vietoris-Rips graphs, cf. [3, Prop. 5.2].

2.2 Local p -adic encodings

In [6, §3.3] a Radon measure on a compact open subset of \mathbb{Q}_p is constructed from a finite ultrametric space. Here, an embedding of the corresponding ultrametric tree into the Bruhat-Tits tree of a suitable p -adic number field necessary for that method is constructed in a more precise manner. This produces a p -adic data encoding, as already observed in [2].

Let $C \in \mathcal{C}(\Gamma_\epsilon)$ be given, and view (C, d_ϵ) as an independent ultrametric space for the moment. The set $\mathcal{B}(C)$ of all non-trivial balls on C is a finite poset with precisely one top element C , and in fact is a tree. Let

$$\rho: \mathcal{B}(C) \rightarrow \mathbb{R}_{>0}, B \mapsto \text{radius of } B$$

whose image $R(C) = \rho(\mathcal{B}(C))$ is a finite ordered set of real numbers. Order this set with a function

$$\varphi: R(X) \rightarrow \mathbb{N}$$

in decreasing order with consecutive natural numbers beginning in 0. Fix a prime number p , and let

$$m = \max \varphi$$

and assign to each $c \in C$ a distinct disc $a_c + p^{(m+1)}\mathbb{Z}_p$ inside the ring \mathbb{Z}_p of p -adic integers inside the field of p -adic numbers \mathbb{Q}_p , where p is bounded from below by the maximal number of children in any ultrametric tree of (C, δ_ϵ) for any $C \in \mathcal{C}(\Gamma_\epsilon)$ plus the number of elements in $\mathcal{C}(\Gamma_\epsilon)$. Assume thereby that all discs in \mathbb{Z}_p have equal radius

$$p^{-(m+1)}$$

for this assignment. The condition about p enables an embedding of any spanning tree of the graph Γ_ϵ into the Bruhat-Tits tree for \mathbb{Q}_p . The latter tree is explained e.g. in [1].

The ultrametric diffusion considered in [6] necessitated the replacement of the p -adic Haar measure on the compact open set obtained by such an embedding as above with a Radon measure ν_C which ensures that the volume of any disc corresponding to a vertex $B \in \mathcal{B}(C)$ in the ultrametric tree for (c, d_ϵ) is equally distributed among the child vertices of B , cf. [6, Lem. 3.8], where such a Radon measure is constructed on the subset

$$\Omega_C = \bigsqcup_{c \in C} (a_c + p^{m+1}\mathbb{Z}_p)$$

of \mathbb{Q}_p .

Definition 2.2. *The measure ν_C is called the equity measure on Ω_C induced by (C, d_ϵ) .*

Now, [6, Lem. 3.9] shows that ν_C is of the form

$$\nu(x) = \phi(|f_C(x)|) |dx|_p$$

where $|dx|_p$ is the Haar measure on \mathbb{Q}_p , $f_C \in \mathbb{Q}_p[X]$ a polynomial nowhere vanishing on Ω_C , and $\phi: p^{\mathbb{Z}} \rightarrow \mathbb{R}_{>0}$ a strictly increasing function. The proof of [6, Lem. 3.9] uses p -adic polynomial interpolation.

2.3 Invariants of a finite metric space

Let $\delta \geq \epsilon$. Define the graph Γ_ϵ^δ whose vertex set is $\mathcal{C}(\Gamma_\epsilon)$, and its edges are pairs (C, C') with $C \neq C'$ and $d(C, C') \leq \delta$. We call this the *coarse ϵ - δ graph*

of (X, d) .

For $x \in X$ define

$$C_\epsilon^\delta(x) = \text{the connected comp. of } \Gamma_\epsilon^\delta \text{ containing } C(x) \in \mathcal{C}(\Gamma_\epsilon)$$

where $C(x) \in \mathcal{C}(\Gamma_\epsilon)$ is the connected component of Γ_ϵ containing $x \in X$. The *local ϵ - δ -genus* is the function

$$g_\epsilon^\delta: X \rightarrow \mathbb{N}, \quad x \mapsto b_1(C_\epsilon^\delta(x))$$

where C_ϵ^δ is viewed as a subgraph of Γ_ϵ^δ .

Lemma 2.3. *It holds true that*

$$0 \leq g_\epsilon^\delta(x) \leq \frac{1}{2} |C_\epsilon^\delta(x)|^2 - \frac{3}{2} |C_\epsilon^\delta(x)| + 1$$

for $\delta \geq \epsilon > 0$. (X, d) is ultrametric, if and only if for all $\delta \geq \epsilon > 0$, the right hand side is an equality.

Proof. The ϵ - δ -genus is non-negative and is maximal, when $C_\epsilon^\delta(x)$ is a complete graph, in which case the first Betti number equals the right hand side of the asserted inequality. The last statement now follows immediately from [3, Lem. 2.2] applied to the quotient metric on the set $\mathcal{C}(\Gamma_\epsilon)$ induced by d . \square

Such a connected component $C = C_\epsilon^\delta(x)$ for $x \in X$ can be viewed as a coarse graph structure on the connected components of Γ_ϵ . Now, replacing, as in the previous subsection, the distance d , restricted to each element of $\mathcal{C}(\Gamma_\epsilon)$ which is contained in C , with an ultrametric, leads to a local tree structure on C . Now, the previous subsection tells us that locally, there is a Radon measure $\nu(x)$ coming from a p -adic differential 1-form ω_ϵ which is algebraic. The local pieces can now be “glued” to a covering of the p -adic points of a Mumford curve \mathcal{C}_ϵ minus the zeros of the regular algebraic differential 1-form ω_ϵ , whose genus equals the first Betti number of C . Notice that the gluing process takes place beyond the mere points whose coordinates are in \mathbb{Q}_p , in the category of p -adic rigid analytic spaces, as laid out e.g. in [7, Ch. 5]. The method from the previous subsection fills a gap of [6, §3.3] by explicitly constructing the equity measure on the Mumford curve, which according to [6, Lem. 3.9] comes from an algebraic differential 1-form ω_ϵ .

Hence, apart from the local ϵ - δ -genus, there is also the equity measure $|\omega_\epsilon|_p$ on each connected component of Γ_ϵ^δ as a further set of invariant of the dataset (X, d) . As another invariant, we suggest also the minimal value δ for which Γ_ϵ^δ is connected, together with the now global ϵ - δ -genus and equity measure for this δ .

3 Experiments

The iris dataset was investigated. For $\epsilon \geq 1.65$, the Vietoris-Rips graph Γ_ϵ becomes connected. Figure 1 shows Vietoris-Rips graphs for four selected values of ϵ . Taking as reference $\epsilon = 0.64$ leads to five clusters (i.e. connected components of Γ_ϵ) having distance matrix as in Table 1. Reference $\epsilon = 0.7$ leads to four clusters with distance matrix as in Table 2. The clusters C_2, C_3 of the first graph merge to cluster C of the second graph. Figure 2 shows the two quotient graphs Γ_ϵ^δ for these values of ϵ , and for the corresponding minimal $\delta > \epsilon$, for which the associated Mumford curve is connected. In the first case, the genus is one, and in the second case it is zero.

Our choice of an ultrametricity is biased by [9, §3.3] who introduced his ultrametricity index also because of the chaining effect problem of the ultrametricity index from [11]. The values of the Murtagh ultrametricity index of the not too small clusters are given in Table 3 and indicate that they are more ultrametric than the whole iris dataset whose Murtagh index is given as 0.0162 in [9, §3.3].

The method of [9, §3.3] of calculating the Murtagh ultrametricity index consists in counting almost ultrametric triangles as follows:

1. Randomly sample the coordinates for triples of three distinct points.
2. Check for possible degenerate triangles and exclude these.
3. The cosine of the angle facing a side of length x is:

$$\frac{y^2 + z^2 - x^2}{2yz}$$

where y, z are the other side lengths.

4. For the two other angles seek an angular difference of at most 2 degrees (0.03490656 radians).

Murtagh's ultrametricity index is then the fraction α of such almost ultrametric triangles.

The 3-adic Radon measure leading to the equity measure on one cluster is now computed as an exemplary study of the iris dataset. The subdominant ultrametric of cluster C_3 in $\Gamma_{0.64}$ can be depicted as the dendrogram w.r.t.

$\epsilon = 0.64$	C_1	C_2	C_3	C_4	C_5
C_1	0	2.08	1.64	3.14	5.46
C_2	2.08	0	0.648	0.735	0.819
C_3	1.64	0.648	0	1.32	4.59
C_4	3.14	0.735	1.32	0	3.79
C_5	5.46	0.819	4.59	3.79	0

Table 1: Cluster distances at $\epsilon = 0.64$.

$\epsilon = 0.7$	C_1	C	C_4	C_5
C_1	0	1.64	3.14	5.46
C	1.64	0	0.735	0.819
C_4	3.14	0.835	0	3.79
C_5	5.46	0.819	3.79	0

Table 2: Cluster distances at $\epsilon = 0.7$.

single-linkage clustering, and is shown in Figure 3. In the case of $p = 3$, assign now to each point of C_3 a 3-adic ball of radius 3^{-2} , centred in

$$0, \quad 1, \quad 1 + 3^2, \quad 1 + 3^2 + 3^3,$$

a set which has the 3-adic tree structure of Figure 3. The equity measure leads to the assignment

$$0 \mapsto \frac{1}{2}, \quad 1 \mapsto \frac{1}{4}, \quad 1 + 3^2 \mapsto \frac{1}{8}, \quad 1 + 3^2 + 3^3 \mapsto \frac{1}{8}$$

leading to the 3-adic interpolation problem

$$f(0) = 1, \quad f(1) = 3, \quad f(1 + 3^2) = 3^2, \quad f(1 + 3^2 + 3^3) = 3^2$$

with the solution

$$f(X) = \frac{31}{9990}X^3 - \frac{1683}{9990}X^2 + \frac{10811}{4995}X + 1 \in \mathbb{Q}_3[X]$$

for the new measure

$$|\omega(x)|_3 = |f(x)|_3 |dx|_3$$

approximating the equity measure for $x \in C_3$ with $\epsilon = 0.64$.

cluster	C_1	C_2	C
Murtagh index	0.026	0.11	0.12

Table 3: The Murtagh ultrametricity index values for the larger clusters in the iris dataset.

4 Conclusion

In the theoretical part of this work, a local ultrametric is associated with a finite metric space (given the meaning of a dataset) via its Vietoris-Rips graph. The equity measure was defined on the local ultrametric parts of the space which, by the result of recent previous work, can be seen as coming from a differential 1-form on a p -adic Mumford curve. These can be viewed as compact algebraic p -adic manifolds, and this approach leads to new invariants for finite metric data by taking a double filtration with ϵ - and δ -balls in the finite metric. In the experimental part of this work, it is conceivable from the findings with the iris dataset, that local ultrametricity could increase within clusters of the Vietoris-Rips graphs in comparison with the ultrametricity of the whole dataset. This resembles *Simpson's paradox* in statistics. Even if this may not be always the case for general datasets, this nevertheless provides a means for classifying different types of datasets. In the iris data case, example Vietoris-Rips graphs were taken and some values of the new invariants calculated. The findings suggest that the double filtration approach can reveal more inherent topological properties of data in their ultrametric approximation via p -adic encoding. This suggests that in the future, a hierarchical or p -adic version of manifold learning could emerge from further investigations in this direction, which is appealing because of its potential for a reduced computational complexity.

Acknowledgements

Data Availability

This research has used publicly available data only. The R code can be made available upon request.

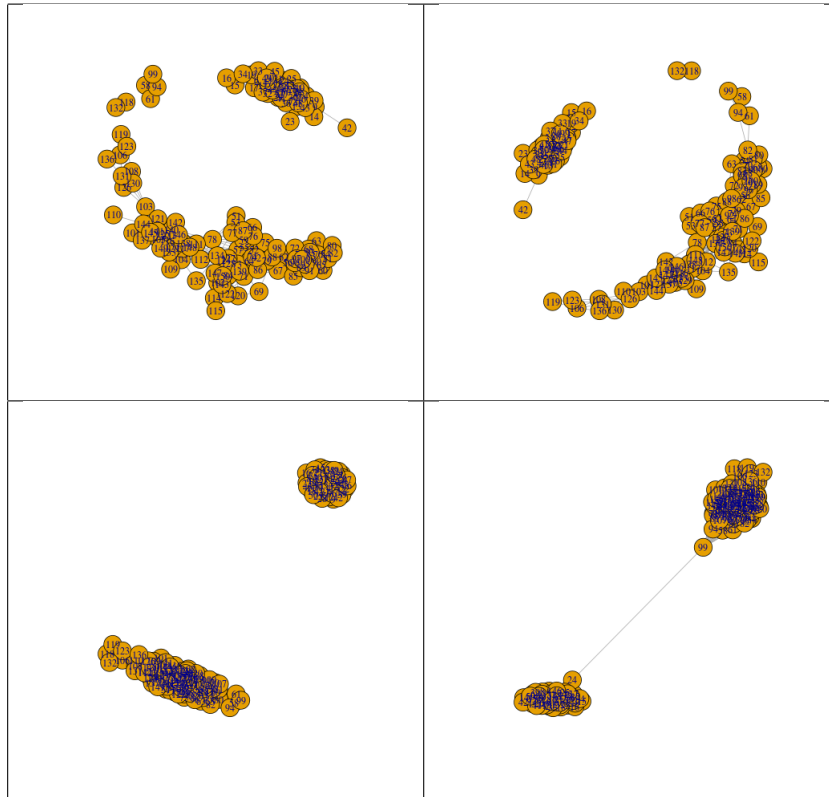


Figure 1: Vietoris-Rips graphs Γ_ϵ for $\epsilon = 0.64$ (top left), $\epsilon = 0.7$ (top right), $\epsilon = 1.640$ (bottom left), $\epsilon = 1.650$ (bottom right). Only the non-singleton connected components are shown.

References

- [1] P.E. Bradley. Degenerating families of dendrograms. *Journal of Classification*, 25(1):27–42, 2006.
- [2] P.E. Bradley. Mumford dendrograms. *The Computer Journal*, 53(4):393–404, 2010.
- [3] P.E. Bradley. Ultrametricity indices for the Euclidean and Boolean hypercubes. *p-Adic Numbers, Ultrametric Analysis, and Applications*, 8(4):298–311, 2016.
- [4] P.E. Bradley. Finding ultrametricity in data using topology. *Journal of Classification*, 34:76–84, 2017.

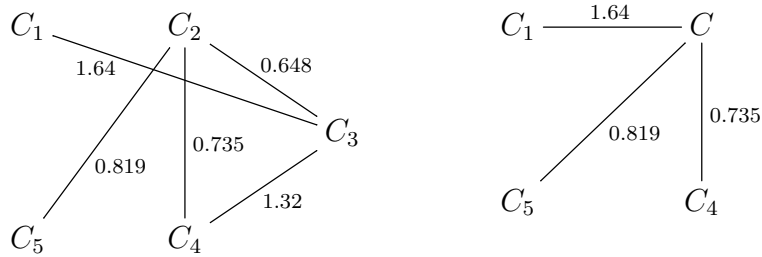


Figure 2: The coarse graphs Γ_ϵ^δ for $(\epsilon, \delta) = (0.64, 1.64)$ (left), and $(\epsilon, \delta) = (0.7, 1.64)$ (right).

- [5] P.E. Bradley. On the logistic behaviour of the topological ultrametricity of data. *Journal of Classification*, 36:266–272, 2019.
- [6] P.E. Bradley and Á. Morán Ledezma. Approximating diffusion on finite multi-topology systems using ultrametrics. in preparation.
- [7] J. Fresnel and M. van der Put. *Rigid Analytic Geometry and Its Applications*. Progress in Mathematics. Birkhäuser, Boston, 2004.
- [8] L. Gerritzen and M. van der Put. *Schottky Groups and Mumford Curves*, volume 817 of *Lecture Notes in Mathematics*. Springer, Heidelberg, New York, 1980.
- [9] F. Murtagh. On ultrametricity, data coding, and computation. *J. Class.*, 21:167–184, 2004.
- [10] F. Murtagh. Sparse p -adic data coding for computationally efficient and effective big data analytics. *p-Adic Numbers, Ultrametric Analysis and Applications*, 8:236–247, 2016.
- [11] R. Rammal, G. Toulouse, and M.A. Virasoro. Ultrametricity for physicists. *Reviews of Modern Physics*, 58(3):765–788, 1986.
- [12] L. Vietoris. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Math. Ann.*, 97(1):454–472, 1927.
- [13] V.S. Vladimirov, I.V. Volovich, and E.I. Zelenov. *p -adic Analysis and Mathematical Physics*. Series on Soviet and East European Mathematics, 1. World Scientific Publishing Co., Inc., River Edge, NJ, 1994.
- [14] A. Zomorodian. Fast construction of the Vietoris-Rips complex. *Comp. & Graph.*, 34(3):263–271, 2010.

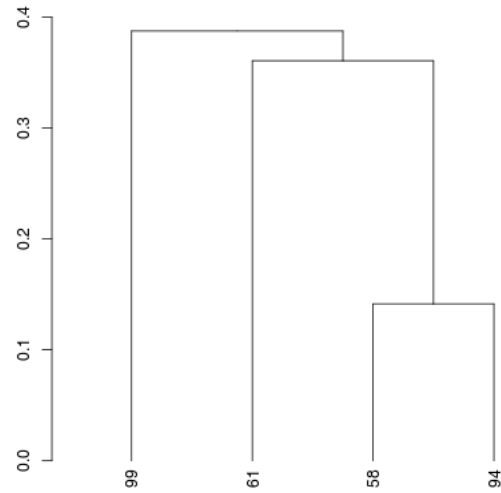


Figure 3: Single-linkage dendrogram of the cluster C_3 in $\Gamma_{0.64}$.

- [15] A.P. Zubarev. On stochastic generation of ultrametrics in high-dimensional Euclidean spaces. *p-Adic Numbers, Ultrametric Analysis, and Applications*, 6(2):155–165, 2014.