Degenerating families of dendrograms

Patrick Erik Bradley

January 10, 2008

Abstract

Dendrograms used in data analysis are ultrametric spaces, hence objects of nonarchimedean geometry. It is known that there exist p-adic representations of dendrograms. Completed by a point at infinity, they can be viewed as subtrees of the Bruhat-Tits tree associated to the p-adic projective line. The implications are that certain moduli spaces known in algebraic geometry are in fact p-adic parameter spaces of dendrograms, and stochastic classification can also be handled within this framework. At the end, we calculate the topology of the hidden part of a dendrogram.

1 Introduction

Dendrograms used in data analysis are ultrametric spaces. Hence they are objects of nonarchimedean geometry, a special instance of which is p-adic geometry. Murtagh (2004b) shows how to associate to a dendrogram a set of p-adic representations of integers. This lies well within the tradition of using ultrametrics in order to describe the hierarchical ordering in classification (Murtagh 2004a and the references therein).

However, there is seemingly a problem in the choice of the prime number p for the p-adic representation of dendrograms by the fact that the geometry of the p-adic number field $\mathbb{Q}_p$ allows only at most p maximal subclusters of any given cluster. We will show that this can be overcome by considering finite field extensions of $\mathbb{Q}_p$, so that the convenient choice $p = 2$ becomes feasible for any dendrogram. This seems to be compliant with the philosophy of allowing any nonarchimedean complete valued field for describing, coding or computing in data analysis. We acknowledge here our inspiration by Murtagh (2004a).

Our point of view is in fact of a geometric nature. For a p-adic geometer, a dendrogram is nothing but the affine p-adic line $\mathbb{A}^1$ with n punctures from which a certain kind of covering of $\mathbb{A}^1$ can be made whose intersection graph is the tree in bijection with the dendrogram from the point of view of data analysis. In this picture, the p-adic numbers forming the punctures are the data. Completing the affine line to the projective line $\mathbb{P}^1$ and then taking an extra puncture ∞, allows us to see the dendrogram as a subtree of the *Bruhat-Tits tree*, which is an important object in the study of p-adic algebraic curves. A first application is in the coding of DNA sequences (Dragovich and Dragovich 2006), which is

1

a special case of p-adic methods for processing strings over a given alphabet, as explained in Bradley (2007b), where also new invariants of time series of dendrograms are developed.

It is an imperative from the geometric viewpoint to study families of dendrograms. For these, there exist already parameter spaces. In fact, it is the *moduli space of genus* 0 *curves with n punctures* $M_{0,n}$ from algebraic geometry which now becomes the central object of interest. Each point of the p-adic version of $M_{0,n}$ is a dendrogram with the extra point ∞. It is then a natural consequence that a stochastic dendrogram is a continuous family of dendrograms together with a probability distribution on it, or, we can make this now more precise, a map from a p-adic set of parameters S to $M_{0,n}$ with a probability distribution on S. We will give an idea of p-adic spaces by explaining the *Berkovich topology* one has on these. Due to the ultrametric property, p-adic spaces in a naïve sense are totally disconnected. This problem can be remedied by introducing extra points which can, in a generalized sense, be viewed as clusters of usual points.

In this framework, collisions of points in their evolution through time can be formally described by considering the compactification $\bar{M}_{0,n}$ by stable trees of projective lines which we call *stable dendrograms*. Time series of dendrograms, on the other hand, yield (analytic) maps $M_{0,m} \rightarrow M_{0,n}$ between the moduli spaces. Further applications of these moduli spaces should be in the study of consensus of dendrograms.

We end by calculating the topology of the *hidden part* of a dendrogram, i.e. the subgraph spanned by vertices corresponding to clusters which do not have singletons as maximal subclusters. This subgraph determines the distribution of the other clusters, which are "near the end" of the dendrogram.

An introduction to p-adic numbers is Gouvêa (1993). Algebraic curves can be learned with a minimum amount of technical requirements in Griffiths (1989). A bird's eye on moduli spaces of curves is found in Mumford (1999, Appendix: Curves and Their Jacobians) A broader introduction to moduli of curves is Harris and Morrison (1998). A non-technical introduction to Berkovich spaces and analysis on the projective line is contained in (Baker 2004; Baker and Rumely 2007). Those who intend an intensive study of these subjects might wish to learn more algebraic geometry which can be found in Mumford (1999).

2   Dendrograms and nonarchimedean geometry

Dendrograms are known to be endowed with a nonarchimedean metric, also called an ultrametric, for which the strict triangle inequality

$$d(x,y) \leq \max\{d(x,z), d(z,y)\}$$

holds. Therefore, it is quite tempting to use p-adic numbers for their description, and in fact, this has recently been done (Murtagh 2004a; Murtagh 2004b). I shall explain this along the example dendrogram of Figure 1, which is a slight modification of Murtagh (2004b, Figure 1). Choose a prime number p, and

Figure 1: A 2-adic dendrogram.

distribute the p numbers $0, \ldots, p - 1$ across the partitioning of the horizontal line segments defined by the intersection points with vertical line segments of the dendrogram. For the top horizontal line segment, one has to introduce one extra vertical line segment going upwards[1], as effected in Figure 1. On going down on a path γ from the top vertical line segment all the way down to one of the points x_i, one picks up the numbers α on the traversed horizontal line segments ℓ and obtains

$$x = \sum_{\nu} \alpha_{\nu} p^{\nu},$$

where $\nu = \nu(\ell)$ runs through all levels of the horizontal parts ℓ of the path γ.

In our example from Figure 1, we assume $p = 2$, and obtain the numbers

$$
\begin{aligned}
&x_1 = 0, &&x_2 = 2^6, &&x_3 = 2^5, &&x_4 = 2^2, \\
&x_5 = 2^2 + 2^4, &&x_6 = 2^2 + 2^3, &&x_7 = 2^0, &&x_8 = 2^0 + 2^1.
\end{aligned}
$$

Note that these dyadic representations differ from the ones in Murtagh (2004b, §2). In any case, each path from the top to a bottom end of the dendrogram corresponds to a p-adic power series representation of an integer number. The choice of the prime p is arbitrary. However, it might seem that the possible number of vertical segments attached to one horizontal line segment allowing a p-adic representation of a dendrogram might be bounded by p. But this is not the case. In fact, one can restrict to the arbitrary choice $p = 2$, if one wishes, and can describe all dendrograms by the help of a little algebra, as will be seen in the following section.

---

[1] The usefulness of this extra detail will become apparent in the following sections.

3 The Bruhat-Tits tree

Let $\mathbb{Q}_p$ be the field of p-adic numbers. It is a complete nonarchimedean normed field whose norm will be denoted by $|\cdot|_p$. Consider the unit disk

$$\mathbb{D} = \{x \in \mathbb{Q}_p \mid |x|_p \leq 1\} = B_1(0).$$

It contains the p maximal smaller disks

$$B_{\frac{1}{p}}(0),\ B_{\frac{1}{p}}(1), \ldots, B_{\frac{1}{p}}(p-1)$$

corresponding to the residue field $\mathbb{F}_p$ of $\mathbb{Q}_p$. This well known fact is actually a consequence of the construction from the previous section.

It is useful to consider the p-adic projective line $\mathbb{P}(\mathbb{Q}_p) = \mathbb{Q}_p \cup \{\infty\}$, in which there is the maximal disk outside $\mathbb{D}$:

$$\{x \in \mathbb{P}(\mathbb{Q}_p) \mid |x|_p \geq p\} = B_p(\infty).$$

Due to the ultrametric topology on the p-adic projective line, the "closure" of an "open" disk depends somewhat on the choice of a point on its "boundary" (Gerritzen 1978, §1.1). Therefore, we make

Definition 3.1. *Let*

$$B = \{x \in \mathbb{P}(\mathbb{Q}_p) \mid |x-a|_p < r\} \quad (resp. \quad B = \{x \in \mathbb{P}(\mathbb{Q}_p) \mid |x-a|_p > r\})$$

for some $a \in \mathbb{Q}_p$ and a p-adic value $r = |\epsilon|_p$, $\epsilon \in \mathbb{Q}_p \setminus \{0\}$, and let $b \in \mathbb{Q}_p$ such that $|a-b|_p = r$. The affinoid closure *of B with respect to ∞ (resp. to b) is the disk*

$$\bar{B} = \{z \in \mathbb{P}(\mathbb{Q}_p) \mid |x-a|_p \leq r\} \quad (resp. \quad \bar{B} = \{z \in \mathbb{P}(\mathbb{Q}_p) \mid |x-b|_p \geq r\}).$$

Using the projective line necessitates the introduction of an equivalence relation on the set of all disks of $\mathbb{P}(\mathbb{Q}_p)$. Namely, disks B_1, B_2 are said to be *equivalent:* $B_1 \sim B_2$, if either $B_1 = B_2$ or the affinoid closure of $\mathbb{P}(\mathbb{Q}_p) \setminus B_2$ with respect to some point $a \in B_2$ equals B_1 (Herrlich 1980, §1). One checks that the relation $\sim$ is indeed an equivalence relation.

The *Bruhat-Tits tree* $\mathscr{T}_{\mathbb{Q}_p}$ is defined by setting its vertices to be the equivalence classes of disks in $\mathbb{P}(\mathbb{Q}_p)$, and its edges are given by maximal inclusion of disks, i.e. an edge $e = ([B_1], [B_2])$ means that B_1 is strictly contained in B_2, and B_1 is a maximal disk with this property, for suitable representative disks. It is a well known fact that $\mathscr{T}_{\mathbb{Q}_p}$ is indeed a tree. This can be seen directly in this way: Each class is obviously represented by a unique disk B which is the closure with respect to $\infty \notin B$, and the disks not containing infinity are preordered by inclusion; so $\mathscr{T}_{\mathbb{Q}_p}$ is a directed acyclic graph, hence a tree by the ultrametric property of $|\cdot|_p$.

The *star* of a vertex v in $\mathscr{T}_{\mathbb{Q}_p}$, denoted as $\mathrm{Star}_{\mathscr{T}_{\mathbb{Q}_p}}(v)$, consists of all edges emanating from v. The edges of any star are in one-to-one correspondence with
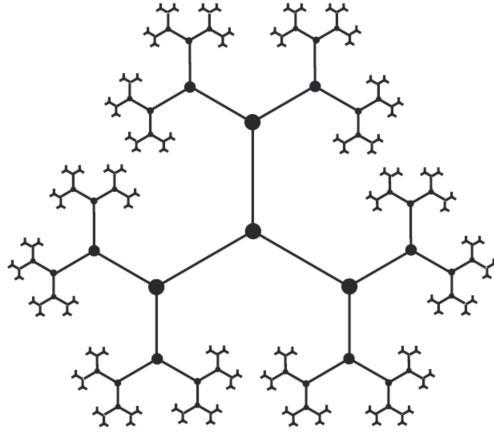
Figure 2: The Bruhat-Tits tree for $\mathbb{Q}_2$.

the points of $\mathbb{P}(\mathbb{F}_p) = \mathbb{F}_p \cup \{\infty\}$, i.e. the $\mathbb{F}_p$-rational points of the projective line over the residue field $\mathbb{F}_p$. Namely, this is true for the vertex $v_\mathbb{D}$ corresponding to the unit disk $\mathbb{D}$, and the group of Möbius transformations acts on $\mathscr{T}_{\mathbb{Q}_p}$ (Herrlich 1980, Bemerkung 5). Thus the Bruhat-Tits tree $\mathscr{T}_{\mathbb{Q}_p}$ is a $p + 1$-regular locally finite tree. An illustration of $\mathscr{T}_{\mathbb{Q}_2}$ from Cornelissen and Kato (2005, Fig. 5) is given in Figure 2.

By construction, the tree $\mathscr{T}_{\mathbb{Q}_p}$ is invariant under transformations of the form $z \mapsto \frac{az+b}{cz+d}$, with $a, b, c, d \in \mathbb{Q}_p$ such that $ad - bc \neq 0$. These transformations are called *projective linear* or *Möbius transformations*, and form the group $\mathrm{PGL}_2(\mathbb{Q}_p)$. The reason for invariance under $\mathrm{PGL}_2(\mathbb{Q}_p)$ is the well known fact that Möbius transformations take equivalent disks to equivalent disks.

As it may happen that a cluster may have more than p maximal subclusters, it would be convenient to be able to represent such dendrograms without enlarging the prime p. So, let $K \supseteq \mathbb{Q}_p$ be a finite extension field of $\mathbb{Q}_p$. The p-adic norm extends, similarly as in the archimedean case, uniquely to an ultrametric norm $|\cdot|_K$ on K, and K is complete with respect to $|\cdot|_K$. Such a field K is called a *p-adic number field*.

For a p-adic number field K, there is in a similar manner as for $\mathbb{Q}_p$ a Bruhat-Tits tree $\mathscr{T}_K$. Again K has a finite residue field with $q = p^m$ elements, and $\mathscr{T}_K$ is $q + 1$-regular. Therefore, in practical applications it should be possible to stick to the prime $p = 2$ and make finite field extensions, if there are clusters with more than 2 children clusters. Again, $\mathrm{PGL}_2(K)$ respects the symmetries of the hierarchical structure of the Bruhat-Tits tree, i.e. $\mathscr{T}_K$ is invariant under projective linear transformations defined over K.

For convenience, we assume now that $K = \mathbb{Q}_p$. However, all what is said in the following is valid also for arbitrary p-adic number fields.

It is well known that any infinite descending chain

$$B_1 \supseteq B_2 \supseteq \dots \tag{1}$$

of strictly smaller disks in $\mathbb{P}(\mathbb{Q}_p)$ converges to a unique point

$$\{x\} = \bigcap_n B_n$$

on the *p*-adic projective line $\mathbb{P}(\mathbb{Q}_p)$. A chain (1) defines a halfline in the Bruhat-Tits tree $\mathscr{T}_{\mathbb{Q}_p}$.

An *end* in a tree is an equivalence class of halflines, where two halflines are said to be equivalent, if they differ only by finitely many edges. It is a fact that the ends of the tree $\mathscr{T}_{\mathbb{Q}_p}$ correspond bijectively to the points in $\mathbb{P}(\mathbb{Q}_p)$, and is not too difficult to check.

The following subtree of the Bruhat-Tits tree is an idea of F. Kato (Kato 2005, §5.4) which turned out useful in the study of discontinuous group actions:

Definition 3.2. *Let $X \subseteq \mathbb{P}(\mathbb{Q}_p)$ be a finite set containing 0, 1 and ∞. Then the smallest subtree $\mathscr{T}^*\langle X \rangle$ of $\mathscr{T}_{\mathbb{Q}_p}$ having X as its set of ends is called the projective dendrogram for X.*

Note that the definition of $\mathscr{T}^*\langle X \rangle$ makes sense, even if X does not contain 0, 1 or ∞.

Example 3.3. (1) Let $x_0, x_1 \in \mathbb{P}(\mathbb{Q}_p)$ be two distinct points, and set $X = \{x_0, x_1\}$. It defines the subtree $\mathscr{T}^*\langle X \rangle$ which is a straight line: the geodesic in $\mathscr{T}_{\mathbb{Q}_p}$ between x_0 and x_1, as illustrated in Figure 3.

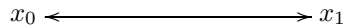$$x_0 \longleftarrow\!\!\!\!\longrightarrow x_1$$

Figure 3: Geodesic line in $\mathscr{T}_{\mathbb{Q}_p}$.

(2) Let $X = \{x_0, x_1, x_2\}$ be a set of three mutually distinct points in $\mathbb{P}(\mathbb{Q}_p)$. Then the subtree $\mathscr{T}^*\langle X \rangle$ is a tripod, as depicted in Figure 4. We denote by $v(x_0, x_1, x_2)$ the unique vertex of $\mathscr{T}^*\langle X \rangle$ whose star has three edges.
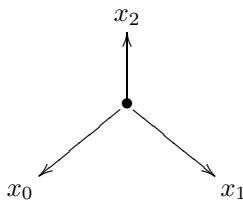

Figure 4: Tripod in $\mathscr{T}_{\mathbb{Q}_p}$.

For a subset X of $\mathbb{P}(\mathbb{Q}_p)$, define $\mathscr{T}\langle X\rangle$ to be the subtree of $\mathscr{T}_{\mathbb{Q}_p}$ that is the smallest subtree among all possible subtrees containing the vertices of the form $v(x_0, x_1, x_2)$ with $x_0, x_1, x_2 \in X$. Notice that this subtree is non-empty if and only if X contains at least three points. We call $\mathscr{T}\langle X\rangle$ the *finite part* of the projective dendrogram $\mathscr{T}^*\langle X\rangle$. We have the obvious inclusion

$$\mathscr{T}\langle X\rangle \longhookrightarrow \mathscr{T}^*\langle X\rangle$$

of trees.

It is useful to not take into account all vertices of the finite part $\mathscr{T} = \mathscr{T}\langle X\rangle$ of a projective dendrogram. Consider all paths $\gamma = [v, w]$ (without backtracking) of maximal length in $\mathscr{T}$ whose vertices in (v, w) have no edges outside γ emanating from them. By replacing every such path γ of $\mathscr{T}$ by a single edge, but of equal length as γ, we obtain a so-called *stable tree* $\mathscr{T}^{\mathrm{stab}}$, whose vertices have the property that at least three edges emanate from each of them. The tree $\mathscr{T}^{\mathrm{stab}}$ is called the *stabilization* of $\mathscr{T}$.

Convention 3.4. *By a (projective) dendrogram $\mathscr{T}^* = \mathscr{T}^*\langle X\rangle$ we will usually mean the tree obtained by identifying the finite part $\mathscr{T}\langle X\rangle$ with its stabilization $\mathscr{T}^{\mathrm{stab}}$.*

A vertex v of $\mathscr{T}\langle X\rangle$ is considered to be a cluster of the points corresponding to the halflines in $\mathscr{T}\langle X\rangle^*$ emanating from v. Fixing the points 0, 1 and ∞ is done for reasons of normalization: two points define a geodesic, three points define a unique vertex in $\mathscr{T}_{\mathbb{Q}_p}$, and the three points 0, 1 and ∞ define the vertex $v_{\mathbb{D}}$ corresponding to the unit disk $\mathbb{D}$.

In this way, the usual dendrogram obtained from $\mathscr{T}^*\langle X\rangle$ is

$$\mathscr{T}^*\langle X\rangle \setminus \text{ the halfline } (v_{\mathbb{D}}, \infty).$$

A "genuine" dendrogram has the property that $X \subseteq \mathbb{Z} \cup \{\infty\}$, or, more generally, $\infty \neq x \in X$ has a finite expansion

$$x = \alpha_0 + \alpha_1 \pi + \cdots + \alpha_m \pi^m, \quad \alpha_\nu \in \{0, \ldots, q-1\},$$

where π is a prime element of $O_K = \{z \in K \mid |z|_K \leq 1\}$, and q the order of the residue field of K. Gouvêa (1993) contains more details on finite field extensions of $\mathbb{Q}_p$.

Remark 3.5. *As noted in Bradley (2007a), the task of hierarchical classification conceptually becomes the finding of a suitable p-adic encoding which reveals the inherent hierarchical structure of data. The reason is that the p-adic dendrogram $\mathscr{T}^*\langle X\rangle$ of a given set $X \subseteq \mathbb{P}^1(\mathbb{Q}_p)$ is uniquely determined by X. Algorithmically, the computation of $\mathscr{T}^*\langle X\rangle$ is much simpler than its classical counterpart (Bradley 2007b, §3.2).*

4 The space of dendrograms

Call $M_{0,n}$ the space of all projective dendrograms for sets of cardinality $n \geq 3$. This space is known also under the name *moduli space for genus 0 curves with n punctures.* The term "genus 0 curve" means nonsingular projective algebraic curve of genus 0, i.e. projective line. By fixing n points $x_1, \ldots, x_n$ on the projective line $\mathbb{P}(\mathbb{Q}_p)$ and then changing these points by a Möbius transformation such that the first three are 0, 1, ∞, we obtain a projective dendrogram as explained in the previous section.

As moduli spaces parametrize objects up to isomorphism, and isomorphisms of punctured curves send punctures to punctures, we indeed have a moduli space $M_{0,n}$ of dendrograms by considering in each isomorphism class a normalized representative.

It is a well established fact that

$$M_{0,n} \cong \left(\mathbb{P}^1 \setminus \{0,1,\infty\}\right)^{n-3} \setminus \Delta,$$

where Δ is the fat diagonal given by $x_i = x_j$, $i \neq j$, and $\mathbb{P}^1$ is the projective line, considered as an algebraic variety (Mumford 1999, Appendix: Lecture II).

One may imagine the space $M_{0,n}$ by fixing three points on $\mathbb{P}^1$ and letting the remaining $n-3$ points vary on the projective line without collision.

In the p-adic setting, a family of dendrograms for n points is given by a map $S \to M_{0,n}$ from some base space S. Each point $s \in S$ corresponds to a dendrogram, and the dendrogram varies in some sense, as s moves along S.

The "geography" of $M_{0,n}$ is as follows: pick a dendrogram x for n points. Moving the points only slightly does not change the finite part of the dendrogram. Moving the points a little more results in changes in the lengths of the edges of x, but the underlying combinatorial structure does not change. The combinatorial tree of x occupies an open subset U of $M_{0,n}$. Moving points of x even more results in edge contractions: by contracting one edge, x moves from U to a neighboring piece V. $M_{0,n}$ is covered by such disjoint open pieces, each belonging to a combinatorial tree with n ends. This is due to the fact that $M_{0,n}$, like many spaces in nonarchimedean geometry, is totally disconnected. This rather uncomfortable fact can be remedied by either resorting to a so-called *Grothendieck topology* or by introducing extra points which then produce a genuine topology, e.g. by considering *Berkovich analytic spaces* (Berkovich 1990). This topology will be explained in the following section.
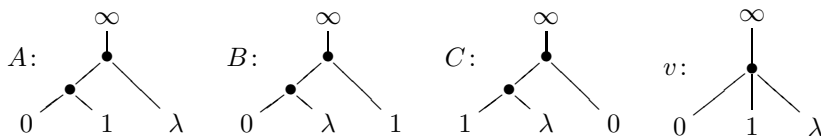

Figure 5: Dendrograms representing $M_{0,4}$.

Figure 5 illustrates the dendrograms represented by the different parts of $M_{0,4}$: one "central" region v (three children) and three "outer" regions A, B, C (at most two children). Any path from A to B or C passes through v, as the edge has to be contracted and then blown up in a different manner.

5   The Berkovich topology on $M_{0,n}$

We begin with the topology on the unit disk $\mathbb{D}$ of a p-adic number field. The *classical* points of $\mathbb{D}$ are its K-rational points. However, Berkovich (1990) defines more points which correspond to multiplicative seminorms on the algebra of power series convergent on nonarchimedean spaces. For the unit disk this amounts to (Berkovich 1990, 1.4.4):

1. the classical points,

2. the disks $\{x \in K \mid |x - a|_K \leq r\}$ in $\mathbb{D}$ with $r = |\epsilon|_K$, $\epsilon \in K \setminus \{0\}$,

3. the disks as in (2), but $0 < r \neq |\epsilon|_K$ for any $\epsilon \in K$,

4. the properly descending chains $B_1 \supset B_2 \supset \ldots$ of disks in $\mathbb{D}$ with $\bigcap B_i = \emptyset$.

The new points corresponding to (2), (3) or (4) are called *generic*, or *generic Berkovich points*. This works also for the affine line K, where one takes the multiplicative seminorms on the polynomial ring $K[T]$ and obtains similarly the types (1) to (4) of points. The analogous result holds for the projective line.

The concept of generic Berkovich points via multiplicative seminorms works also in higher dimension, and the result is that p-adic manifolds are locally contractible (Berkovich 1999). In any case, by that concept, the data domain can be viewed as a contiunuum.

Endowing our space of dendrograms $M_{0,n}$ with the Berkovich topology gives us now a framework for considering continuously varying families of dendrograms. For example, a stochastic classification of n points (including ∞) is nothing but a probability distribution on $M_{0,n}$, possibly with compact support. Or the problem of adding a new datapoint to a given classification $x \in M_{0,n}$ means finding a probability distribution on the fiber $\pi^{-1}(x)$, where $\pi \colon M_{0,n+1} \to M_{0,n}$ is the map which forgets the $(n+1)$-th puncture on the p-adic projective line. A similar thing applies also to a family $S \to M_{0,n}$, where a distribution has to be found on the fiber product $S \times_{M_{0,n}} M_{0,n+1}$ with the map π.

6   Allowing collisions

So far, our dendrograms for n points can vary continuously in families, but collisions of points are strictly excluded. In order to allow collisions, one compactifies the space $M_{0,n}$ to $\bar{M}_{0,n}$. We call the points of $\partial \bar{M}_{0,n}(\mathbb{Q}_p)$ *stable trees of dendrograms* or, by abuse of language, simply *stable*. In fact, these are the so-called *stable n-pointed trees of projective lines* (Gerritzen, Herrlich and van der

Put 1988). Such are algebraic curves C which are unions of projective lines L together with n points $X = \{x_1, \ldots, x_n\} \subseteq C$ and have the defining properties:

1. every singular point is an ordinary double point,

2. the intersection graph of the projective lines L is a tree,

3. every projective line L of which C is composed contains at least three points which are either singular points of C or lie in X,

4. X consists of regular points of C.

In some sense, we can view the points of the boundary $\partial M_{0,n}(\mathbb{Q}_p)$ as dendrograms of dendrograms. We indeed have such applications in mind as classifications of classifications.

In order to understand what happens if a dendrogram $x \in M_{0,n}$ moves to the boundary, consider a dendrogram with four distinct ends $0, 1, \infty, \lambda$, considered as points on the projective line L. The effect of λ moving towards one of the other three points x is that, upon collision, another projective line L' is formed which intersects the original line L and on which λ and the point x are again distinct. Such a configuration corresponding to a point of $\partial M_{0,4}$ is given in Figure 6. In any case, the resulting tree of dendrograms is indeed stable also for $n \geq 4$.
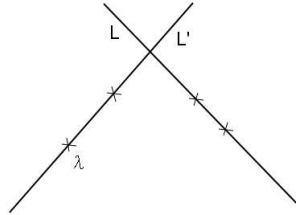

Figure 6: A stable 4-pointed tree of projective lines.

Note that the tree with ends corresponding to a stable dendrogram does geometrically not differ from a projective dendrogram in $M_{0,n}$, if one forms a dendrogram for the punctures on each of the projective lines. The difference is that different parts of that tree correspond to different projective lines. This is useful for distinguishing points which are otherwise identified by collisions.

7   Finite families of dendrograms

Assume a finite family X of datasets $X_1, \ldots, X_m$ each consisting of n (classical) points of the p-adic projective line:

$$X_j = \{x_{1j}, \ldots, x_{nj}\},$$

and assume at the moment that they are all different. For example, X could be a time series $x_i(t_j) = x_{ji}$ of positions of n not colliding particles never at the same place. Thus X is the union of the X_i and represents an element of $M_{0,mn}$, if we assume $x_{11} = 0$, $x_{12} = 1$ and $x_{13} = \infty$. By restricting to the points of X_j (e.g by taking the points at time t_j), we obtain a map

$$\pi_j(X) \colon M_{0,mn} \to M_{0,n}$$

which is the composition of the two maps

$$(0, 1, \infty, x_{14}, \dots, x_{nm}) \mapsto (x_{1j}, \dots, x_{nj}), \tag{2}$$
$$(x_1, \dots, x_n) \mapsto (0, 1, \infty, x'_4, \dots, x'_n), \tag{3}$$

i.e. the canonical projection onto X_j followed by a Möbius transformation $\alpha \in \mathrm{PGL}_2(K)$ (cf. Section 3) which sends the first three points of X_j to 0, 1, and ∞. Note that the Möbius transformation $\alpha = \alpha_X$ is uniquely determined by X and can be easily computed.

If we now allow collisions of datapoints, then we obtain a map

$$\bar{\pi}_j(X) \colon \bar{M}_{0,mn} \to \bar{M}_{0,n},$$

which we will not make explicit. Instead we note that if the number of distinct points of X is k, then we have maps as before

$$\pi_j(X) \colon M_{0,k} \to M_{0,n_j},$$

where n_j is the number of distinct points in X_j. The $\pi_j(X)$ are again canonical projections followed by Möbius transformations, and are closely related to the maps $\bar{\pi}_j(X)$.

The advantage of this moduli space approach to finite families lies in the feasibility of handling situations where one has a continuous family of such X. Moreover, the Möbius transformation α_X varies continuously with X.

Again, as in Section 5, one can enrich the families by probability distributions in order to obtain stochastic classifications.

8 Hidden vertices

Definition 8.1. *Let $\mathscr{T}^* = \mathscr{T}^*\langle X \rangle$ be a projective dendrogram for X. A vertex v of $\mathscr{T} = \mathscr{T}\langle X \rangle$ is called* hidden, *if* $\mathrm{Star}_{\mathscr{T}}(v) = \mathrm{Star}_{\mathscr{T}^*}(v)$. *The subgraph Γ^h of $\mathscr{T}$ spanned by all its hidden vertices is called the* hidden subgraph of $\mathscr{T}$.

The quantity b_0^h, defined as the number of connected components of Γ^h, measures how the clusters corresponding to non-hidden vertices are spread. As Γ^h is a subgraph of a tree, this number equals also the Euler characteristic $\chi(\Gamma^h)$.

Definition 8.2. *Let v be a vertex of a graph Γ. The number $\operatorname{ord}_\Gamma(v) = \#\operatorname{Star}_\Gamma(x)$ is called the* order *of v in Γ. If $\operatorname{ord}_\Gamma(v) = 1$, then v is called a* tip *of Γ.*

By our convention, any vertex v of a dendrogram has order either 1 or greater than 2.

Theorem 8.3. *Let $\mathscr{T}^* = \mathscr{T}^*\langle X \rangle$ be a (projective) dendrogram with $\#X = n$. Then $v^h = \#\operatorname{Vert}(\Gamma^h)$ is bounded from above:*

$$v^h \leq \frac{n - b_0^h + 1}{2}.$$

Proof. Case: Γ^h connected. If Γ^h is connected, then either $b_0^h = 1$ or $\Gamma^h = \emptyset$. We have for the number t^h of tips of Γ^h:

$$4t^h \leq n, \tag{4}$$

because each tip v in Γ^h must have at least two edges in $\mathscr{T} \setminus \Gamma^h$, and, again for reasons of order, there must be at least two ends in $\mathscr{T}^*$ emanating from each edge in $\operatorname{Star}_{\mathscr{T}}(v) \setminus \operatorname{Star}_{\Gamma^h}(v)$. This is illustrated in Figure 7, where v is a tip in Γ^h, and e the unique edge in $\operatorname{Star}_{\Gamma^h}(v)$.
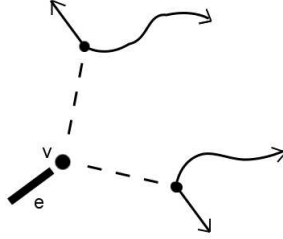

Figure 7: A hidden tip in a projective dendrogram.

Now, the order in Γ^h of any vertex v is 0, 1 or ≥ 3. In the first case, $t^h = 0$, and then

$$v^h = 1 \leq \frac{n}{6} \leq \frac{n}{2},$$

where the first inequality follows in a similar way as (4). Assume now that Γ^h has an edge. Removing from Γ^h all tips and their adjacent edges, and then repeating the process iteratively yields a strictly descending sequence of trees whose limit is either a vertex or empty. In any case, the number of tips removed in the ν-th step is at most $\frac{t^h}{2^\nu}$. Hence, for some $m \in \mathbb{N}$ it holds true that

$$v^h \leq \sum_{\nu=0}^{m} \frac{t^h}{2^\nu} \leq 2t^h \leq \frac{n}{2}$$

which is the bound in case $b_0^h = 1$.

General case. In the general case, one must remove at least $b_0^h - 1$ ends in order to make Γ^h connected. Then, after stabilizing, the resulting tree will have at least v^h vertices. Hence,

$$v^h \leq \frac{n - b_0^h + 1}{2}$$

by the previous case. □

Corollary 8.4. *For X with $n = \#X$, there is a bound for the number of connected components of Γ^h:*

$$b_0^h \leq \frac{n+1}{3}.$$

Proof. We may assume that Γ^h contains no edges. Then $b_0^h = v^h$, and

$$v^h \leq \frac{n - v^h + 1}{2},$$

from which the asserted bound follows. □

The bound in Corollary 8.4 is not sharp, however.

Theorem 8.5. *For the number of connected components of Γ^h, there is the following sharp bound:*

$$b_0^h \leq \frac{n-3}{3},$$

where n is the cardinality of X.
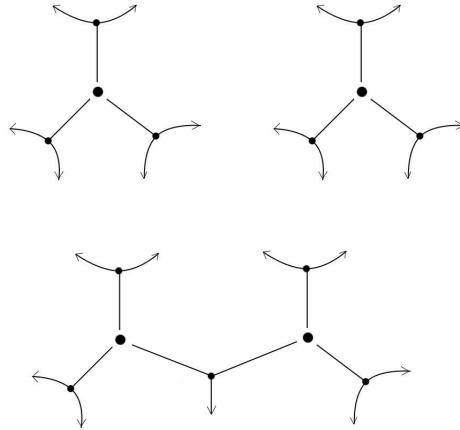
Figure 8: Glueing trees along a vertex and removing three ends.

Proof. We may assume that Γ^h has no edges. By an inductive glueing of trees as in Figure 8 we obtain that for each additional connected component, one has to subtract three ends, in order to produce a dendrogram having as few ends as possible. Thus,

$$b_0^h \leq \frac{n + 3(b_0^h - 1)}{6} = \frac{n-3}{6} + \frac{b_0^h}{2},$$

from which the bound follows. Now, if n is a multiple of 3, then $b_0^h = \frac{n-3}{3}$ by construction. Therefore, in the general case,

$$b_0^h = \left\lfloor \frac{n-3}{3} \right\rfloor$$

can be constructed. This means that the bound is sharp. $\qquad\square$

9 Conclusion

We have given a geometric foundation for an ultrametric approach towards classification. By extending usual dendrograms by an additional point ∞, they can be considered as points of the moduli space $M_{0,n}$ for the projective line with n punctures. The Berkovich topology allows to consider stochastic classification as giving a continuous family of dendrograms with a probabiliy distribution on it. The points on the boundary of $M_{0,n}$ arise from collisions of continuously evolving datapoints and are interpreted as dendrograms of dendrograms. Time sections of time series are given by maps $M_{0,m} \to M_{0,n}$. Finally, the topology of dendrograms is studied, resulting in bounds for the number of hidden vertices and the Euler characteristic of the hidden graph which separates those clusters containing datapoints as maximal subclusters. The consequence of using p-adic methods is the shift of focus from imposing a hierarchic structure on data to finding a p-adic encoding which reveals the inherent hierarchies.

Acknowledgements

References

BAKER, M. (2004), "Analysis and dynamics on the Berkovich projective line." *Lecture notes.* http://www.math.gatech.edu/~mbaker/pdf/BerkNotes.pdf

BAKER, M., and RUMELY, R (2007) "Potential theory on the Berkovich projective line." *Book in preparation.* http://www.math.gatech.edu/~mbaker/pdf/BerkBook.pdf

BERKOVICH, V.G. (1990), *Spectral theory and analytic geometry over non-archimedean fields.* Number 33 in Mathematical Surveys and Monographs. American Mathematical Society.

BERKOVICH, V.G. (1999), "Smooth p-adic analytic spaces are locally contractible." *Inventiones Mathematicae*, 137, 1–84.

BOCK, H.H. (1974), *Automatische Klassifikation.* Studia Mathematica/Mathematische Lehrbücher, Band XXIV. Vandenhoek & Ruprecht in Göttingen.

BRADLEY, P.E. (2007), "Families of dendrograms." To appear in: *Proceedings of the 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications, Freiburg im Breisgau.* Springer series Studies in Classification, Data Analysis, and Knowledge Organization, 2007.

BRADLEY, P.E. (2007), "Mumford dendrograms." To appear in: *Computer Journal.*

CORNELISSEN, G., and KATO, F. (2005), "The p-adic icosahedron." *Notices of the AMS*, 52, 720–727.

DRAGOVICH, B., and DRAGOVICH, A. (2006), "A p-adic model of DNA-sequence and genetic code." *Preprint.* arXiv:q-bio.GN/0607018.

GERRITZEN, L. (1978), "Unbeschränkte Steinsche Gebiete von $\mathbb{P}^1$ und nichtarchimedische automorphe Formen." *Journal für die Reine und Angewandte Mathematik*, 297, 21–34.

GERRITZEN, L., HERRLICH, F., and VAN DER PUT, M. (1988), "Stable n-pointed trees of projective lines." *Indagationes mathematicae. Proceedings A*, 91, 131–163.

GOUVÊA, F.Q. (1993), *p-adic numbers. An introduction.* Universitext. Springer, Berlin.

GRIFFITHS, P.A. (1989), *Introduction to algebraic curves.* Translations of Mathematical Monographs, 76. American Mathematical Society.

HARRIS, J., and MORRISON, I. (1998), *Moduli of curves.* Graduate Texts in Mathematics, 187. Springer, Berlin.

HERRLICH, F. (1980), "Endlich erzeugbare p-adische diskontinuierliche Gruppen." *Archiv der Mathematik*, 35, 505–515.

KATO, F. (2005), "Non-archimedean orbifolds covered by Mumford curves." *Journal of Algebraic Geometry*, 14, 1–34.

MUMFORD, D. (1999), *The red book of varieties and schemes.* Second, expanded edition. Lecture Notes in Mathematics, 1358. Springer, Berlin.

MURTAGH, F. (2004), "On ultrametricity, data coding, and computation." *Journal of Classification*, 21, 167–184.

MURTAGH, F. (2004), "Thinking ultrametrically." In: *Classification, Clustering and Data Mining Applications*, pages 3–14. Springer, Berlin.

Universität Karlsruhe, Institut für Industrielle Bauproduktion, Fakultät für Architektur, Englerstr. 7, D-76128 Karlsruhe, Germany
E-mail: `bradley@ifib.uni-karlsruhe.de`