S.V.Kozyrev
Steklov Mathematical Institute

Multidimensional clustering and hypergraphs

Data analysis, taxonomy for reticulate evolution and geometry of Bruhat–Tits buildings

Talk given at The International Workshop
on p-Adic Methods for Modeling of Complex Systems,
ZiF, Bielefeld, April 15 – 19, 2013

S.Albeverio, S.V.Kozyrev, Clustering by hypergraphs and dimensionality of cluster systems, *p*-Adic Numbers, Ultrametric Analysis and Applications. 2012. V.4. No.3. P.167–178. arXiv:1204.5952v1

Clustering – construction of a tree of clusters
with a hierarchy (partial order)
starting from a metric on a set of points.

Data analysis, Applications to taxonomy (tree of life).

Problem: assume we have instead of one metric
a family of metrics depending on a set of parameters
(this is a typical situation).
We have a family of clusterings.
Can we describe this family by a single mathematical object?

p-Adic geometry — trees of balls.
Multidimensional p-adic geometry —
Bruhat–Tits buildings.

Clustering (standard definitions)

Let (M, ρ) be an arbitrary metric space.
A sequence of points $a = x_0, x_1, \ldots, x_{n-1}, x_n = b$ in (M, ρ) is called an ε-chain connecting two points a and b if $\rho(x_k, x_{k+1}) \leq \varepsilon$ for all $0 \leq k < n$.

If there exists an ε-chain connecting a and b then a and b are ε-connected.

The chain distance between a and b:
$d(a, b) = \inf(\varepsilon\colon\ a,\ b\ \varepsilon\text{-connected})$.
All the properties of an ultrametric excluding non–degeneracy

A cluster $C(i, R)$ in a metric space (M, ρ) is a ball with the center i and radius R with respect to the chain distance, i.e. the set $\{j \in M \colon d(i, j) \leq R\}$.

A clustering of a metric space M is a cluster set:
i) every element in M belongs to some cluster;
ii) for any pair a, b of elements in M there exists a minimal cluster $\sup(a, b)$ containing both elements;
iii) for arbitrary embedded clusters $A \subset B$ every increasing sequence of embedded clusters $\{A_i\}$,
$A \subset \cdots \subset A_i \subset A_{i+1} \subset \cdots \subset B$ is finite.

Clustering — partially ordered tree of clusters (dendrogram):
(vertices are clusters);
partial order by inclusion of clusters;
edge connects two clusters nested without intermediaries.

Multidimensional generalization of clustering

Family of metrics
(complex systems, bioinformatics, classification) —
family of clusterings —
hypergraph clustering.

Multidimensional structure of the data —
Multidimensional structure of the clustering.
Relation to affine Bruhat–Tits buildings.

Taxonomy –
discussion of phylogenetic networks.

Hypergraphs

A hypergraph is a set Γ
with a selected system of finite sets E consisting of subsets containing two or more elements of Γ.

The elements of Γ are called hypergraph vertices,
the sets in E are called hypergraph edges.

If all the edges in E are of cardinality two,
then the hypergraph is a graph.

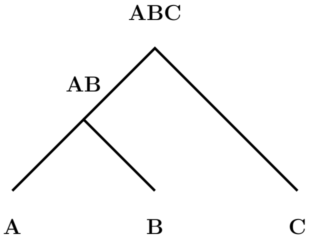
Examples of hypergraph clustering

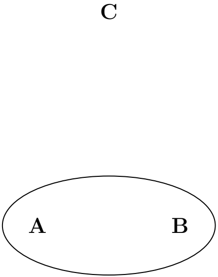
Example 1: The case of a set of three points A, B, C in a two–dimensional real plane $\mathbb{R}^2$ with the standard Euclidean metric. Parameters defining the metric are coordinates of the points in the plane.

Cluster tree $\mathcal{A}_1$:
The cluster set A, B, C, AB, ABC (vertices of the cluster tree), edges join the vertices in accordance with the growth of the clusters – the cluster tree contains the edges

$$(A, AB), \quad (B, AB), \quad (AB, ABC), \quad (C, ABC).$$

We denote by ABC the cluster containing A, B and C.

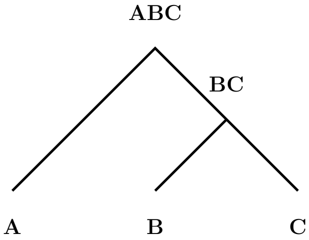

ABC
AB
A
B
C


C
A
B

Cluster tree $\mathcal{B}_1$:

Variation of the metric (motion of the points in the plane $\mathbb{R}^2$) – replacing the above cluster set by the set A, B, C, BC, ABC with the corresponding edges

$$(B, BC), \quad (C, BC), \quad (BC, ABC), \quad (A, ABC).$$



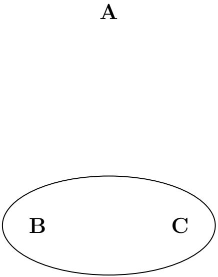
ABC

BC

A

B

C



A

B

C

Cluster network $\mathcal{C}_1$: The union of the trees of clusters $\mathcal{A}_1$ and $\mathcal{B}_1$ (where we identify the clusters which coincide as sets).
The vertex set of the hypergraph $\mathcal{C}_1$ contains the clusters

$$A, \quad B, \quad C, \quad AB, \quad BC, \quad ABC$$

the set of 2-edges (two-point edges) of $\mathcal{C}_1$

$$(A, AB), (B, AB), (AB, ABC), (C, ABC),$$

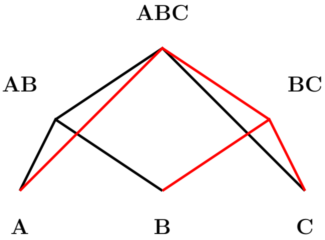
$$(B, BC), \quad (C, BC), \quad (BC, ABC), \quad (A, ABC).$$

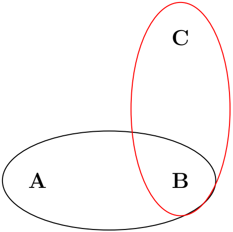
Hyperedges of $\mathcal{C}_1$: the 3-edges

$$(A, AB, ABC), \quad (C, BC, ABC)$$

and the 4-edge $(B,\ AB,\ BC,\ ABC)$.

The partial order of vertices is given by inclusion of clusters.


ABC
AB
BC
A
B
C


C
A
B

Schematically the structure of $\mathcal{C}_1$ is described by the table

$$\begin{array}{|c|c|c|}\hline A & & \\ \hline AB & ABC & \\ \hline B & BC & C \\ \hline\end{array},$$

where the matrix elements are vertices of $\mathcal{C}_1$, 2-edges connect all the neighbor vertices in the table and the pairs (C, ABC), (A, ABC).

Edges of the hypergraph $\mathcal{C}_1$ (union of the clustering trees $\mathcal{A}_1$ and $\mathcal{B}_1$) describe the growth of clusters starting from some vertex. Higher-order edges correspond to cycles in this graph.

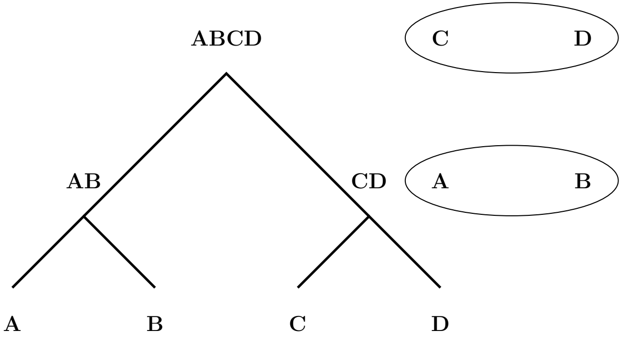
Example 2: A set of four points A, B, C, D located in the plane $\mathbb{R}^2$ at the vertices of some quadrangle.

Cluster tree $\mathcal{A}_2$:
Clustering with respect to the plane metric gives the clusters

$$A, \quad B, \quad C, \quad D, \quad AB, \quad CD, \quad ABCD.$$

The set of 2-edges contains the edges

$$(A, AB), (B, AB), (C, CD), (D, CD), (AB, ABCD), (CD, ABCD).$$



ABCD

C D

AB

CD

A B

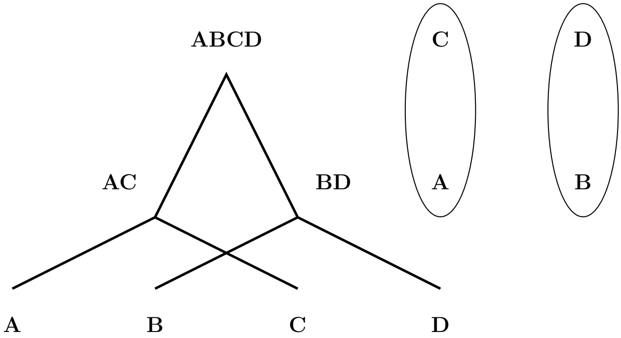
A B C D

Cluster tree $\mathcal{B}_2$:
Deformation of the mentioned quadrangle gives the cluster set

$$A, \quad B, \quad C, \quad D, \quad AC, \quad BD, \quad ABCD$$

with the 2-edges

$$(A, AC), (C, AC), (B, BD), (D, BD), (AC, ABCD), (BD, ABCD).$$


ABCD
AC
BD
A
B
C
D
C
A
D
B

Cluster network $\mathcal{C}_2$: Union of the trees $\mathcal{A}_2$ and $\mathcal{B}_2$ of clusters. Contains the unions of the vertex sets and the 2-edges sets in $\mathcal{A}_2$ and $\mathcal{B}_2$ and the four 4-edges

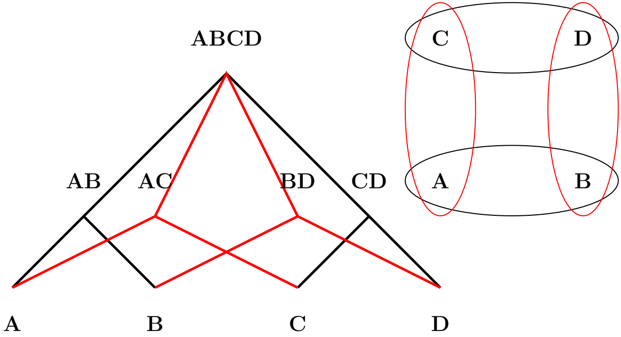
$$(A, AB, AC, ABCD), (B, AB, BD, ABCD),$$

$$(C, AC, CD, ABCD), (D, BD, CD, ABCD).$$

Such a hypergraph can be represented schematically by the table

$$\begin{array}{|c|c|c|}\hline A & AC & C \\ \hline AB & ABCD & CD \\ \hline B & BD & D \\ \hline \end{array}.$$

The matrix entries are the hypergraph vertices, the 2-edges join the neighboring vertices (in the horizontal and vertical directions), and the 4-edges correspond to the small 2×2 squares containing the matrix corners and the cluster $ABCD$.
The 4-edges describe the histories of the growth of one-point clusters with respect to the different clustering trees.


ABCD
AB
AC
BD
CD
A
B
C
D
C
D
A
B

p-**Adic case**

Multidimensional metric in $\mathbb{Q}_p^d$

$$d_{q_1,\ldots,q_d}(x,y) = \max_{i=1,\ldots,d}(q_i|x_i - y_i|_p),$$

$$p^{-1} < q_1 < \cdots < q_d \leq 1.$$

The dilations $p^k\mathbb{Z}_p^d$, $k \in \mathbb{Z}$ (and their translations) are balls with respect to all above ultrametrics.

The set of intermediary $d_{q_1,\ldots,q_d}$–balls between $p\mathbb{Z}_p^d$ and $\mathbb{Z}_p^d$ – the sequence of products

$$B_a = \mathbb{Z}_p \times \cdots \times \mathbb{Z}_p \times p\mathbb{Z}_p \times \cdots \times p\mathbb{Z}_p,$$

with a components $\mathbb{Z}_p$ and $d-a$ components $p\mathbb{Z}_p$, $a = 0, \ldots, d$.

The sequence $\{B_a\}$, $a = 0, \ldots, d$ of balls is a complete flag, where we consider the natural correspondence between the a-dimensional spaces and $B_a/p\mathbb{Z}_p^d$.

Recall that a flag is an increasing sequence

$$V_0 \subset V_1 \subset \cdots \subset V_N$$

of subspaces of a finite–dimensional vector space V.
A flag in the space of dimension d is complete if it contains spaces of all dimensions $0, 1, \ldots, d$.

Taking reorderings of q_i we get the different trees of clusters.
$p\mathbb{Z}_p^d$ and $\mathbb{Z}_p^d$ are balls for all such metrics, this implies cycles in the union of the corresponding cluster trees.

This union is an apartment in the spherical Bruhat–Tits building.

In general, we consider a metric $d_{q_1,\dots,q_d}$ where $q_i \neq 0$ (not necessarily between p^{-1} and 1) and a metric

$$s^A_{q_1,\dots,q_d}(x, y) = d_{q_1,\dots,q_d}(Ax, Ay),$$

A is a $d \times d$ matrix, matrix elements $A_{ij} \in \mathbb{Z}_p$, $|\det A|_p = 1$.

Hypergraphs of balls for the family of metrics $s^A_{q_1,\dots,q_d}$ — related to the affine Bruhat–Tits building (since balls are lattices in $\mathbb{Q}_p^d$).

Affine Bruhat–Tits buildings

The affine Bruhat–Tits building is a hypergraph (simplicial complex). Vertices are equivalence classes of lattices.

A lattice in $\mathbb{Q}_p^d$ is a $\mathbb{Z}_p$–module of the form

$$\oplus_{i=1}^{d}\mathbb{Z}_p e_i,$$

where $\{e_i\}$ is an arbitrary basis in $\mathbb{Q}_p^d$. Lattice is a ball with respect to some metric $s^A_{q_1,\dots,q_d}$.

Two lattices are equivalent if one is a scalar multiple of the other.

Two lattices L_1 and L_2 are adjacent (connected by an edge) if some representatives from equivalence classes L_1 and L_2 satisfy

$$pL_1 \subset L_2 \subset L_1.$$

k–Simplices are defined as equivalence classes of $k+1$ mutually adjacent lattices, i.e. the chains

$$pL_{k+1} \subset L_1 \subset L_2 \subset \cdots \subset L_{k+1}.$$

Here $1 \leq k \leq d-1$.

Apartment in the defined above building is a set of vertices and simplices corresponding to a fixed basis $\{e_i\}$ in $\mathbb{Q}_p^d$ which contains the lattices $\oplus_{i=1}^{d} \mathbb{Z}_p p^{a_i} e_i$, $a_i \in \mathbb{Z}$.

This definition can not be directly generalized for the case of general families of ultrametrics (since it is valid for equivalence classes of lattices, i.e. balls), but if we will consider instead of equivalence classes the balls itself we can propose a natural generalization.

General hypergraphs of clusters

Let X be a locally compact ultrametric space with some finite family of ultrametrics $\mathbf{s}$ defined on X. Moreover, let, for any pair of metrics $s, r \in \mathbf{s}$, any s–ball be a finite union of r–balls.

The family $\mathbf{s}$ of ultrametrics on X is compatible, if for any two balls, an s–ball I and an r–ball J, $s, r \in \mathbf{s}$, the intersection $I \bigcap J$ is a ball with respect to some ultrametric $t \in \mathbf{s}$.

The hypergraph $\mathcal{C}(X, \mathbf{s})$ as a graph is a union of the trees $\mathcal{T}(X, s)$ of s–balls, $s \in \mathbf{s}$.
The set of vertices of $\mathcal{C}(X, \mathbf{s})$ is the union of the sets of s–balls, $s \in \mathbf{s}$, edges connect s–balls (with the same s) nested without intermediaries. The partial order is by the inclusion of subsets in X. If some s–ball coincides with some r–ball as a set, they define the same vertex in $\mathcal{C}(X, \mathbf{s})$.

Hyperedges $\mathcal{E}$ in $\mathcal{C}(X, \mathbf{s})$:

Fix a subfamily $\mathbf{r} \subset \mathbf{s}$ of ultrametrics on X. Let us fix some $\mathbf{r}$–ball I (i.e. I is an s–ball with respect to all $s \in \mathbf{r}$).

Let J be a smallest $\mathbf{r}$–ball which is strictly greater than I.

We define $\mathcal{E}$ as a family $\{K : I \subset K \subset J\}$ of s–balls for $s \in \mathbf{r}$ (i.e. any of K is an s–ball for some $s \in \mathbf{r}$). In particular $I, J \in \mathcal{E}$.

Dimension of a hypergraph of clusters

The idea – to generalize the definition of dimension (the number of p-adic parameters) to general hypergraphs of clusters.

Let $\mathcal{E}$ be a $\mathbf{r}$–hyperedge in $\mathcal{C}(X, \mathbf{s})$, $\mathbf{r} \subset \mathbf{s}$, with the minimal $\mathbf{r}$–ball I and the maximal $\mathbf{r}$–ball J.

The A–dimension of the hyperedge $\mathcal{E}$ is the maximum of lengths of increasing paths in $\mathcal{E}$ from I to J (with respect to the partial order in $\mathcal{E}$).

The B–dimension of the hyperedge $\mathcal{E}$ is the number of balls J_k, where $I \subset J_k \subset J$ and J_k is a maximal subball of J with respect to some metric $r \in \mathbf{r}$.

In the p-adic case both above dimensions are equal to the number of p-adic parameters.

The introduced dimensions are not equal to the VC (or Vapnik – Chervonenkis, or combinatorial) dimension.

Applications to data analysis

Clustering – a tool of data analysis with applications to bioinformatics and taxonomy.

The set X of data may be generated in a complex way, there may be some independent contributions.

In mathematics independence is described by dimensionality.

There should be some way to describe independencies in data at the level of graphs (and hypergraphs) of clusters.

Classification trees (trees of clusters) describe the diversity of data, the multidimensional generalization should describe the situation where we have independent sources of diversity.

Dimension of a hyperedge of a hypergraph of clusters will describe the number of sources of diversity.

Applications to taxonomy

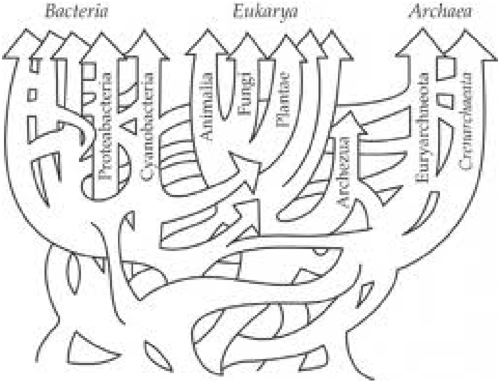
The metric for the clustering procedure – sum of contributions from the different genetic markers

$$d(X, Y) = \sum_{j=1}^{N} w_j d_j(X_j, Y_j),$$

$w_j \geq 0$ are weights, X and Y are genomes,
X_j are Y_j are genetic markers,
d_j – the distance for the j-th genetic marker.

Different sets of weights – different classification trees.
Each d_j generates the corresponding tree,
Union of the trees – the forest of life.

Evolution is *reticulate*
– some parts of a genome may have the different origin (hybridization, horizontal gene transfer).
Phylogenetic networks instead of trees



Discussion of forest of life and application of (non–tree) networks for description of reticulate evolution

K.Vetsigian, C.Woese, N.Goldenfeld, Collective evolution and the genetic code, PNAS, 2006, vol. 103, no. 28, P.10696–10701.

E.V. Koonin, Yu. Wolf, G. Karev (Editors), Power Laws, Scale-Free Networks and Genome Biology, Springer, 2006.

E.V.Koonin, The Logic of Chance: The Nature and Origin of Biological Evolution. FT Press, 2011.

Mathematical methods of analysis of phylogenetic networks

D.H. Huson, R. Rupp, C. Scornavacca, Phylogenetic Networks, Cambridge University Press, 2010.

A. Dress, K.T. Huber, J. Koolen, V. Moulton, A. Spillner, Basic Phylogenetic Combinatorics, Cambridge University Press, 2012.

Description of Horizontal Gene Transfer:
transfer of a genetic marker from one specie to the other:
we use the same formula for the metric

$$d(X, Y) = \sum_{j=1}^{N} w_j d_j(X_j, Y_j),$$

where X and Y are genomes,
X_j are Y_j are genetic markers.

If some genetic marker Y_k is absent in the genome X we add a symbol Ω (the blank space symbol), and define the corresponding distance

$$d_k(\Omega, Y_k) = D$$

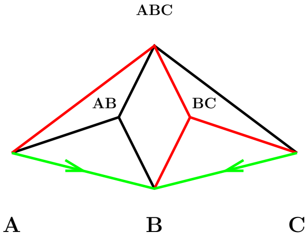
to be sufficiently large (as for completely different genetic markers).

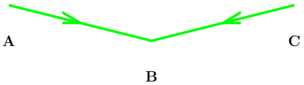
For reticulate evolution one might construct a family of classification trees using the clustering procedure.

Since we have several genetic markers with the different evolution histories the classification tree is non–unique.

One can combine the classification trees into a single network using the described procedure and use this network as a phylogenetic network for reticulate evolution.

Example: HGT between the species A and C, we get the hybrid B. Taking the different metrics, we obtain the classification graph, see below (in black and red). The picture biologists would like to get is depicted in green.
In general it is not clear how to construct a phylogenetic network for the case of reticulate evolution, but the classification network can be constructed automatically.


ABC
AB
BC
A
B
C


A
C
B

Instead of reproduction of the detailed genetic history of populations (which is not possible) we use classification networks to describe the evolution of ensembles of genes (in general, of genetic markers).

The minimal nodes of the classification network correspond to some species or individuals, but non–minimal nodes of the network are not necessarily correspond to some ancestor species of the minimal nodes (**ancestors exist only as ensembles of genes**, these genes may belong to the different populations or species).

Since we use a general classification approach and non–minimal nodes of the network do not necessarily correspond to the ancestral populations the obtained phylogenetic networks are Linnaean but not necessarily Darwinian (for the Darwinian tree of life the non–minimal nodes describe the ancestors).

Conclusion

Clustering with respect to a family of metrics

Hypergraph of clusters

p-Adic case – relation to spaces of flags
and Bruhat–Tits buildings

Dimensionality of cluster systems
– number of sources of diversity in data

Description of taxonomy for reticulate evolution
— Linnaeus instead of Darwin