Topological Analysis of Syntactic Structures

Matilde Marcolli

MAT1509HS: Mathematical and Computational Linguistics

University of Toronto, Winter 2019, T 4-6 and W 4, BA6180

Main Reference

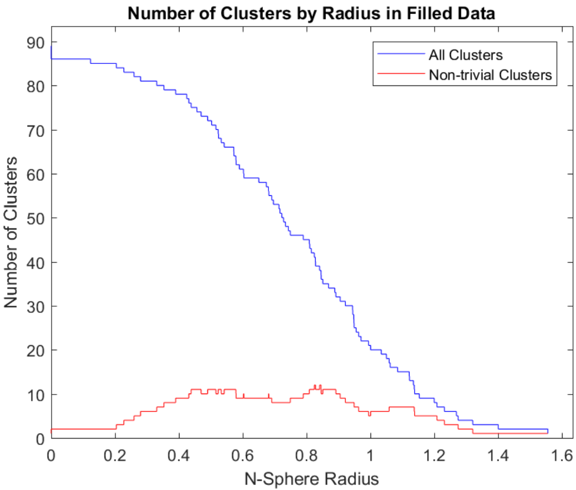
- Alexander Port, Taelin Karidi, Matilde Marcolli, *Topological Analysis of Syntactic Structures*, arXiv:1903.05181

Outline

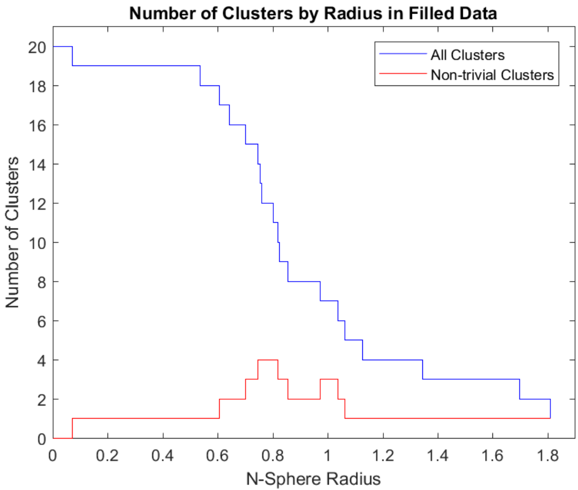
- databases of syntactic structures: SSWL (syntactic structures of world languages) and LanGeLin collaboration (York University)
- view datapoints as
  1. syntactic parameters (with values for languages as coordinates)
  2. languages (with their syntactic parameters as coordinates)
- search for structure of dependencies between syntactic parameters
- search for structures of relatedness between languages at syntactic level
- dimensional analysis
- topological data analysis

Clustering structure

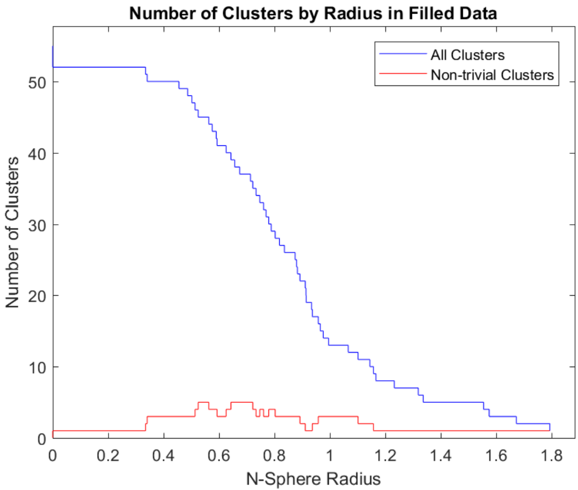
- different language families have different clustering structure
- families with less clustering: syntactic parameters more homogeneously distributed across languages in the family
- Indo-European family has largest number of non-trivial clusters among families in the database (also most extensively represented)
- also difference between SSWL and LanGeLIn data in terms of clusters
- in the SSWL there are singletons (only one data point) for every radius
- LanGeLin data: starting from a certain radius no more singletons



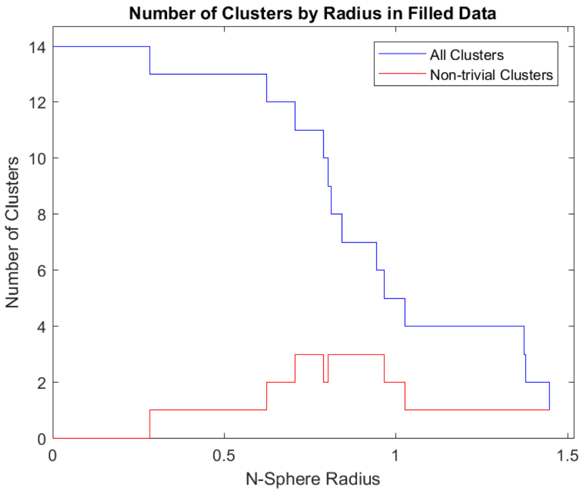
Number of clusters by radius for the SSWL data for the Indo-European languages



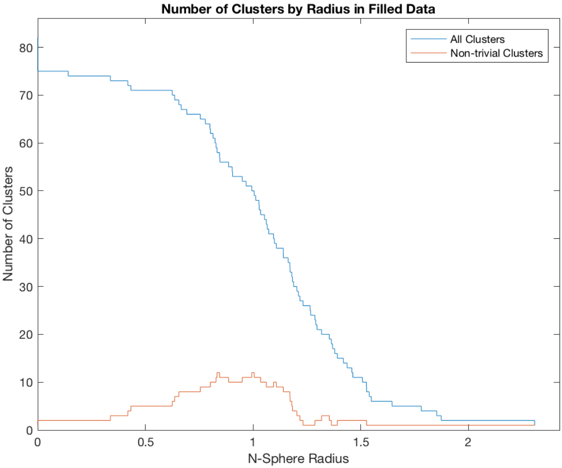
Number of clusters by radius for the SSWL data for the Austronesian languages



Number of clusters by radius for the SSWL data for the Niger-Congo languages



Number of clusters by radius for the SSWL data for the Afro-Asiatic languages



Number of clusters by radius for the LanGeLin data (mostly Indo-European languages)

Example: Clustering and the Greek-Italian Microvariations

- different clustering in SSWL and LanGeLin: Hellenic languages
  - in SSWL certain Hellenic languages (Cappadocian Greek, Modern Greek) remain singletons for a long range of radii and join other clusters very late in the persistence scale
  - in LanGeLin the Hellenic languages join clusters very early in the persistence diagram
- LanGeLin data include a range of Southern-Italian dialect that are either Romance or Hellenic (Salento Greek, Calabrian Greek A, Calabrian Greek B)

• Microvariations: languages either genealogically very closely related or in distinct genealogical groups but in close geographic proximity and interaction

• These Italian-Greek Microvariations studied at length in

- C. Guardiano, D. Michelioudakis, A. Ceolin, M. Irimia, G. Longobardi, N. Radkevich, I. Sitaridou, G. Silvestri, *South by Southeast. A Syntactic Approach to Greek and Romance Microvariation*, L'Italia Dialettale, Vol. 77 (2016) 95–166.

Principal Component Analysis

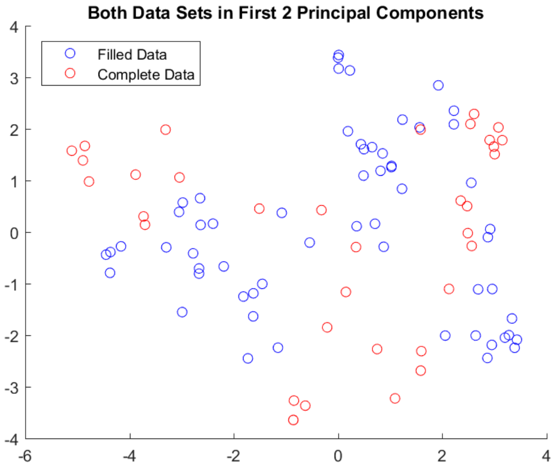
- in order to be able to perform persistent homology computations need to reduce the dimensionality of data
- use persistent component analysis: projection onto first few principal components
- data set: M vectors X_j of length N, matrix $N \times M$, search for linear combinations $\sum_j a_j X_j$ of columns with *maximum variance* $\mathrm{var}(Xa) = a^t S a$, with $S = X^t X$ variational problem is largest eigenvalue $Sa = \lambda a$:

$$\mathrm{argmax}_{\|a\|=1} \{a^t X^t X a\}$$

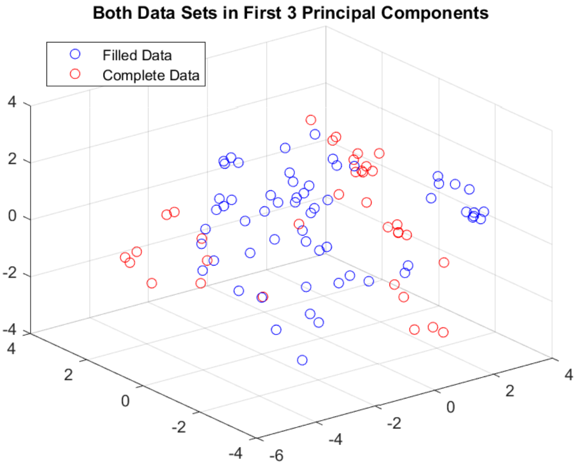
- successive components: variational problem on complement of previous components
- quality of the lower dim approximation measured by variability associated with the set of retained PC: the proportion of total variance retained

$$\pi_j = \frac{\lambda_j}{\sum_j \lambda_j} = \frac{\lambda_j}{\mathrm{Tr}(S)}$$

- work mostly with PCA 60%
- but there are some dependences of the clustering structure on the PCA variance level
- these introduce some differences in smaller structures in the tree diagram of how clusters merge as a function of the scale parameter in the persistent H_0
- specific examples of how this can alter some proximity relations between languages in subfamilies and cause some misplacements in the tree topology
- principal component structure for SSWL and LanGeLin data also very different: fewer *linear* relations between SSWL data (but expect more relations)
- principal components can be seen as an assignment (based on language data) of real weight to each syntactic parameter/variable: linguistic interpretation of these weights?



$2D$ scatters plot of the data in the first 2 principal components



3D scatters plot of the data in the first 3 principal components

Dimensionality Analysis

Algorithm

- $D \subseteq \mathbb{R}^d$ data set
- choose $p \in \{1, \dots, |D|\}$ and point $\vec{x}_1^{(p)}$
- sort D into vector $\{\vec{x}_i^{(p)}\}_{i=1}^{|D|}$ entries ordered by distance from chosen point

$$d(\vec{x}_i^{(p)}, \vec{x}_1^{(p)}) \leq d(\vec{x}_j^{(p)}, \vec{x}_1^{(p)}) \quad \text{for} \quad i \leq j$$

- for $s \in \{1, \dots, |D|\}$ a number of nearest neighbors set

$$X^{(p,s)} = \frac{1}{d(\vec{x_{s+1}}^p, \vec{x_1}^p)} \begin{bmatrix} \vec{x}_2^{(p)} - \vec{x}_1^{(p)} \\ \vec{x}_3^{(p)} - \vec{x}_1^{(p)} \\ \vdots \\ \vec{x}_{s+1}^{(p)} - \vec{x}_1^{(p)} \end{bmatrix}$$

- data spread out: shift points s-nearest neighbors so fit into d-dimensional unit ball at selected point

Best Fit Algorithm

- for a weight matrix $W^{(p,s)} \in M_{s\times s}(\mathbb{R})$

1. weighted covariant matrix:

$$C^{(p,s)} = \frac{1}{s} (\vec{x}^{(p,s)})^T W^{(p,s)} \vec{x}^{(p,s)}$$

2. compute its eigenvalues $\lambda_1^{(p,s)} \geq \cdots \geq \lambda_d^{(p,s)}$ and corresponding eigenvectors $\{\vec{v}_{\lambda_1}^{(p,s)}, \ldots, \vec{v}_{\lambda_d}^{(p,s)}\}$
3. eigenbasis matrix $V^{(p,s)} \in M_{d\times d}(\mathbb{R})$ columns $\vec{v}_{\lambda_i}^{(p,s)}$
4. choose dimension of fit $f \in \{1, \ldots, s\}$ and set $P^{(f)} \in M_{d\times d}(\mathbb{R})$ diagonal

$$P_{ii}^{(f)} = \begin{cases} 0, & \text{if } i \leq f \\ 1, & \text{if } i > f \end{cases}$$

5. (p, s, f)-error $W^{(p,s)} X^{(p,s)} (V^{(p,s)^t})^{-1} P^{(f)} V^{(p,s)^t}$

Weight Matrix

- $W^{(p,s)} \in M_{s\times s}(\mathbb{R})$ be the weight matrix such that
  1. $W^{(p,s)}$ is diagonal
  2. $\det(W^{(p,s)}) = 1$
  3. $W^{(p,s)}_{i,i} \sim \exp\left(\frac{-d(\vec{x}^{(p)}_{i+1} - \vec{x}^{(p)}_1)^2}{\alpha}\right)$
- range α small local behavior, range α large global behavior

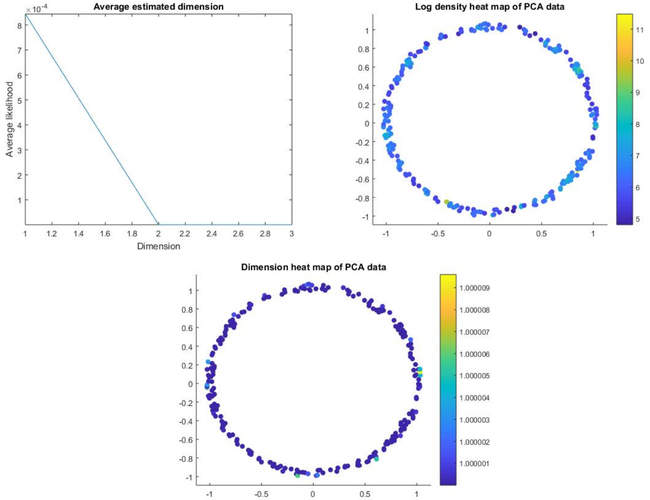
Dimension Test

- (p, s, f)-error given by

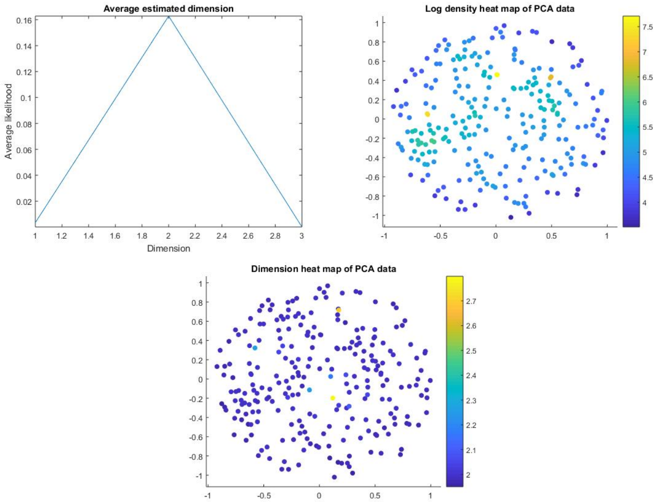
$$W^{(p,s)} X^{(p,s)} ({V^{(p,s)}}^t)^{-1} P^{(f)} {V^{(p,s)}}^t$$

- magnitude of the i-th row of $X^{(p,s)} ({V^{(p,s)}}^t)^{-1} P^{(f)} {V^{(p,s)}}^t$ is orthonormal error of the i-th nearest neighbor
- compute all the (p, s, f)-errors on selected data and also on balls and spheres of $dim \leq d$ for comparison
- run paired T-test between the selected points and balls/spheres database to get maximum likelihood value
- best estimate of dimension at that point

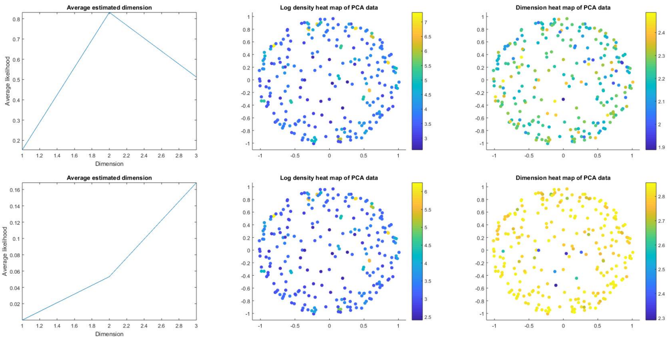
Spheres and balls



dimension estimate and density and dimension heat maps for 1-sphere with $\alpha = 1/3$

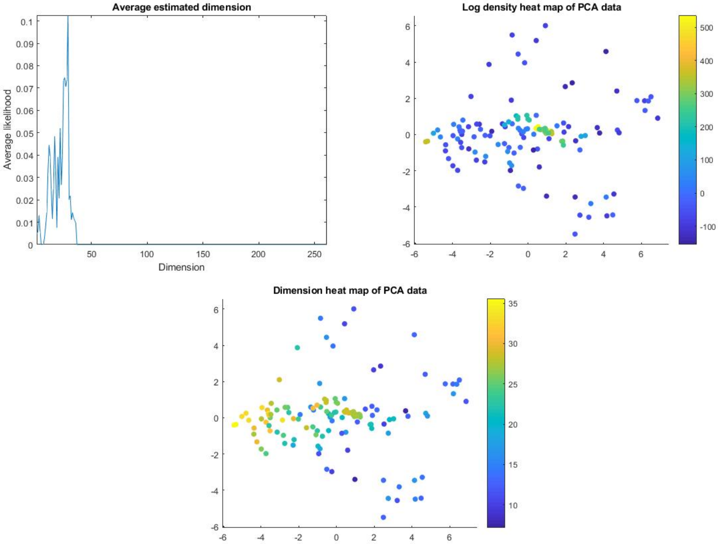


dimension estimate and density and dimension heat maps for 2-ball with $\alpha = 1/3$

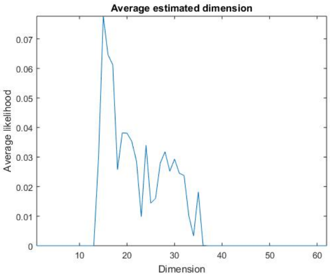


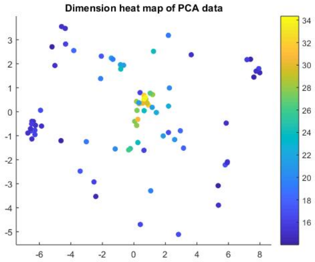
dimension estimate and density and dimension heat maps for 2-sphere with $\alpha = 1/10$ and $\alpha = 1/3$ (sometimes large α favor dimension of ambient space: mostly use $\alpha = 1/3$)

Dimension of SSWL syntactic variables (116 dim ambient space)



Dimension of LanGeLin syntactic variables (83 dim ambient space)



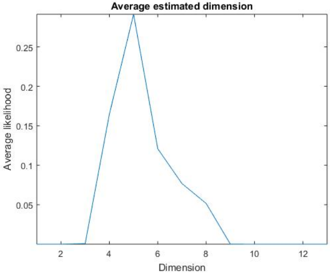


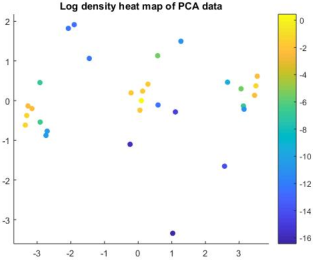
- SSWL syntactic variables estimate $d \sim 30$
- LanGeLin syntactic parameters estimate $d \sim 15$

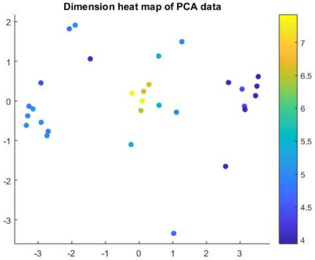
Relations between syntactic parameters

- *universal* relations that hold across language families
- *family-specific* relations that only hold within a given language family
- expect further dimension drop when estimating dimension of syntactic parameters with the parameters evaluated only on languages of a particular family
- certain families have more family-specific relations (bigger drop in dimension)
- example: SSWL syntactic variables have more family-specific relations for the Niger-Congo family than for the Indo-European family

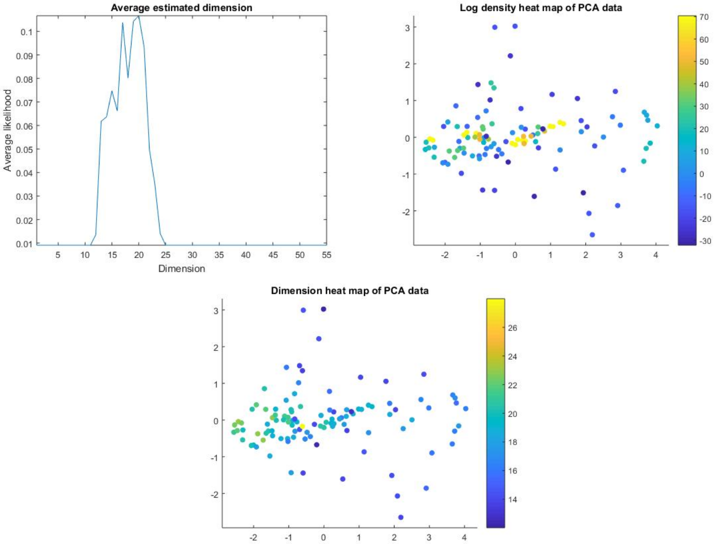
Family-specific relations for LanGeLin parameters: Romance languages $d \sim 5$ (while universal relations $d \sim 15$)



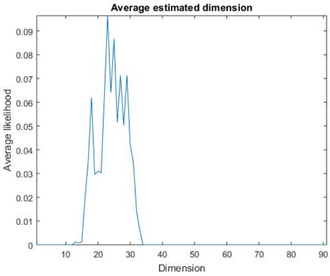


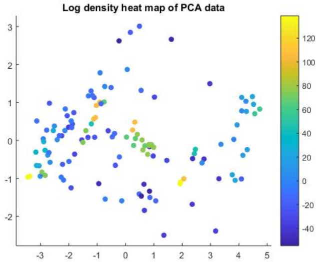


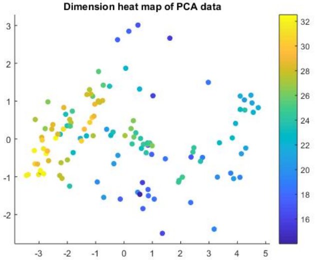
Family-specific relations for SSWL syntactic variables: Niger-Congo languages $d \sim 20$ (universal relations $d \sim 30$)



Family-specific relations for SSWL syntactic variables: Indo-European languages $d \sim 23$ (universal relations $d \sim 30$)

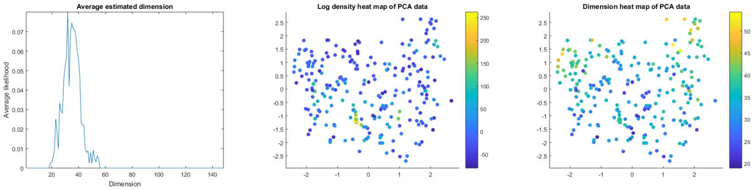






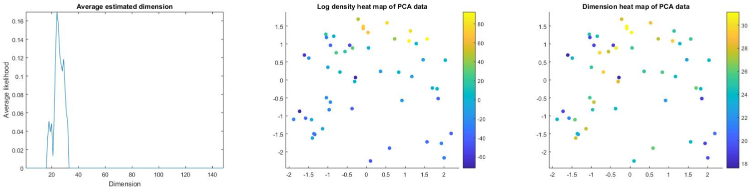
Dimensional Analysis with Languages as Data
(and parameters as coordinates)

- provides a measure of how spread-out syntactic features are across languages in a given family
- how diversified (syntactically) a historical language family

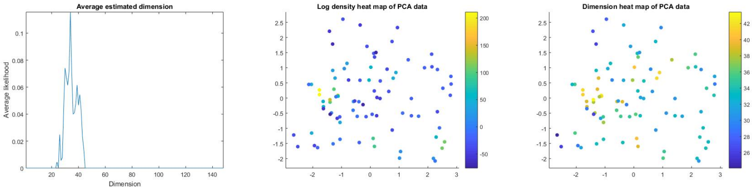


SSWL data across all language families dimension $d \sim 38$

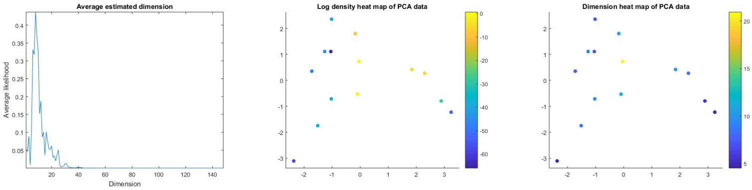
Niger-Congo languages SSWL data: $d \sim 23$



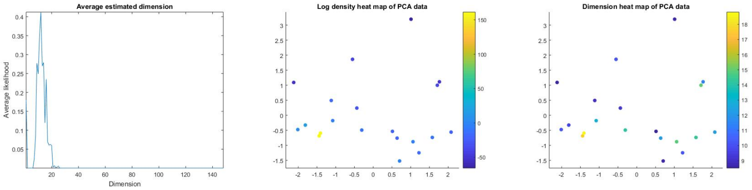
Indo-European languages SSWL data: $d \sim 38$



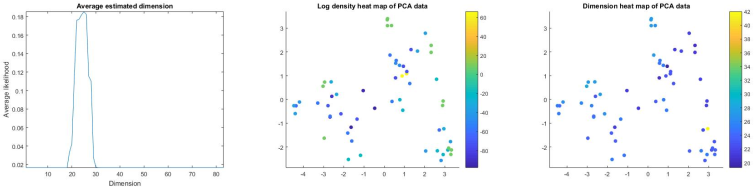
Afro-Asiatic languages SSWL data: $d \sim 8$



Austronesian languages SSWL data: $d \sim 12$



Dimensionality by languages: LanGeLin data $d \sim 25$



Question can apply this kind of dimensionality analysis by language families to some historical linguistics problem?

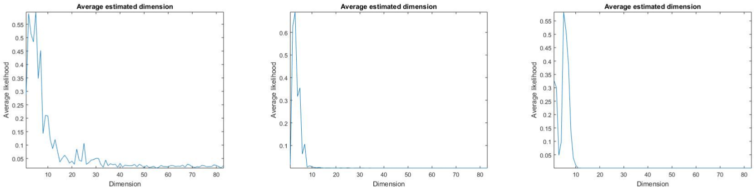
The Ural-Altaic Question

- Uralic languages: Estonian, Finnish, Hungarian, Udmurt, Yukaghir, Khanty
- Altaic languages (more hypothetical): Turkish, Buryat, Yakut, Even, Evenki, Karachay, Tatar, Tuvan, Uyghur
- previously proposed as Altaic: Korean and Japanese (now largely discarded hypothesis)
- very hypothetical Ural-Altaic family: proposed unified historical origin of the Uralic and Altaic families
- hypothesis with a long contested history...

- some bibliographic references
  - A. Marcantonio, *The Uralic Language Family: Facts, Myths and Statistics*, Publications of the Philological Society, Vol. 35, Blackwell, 2002
  - D. Sinor, *The Problem of the Ural-Altaic relationship*, in "The Uralic Languages: Description, History and Modern Influences", Brill 1988, pp. 706–741.

Dimensional Analysis of the Uralic and Altaic Languages

LanGeLin data

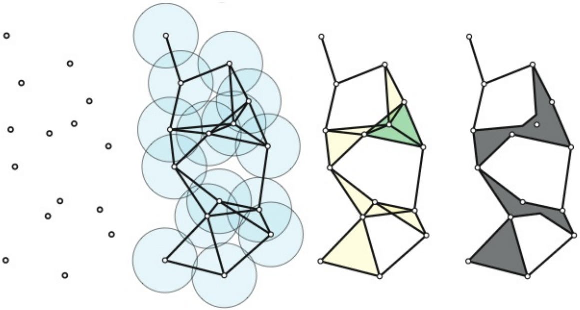


Uralic languages; Altaic languages; Uralic and Altaic languages

- when considered together the Ural-Altaic family has two distinct peaks of estimated dimension
- these roughly reflect the separate Uralic and Altaic cases
- supports the claim of distinct families (no Ural-Altaic)
- but... we'll see later that some mixing of Uralic and Altaic languages happens in the clustering of persistent connected components in the barcode diagrams of persistent H_0

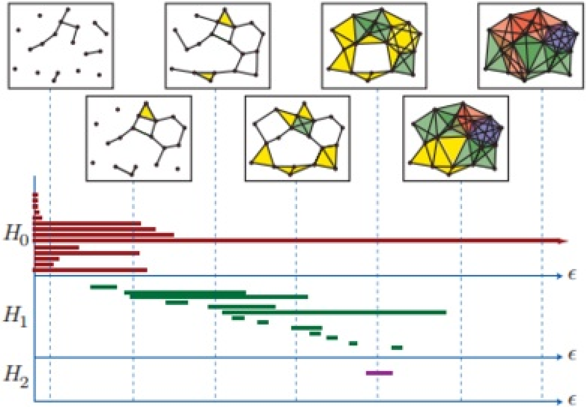
Vietoris–Rips complexes

- set $X = \{x_\alpha\}$ of points in Euclidean space $\mathbb{E}^N$, distance $d(x,y) = \|x-y\| = (\sum_{j=1}^N (x_j - y_j)^2)^{1/2}$
- Vietoris-Rips complex $R(X,\epsilon)$ of scale ϵ over field $\mathbb{K}$:
- $R_n(X,\epsilon)$ is $\mathbb{K}$-vector space spanned by all unordered $(n+1)$-tuples of points $\{x_{\alpha_0}, x_{\alpha_1}, \dots, x_{\alpha_n}\}$ in X where all pairs have distances $d(x_{\alpha_i}, x_{\alpha_j}) \le \epsilon$



Barcode Diagrams

- inclusion maps $R(X, \epsilon_1) \hookrightarrow R(X, \epsilon_2)$ for $\epsilon_1 < \epsilon_2$ induce maps in homology by functoriality $H_n(X, \epsilon_1) \to H_n(X, \epsilon_2)$
- barcode diagrams: births and deaths of persistent generators



Persistent Components Tree

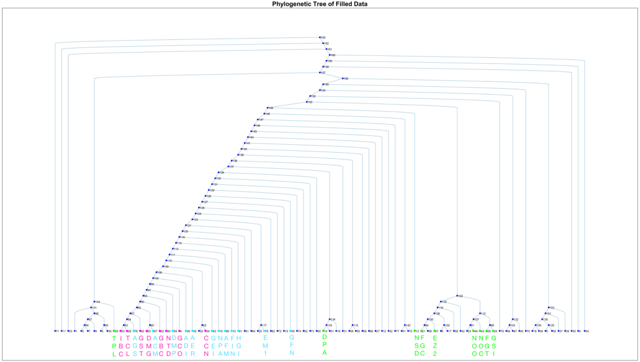
- for very small $\epsilon > 0$ each point a singleton component
- for very large $\epsilon > 0$ all points have joined into the same persistent connected component
- in between the components join in a certain order as function of ϵ that reflects the barcode diagram
- construct a tree that follows the merging of connected components as ϵ grows

Tree construction algorithm

- construct a PCA basis and takes up to chosen percent variance we choose (usually take 60% sometimes compare result with 80% to detect effects of PCA variance)
- compute pairwise Euclidean distances and find critical radius where only one connected component left.
- computes all clusters in small incremental radii and assemble persistent components tree based on inclusions:
  - C_r the set of all clusters at radius r and $C = \sqcup_r C_r$
  - in C cluster C_i is a child of cluster C_j if $C_i \subseteq C_j$ and there is no cluster C_k such that $C_i \subseteq C_k \subseteq C_j$

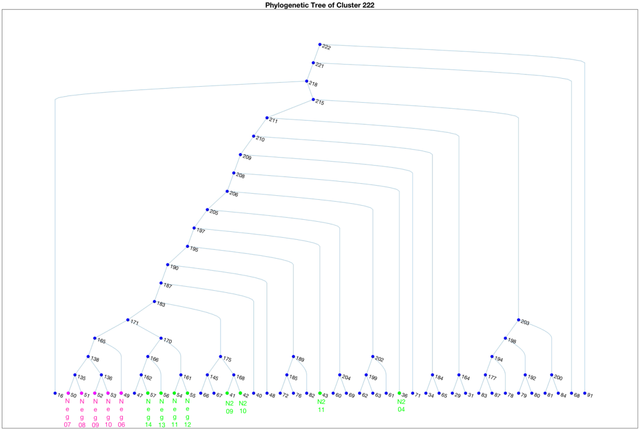
First Goal: use this tree of persistent connected component (hierarchical clustering structure) to describe relations ("tendency to align" type relations) among syntactic parameters, compare with other methods (heat kernel) of detecting relations

Persistent Components Tree of LanGeLin parameters compared to heat kernel clusters



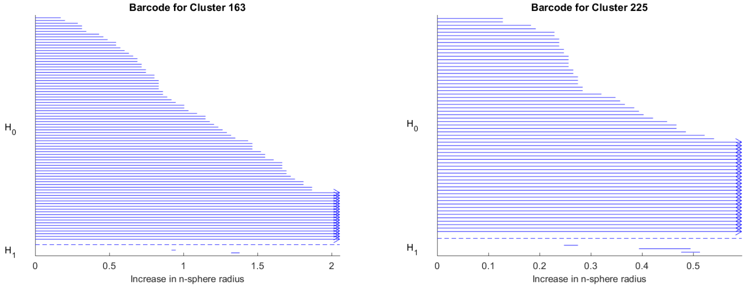
pink-colored and blue-colored: same as in two main sub-clusters of first cluster with heat-kernel method; green-colored second smaller cluster in heat-kernel method

Persistent Components Tree of SSWL syntactic variables compared to heat ker



pink-colored: first heat kernel cluster; green-colored: second

Persistent H_1 relations of syntactic parameters



persistent H_1 (truncated calculation of barcode diagrams) for LanGeLin and SSWL data

Question: interpretation of H_1 relations, explicit loop generators, meaning of homologous loops

Language Relatedness via Persistent Components Trees

- Main Question: how close is the persistent components tree for the syntactic data of languages in a given family to a phylogenetic tree of historical linguistic development of that family?

- Quick Answer: they are close to phylogenetic trees
...but not too close!

- in a phylogenetic tree the inner nodes represent branching between languages and ancient languages or proto-languages; in persistent component trees inner nodes only represent hierarchical clustering of languages according to the structure of their syntactic parameters but do not stand for other ancestor languages

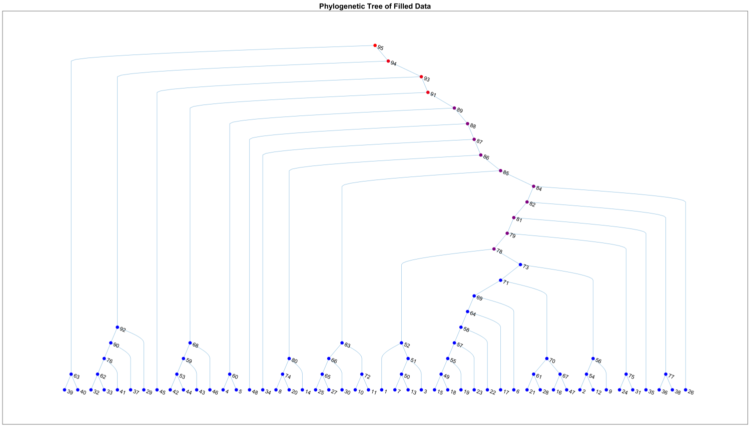
- still there is some strong correlation between the persistent components trees and phylogenetic trees

- Note: here the data set is read with languages as data points and syntactic parameters as their coordinates (transpose of previous for analysis of syntactic relations)

Filtering SSWL data

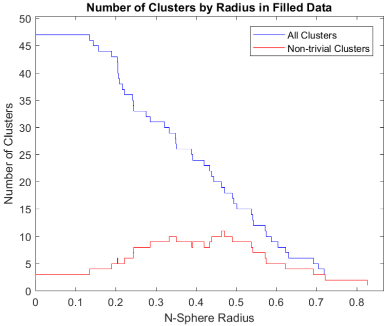
- observed in previous tree constructions based on SSWL data: many languages have incomplete mapping of parameters, this affects tree construction algorithms
- restrict of smaller trees using only parameters that are completely mapped for all languages in the subfamily
- if considering languages across family filtering procedure:
  1. retain only languages that are at least 50% complete in parameters values
  2. retain only parameters that are complete for all languages in this set
- this procedure ensures remaining set of completely mapped parameters not too small
- but it works best on language families that have enough well mapped languages
- Indo-European languages predominant

Persistent Components Tree (Indo-European + Ural-Altaic) filtered

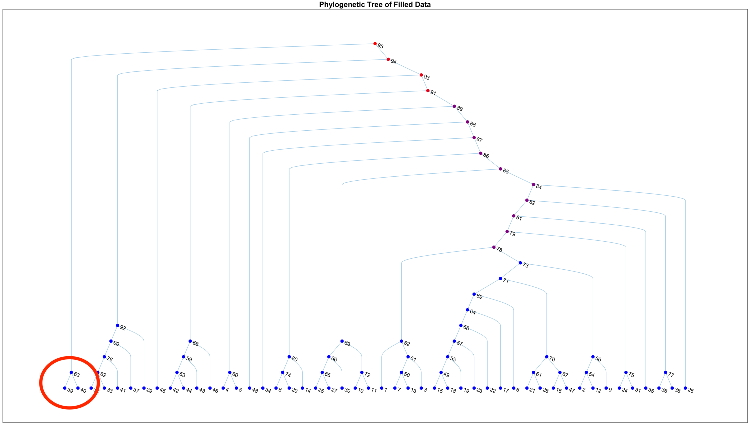


shows a mix of clusters that resemble historical phylogenetic relatedness and others that do not

Cluster structure of this Persistent Components Tree (filtered SSWL)

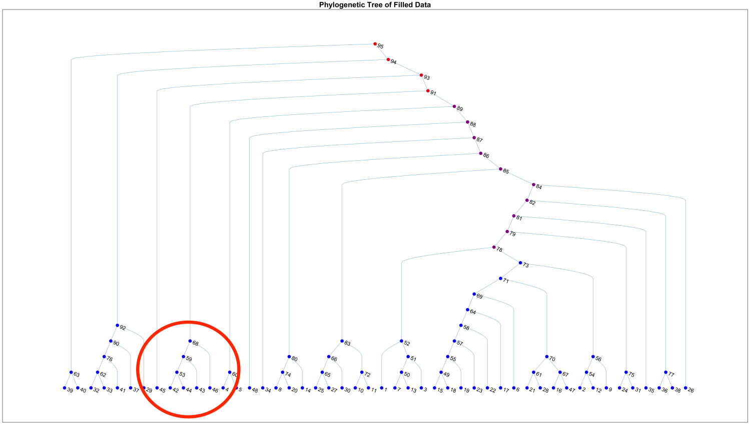


Example: cluster N. 63



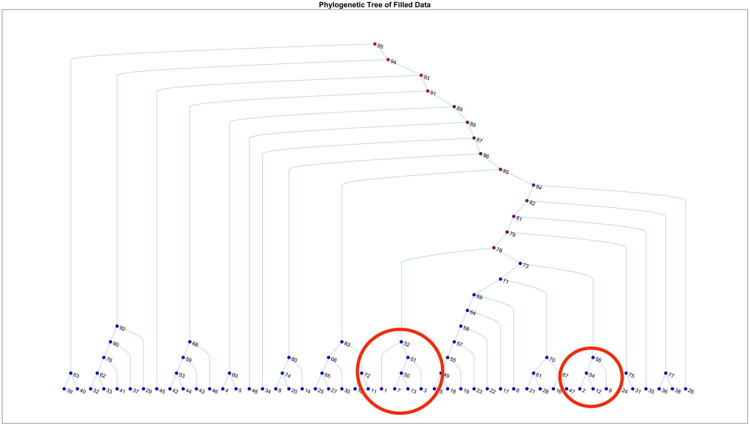
Korean and Japanese cluster together and excluded from Ural-Altaic hypothesis: good

Example: cluster N. 68



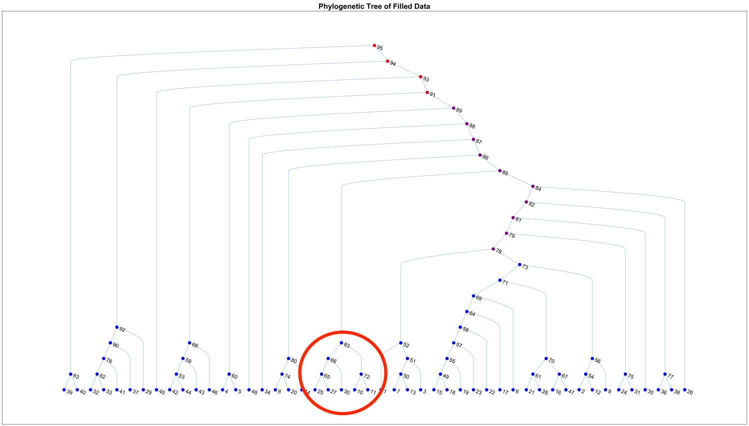
Indo-Iranic IE languages: Hindi, Panjabi, Pashto, Nepali: good

Example: clusters N. 52 and N. 56



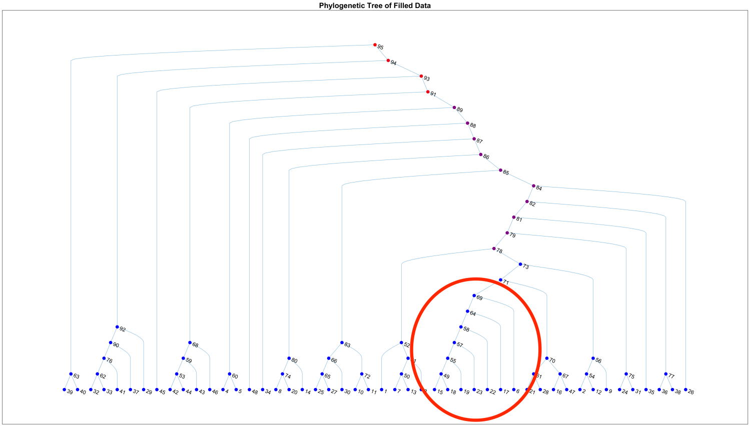
West-Germanic (Afrikaans, Dutch, German, West Flemish) and North-Germanic (Danish, Norwegian, Swedish), but English misplaced and Icelandic and Faroese misplaced

Example: cluster N. 83



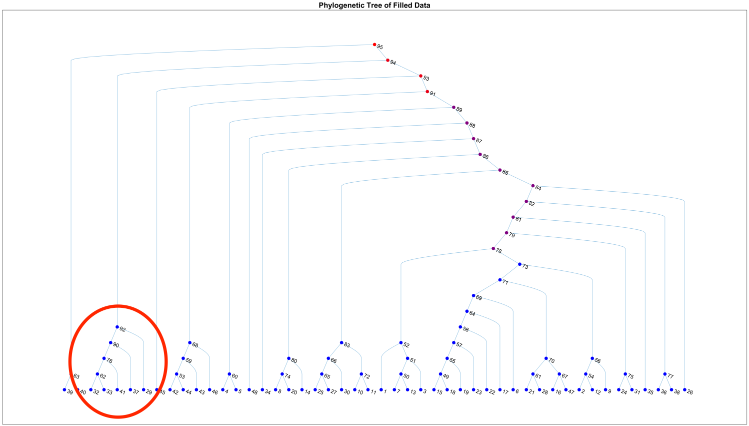
Ancient IE languages tend to group together away from their modern descendants: Latin, Ancient Greek, and Homeric Greek (sub-cluster 66) and Old English and Old Saxon (sub-cluster 72)

Example: cluster N. 69



Romance languages: Portuguese, Brazilian Portuguese, Catalan, Italian, Napoletano Antico, Sicilian, Spanish, French (other Romance languages misplaced in nearby cluster 70 with Albanian and some Hellenic)

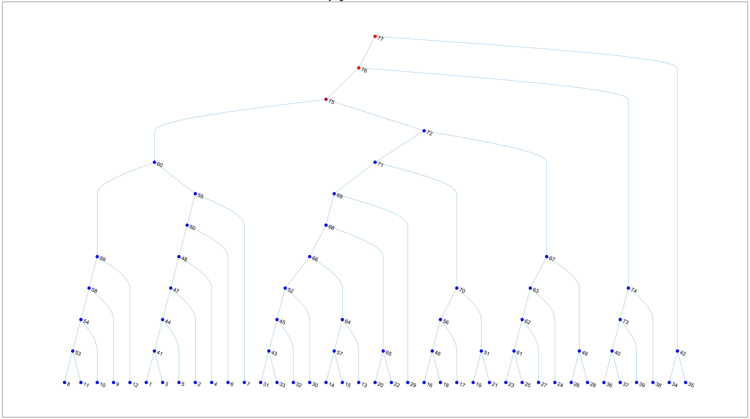
Example: cluster N. 92



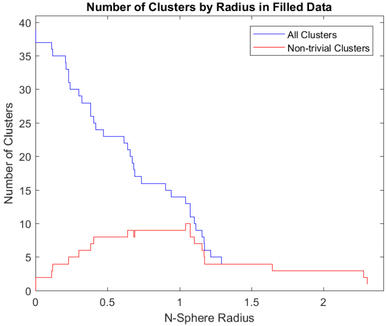
Mixed cluster with Eastern and Western Armenian, Turkish, Hungarian, and Cappadocian Greek: bad mix of IE and UA languages

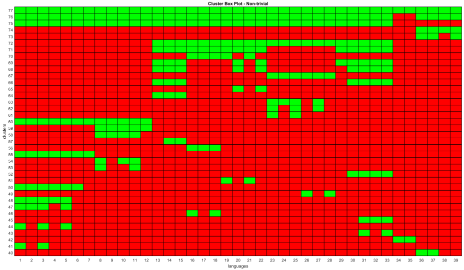
Indo-European Persistent Components Tree: LanGeLin data

Phylogenetic Tree of Filled Data

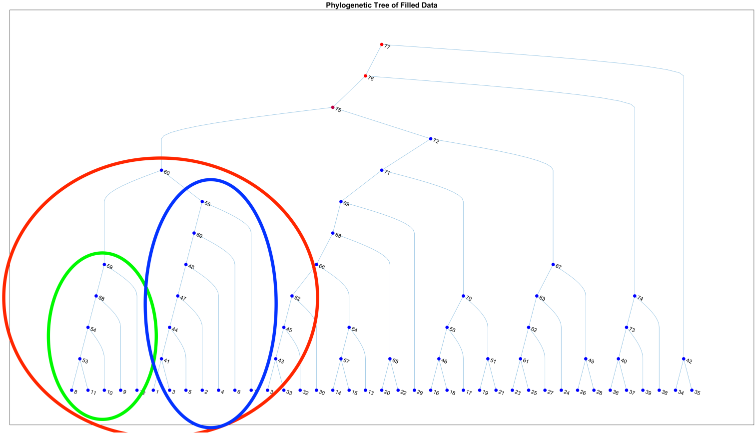


Cluster structure for this tree



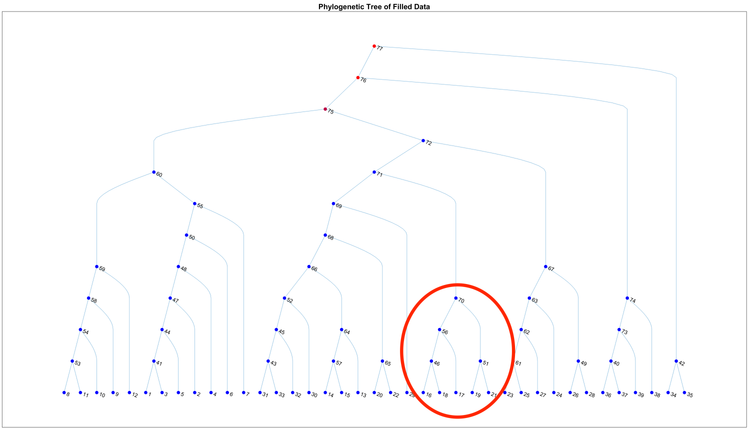


Example cluster N. 60



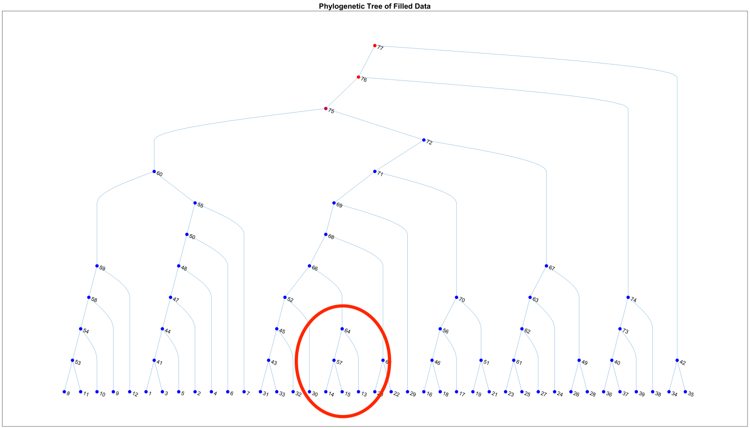
modern Romance languages: Italian, Spanish, French, Portuguese, and Romanian (sub-cluster 59) and Romance Southern Italian dialects: Ragusa, Mussomeli, Aidone, Southern Calabrese, Salentino, Northern Calabrese, Campano (sub-cluster 55)

Example cluster N. 70



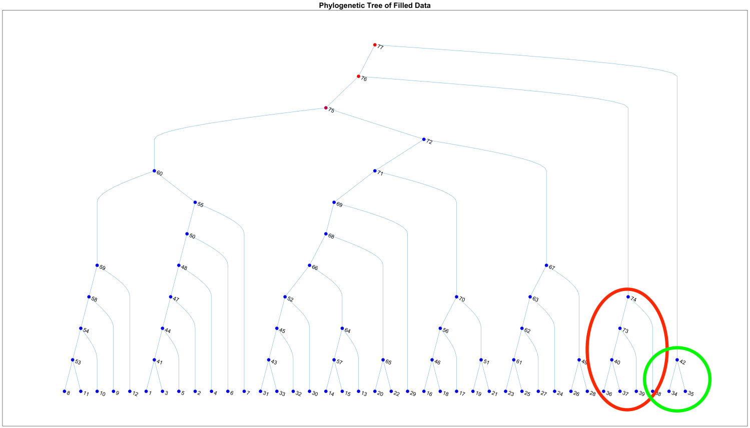
Hellenic languages: Salento Greek, Calabrian Greek A, Calabrian Greek B, Modern Greek, Cypriot Greek

Example cluster N. 64



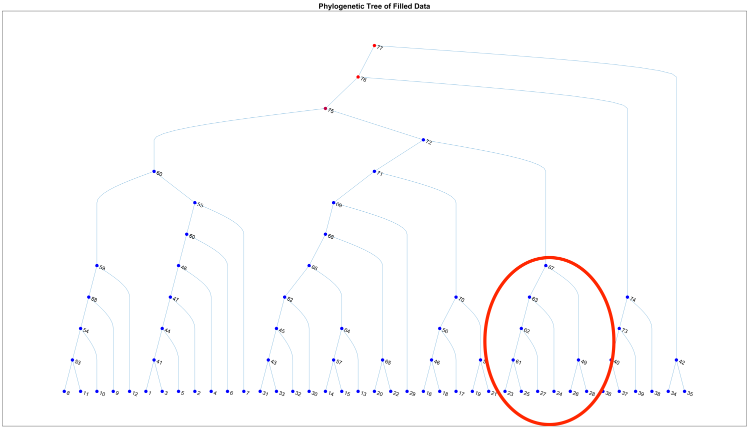
Ancient languages tend to group together: Latin, Classical Greek, New Testament Greek (nearby cluster 65 with Romeyka Pontic Greek grouped with Gothic)

Example clusters N. 74 and N. 42



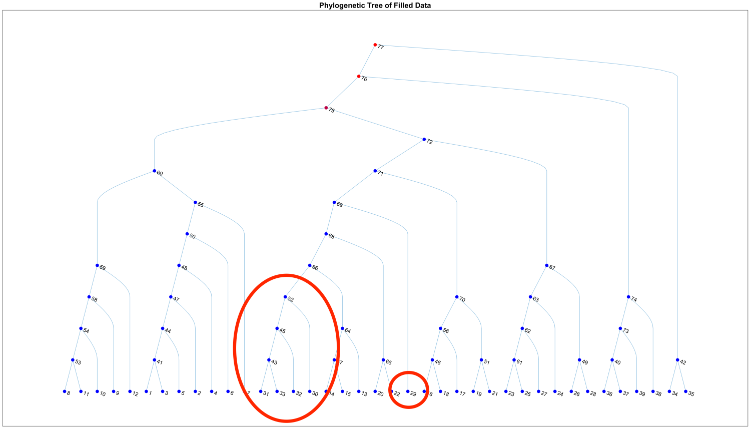
Indo-Iranian languages: Marathi, Hindi, Farsi, and Pashto; and Celtic languages: Irish and Welsh

Example cluster N. 67



Germanic languages: Old English, English, Dutch, Danish, Icelandic, and Norwegian (but Icelandic grouped with West-Germanic)

Example cluster N. 52

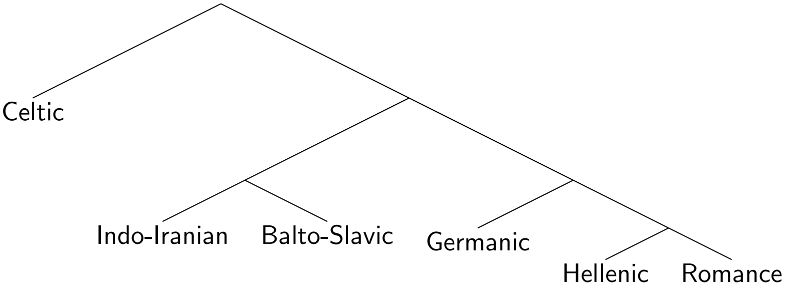


Slavic languages: Serb-Croatian, Slovenian, Polish, Russian; but Bulgarian as singleton 29 (structure 69 involving Slavic, Hellenic and Gothic: also occurs as an H_1-structure)

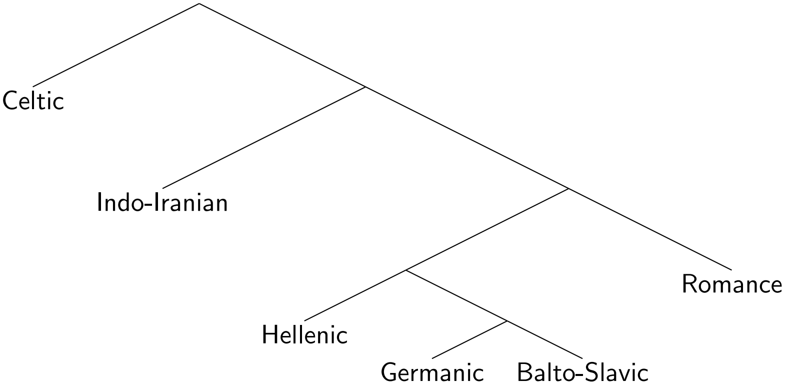
Observations

- generally persistent component clustering of LanGeLin data has closer correlation to phylogenetic trees of historical relatedness than for SSWL data (even after filtering)
- closer look at some subfamilies reveal misplacements
- misplacements within smaller subfamilies also affected by changing the PCA variance level
- on examples where phylogenetic algebraic geometry method (applied to same data) gives correct historical tree, persistent components tree is somewhat different
- not a problem of the data: persistent components merging measures something different than historical branching, though related

- historical main branching structure of the IE family obtained by LanGeLin collaboration with phylogenetic tree reconstruction algorithms:



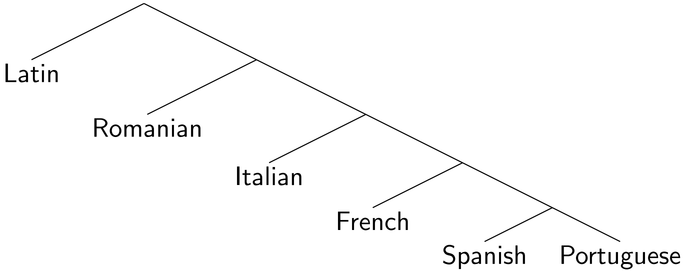
- branching structure of persistent connected components tree (same LanGeLin data):



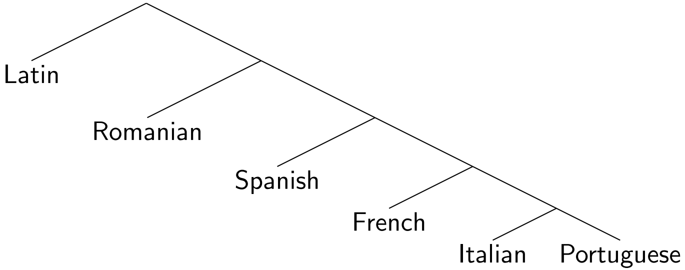
- related but not matching historical phylogenetic tree

Example of Romance languages

• correct historical phylogenetic tree (obtained from same SSWL and LanGeLin data by phylogenetic algebraic geometry)



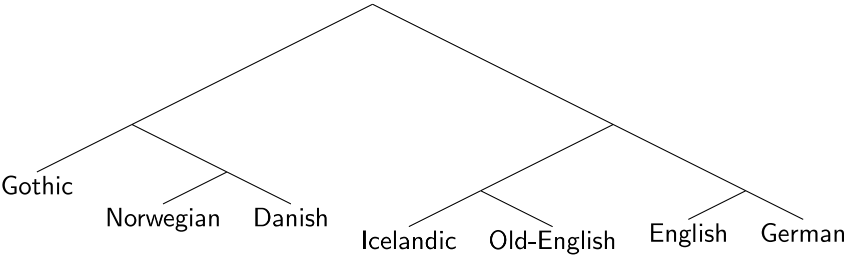
• persistent components tree (same for SSWL and for LanGeLin data)



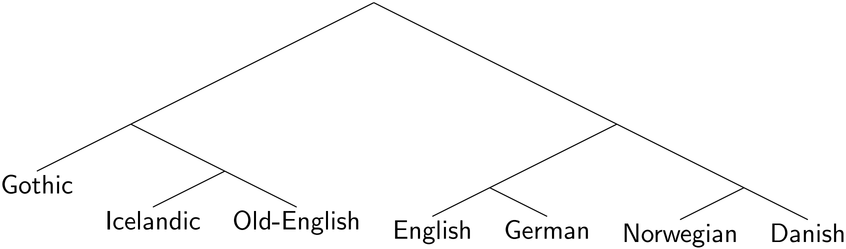
• positions of Italian and Spanish inverted

Example of Germanic languages

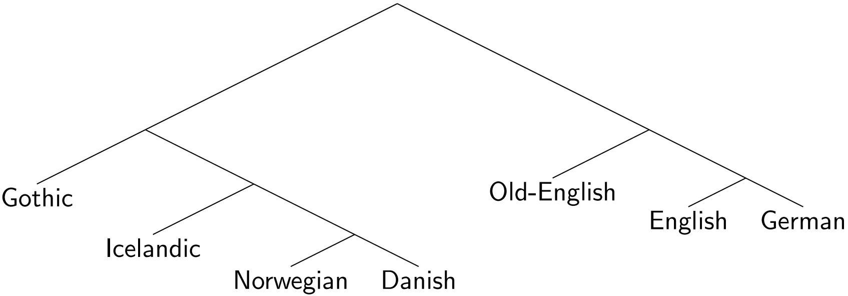
- persistent components tree (PCA variance 60%)



- persistent components tree (PCA variance 80%)

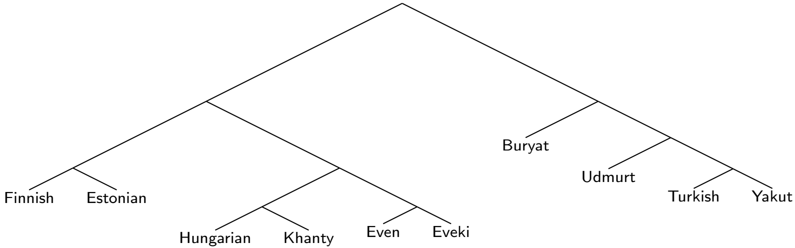


• correct historical phylogenetic tree obtained from same data via phylogenetic algebraic geometry



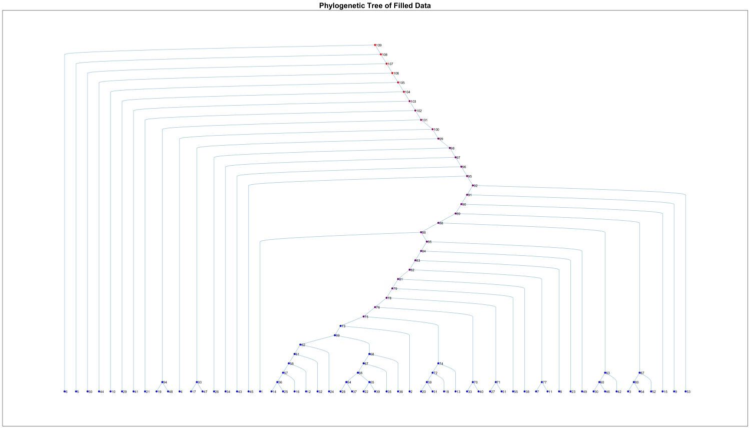
Persistent Connected Components and the Ural-Altaic Question

- LanGeLin data
- Japanese and Korean cluster together away from the Uralic and Altaic languages (excluded from the Ural-Altaic hypothesis)
- Uralic and Altaic languages persistent components tree

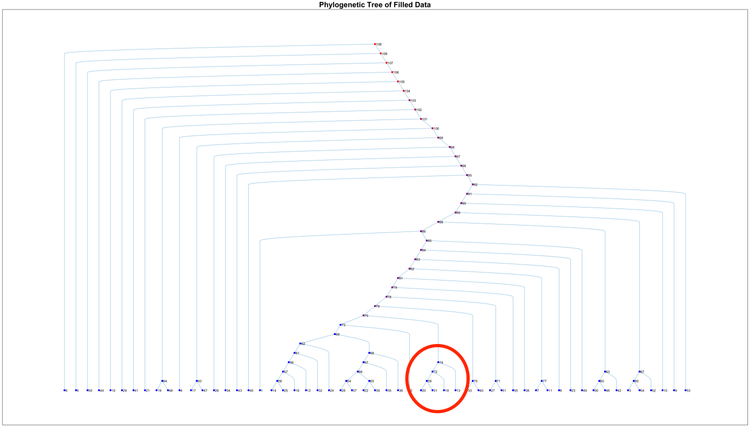


- Uralic language Udmurt placed in sub-cluster with Altaic; sub-cluster formed by Altaic languages Even and Eveki within a cluster of Uralic languages Estonian, Finnish, Hungarian, Khanty: persistent components mix Uralic and Altaic

Persistent components tree of Niger-Congo languages

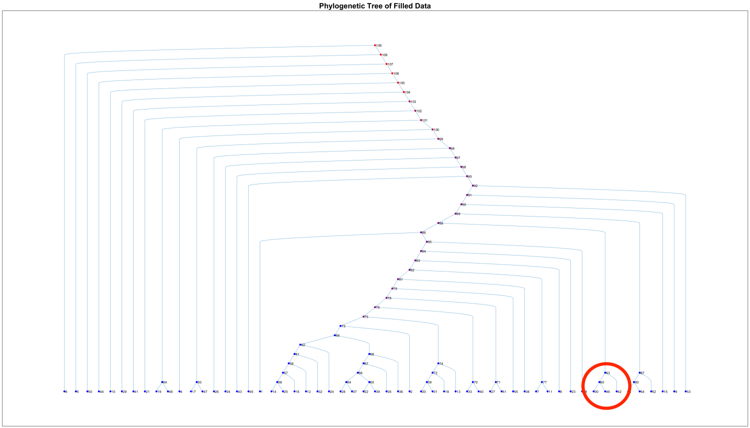


Example: cluster N. 74



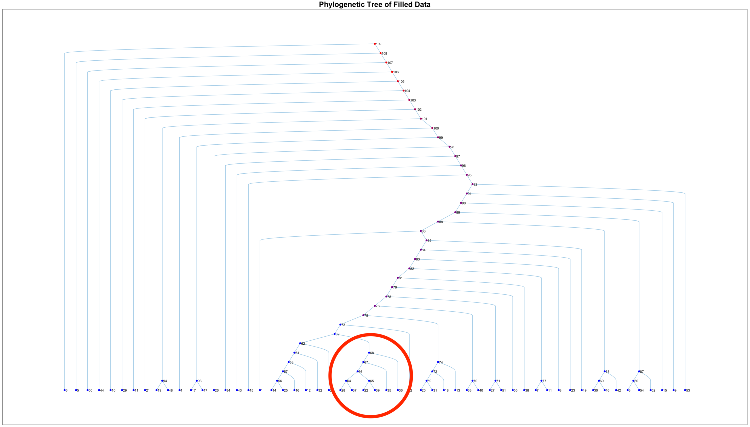
Atlantic-Congo Gur languages: Hanga, Konni, Gurene, Farefari

Example: cluster N. 63



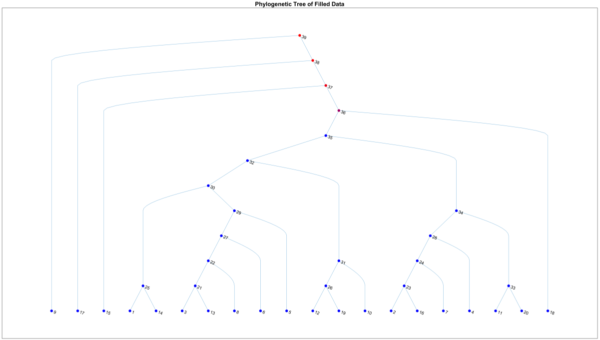
Southern Bantoid languages: Kom, Nweh (Grassfields group), Tuki (Mbam group)

Example: cluster N. 68



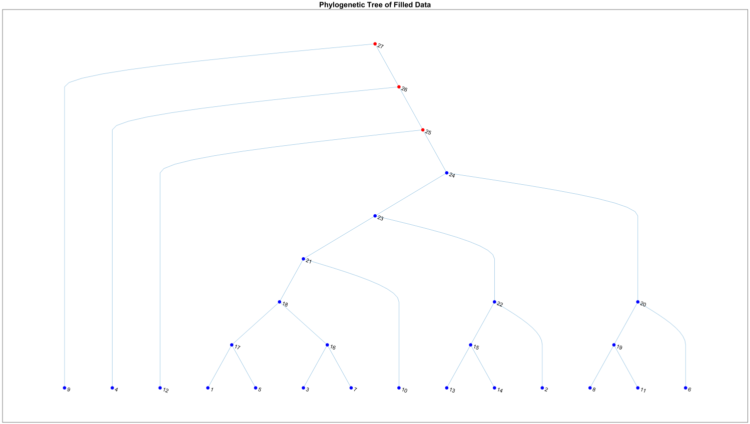
Mixes different branches: Mankanya (Bak group), Ndut (Senegambian Cangin), Kindendeule (Southern Bantoid), Naki (Southern Bantoid East Beboid), Medumba (Southern Bantoid Grassfields group), Igala (Volta-Niger Yoruboid)

Persistent components tree of Austronesian languages

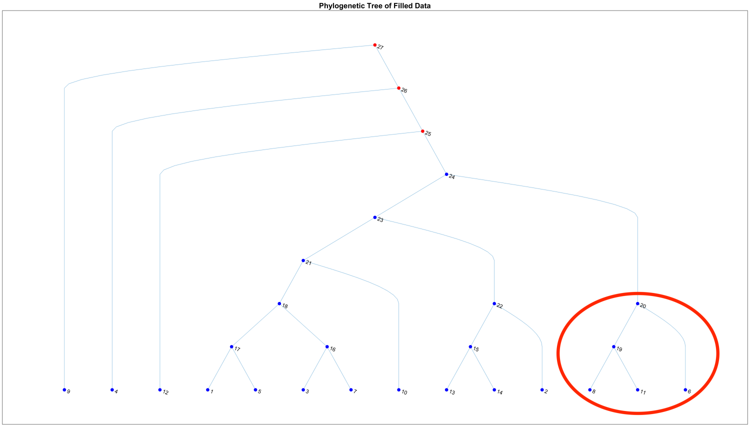


Prevalently structures of Malayo-Polynesian languages (some Formosan languages) but a lot of mixing of different groups of Malayo-Polynesian; Malagasy (Malayo-Polynesian East Barito of Madagascar) last to merge with the tree

Persistent components tree of Afro-Asiatic languages

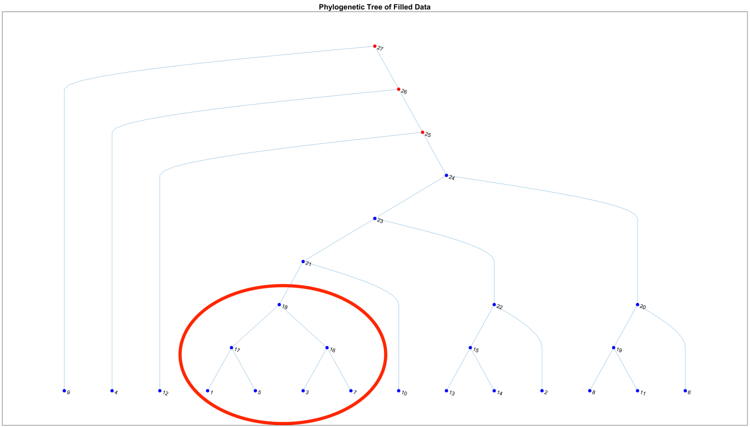


Example: cluster N. 20



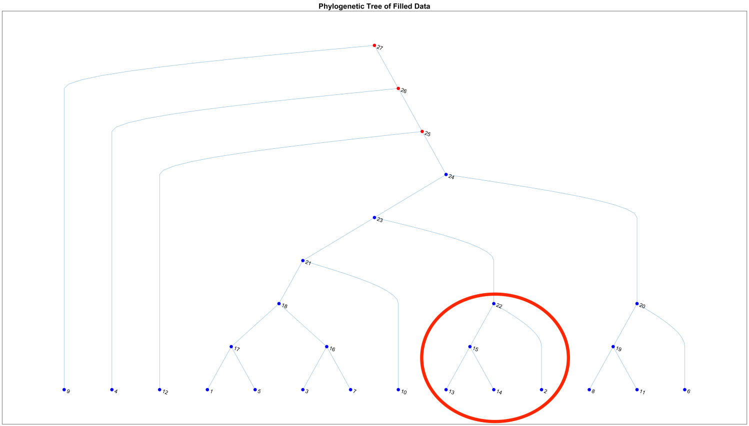
Bole, Hausa, Miya (Chadic languages) and Moroccan Arabic (influence of Chadic Berber languages)

Example: cluster N. 18



Biblical Hebrew, Gulf Arabic, Egyptian and Lebanese Arabic (Semitic languages), but Modern Hebrew misplaced

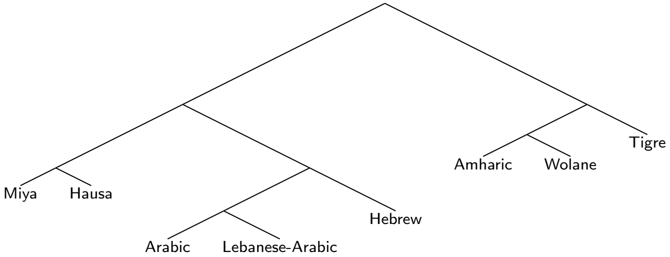
Example: cluster N. 22



Tigre, Wolane (South Semitic Ethiopic languages, but Amharic misplaced); Senaya (Central Semitic Aramaic)

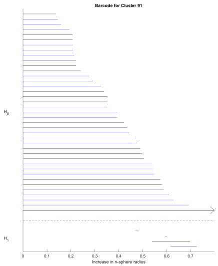
Compare persistent components to historical phylogenetic tree

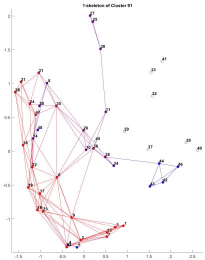
• relative position of some languages in this family according to historical tree

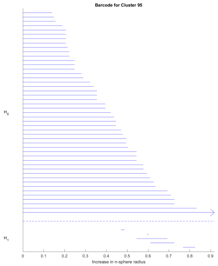


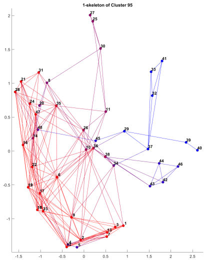
Persistent First Homology of Language Families

- filtered SSWL data: Indo-European + Ural-Altaic languages

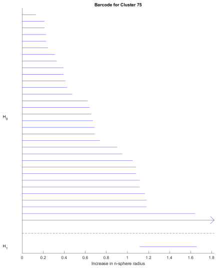


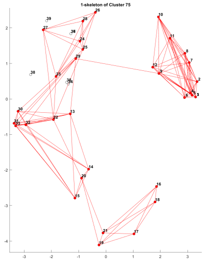


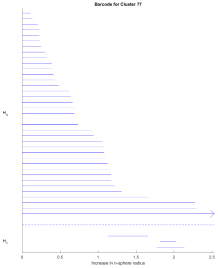


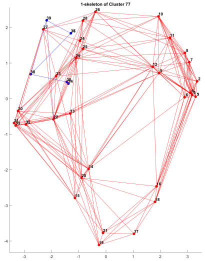


- LanGeLin data: Indo-European + Ural-Altaic languages



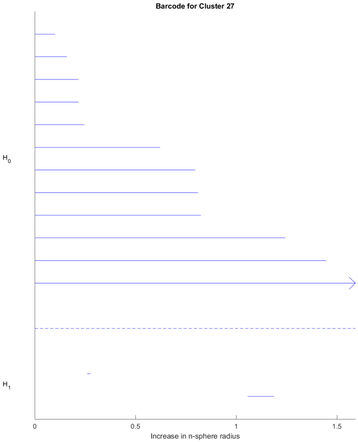


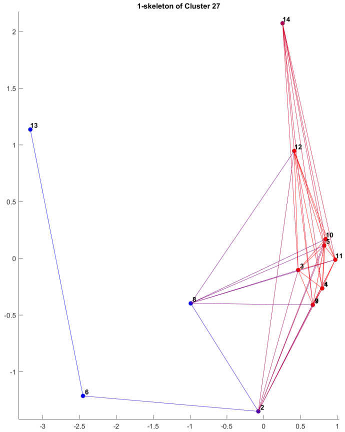




Nontrivial H_1 within a single subfamily

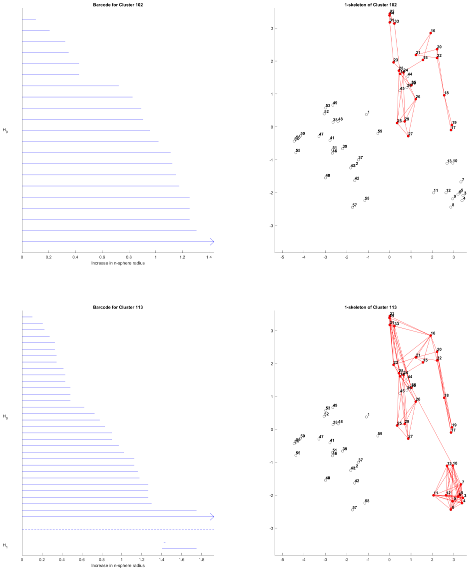
- Romance language family (SSWL data)





Gothic–Slavic–Greek loop (historical linguistic explanation)

- LanGeLin data



Identify cycle representative for persistent H_1-generator

1. identify first cluster of persistent components where new H_1-generator appears
2. list languages (vertices) added and all new cycles added in Vietoris-Rips 1-skeleton
3. in turn remove the languages belonging to one of the new cycles and recompute
4. if new generator disappears have a cycle representative
5. homologous cycles (remove all at once)

Gothic–Slavic–Greek loop: forms in the Indo-European languages between New Testament Greek, Romeyka Pontic Greek, Gothic, and Slavic languages (need to remove all Slavic languages together: homologous cycles)

Possible historical linguistic explanation for this H_1-generator

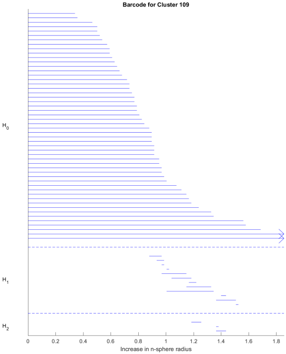
- influence (also at syntactic level) between Greek languages and South Slavic languages
- syntactic influence of New Testament Greek on Gothic (observed calques of Greek constructions in Gothic syntax)
- Proto-Slavic borrowing (influence of Gothic mostly lexical, but indications of morpho-syntactic borrowing as well)

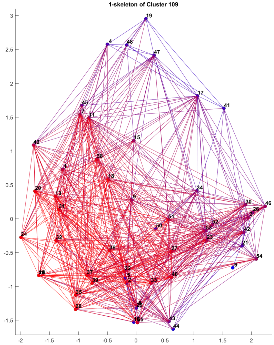
• Some References:

- O. Mišeska-Tomić, *Balkan Sprachbund. Morpho-syntactic Features*, Dordrecht, Springer 2006
- J.D. Gliesche, *Gothic Syntax*, lecture notes `http://users.clas.ufl.edu/drjdg/oe/pubs/gothicsyntax.pdf`
- R. Genis, *Comparing verbal aspect in Slavic and Gothic*, Amsterdam contributions to Scandinavian studies; No. 8, (2012) 59–80.

• other H_1-generators may reflect only homoplasy phenomena

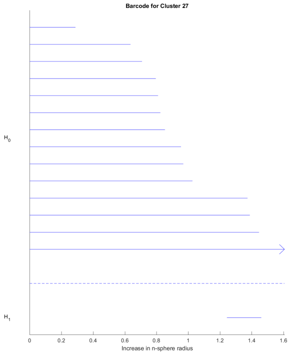
Persistent First Homology of the Niger-Congo languages

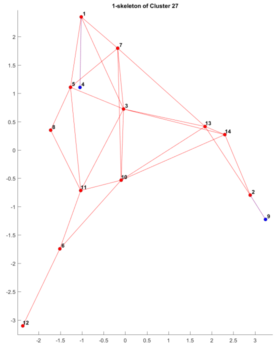




More persistent homology in high clusters than other language families (not seen in previous work where only a few clusters analyzed)

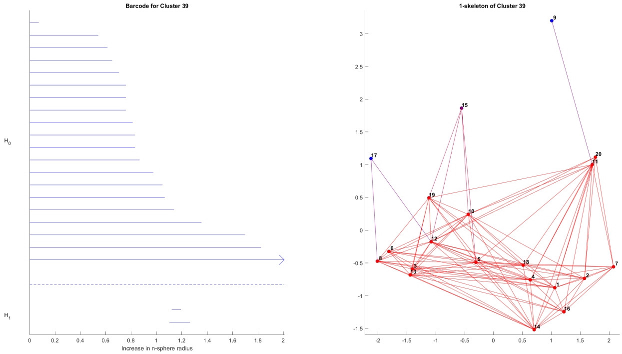
Persistent First Homology of the Afro-Asiatic languages





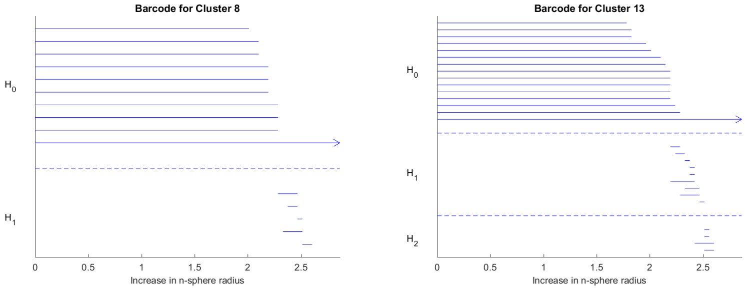
single persistent H_1-generator added only in the top cluster

Persistent First Homology of the Austronesian languages



two small persistent H_1-generators also added in top cluster

Comparison with homology of random simplicial sets



main differences: in random case H_1 occurs already in small clusters, shorter persistence, more stacking of many generators

Conclusions: more work (computationally hard) to identify representative cycles for all these H_1-generators for different language families, separate those that appear caused by homoplasy from those with historical linguistic significance